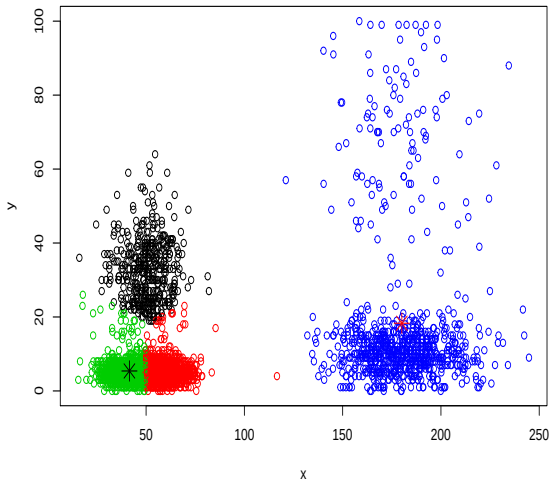


Introdução à Ciência de Dados

Fundamentos e Aplicações

Versão parcial preliminar

abril 2020

Pedro A. Morettin

Julio M. Singer

Departamento de Estatística
Universidade de São Paulo
Rua do Matão, 1010
São Paulo, SP 05508-090
Brasil

Conteúdo

Capítulo 1

Estatística, Ciência de Dados e Megadados

> Statistical learning theory does not belong to any specific branch of science: It has its own goals, its own paradigm, and its own techniques. Statisticians (who have their own paradigm) never considered this theory as part of statistics.
>
> V. Vapnik

1.1 Introdução

Atualmente, os termos *Data Science* (Ciência de Dados) e *Big Data* (Megadados)[1] são utilizados em profusão, como se fossem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos. Na década de 1980, numa palestra na Universidade de Michigan, EUA, C.F. Jeff Wu já sugeria que se adotassem os rótulos *Statistical Data Science*, ou simplesmente, *Data Science*, em lugar de *Statistics*, para dar maior visibilidade ao trabalho dos estatísticos. Talvez seja Tukey (1962, 1977), sob a denominação *Exploratory Data Analysis* (**Análise Exploratória de Dados**), o primeiro a dar importância ao que hoje se chama Ciência de Dados, sugerindo que se desse mais ênfase ao uso de tabelas, gráficos e outros dispositivos para uma análise preliminar de dados, antes que se passasse a uma **análise confirmatória**, que seria a **inferência estatística**. Outros autores, como Chambers (1993), Breiman (2001) e Cleveland (1985, 1993, 2001), também enfatizaram a preparação, apresentação e descrição dos dados como atividades que devem preceder a inferência ou modelagem.

Basta uma procura simples na Internet para identificar novos centros de Ciências de Dados (CD) em várias universidades ao redor do mundo, com programas de mestrado, doutorado e mesmo graduação. O interessante

[1] Para esclarecimento do significado dos termos cunhados em inglês, optamos pela tradução oriunda do **Glossário Inglês-Português de Estatística** produzido pela Associação Brasileira de Estatística e Sociedade Portuguesa de Estatística, disponível em `http://glossario.spestatistica.pt/`.

é que muitos desses programas estão alojados em escolas de Engenharia, Bioestatística, Ciência da Computação, Administração, Economia etc., e não em departamentos de Estatística. Paradoxalmente, há estatísticos que acham que Estatística é a parte menos importante de CD! Certamente isso é um equívoco. Como ressalta Donoho (2017), se uma das principais características de CD é analisar grandes conjuntos de dados (Megadados), há mais de 200 anos os estatísticos têm se preocupado com a análise de vastos conjuntos de dados provenientes de censos, coleta de informações meteorológicas, observação de séries de índices financeiros etc., que têm essa característica.

Outro equívoco consiste em imaginar que a Estatística Clássica (frequentista, bayesiana etc.) trata somente de pequenos volumes de dados, conhecidos como *Small Data*. Essa interpretação errônea vem do fato de que muitos livros didáticos apresentam conjuntos de dados, em geral de pequeno ou médio porte, para que as metodologias apresentadas possam ser aplicadas pelos leitores, mesmo utilizando calculadoras ou aplicativos estatísticos (pacotes). Nada impede que essas metodologias sejam aplicadas a grandes volumes de dados a não ser pelas dificuldades computacionais inerentes. Talvez seja este aspecto computacional, aquele que mascara os demais componentes daquilo que se entende por CD, pois em muitos casos, o interesse é dirigido apenas para o desenvolvimento de algoritmos cuja finalidade é aprender a partir dos dados, omitindo-se características estatísticas.

Em particular, Efron e Hastie (2016) ressaltam que tanto a teoria bayesiana quanto a frequentista têm em comum duas características: a **algorítmica** e a **inferencial**. Como exemplo, citam o caso da média amostral de um conjunto de dados $x_1, \ldots, x_n$, como estimador da média μ de uma população da qual se supõe que a amostra tenha sido obtida. O algoritmo para cálculo é

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.1}$$

Resta saber quão acurado e preciso é este estimador. Como sabemos, admitindo-se que a amostra tenha sido colhida segundo um procedimento adequado, $\mathrm{E}(\overline{x}) = \mu$, ou seja, o estimador $\overline{x}$ é não enviesado e o seu erro padrão é

$$\mathrm{ep} = [\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2]^{1/2}, \tag{1.2}$$

que mostra a sua consistência e estabelece as bases para uma inferência estatística (frequentista!) adequada. Mais adiante, esses autores mencionam que:

Optimality theory, both for estimating and for testing, anchored statistical practice in the twentieth century. The larger datasets and more complicated inferential questions of the current era have strained the capabilities of that theory. Computer-age statistical inference, as we will see, often displays an unsettling ad hoc character. Perhaps some contemporary Fishers

and Neymans will provide us with a more capacious optimality theory equal to the challenges of current practice, but for now that is only a hope.

Blei e Smyth (2017) discutem Estatística e CD sob três perspectivas: estatística, computacional e humana. Segundo os autores, CD é uma filha da estatística e da ciência da computação. A estatística serviria à CD guiando a coleta e análise de dados complexos. A computação, desenvolvendo algoritmos que distribuem conjuntos enormes de dados por múltiplos processadores (proporcionando velocidade de cálculo) e equipamentos com grande capacidade de memória (permitindo seu armazenamento). Sob a perspectiva humana, a CD contempla modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas, entender o domínio desses problemas, decidir quais dados obter, como processá-los, explorá-los e visualizá-los, selecionar um modelo estatístico e métodos computacionais apropriados, além de comunicar os resultados da análise de uma forma inteligível para aqueles que propuseram os problemas.

Donoho (2017) discute várias **memes** (uma ideia ou símbolo transmitido pelas chamadas **mídias sociais**) sobre Megadados (MD) e CD. Algumas **memes** são vídeos ou expressões verbais, outras, como MD e CD, são providas de um conteúdo filosófico mais profundo. Por exemplo, sobre a *Big Data Meme*, diz que se pode rejeitar o termo MD como um critério para uma distinção séria entre Estatística e CD, para isso evocando o que dissemos acima sobre análise de dados de censos e o fato de pesquisadores na área de inferência estatística terem buscado o entendimento científico de MD por décadas.

Um dos aspectos tradicionalmente neglicenciados por estatísticos é aquele em que os dados têm natureza não ortodoxas como imagens, sons etc. Nesse caso, algoritmos computacionais são essenciais para seu tratamento que, por sua vez, não pode prescindir do apoio de um estatístico.

1.2 Aprendizado com Estatística

Outros termos muito utilizados hoje em dia são *Statistical Learning* (**Aprendizado com Estatística**) e *Machine Learning* (**Aprendizado com Máquina ou Automático**) e consistem na utilização de modelos estatísticos acoplados a algoritmos computacionais desenvolvidos para extrair informação de conjuntos de dados contendo, em geral, muitas unidades amostrais e muitas variáveis.

O Aprendizado com Estatística (AE) pode ser **supervisionado** ou **não supervisionado**. No AE supervisionado, o objetivo é prever o valor de uma variável resposta (*output*) a partir de variáveis preditoras (*input*). A variável reposta pode ser quantitativa ou qualitativa (ver Capítulo 3). No caso de variáveis respostas quantitativas, um dos modelos estatísticos mais utilizados no AE é o de **regressão**; quando a variável resposta é qualitativa, utilizam-se geralmente modelos de **regressão logística** para a análise (ver Capítulo 6). Adicionalmente, para variáveis qualitativas (categóricas ou

discretas), podem ser utilizados modelos de classificação como máquinas de suporte vetorial (*support vector machines*), árvores de decisão, método do k-ésimo vizinho mais próximo, função discriminante linear de Fisher etc. (ver Capítulo 10).

No caso de AE não supervisionado, temos apenas um conjunto de variáveis preditoras (*inputs*) e o objetivo é descrever associações e padrões entre essas variáveis. Nesse caso, não há uma variável resposta. Dentre as técnicas mais utilizadas nesta situação temos aquela de obter a melhor representação dos dados. Há várias maneiras de conseguir tal representação. Entre elas, incluímos a **análise de componentes principais** e a **análise de componentes independentes** (ambas proporcionando a redução da dimensionalidade dos dados) e a análise de conglomerados (ou agrupamentos).

1.3 Aprendizado automático

Inteligência Artificial (IA) é um tópico de extremo interesse e que aparece frequentemente nas mídias escritas e faladas. Normalmente o termo suscita questões do tipo: no futuro computadores tornar-se-ão inteligentes e a raça humana será substituída por eles? Todos perderemos nossos empregos, porque seremos substituídos por robôs inteligentes? Pelo menos até o presente esses receios são infundados. Nesse contexto, veja Jordan (2019). Segundo esse autor, o que é rotulado hoje como IA, nada mais é daquilo que chamamos de Aprendizado Automático (*machine learning*).

Acredita-se que o artigo de Turing (1950) seja o primeiro a tratar do tema. A primeira frase do artigo diz:

I propose to consider the question, "Can machines think?"

Segue-se discussão sobre o que se entende por "máquina", por "pensar"e por um jogo, chamado "jogo da imitação". Turing também discute condições para considerar uma máquina inteligente, as chamadas de *Turing test*. A primeira página do artigo está na Nota de Capítulo 1.

O tópico de IA foi tratado a seguir por McCarthy et al. (1955), na forma de uma proposta para um projeto de pesquisa no Dartmouth College. Cópia da primeira página do original encontra-se na Nota de Capítulo 2. Entre os signatários, encontra-se Shannon, precursor da Teoria da Informação.

De modo informal, a IA é um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos (Chollet, 2018). A IA está intimamente ligada ao desenvolvimento da computação (ou programação de computadores) e até a década de 1980, a IA era entendida como na programação clássica, baseada em um sistema computacional (SC) (um computador ou um conglomerado (*cluster*) de computadores ou nuvem etc.) no qual se alimentam dados e uma regra de cálculo para se obter uma reposta. Por exemplo, num problema de regressão a ser resolvido por meio do método de mínimos quadrados para obtenção dos estimadores dos parâmetros, a regra de cálculo é um algoritmo que pode ser programado em alguma linguagem (Fortran, C, R, Python etc.). A maioria dos pacotes computacionais

existentes funciona dessa maneira.

A partir da década de 1990, o aprendizado automático (AA) criou um novo paradigma para analisar dados oriundos de reconhecimento de imagens, voz, escrita etc., dificilmente solucionáveis sem o recente avanço na capacidade computacional. A ideia subjacente é **treinar** um SC programando-o para ajustar diferentes modelos por meio dos algoritmos associados (muitas vezes bastante complexos) repetidamente na análise de um conjunto de dados. Nesse processo, diferentes modelos são ajustados a conjuntos de dados (chamados **dados de treinamento**) de modo que o melhor (segundo algum critério pré-estabelecido) deles seja encontrado. A seleção do modelo mais adequado para a análise é baseada na avaliação do desempenho dos competidores em um novo conjunto de observações chamado de **dados de teste**. Convém ressaltar que o objetivo do aprendizado automático não é selecionar o melhor **ajuste** do modelo (por exemplo, as estimativas dos coeficientes de um modelo de regressão), mas sim o melhor modelo (por exemplo, que termos polinomiais devem ser incluídos em um modelo de regressão). O modelo selecionado deve ser ajustado ao conjunto de dados completo (treinamento e teste) para se obter o ajuste (estimativas dos coeficientes de um modelo de regressão, por exemplo) que será empregado para previsão de novos dados (consistindo apenas dos valores dos preditores).

Quando esses dois conjuntos de dados (treinamento e teste) não estão definidos *a priori*, o que é mais comum, costuma-se dividir o conjunto disponível em dois, sendo um deles destinado ao treinamento do SC com o outro servindo para teste. Calcula-se então alguma medida do erro de previsão obtido ao se aplicar o resultado do ajuste do modelo de treinamento aos dados de teste. Essa subdivisão (em conjuntos de treinamento e de teste diferentes) é repetida várias vezes, ajustando o modelo a cada conjunto de dados de treinamento, utilizando os resultados para previsão com os dados de teste e calculando a medida adotada para o erro de previsão. A média dessa medida é utilizada como avaliação do desempenho do modelo proposto. Para comparar diferentes modelos, repete-se o processo com cada um deles e aquele que produzir a menor média do erro de previsão é o modelo a ser selecionado. Esse processo é conhecido como **validação cruzada** (ver a Nota de Capítulo 2 no Capítulo 8).

Muitos métodos de aprendizado automático são implementados por meios de redes neurais e no Capítulo 13 daremos algumas ideias sobre elas.

1.4 Uma cronologia do AE

Embora a terminologia Aprendizado com Estatística seja recente, a maioria dos conceitos foi desenvolvida a partir do século 19. Essencialmente, AE e AM (ou AA) tratam dos mesmos tópicos, mas utilizaremos o termo AE quando os métodos de AM são tratados com metodologias estatísticas apropriadas.

1.4.1 Probabilidades

As origens da teoria de probabilidades remontam a 1654, com Fermat (1601–1665), Pascal (1632–1662), que trataram de jogos de dados, baralhos etc. Huygens (1629–1695) escreveu o primeiro livro sobre probabilidades em 1657. A primeira versão do Teorema de Bayes (Bayes, 1702–1761) foi publicada em 1763.

1.4.2 Estatística

Gauss (1777–1856) propôs o método de mínimos quadrados (MQ) na última década do Século 18 (1795) e usou-o regularmente depois de 1801 em cálculos astronômicos. Todavia, foi Legendre (1752–1833) quem primeiro publicou sobre o método no apêndice de seu livro "Nouvelles Methodes pour la Détermination des Orbites des Comètes", sem justificação. Gauss (1809) deu a justificativa probabilística do método em "The Theory of the Motion of Heavenly Bodies". Basicamente, eles implementaram o que é hoje chamado de Regressão Linear.

Laplace (1749–1761) desenvolveu o Teorema de Bayes independentemente, em 1774. Em 1812 e 1814 deu a interpretação bayesiana para probabilidade e fez aplicações científicas práticas. Há autores que julgam que a chamada Inferência Bayesiana dever-se-ia chamar Inferência Laplaciana, devido às suas contribuições na área (lembremos da Aproximação de Laplace, que se usava para obter distribuições a posteriori, antes do adventos de métodos MCMC e filtros de partículas).

Contribuições de Jeffrey (1939) podem ser consideradas como um reinício da Inferência Bayesiana, juntamente com as obras de de Finetti, Savage e Lindley.

A Inferência Frequentista (testes de hipóteses, estimação, planejamento de experimentos e amostragem) foi iniciada por Fisher (1890–1962) e Neyman (1894–981). Fisher, em 1936, propõe a técnica de Análise Discriminante Linear e seus dois livros "Statistical Methods for Research Workers", de 1925 e "The Design of Experiments", de 1935 são marcos da teoria frequentista.

Segundo Stigler, o artigo de Fisher (1922), "On the mathematical foundation of theoretical statistics", publicado na *Phil. Trans. Roya Society, A* foi o artigo mais influente sobre Teoria Estatística no Século 20".

Por sua vez, Neyman e Pearson (1933), publicaram os dois artigos fundamentais sobre testes de hipóteses, consubstanciados no excelente livro de Lehmann de 1967.

A partir da década de 1940 começaram a aparecer abordagens alternativas ao modelo de regressão linear, como a **Regressão Logística**, os **Modelos Lineares Generalizados** (Nelder e Wedderburn, 1970), além dos **Modelos Aditivos Generalizados** (Hastie e Tibshirani, 1986).

Em 1969, Efrom introduz a técnica **Bootstrap** e em 1970, Hoerl e Kennard introduzem a **Regressão Ridge**. Até o final da década de 1970, os métodos lineares predominaram. A partir da década de 1980, os avanços computacionais possibilitaram a aplicação de métodos não lineares, como

o **CART** (Classification and Regresion Trees, Friedman et al., 1984). Em 1996, Tibshirani introduz o método de regularização **LASSO**, que juntamente com os métodos **Ridge**, **Elastic net** e outras extensões passam a ser usados em conjunção com modelos de regressão, por exemplo, com o intuito de prevenir o fenômeno de sobreajuste (**overfitting**), mas que também funcionam como métodos de seleção de modelos.

1.4.3 Estatística e computação

Os avanços no AE estão diretamente relacionados com avanços na área computacional. Até 1960, os métodos estatísticos tinham que ser implementados em máquinas de calcular manuais ou elétricas. A partir de 1960 até 1980, apareceram as máquinas eletrônicas e os **grandes computadores**, como o IBM 1620, CDC 360, VAX etc, que trabalhavam com cartões e discos magnéticos. A linguagem FORTRAN predominava.

A partir de 1980 aparecem os computadores pessoais, supercomputadores, computação paralela, *clouds*, linguagens C, C_+, S e os pacotes estatísticos SPSS, BMDP, SAS, SPlus (que utiliza a linguagem S, desenvolvida por Chambers, do Bell Labs), MatLab etc. A partir de 1984 surge a linguagem R (que na realidade é basicamente a linguagem S com algumas modificações) e o repositório CRAN, de onde pacotes para análises estatísticas podem ser obtidos livremente; essa linguagem passa a ser a linguagem preferida dos estatísticos.

Métodos de AE não usualmente considerados em programas de graduação e pós-graduação em Estatística surgiram recentemente (e não tão recentemente, veja a citação de Vapnik no início desse capítulo), estão atraindo a atenção e são englobados no que hoje chamamos de Ciência de Dados. Tais métodos incluem *Support Vector Machines*, *Decision Trees*, *Random Forests*, *Bagging*, *Boosting* etc. Outros métodos mais tradicionais, mas que voltaram a estar em evidência, como Redução da Dimensionalidade (incluindo Análise de Componentes Principais, Análise Fatorial, Análise de Componentes Independentes) e Análise de Agrupamentos já fazem parte de métodos estudados em cursos de estatística.

1.5 Notação e tipos de dados

Vamos introduzir, agora, a notação usada no livro. Denotemos por $\mathbf{X}$, uma matriz com dimensão $n \times p$, contendo os dados; n indica o número de unidades amostras (indivíduos, por exemplo) e p o número de variáveis. Especificamente,

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

As linhas de $\mathbf{X}$ são denotadas por $\mathbf{x}_1, \ldots, \mathbf{x}_n$, sendo que cada $\mathbf{x}_i$ é um vetor com dimensão $p \times 1$. As colunas de $\mathbf{X}$ são denotadas por $\mathbf{x}_1^*, \ldots, \mathbf{x}_p^*$, sendo que cada $\mathbf{x}_j^*$ é um vetor com dimensão $n \times 1$.

Então

$$\mathbf{X} = [\mathbf{x}_1^*, \mathbf{x}_2^*. \ldots, \mathbf{x}_p^*] = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}.$$

Também, denotemos por y_i a i-ésima componente do vetor $\mathbf{y} = (y_1, \ldots, y_n)^\top$. No caso de AE supervisionado, y_i é a resposta aos preditores $\mathbf{x}_i$, num problema de regressão e corresponde ao rótulo da i-ésima classe, num problema de classificação. Consequentemente os dados são os pares $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

Podemos ter as seguintes estruturas de dados:

a) grande número de unidades amostrais e pequeno número de variáveis, $n >> p$;

b) pequeno número de unidades amostrais e grande número de variáveis, $p >> n$;

c) grande número de unidades amostrais e grande número de variáveis, n e p grandes.

Todos esses casos podem ser cognominados **megadados** (*big data*). Quando $n << p$, os dados têm **alta dimensão** (*high dimension*) e requerem procedimentos especiais. Por outro lado, megadados podem ser classificados como:

a) **Dados estruturados**: em que a informação se ajusta às estruturas usuais de bases de dados, relativamente fáceis de armazenar e analisar. Exemplos usuais de dados numéricos ou não, que podem ser dispostos em matrizes de dados.

b) **Dados não estruturados**: tudo o que não se encaixa no item anterior, como arquivos de textos, páginas da *web*, emails, midias sociais etc.

Os megadados podem ser descritos pelos quatro V sugeridos em `\http://www.ibmbigdatahub.com`, nomeadamente, **Volume** (escala dos dados), **Variedade** (formas diferentes de dados), **Velocidade** (análise de *streaming data*) e **Veracidade** (incerteza sobre os dados). Uma representação pictórica dessas estruturas de dados está apresentada na Figura 1.1.

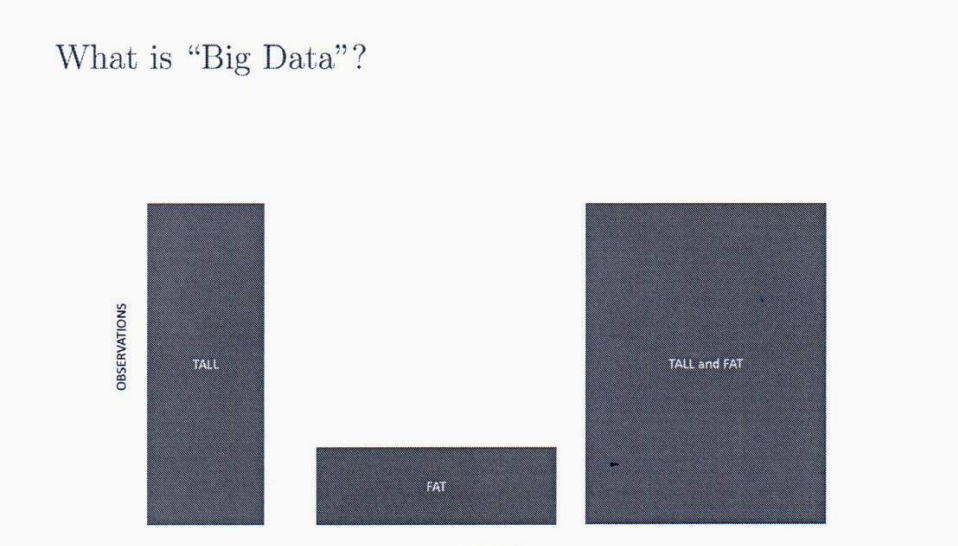


Figura 1.1: Estruturas de dados

Megadados implicam megamodelos, que contêm um grande número de parâmetros a serem estimados, como em modelos de regressão múltipla em que o número de variáveis, p, é grande. O ajuste modelos lineares a dados de alta dimensão pode ser tratado por meio técnicas de redução da dimensionalidade (ACP, AF, ACI), regularização ou métodos bayesianos. Para modelos não lineares, árvores de decisão e redes neurais são técnicas mais adequadas.

1.6 O modelo geral para o Aprendizado Estatístico

1.6.1 Regressão

Dados o vetor $\mathbf{y}$ com os valores da variável resposta e a matriz $\mathbf{X}$ com os correspondentes valores dos preditores, o modelo geral para regressão em AE supervisionado tem a forma

$$y_i = f(\mathbf{x}_i) + e_i, \quad i = 1, \ldots, n, \tag{1.3}$$

com $\mathrm{E}(e_i) = 0$ e f denotando uma função desconhecida, chamada de **informação sistemática**. O objetivo do AE é encontrar métodos para estimar f.

Estimada f, podemos usar o modelo (1.3) para fazer previsões ou inferência sobre a população de onde os dados foram extraídos. O **previsor** de y_i é $\widehat{y}_i = \widehat{f}(\mathbf{x}_i)$ e a acurácia de $\widehat{\mathbf{y}}$ como previsor de $\mathbf{y}$ depende dos seguintes dois erros (James et al., 2017):

a) **erro redutível**, introduzido pelo previsor de f; assim chamado porque podemos melhorar a acurácia de $\widehat{f}$ usando técnicas de AE mais apropriadas;

b) **erro irredutível**, que depende de e_i e não pode ser previsto por $\mathbf{X}$, mesmo usando o melhor previsor de f.

Supondo $\widehat{f}$ e $\mathbf{X}$ fixos, pode-se verificar que

$$\begin{aligned} E(y_i - \widehat{y}_i)^2 &= \mathrm{E}[f(\mathbf{x}_i) + e_i - \widehat{f}(\mathbf{x}_i)]^2 \\ &= \mathrm{E}[f(\mathbf{x}_i) - \widehat{f}(\mathbf{x}_i)]^2 + \mathrm{Var}(e_i), \end{aligned} \tag{1.4}$$

para $i = 1, \ldots, n$. O primeiro termo do segundo membro de (1.4) mede o efeito do erro redutível e o segundo termo, o efeito do erro irredutível. Consequentemente, o objetivo é minimizar o primeiro.

Muitas vezes, o interesse pode não ser fazer previsões, mas entender como a resposta é afetada pela variação dos preditores ou identificar os mais importantes na relação entre a resposta e cada um dos deles.

Para estimar f podemos usar **métodos paramétricos** ou **métodos não paramétricos**.

No primeiro caso, fazemos alguma suposição sobre a forma de f como no modelo de regressão múltipla usual com p variáveis. Nesse caso, o problema é mais simples, pois temos que estimar um número finito de parâmetros. Selecionado o modelo, devemos ajustá-lo aos dados de treinamento, ou seja, devemos **treinar** o modelo. No caso de modelos de regressão, o método mais usado na estimação é o de Mínimos Quadrados (MQ) mas há outros métodos disponíveis, como SVM (support vector machines). O ajuste de um modelo de regressão por MQ, por exemplo, pode ser pobre, como no Exemplo 6.7 do Capítulo 6 (veja a Figura 6.22). Nesse caso, pode-se tentar ajustar modelos mais flexíveis, escolhendo outras formas funcionais para f, incluindo aí modelos não lineares. Todavia, modelos mais flexíveis envolvem a estimação de um número muito grande de parâmetros, o que pode gerar um problema de sobreajuste (*overfitting*).

No segundo caso, não fazemos nenhuma hipótese sobre a forma funcional de f e como o problema envolve a estimação de grande número de parâmetros, necessitamos um número grande de observações para obter estimadores acurados de f. Vários métodos podem ser usados com essa finalidade, dentre os quais destacamos aqueles que utilizam:

- kernels;
- polinômios locais (*e.g.*, Lowess)
- splines
- polinômios ortogonais (*e.g.*, Chebyshev)
- outras bases ortogonais (*e.g.*, Fourier, ondaletas)

Métodos menos flexíveis (*e.g*, regressão linear) ou mais restritivos em geral são menos acurados e mais fáceis de interpretar. Por outro lado, métodos

mais flexíveis (*e.g*, splines) são mais acurados e mais difíceis de interpretar. Para cada conjunto de dados, um método pode ser preferível a outros, dependendo do objetivo da análise. A escolha do método talvez seja a parte mais difícil do AE.

No caso de modelos de regressão, a medida mais usada para avaliação da qualidade do ajuste é o **Erro Quadrático Médio** (EQM), definido por

$$\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \widehat{f}(\mathbf{x}_i)]^2, \tag{1.5}$$

em que $\widehat{f}(\mathbf{x}_i)$ é o preditor de y para a i-ésima observação. O EQM calculado no conjunto de treinamento que produz $\widehat{f}$ é chamado **EQM de treinamento**. Em geral, estamos mais interessados na acurácia do ajuste para os dados de teste e nesse caso podemos calcular o **EQM de teste**,

$$\text{Média}[(y_0 - \widehat{f}(\mathbf{x}_0)]^2, \tag{1.6}$$

que é o erro de previsão quadrático médio para as observações do conjunto de dados de teste, em que o elemento típico é denotado por $(\mathbf{x}_0, y_0)$. A ideia é ajustar diferentes modelos aos dados de treinamento, obtendo diferentes estimativas $\widehat{f}$ por meio da minimização de (1.5), calcular o correspondente EQM de teste via (1.6) e escolher o modelo para o qual esse valor seja mínimo. Muitas vezes, usa-se **validação cruzada** (VC) nesse processo (veja a Nota de Capítulo 2 do Capítulo 8).

Para os dados do conjunto de teste, $(\mathbf{x}_0, y_0)$,

$$\text{E}[y_0 - \widehat{f}(\mathbf{x}_0)]^2 = \text{Var}[\widehat{f}(\mathbf{x}_0)] + [\text{Vies}(\widehat{f}(\mathbf{x}_0))]^2 + \text{Var}(e_0). \tag{1.7}$$

Em resumo, procuramos selecionar o modelo que produza simultaneamente baixo viés e baixa variância que atuam em sentidos opostos. Na prática, podemos estimar (1.7) para os dados do conjunto de teste por meio de (1.6). Também é possível estimar $\text{Var}[\widehat{f}(\mathbf{x}_0)]$, mas como f é desconhecida não há como estimar o viés de $\widehat{f}(\mathbf{x}_0)$ dado que $\text{Var}(e_0)$ também não é conhecida. Em geral, métodos de AE mais flexíveis têm viés baixo e variância grande. Na maioria dos casos, o EQM de treinamento é menor que o EQM de teste e o gráfico do EQM de teste em função do número de parâmetros de diferentes modelos, em geral, apresenta uma forma de U, resultante da competição entre viés e variância.

1.6.2 Classificação

Problemas de **classificação** são aqueles em que as respostas $y_1, \ldots, y_n$ são qualitativas. Formalmente, seja $(\mathbf{x}, y)$, com $\mathbf{x} \in \mathbb{R}^p$ e $y \in \{-1, 1\}$. Um **classificador** é uma função $g : \mathbb{R}^p \to \{-1, 1\}$ e a **função erro** ou **risco** é a probabilidade de erro, $L(g) = P\{g(X) \neq Y\}$.

Obtendo-se um estimador de g, digamos $\widehat{g}$, sua acurácia pode ser medida pelo estimador de $L(g)$, chamado de **taxa de erro de treinamento**,

que é a proporção de erros gerados pela aplicação de $\widehat{g}$ às observações de treinamento, ou seja,

$$\widehat{L}(\widehat{g}) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{y}_i), \tag{1.8}$$

O interesse está na **taxa de erro de teste**

$$\text{Média}[I(y_0 \neq \widehat{y_0})], \tag{1.9}$$

para as observações de teste $(\mathbf{x}_0, y_0)$. Um bom classificador tem (1.9) pequeno. Pode-se provar que (1.9) é minimizado, em média, por um classificador que associa cada observação à classe mais provável, dados os preditores; ou seja, por aquele que maximiza

$$P(y = j|\mathbf{x} = \mathbf{x}_0). \tag{1.10}$$

Tal classificador é chamado de **classificador de Bayes**.

No caso de duas classes, a ideia é classificar a observação teste na classe -1 se $P(y = -1|\mathbf{x} = \mathbf{x}_0) > 0,5$ ou na classe 1, em caso contrário. O classificador de Bayes produz a menor taxa de erro, dada por $1 - \max_j P(y = j|\mathbf{x} = \mathbf{x}_0)$. A taxa de erro de Bayes global é $1 - \text{E}(\max_j P(y = j|\mathbf{x} = \mathbf{x}_0)$, em que $\text{E}(\cdot)$ é caculada sobre todos os valores de $\mathbf{x}$. O classificador de Bayes não pode ser calculado na prática, pois não temos conhecimento da distribuição condicional de y, dado $\mathbf{x}$. Uma alternativa é estimar a distribuição condicional e, então, estimar (1.10).

O classificador do K**-ésimo vizinho mais próximo** (*K-nearest neighbors, KNN*) estima tal distribuição por mrio do seguinte algoritmo:

i) Escolha $K > 0$ inteiro e uma observação teste $\mathbf{x}_0$.

ii) Identifique os K pontos do conjunto de treinamento mais próximos de $\mathbf{x}_0$; chame-os de $\mathcal{N}$.

iii) Estime a probabilidade condicional da classe j por meio de

$$P(y = j|\mathbf{x} = \mathbf{x}_0) = \frac{1}{K}\sum_{i\in\mathcal{N}} I(y_i = j). \tag{1.11}$$

iv) Classifique $\mathbf{x}_0$ na classe com a maior probabilidade condicional.

A escolha de K crucial e o resultado depende dessa escolha. Tratamos desse problema no Capítulo 10.

1.7 Este livro

Um dos maiores problemas oriundos da disseminação indiscriminada das técnicas utilizadas em CD é a confiança exagerada nos resultados obtidos da aplicação de algoritmos computacionais. Embora sejam essenciais em muitas situações, especialmente com megadados, sua utilização sem o concurso

dos princípios do pensamento estatístico, fundamentado nas características de aleatoriedade e variabilidade inerentes a muitos fenômenos, pode gerar conclusões erradas ou não sustentáveis. Também lembramos que CD só faz sentido a partir de um problema em que as questões a serem respondidas estejam claramente especificadas.

Independentemente do volume de dados disponíveis para análise, Ciência de Dados é uma atividade multidisciplinar que envolve

i) um problema a ser resolvido com questões claramente especificadas;
ii) um conjunto de dados (seja ele volumoso ou não);
iii) os meios para sua obtenção;
iv) sua organização;
v) a especificação do problema original em termos das variáveis desse conjunto de dados;
vi) a descrição e resumo dos dados à luz do problema a ser resolvido;
vii) a escolha das técnicas estatísticas apropriadas para a resolução desse problema;
viii) os algoritmos computacionais necessários para a implementação dessas técnicas;
ix) a apresentação dos resultados.

Obviamente, a análise de problemas mais simples pode ser conduzida por um estatístico (sempre em interação com investigadores da área em que o problema se insere). Em problemas mais complexos, especialmente aqueles com grandes conjuntos de dados que possivelmente contenham imagens, sons etc., só **uma equipe** com profissionais de diferentes áreas poderá atacá-los adequadamente. Em particular, essa equipe deve ser formada, pelo menos, por um profissional de alguma área do conhecimento em que o problema a ser resolvido se situa, por um estatístico, por um especialista em banco de dados, por um especialista em algoritmos computacionais e possivelmente por um profissional da área de comunicação. Se por um lado os aspectos computacionais são imprescindíveis nesse contexto, por outro, uma compreensão dos conceitos básicos de Estatística deve constar da formação de todos os membros da equipe.

Embora todos os aspectos envolvidos em Ciência dos Dados sejam abordados neste livro, o foco será na metodologia estatística necessária em qualquer análise, envolvendo a coleta, organização, resumo e análise dos dados. Para um melhor entendimento das técnicas apresentadas, em geral, usamos conjuntos de dados não muito volumosos, de modo que os leitores poderão reanalisá-los usando desde uma calculadora até programas sofisticados. Desde que tenham acesso e aptidão para lidar com *software* adequado, os leitores não terão dificuldades em analisar grandes conjuntos de dados sob as mesmas perspectivas apresentadas no texto.

A (bem-vinda) popularização das técnicas utilizadas naquilo que se chama Ciência de Dados não está isenta de problemas. Nem sempre os profissionais

que se aventuram por essa seara têm o conhecimento básico dos métodos estatísticos que fundamentam os algoritmos mais empregados na análise de dados. Nosso objetivo é preencher essa lacuna, apresentando conceitos e métodos da Análise Exploratória de Dados necessários para a análise de dados e indicando como são empregados nos problemas práticos com que "cientistas de dados" são usualmente desafiados. O Capítulo 2 é dedicado à preparação dos dados, geralmente obtidos de forma inadequada para análise. Os conceitos subjacentes a esses métodos são detalhados nos Capítulos 3 a 5. A utilização desses métodos no aprendizado estatístico supervisionado ou não é apresentada nos Capítulos 6 a 12.

O texto poderá ser utilizado em programas de bacharelado em Estatística (especialmente na disciplina de Estatística Descritiva com os capítulos 1 a 7 e na disciplina de Estatística Aplicada, com a inclusão dos capítulos restantes). Além disso, servirá para "cientistas de dados" que tenham interesse nos aspectos que fundamentam quaisquer análises de dados.

Embora muitos cálculos necessários para uma análise estatística possam ser concretizados por meio de calculadoras, o recurso a pacotes computacionais é necessário tanto para as análises mais sofisticadas quanto para análises extensas. Neste livro usaremos preferencialmente o repositório de pacotes estatísticos `R`, obtido livremente em *Comprehensive R Archive Network*, `CRAN`, no sítio

`http://CRAN.R-project.org`.

Dentre os pacotes estatísticos disponíveis na linguagem `R`, aqueles mais utilizados neste texto são: `adabag, caret, cluster, e1071, forecast, ggplot2, MASS, mgcv, nlme, randomForests, xgboost`. As funções de cada pacote necessárias para a realização das análises serão indicadas ao longo do texto.
Pacotes comerciais alternativos incluem SPlus, Minitab, SAS, MatLab etc.

1.8 Conjuntos de dados

Alguns conjuntos de dados analisados são dispostos ao longo do texto; outros são apresentados em formato Excel em arquivos disponíveis no formato

`http://www.ime.usp.br/~jmsinger/MorettinSinger/arquivo.xls`

Por exemplo, no sitio
`http://www.ime.usp.br/~jmsinger/MorettinSinger/coronarias.xls`
encontramos uma planilha com dados de um estudo sobre obstrução coronariana; quando pertinentes, detalhes sobre as variáveis observadas no estudo estarão na aba intitulada"descricao"; os dados estão dispostos na aba intitulada "dados". Conjuntos de dados também poderão ser referidos por meio de seus endereços `URL`. Quando necessário, indicaremos os sítios em que se podem obter os dados utilizados nas análises.

Na Tabela 1.1 listamos os principais conjuntos de dados e uma breve descrição de cada um deles.

Tabela 1.1: Conjuntos de dados para alguns exemplos e exercícios do livro

Rótulo	Descrição
adesivo	Resistencia de adesivos dentários
arvores	Concentração de elementos químicos em cascas de árvores
bezerros	Medida longitudinal de peso de bezerros
ceagfgv	Questionário respondido por 50 alunos da FGV-SP
coronarias	Fatores de risco na doença coronariana
cifose	Dados de crianças submetidas a cirurgia de coluna
disco	Deslocamento do disco temporomandibular
distancia	Distância para distinguir objeto em função da idade
empresa	Dados de funcionários de uma empresa
endometriose	Dados de um estudo sobre endometriose
endometriose2	Dados de um estudo sobre endometriose (1500 pacientes)
entrevista	Comparação intraobservadores em entrevista psicológica
esforco	Respostas de cardíacos em esteira ergométrica
esteira	Medidas obtidas em testes ergométricos (parcial)
figado	Relação entre volume e peso do lobo direito de fígados em transplantes intervivos
figadodiag	Medidas radiológicas e intraoperatórias de alterações anatômicas do fígado
hiv	Dados de sobrevivência de pacientes HIV
inibina	Utilização de inibina como marcador de reserva ovariana
iris	Medidas de comprimento e largura de pétalas e sépalas para três espécies de Iris: setosa, virgínica e versicolor
lactato	Concentração de lactato de sódio em atletas
manchas	Número de manchas solares
morfina	Dados de um estudo sobre concentração de morfina em cabelos
municipios	Populações dos 30 maiores municípios do Brasil
neonatos	Pesos de recém nascidos
palato	Dados de um estudo sobre efeito de peróxido de hidrogênio na em palatos de sapos
placa	Índice de remoção de placa dentária
poluicao	Concentração de poluentes em São Paulo
precipitacao	Precipitação em Fortaleza, CE, Brasil
producao	Dados hipotéticos de produção de uma empresa
profilaxia	pH da placa bacteriana sob efeito de enxaguatório
regioes	Dados populacionais de estados brasileiros
rehabcardio	Dados sobre rehabilitação de pacientes de infartos
rotarod	Tempo com que ratos permanecem em cilindro rotativo
salarios	Salários de profissionais em diferentes países
sondas	Tempos de sobrevivência de pacientes de câncer com diferentes tipos de sondas
suicidios	Frequência de suicídios por enforcamento em São Paulo
temperaturas	Temperaturas mensais em Ubatuba e Cananéia
tipofacial	Classificação de tipos faciais
veiculos	Características de automóveis nacionais e importados
vento	Velocidade do vento no aeroporto de Philadelphia

1.9 Notas do Capítulo

1) Apresentamos, a seguir, a primeira página do artigo de Alan Turing, publicado na revista Mind, em 1950.

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

2) Apresentamos, abaixo, a primeira página do Projeto de IA de Dartmouth, publicado originalmente em 1955, e reproduzido na revista AI Magazine, de 2006.

AI Magazine Volume 27 Number 4 (2006) (© AAAI)

Articles

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky,
Nathaniel Rochester,
and Claude E. Shannon

■ The 1956 Dartmouth summer research project on artificial intelligence was initiated by this August 31, 1955 proposal, authored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The original typescript consisted of 17 pages plus a title page. Copies of the typescript are housed in the archives at Dartmouth College and Stanford University. The first 5 papers state the proposal, and the remaining pages give qualifications and interests of the four who proposed the study. In the interest of brevity, this article reproduces only the proposal itself, along with the short autobiographical statements of the proposers.

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1. Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2. How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalization consists of admitting a new

PARTE I: ANÁLISE EXPLORATÓRIA DE DADOS

A primeira parte deste texto é dedicada à discussão de alguns conceitos básicos, como distribuição de frequências, variabilidade e associação entre variáveis, além de métodos de resumo de dados por meio de tabelas e gráficos. Para efeito didático, discutimos separadamente os casos de uma, duas ou mais que duas variáveis. Consideramos técnicas de regressão, essenciais para o entendimento da associação entre uma ou mais variáveis explicativas e uma variável resposta. Nesse contexto, incluímos análise de sobrevivência, em que a variável resposta é o tempo até a ocorrência de um evento de interesse. Os conceitos e técnicas aqui abordados servem de substrato e são imprescindíveis para a compreensão e aplicação adequada das técnicas estatísticas de análise apresentadas nas Partes II e III.

Capítulo 2

Preparação dos dados

> A tarefa de converter observações em números é a mais difícil de todas, a última e não a primeira coisa a fazer, e pode ser feita somente quando você aprendeu bastante sobre as observações.
>
> Lewis Thomas

2.1 Considerações preliminares

Em praticamente todas as áreas do conhecimento, dados são coletados com o objetivo de obtenção de informação. Esses dados podem representar uma população (como o censo demográfico) ou uma parte (amostra) dessa população (como aqueles oriundos de uma pesquisa eleitoral). Eles podem ser obtidos por meio de estudos observacionais (como aqueles em que se examinam os registros médicos de um determinado hospital), de estudos amostrais (como pesquisas de opinião) ou experimentais (como ensaios clínicos).

Mais comumente, os dados envolvem valores de várias variáveis, obtidos da observação de unidades de investigação que constituem uma amostra de uma população. A análise de dados amostrais possibilita que se faça inferência sobre a distribuição de probabilidades das variáveis de interesse, definidas sobre a população da qual a amostra foi (ao menos conceitualmente) colhida. Nesse contexto, a Estatística é uma ferramenta importante para organizá-los, resumi-los, analisá-los e utilizá-los para tomada de decisões. O ramo da Estatística conhecido como **Análise Exploratória de Dados** se ocupa da organização e resumo dos dados de uma amostra ou, eventualmente, de toda a população e o ramo conhecido como **Inferência Estatística** se refere ao processo de se tirar conclusões sobre uma população com base em uma amostra dela.

A abordagem estatística para o tratamento de dados envolve

i) O planejamento da forma de coleta em função dos objetivos do estudo.
ii) A organização de uma planilha para seu armazenamento eletrônico; no caso de megadados, a organização de um banco de dados (*data warehouse*) pode ser necessária (ver Nota de Capítulo 1).
iii) O seu resumo por meio de tabelas e gráficos.

iv) A identificação e correção de possíveis erros de coleta e/ou digitação.

v) A proposta de modelos probabilísticos baseados na forma de coleta dos dados e nos objetivos do estudo; a finalidade desses modelos é relacionar a amostra (se for o caso) à população para a qual se quer fazer inferência.

vi) A proposta de modelos estruturais para os parâmetros do modelo probabilístico com a finalidade de representar relações entre as características (variáveis) observadas.

vii) A avaliação do ajuste do modelo aos dados por meio de técnicas de diagnóstico e/ou simulação.

viii) A reformulação e reajuste do modelo à luz dos resultados do diagnóstico e/ou simulação.

ix) A tradução dos resultados do ajuste em termos não técnicos.

O item i), por exemplo, pode ser baseado em uma hipótese formulada por um cientista. Numa tentativa de comprovar a sua hipótese, ele identifica as variáveis de interesse e planeja um experimento (preferencialmente com o apoio de um estatístico) para a coleta dos dados que serão armazenados numa planilha. Um dos objetivos deste livro é abordar detalhadamente os itens ii), iii), iv) e viii), que constituem a essência da Estatística Descritiva, com referências eventuais aos itens v), vi), vii), viii) e ix), que formam a base da Inferência Estatística. Esses itens servem de fundamento para as principais técnicas utilizadas em Ciência de Dados, cuja apresentação constitui outro objetivo do texto.

Exemplo 2.1: Se quisermos avaliar a relação entre o consumo (variável C) e renda (variável Y) de indivíduos de uma população, podemos escolher uma amostra[1] de n indivíduos dessa população e medir essas duas variáveis nesses indivíduos, obtendo-se o conjunto de dados $\{(Y_1, C_1), \ldots, (Y_n, C_n)\}$.

Para saber se existe alguma relação entre C e Y podemos construir um gráfico de dispersão, colocando a variável Y no eixo das abscissas e a variável C no eixo das ordenadas. Obteremos uma nuvem de pontos no plano (Y, C), que pode nos dar uma ideia de um **modelo** relacionando Y e C. No Capítulo 4 trataremos da análise de duas variáveis e, no Capítulo 6, estudaremos os chamados modelos de regressão, que são apropriados para o exemplo em questão. Em Economia, sabe-se, desde Keynes, que o gasto com o consumo de pessoas (C) é uma função da renda pessoal disponível (Y), ou seja

$$C = f(Y),$$

para alguma função f.

[1]Em geral, a amostra deve ser obtida segundo alguns critérios que servirão para fundamentar os modelos utilizados na inferência; mesmo nos casos em que esses critérios não são seguidos, as técnicas abordadas neste texto podem ser utilizadas para o entendimento das relações entre as variáveis observadas. No Capítulo 3 definiremos formalmente o que se chama uma amostra aleatória simples retirada de uma população.

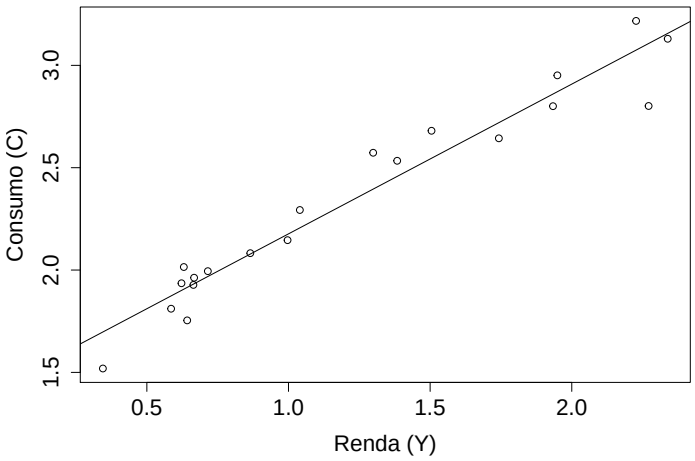


Figura 2.1: Relação entre renda e consumo de 20 indivíduos.

Para se ter uma ideia de como é a função f para essa comunidade, podemos construir um gráfico de dispersão entre Y e C. Com base em um conjunto de dados hipotéticos com $n = 20$, esse gráfico está apresentado na Figura 2.1 e é razoável postular o modelo

$$C_i = \alpha + \beta Y_i + e_i, \quad i = 1, \ldots, n, \tag{2.1}$$

em que $(Y_i, C_i),\ i = 1, \ldots, n$ são os valores de Y e C efetivamente observados e $e_i,\ i = 1, \ldots, n$ são variáveis não observadas, chamadas **erros**. No jargão econômico, o parâmetro α é denominado **consumo autônomo** e β representa a **propensão marginal a consumir**. A reta representada no gráfico foi obtida por meio dos métodos discutidos no Capítulo 6. Nesse caso, obtemos $\alpha = 1,44$ e $\beta = 0,73$, aproximadamente. Para diferentes comunidades (populações) poderemos ter curvas (modelos) diferentes para relacionar Y e C.

Exemplo 2.2: Os dados da Tabela 2.1 foram extraídos de um estudo realizado no Instituto de Ciências Biomédicas da Universidade de São Paulo com o objetivo de avaliar a associação entre a infecção de gestantes por malária e a ocorrência de microcefalia nos respectivos bebês. O dicionário das variáveis observadas está indicado na Tabela 2.1.

Tabela 2.1: Dicionário para as variáveis referentes ao Exemplo 2.2

Rótulos	**Variável**	**Unidade de medida**
id	identificador de paciente	
idade	Idade da mãe	anos
nmal	Quantidade de malárias durante a gestação	número inteiro
parasit	Espécie do parasita da malária	0: não infectada
		1: P. vivax
		2: P. falciparum
		3: malária mista
		4: indeterminado
ngest	Paridade (quantidade de gestações)	Número inteiro
idgest	Idade gestacional no parto	semanas
sexrn	Sexo do recém-nascido	1: masculino
		2: feminino
pesorn	Peso do recém-nascido	g
estrn	Estatura do recém-nascido	cm
pcefal	Perímetro cefálico do recém-nascido	cm
Obs:	Observações omissas são representadas por um ponto	

A disposição dos dados do Exemplo 2.2 no formato de uma planilha está representada na Figura 2.2.[2]

Neste livro estamos interessados na análise de conjuntos de dados, que podem ser provenientes de populações, amostras ou de estudos observacionais. Para essa análise usamos tabelas, gráficos e diversas medidas de posição (localização), variabilidade e associação, com o intuito de resumir e interpretar os dados.

2.2 Planilhas de Dados

Planilhas (usualmente eletrônicas) são matrizes em que se armazenam dados com o objetivo de permitir sua análise estatística. Em geral, cada linha da matriz de dados corresponde a uma unidade de investigação (*e.g.* unidade amostral) e cada coluna, a uma variável. Uma planilha bem elaborada contribui tanto para o entendimento do processo de coleta de dados e especificação das variáveis sob investigação quanto para a proposta de uma análise estatística adequada. A primeira etapa para a construção de uma planilha de dados consiste na elaboração de um dicionário com a especificação das variáveis, que envolve

[2]Note que representamos tabelas e planilhas de forma diferente. Planilhas são representadas no texto como figuras para retratar o formato com que são apresentadas nos *software* mais utilizados, como o **Excel**. Veja a Seção 2.2.

id	idade	nmal	parasit	ngest	idgest	sexrn	pesorn	estrn	pcefal
1	25	0	0	3	38	2	3665	46	36
2	30	0	0	9	37	1	2880	44	33
3	40	0	0	1	41	1	2960	52	35
4	26	0	0	2	40	1	2740	47	34
5	.	0	0	1	38	1	2975	50	33
6	18	0	0	.	38	2	2770	48	33
7	20	0	0	1	41	1	2755	48	34
8	15	0	0	1	39	1	2860	49	32
9	.	0	0	.	42	2	3000	50	35
10	18	0	0	1	40	1	3515	51	34
11	17	0	0	2	40	1	3645	54	35
12	18	1	1	3	40	2	2665	48	35
13	30	0	0	6	40	2	2995	49	33
14	19	0	0	1	40	1	2972	46	34
15	32	0	0	5	41	2	3045	50	35
16	32	0	0	8	38	2	3150	44	35
17	18	0	0	2	40	1	2650	48	33.5
18	18	0	0	1	41	1	3200	50	37
19	19	0	0	1	39	1	3140	48	32
20	18	0	0	2	40	1	3150	47	35
21	27	0	0	3	40	1	4185	52	35.5
22	26	0	0	3	40	2	4070	52	35
23	.	0	0	.	40	1	3950	50	37
24	19	0	0	1	40	1	3245	51	33
25	23	0	0	.	41	1	3010	49	35
26	.	0	0	.	40	2	3260	50	33
27	20	1	1	2	40	2	3450	49	33
28	19	0	0	3	40	2	2765	48	32
29	22	0	0	4	40	1	4190	50	34
30	32	0	0	4	42	2	4035	51	34
31	33	0	0	5	39	2	3620	51	33
32	30	3	3	5	38	1	3230	48	34
33	36	0	0	7	39	2	3185	50	38
34	.	0	0	.	39	2	2950	47	33

Figura 2.2: Planilha com dados referentes ao Exemplo 2.2.

i) sua definição operacional (ver Nota de Capítulo 2);

ii) a atribuição de rótulos (mnemônicos com letras minúsculas e sem acentos para facilitar a digitação e leitura por pacotes computacionais);

iii) a especificação das unidades de medida ou definição de categorias; para variáveis categorizadas, convém atribuir valores numéricos às categorias com a finalidade de facilitar a digitação e evitar erros (veja a variável "Sexo do recém-nascido" na Tabela 2.1);

iv) a atribuição de um código para valores omissos (*missing*);

v) a indicação de como devem ser codificados dados abaixo do limite de detecção (*e.g.*, $< 0,05$ ou $0,025$ se considerarmos que medidas abaixo

do limite de detecção serão definidas como o ponto médio entre $0,00$ e $0,05$);

vi) a especificação do número de casas decimais (correspondente à precisão do instrumento de medida) - ver Notas de Capítulo 3 e 4;

vii) quando possível, indicar limites (inferior ou superior) para facilitar a identificação de erros; por exemplo, o valor mínimo para aplicação num fundo de ações é de R$ 2000,00;

viii) mascarar (por meio de um código de identificação, por exemplo) informações sigilosas ou confidenciais como o nome de pacientes de ensaios clínicos.

Algumas recomendações para a construção da planilha de dados são:

i) não utilizar limitadores de celas (*borders*) ou cores;

ii) reservar a primeira linha para os rótulos das variáveis;

iii) não esquecer uma coluna para a variável indicadora das unidades de investigação (evitar informações confidenciais como nomes de pacientes); essa variável é útil para a correção de erros identificados na análise estatística além de servir como elo de ligação entre planilhas com diferentes informações sobre as unidades de investigação;

iv) escolher ponto ou vírgula para separação de casas decimais[3];

v) especificar o número de casas decimais (ver Nota de Capítulo 3);

vi) formatar as celas correspondentes a datas para manter a especificação uniforme (dd/mm/aaaa ou mm/dd/aaaa, por exemplo).

Exemplo 2.3: Na Tabela 2.2 apresentamos dados provenientes de um estudo em que o objetivo é avaliar a variação do peso (kg) de bezerros submetidos a uma determinada dieta entre 12 e 26 semanas após o nascimento.

Dados com essa natureza são chamados de dados longitudinais por terem a mesma característica (peso, no exemplo) medida ao longo de uma certa dimensão (tempo, no exemplo). De acordo com nossa especificação, há nove variáveis na Tabela 2.2, nomeadamente, Animal, Peso na 12a semana, Peso na 14a semana etc. Para efeito computacional, no entanto, esse tipo de dados deve ser disposto numa planilha com formato diferente (às vezes chamado de **formato longo**) como indicado na Figura 2.3. Nesse formato apropriado para dados longitudinais (ou mais geralmente, para medidas repetidas), há apenas três variáveis, a saber, Animal, Semana e Peso. Note que a mesma unidade amostral (animal) é repetida na primeira coluna para caracterizar a natureza longitudinal dos dados. Ele é especialmente adequado para casos em que as unidades de investigação são avaliadas em instantes diferentes.

Na Figura 2.4 apresentamos um exemplo em que o diâmetro da aorta (mm) de recém nascidos pré termo, com peso adequado (AIG) ou pequeno

[3]Embora a norma brasileira ABNT indique a vírgula para separação de casas decimais, a maioria dos pacotes computacionais utiliza o ponto com essa função; por essa razão é preciso tomar cuidado com esse detalhe na construção de planilhas a serem analisadas computacionalmente. Em geral adotaremos a norma brasileira neste texto.

Tabela 2.2: Peso de bezerros (kg)

	Semanas após nascimento							
animal	12	14	16	18	20	22	24	26
1	54.1	65.4	75.1	87.9	98.0	108.7	124.2	131.3
2	91.7	104.0	119.2	133.1	145.4	156.5	167.2	176.8
3	64.2	81.0	91.5	106.9	117.1	127.7	144.2	154.9
4	70.3	80.0	90.0	102.6	101.2	120.4	130.9	137.1
5	68.3	77.2	84.2	96.2	104.1	114.0	123.0	132.0
6	43.9	48.1	58.3	68.6	78.5	86.8	99.9	106.2
7	87.4	95.4	110.5	122.5	127.0	136.3	144.8	151.5
8	74.5	86.8	94.4	103.6	110.7	120.0	126.7	132.2
9	50.5	55.0	59.1	68.9	78.2	75.1	79.0	77.0
10	91.0	95.5	109.8	124.9	135.9	148.0	154.5	167.6
11	83.3	89.7	99.7	110.0	120.8	135.1	141.5	157.0
12	76.3	80.8	94.2	102.6	111.0	115.6	121.4	134.5
13	55.9	61.1	67.7	80.9	93.0	100.1	103.2	108.0
14	76.1	81.1	84.6	89.8	97.4	111.0	120.2	134.2
15	56.6	63.7	70.1	74.4	85.1	90.2	96.1	103.6

animal	semana	peso
1	12	54.1
1	14	65.4
⋮	⋮	⋮
1	24	124.2
1	26	131.3
2	12	91.7
2	14	104.0
⋮	⋮	⋮
2	26	176.8
⋮	⋮	⋮
15	12	56.6
⋮	⋮	⋮
15	26	103.6

Figura 2.3: Planilha computacionalmente adequada para os dados do Exemplo 2.3.

(PIG) para a idade gestacional foi avaliado até a 40a semana pós concepção. Note que o número de observações pode ser diferente para as diferentes unidades de investigação. Esse formato também é comumente utilizado para armazenar dados de **séries temporais**.

grupo	ident	sem	diam
AIG	2	30	7.7
AIG	2	31	8.0
⋮	⋮	⋮	⋮
AIG	2	36	9.8
AIG	12	28	7.1
AIG	12	29	7.1
⋮	⋮	⋮	⋮
AIG	12	30	9.4
⋮	⋮	⋮	⋮
PIG	17	33	7.5
PIG	17	34	7.7
PIG	17	36	8.2
PIG	29	26	6.3
PIG	29	27	6.5
⋮	⋮	⋮	⋮
PIG	29	31	7.2
PIG	29	32	7.2

Figura 2.4: Planilha com diâmetro da aorta (mm) observado em recém nascidos pré termo.

2.3 Construção de tabelas

A finalidade primordial de uma tabela é resumir a informação obtida dos dados. Sua construção deve permitir que o leitor entenda esse resumo sem a necessidade de recorrer ao texto. Algumas sugestões para construção de tabelas estão apresentadas a seguir.

1) Não utilize mais casas decimais do que o necessário para não mascarar as comparações de interesse. A escolha do número de casas decimais depende da precisão do instrumento de medida e/ou da importância prática dos valores representados. Para descrever a redução de peso após um mês de dieta, por exemplo, é mais conveniente representá-lo como 6 kg do que como 6,200 kg. Por outro lado, quanto mais casas decimais forem incluídas, mais difícil é a comparação. Por exemplo, compare a Tabela 2.3 com a Tabela 2.4.

   Observe que calculamos porcentagens em relação ao total de cada linha. Poderíamos, também, ter calculado porcentagens em relação ao total de cada coluna ou porcentagens em relação ao total geral (50). Cada uma dessas maneiras pode ser útil em determinada situação; por exemplo, determinar se há alguma dependência entre as duas variáveis estado civil e bebida preferida, avaliada em 50 indivíduos.

Tabela 2.3: Número de alunos

Estado civil	Bebida preferida			Total
	não alcoólica	cerveja	outra alcoólica	
Solteiro	19 (53%)	7 (19%)	10 (28%)	36 (100%)
Casado	3 (25%)	4 (33%)	5 (42%)	12 (100%)
Outros	1 (50%)	0 (0%)	1 (50%)	2 (100%)
Total	23 (46%)	11 (22%)	16 (32%)	50 (100%)

Tabela 2.4: Número de alunos (e porcentagens com duas casas decimais)

Estado civil	Bebida preferida			Total
	não alcoólica	cerveja	outra alcoólica	
Solteiro	19 (52,78%)	7 (19,44%)	10 (27,78%)	36 (100,00%)
Casado	3 (25,00%)	4 (33,33%)	5 (41,67%)	12 (100,00%)
Outros	1 (50,00%)	0 (0,00%)	1 (50,00%)	2 (100,00%)
Total	23 (46,00%)	11 (22,00%)	16 (32,00%)	50 (100,00%)

2) Proponha um título autoexplicativo e inclua as unidades de medida. O título deve dizer o que representam os números do corpo da tabela e, em geral, não deve conter informações que possam ser obtidas diretamente dos rótulos de linhas e colunas. Compare o título da Tabela 2.5 com: Intenção de voto (%) por candidato para diferentes meses.

3) Inclua totais de linhas e/ou colunas para facilitar as comparações. É sempre bom ter um padrão contra o qual os dados possam ser avaliados.

4) Não utilize abreviaturas ou indique o seu significado no rodapé da tabela (*e.g.* Desvio padrão em vez de DP); se precisar utilize duas linhas para indicar os valores da coluna correspondente.

5) Ordene colunas e/ou linhas quando possível. Se não houver impedimentos, ordene-as segundo os valores, crescente ou decrescentemente. Compare a Tabela 2.5 com a Tabela 2.6.

Tabela 2.5: Intenção de voto (%)

Candidato	janeiro	fevereiro	março	abril
Nononono	39	41	40	38
Nananana	20	18	21	24
Nenenene	8	15	18	22

Tabela 2.6: Intenção de voto (%)

Candidato	janeiro	fevereiro	março	abril
Nananana	20	18	21	24
Nononono	39	41	40	38
Nenenene	8	15	18	22

6) Tente trocar a orientação de linhas e colunas para melhorar a apresentação. Em geral, é mais fácil fazer comparações ao longo das linhas do que das colunas.

7) Altere a disposição e o espaçamento das linhas e colunas para facilitar a leitura. Inclua um maior espaçamento a cada grupo de linhas e/ou colunas em tabelas muito extensas.

8) Não analise a tabela descrevendo-a, mas sim comentando as principais tendências sugeridas pelos dados. Por exemplo, os dados apresentados na Tabela 2.3 indicam que a preferência por bebidas alcoólicas é maior entre os alunos casados do que entre os solteiros; além disso, há indicações de que a cerveja é menos preferida que outras bebidas alcoólicas, tanto entre solteiros quanto entre casados.

2.4 Construção de gráficos

A seguir apresentamos algumas sugestões para a construção de gráficos, cuja finalidade é similar àquela de tabelas, ou seja, resumir a informação obtida dos dados; por esse motivo, convém optar pelo resumo em forma de tabela ou de gráfico.

1) Proponha um título autoexplicativo.
2) Escolha o tipo de gráfico apropriado para os dados.
3) Rotule os eixos apropriadamente, incluindo unidades de medida.
4) Procure escolher adequadamente as escalas dos eixos para não distorcer a informação que se pretende transmitir. Se o objetivo for comparar as informações de dois ou mais gráficos, use a mesma escala.
5) Inclua indicações de "quebra" nos eixos para mostrar que a origem (zero) está deslocada.
6) Altere as dimensões do gráfico até encontrar o formato adequado.
7) Inclua uma legenda.
8) Tome cuidado com a utilização de áreas para comparações, pois elas variam com o quadrado das dimensões lineares.
9) Não exagere nas ilustrações que acompanham o gráfico para não o "poluir" visualmente, mascarando seus aspectos mais relevantes.

2.5 Notas de capítulo

1) **Bancos de dados**

Projetos que envolvem grandes quantidades de dados, em geral provenientes de diversas fontes e com diversos formatos necessitam a construção de bancos de dados (*data warehouses*), cuja finalidade é prover espaço suficiente para sua armazenagem, garantir sua segurança, permitir a inclusão por meio de diferentes meios e proporcionar uma interface que permita a recuperação da informação de forma estruturada para uso por diferentes pacotes de análise estatística.

Bancos de dados têm se tornado cada vez maiores e mais difíceis de administrar em função da crescente disponibilidade de sistemas analíticos em que os dados são oriundos de diferentes sistemas de transações. Em geral, esses bancos de dados envolvem a participação de profissionais de áreas e instituições diversas. Por esse motivo, os resultados de sua implantação são lentos e às vezes inexistentes. Conforme pesquisa elaborada pelo Grupo Gartner (2005), 50% dos projetos de bancos de dados tendem a falhar por problemas em sua construção. Uma das causas para esse problema, é o longo tempo necessário para o seu desenvolvimento, o que gera uma defasagem na sua operacionalidade. Muitas vezes, algumas de suas funcionalidades ficam logo obsoletas enquanto novas demandas estão sendo requisitadas. Duas razões para isso são a falta de sincronização entre os potenciais usuários e os desenvolvedores do banco de dados e o fato de que técnicas tradicionais usadas nesse desenvolvimento não permitem a rápida disponibilidade de suas funções. Para contornar esses problemas, sugere-se uma arquitetura modular cíclica em que o foco inicial é a criação dos principais elementos do sistema, deixando os detalhes das características menos importantes para uma segunda fase em que o sistema se torna operacional. No Módulo 1, são projetados os sistemas para inclusão e armazenagem dos dados provenientes de diferentes fontes. A detecção, correção de possíveis erros e homogeneização dos dados é realizada no Módulo 2. Como esperado, dados obtidos por diferentes componentes do projeto geralmente têm codificação distinta para os mesmos atributos, o que requer uniformização e possível indicação de incongruências que não podem ser corrigidas. No Módulo 3, os dados tratados no módulo anterior são atualizados e inseridos numa base de dados históricos, devidamente padronizada. Nesse módulo, a integridade dos dados recém obtidos é avaliada comparativamente aos dados já existentes para garantir a consistência entre eles. O foco do Módulo 4 é a visualização, análise e exportação dos dados. Esse módulo contém as ferramentas que permitem a geração de planilhas de dados apropriadas para a análise estatística.

Detalhes sobre a construção de bancos de dados podem ser encontrados em Rainardi (2008), entre outros. Para avaliar das dificuldades

de construção de um banco de dados num projeto complexo, o leitor poderá consultar Ferreira et al. (2017).

2) **Definição operacional de variáveis**

Para efeito de comparação entre estudos, a definição das variáveis envolvidas requer um cuidado especial. Por exemplo em estudos cujo objetivo é avaliar a associação entre renda e gastos com lazer, é preciso especificar se a variável "Renda" se refere à renda familiar total ou *per capita*, se benefícios como vale transporte, vale alimentação ou bônus estão incluídos etc. Num estudo que envolva a variável "Pressão arterial", um exemplo de definição operacional é: "média de 60 medidas com intervalo de 1 minuto da pressão arterial diastólica (mmHg) obtida no membro superior direito apoiado à altura do peito com aparelho automático de método oscilométrico (Dixtal, São Paulo, Brasil)". Num estudo cujo objetivo é comparar diferentes modelos de automóveis com relação ao consumo de combustível, uma definição dessa variável poderia ser "número de quilômetros percorridos em superfície plana durante 15 minutos em velocidade constante de 50 km/h e sem carga por litro de gasolina comum (km/L)."

Neste texto, não consideraremos definições detalhadas por razões didáticas.

3) **Ordem de grandeza, precisão e arredondamento de dados quantitativos**

A precisão de dados quantitativos contínuos está relacionada com a capacidade de os instrumentos de medida distinguirem entre valores próximos na escala de observação do atributo de interesse. O número de dígitos colocados após a vírgula indica a precisão associada à medida que estamos considerando. O volume de um certo recipiente expresso como 0,740 L implica que o instrumento de medida pode detectar diferenças da ordem de 0,001 L (= 1 mL, ou seja 1 mililitro); se esse volume for expresso na forma 0,74 L, a precisão correspondente será de 0,01 L (= 1 cL, ou seja 1 centilitro).

Muitas vezes, em função dos objetivos do estudo em questão, a expressão de uma grandeza quantitativa pode não corresponder à precisão dos instrumentos de medida. Embora com uma balança suficientemente precisa, seja possível dizer que o peso de uma pessoa é de 89,230 kg, para avaliar o efeito de uma dieta, o que interessa saber é a ordem de grandeza da perda de peso após três meses de regime, por exemplo. Nesse caso, saber se a perda de peso foi de 10,230 kg ou de 10,245 kg é totalmente irrelevante. Para efeitos práticos, basta dizer que a perda foi da ordem de 10 kg. A ausência de casas decimais nessa representação indica que o próximo valor na escala de interesse seria 11 kg, embora todos os valores intermediários com unidades de 1 g sejam mensuráveis.

Para efeitos contábeis, por exemplo, convém expressar o aumento das exportações brasileiras num determinado período como R$ 1 657 235 458,29; no entanto, para efeitos de comparação com outros períodos, é mais conveniente dizer que o aumento das exportações foi da ordem de 1,7 bilhões de reais. Note que nesse caso, as grandezas significativas são aquelas da ordem de 0,1 bilhão de reais (= 100 milhões de reais).

Nesse processo de transformação de valores expressos com uma determinada precisão para outros com a precisão de interesse é preciso arredondar os números correspondentes. Em termos gerais, se o dígito a ser eliminado for 0, 1, 2, 3 ou 4, o dígito precedente não deve sofrer alterações e se o dígito a ser eliminado for 5, 6, 7, 8 ou 9, o dígito precedente deve ser acrescido de uma unidade. Por exemplo, se desejarmos reduzir para duas casas decimais números originalmente expressos com três casas decimais, 0,263 deve ser transformado para 0,26 e 0,267 para 0,27. Se desejarmos uma redução mais drástica para apenas uma casa decimal, tanto 0,263 quanto 0,267 devem ser transformados para 0,3.

É preciso tomar cuidado com essas transformações quando elas são aplicadas a conjuntos de números cuja soma seja prefixada (porcentagens, por exemplo) pois elas podem introduzir erros cumulativos. Discutiremos esse problema ao tratar de porcentagens e tabulação de dados. É interessante lembrar que a representação decimal utilizada nos EUA e nos países da comunidade britânica substitui a vírgula por um ponto. Cuidados devem ser tomados ao se fazerem traduções, embora em alguns casos, esse tipo de representação já tenha sido adotada no cotidiano (veículos com motor 2.0, por exemplo, são veículos cujo volume dos cilindros é de 2,0 L).

Finalmente, convém mencionar que embora seja conveniente apresentar os resultados de uma análise com o número de casas decimais conveniente, os cálculos necessários para sua obtenção devem ser realizados com maior precisão para evitar propagação de erros. O arredondamento deve ser concretizado ao final dos cálculos.

4) **Proporções e porcentagens**

Uma proporção é um quociente utilizado para comparar duas grandezas através da adoção de um padrão comum. Se 31 indivíduos, num total de 138, são fumantes, dizemos que a proporção de fumantes entre esses 138 indivíduos é de 0,22 (= 31/138). O denominador desse quociente é chamado de base e a interpretação associada à proporção é que 31 está para a base 138 assim como 0,22 está para a base 1,00. Essa redução a uma base fixa permite a comparação com outras situações em que os totais são diferentes. Consideremos, por exemplo, um outro conjunto de 77 indivíduos em que 20 são fumantes; embora o número de fumantes não seja comparável com o do primeiro grupo, dado que as bases são diferentes, pode-se dizer que a proporção de fumantes desse segundo grupo, 0,26 (= 20/77) é maior que aquela associada ao

primeiro conjunto.

Porcentagens, nada mais são do que proporções multiplicadas por 100, o que equivale a fazer a base comum igual a 100. No exemplo acima, podemos dizer que a porcentagem de fumantes é de 22% (=100 x 31/138) no primeiro grupo e de 26% no segundo. Para efeito da escolha do número de casas decimais, note que a comparação entre essas duas porcentagens é mais direta do que se considerássemos suas expressões mais precisas (com duas casas decimais), ou seja 22,46% contra 25,97%.

A utilização de porcentagens pode gerar problemas de interpretação em algumas situações. A seguir consideramos algumas delas. Se o valor do IPTU de um determinado imóvel cobrado foi de R$ 500,00 em 1998 e de R$ 700,00 em 1999, podemos dizer que o valor do IPTU em 1999 é 140% (= 100 x 700/500) do valor em 1998, mas o aumento foi de 40% [= 100 x (700-500)/500]. Se o preço de uma determinada ação varia de R$ 22,00 num determinado instante para R$ 550,00 um ano depois, podemos dizer que o aumento de seu preço foi de 2400% [=100x(550-22)/22] nesse período. É difícil interpretar porcentagens "grandes" como essa. Nesse caso é melhor dizer que o preço dessa ação é 25 (= 550/22) vezes seu preço há um ano. Porcentagens calculadas a partir de bases de pequena magnitude podem induzir conclusões inadequadas. Dizer que 43% dos participantes de uma pesquisa preferem um determinado produto tem uma conotação diferente se o cálculo for baseado em 7 ou em 120 entrevistados. É sempre conveniente explicitar a base relativamente à qual se estão fazendo os cálculos.

Para se calcular uma porcentagem global a partir das porcentagens associadas às partes de uma população, é preciso levar em conta sua composição. Suponhamos que numa determinada faculdade, 90% dos alunos que usam transporte coletivo sejam favoráveis à cobrança de estacionamento no campus e que apenas 20% dos alunos que usam transporte individual o sejam. A porcentagem de alunos dessa faculdade favoráveis à cobrança do estacionamento só será igual à média aritmética dessas duas porcentagens, ou seja 55%, se a composição da população de alunos for tal que metade usa transporte coletivo e metade não. Se essa composição for de 70% e 30% respectivamente, a porcentagem de alunos favoráveis à cobrança de estacionamento será de 69% (= 0,9x70% + 0,20x30% ou seja, 90% dos 70% que usam transporte coletivo + 20% dos 30% que utilizam transporte individual). Para evitar confusão, ao se fazer referência a variações, convém distinguir porcentagem e ponto percentual. Se a porcentagem de eleitores favoráveis a um determinado candidato aumentou de 14% antes para 21% depois da propaganda na televisão, pode-se dizer que a preferência eleitoral por esse candidato aumentou 50% [= 100 x (21-14)/14] ou foi de 7 pontos percentuais (e não de 7%). Note que o que diferencia esses dois enfoques é a base em relação à qual se calculam as porcentagens;

no primeiro caso, essa base é a porcentagem de eleitores favoráveis ao candidato antes da propaganda (14%) e no segundo caso é o total (não especificado) de eleitores avaliados na amostra (favoráveis ou não ao candidato).

Uma porcentagem não pode diminuir mais do que 100%. Se o preço de um determinado produto decresce de R$ 3,60 para R$ 1,20, a diminuição de preço é de 67% [= 100 x (3,60 - 1,20)/3,60] e não de 200% [= 100 x (3,60 - 1,20)/1,20]. Aqui também, o importante é definir a base: a ideia é comparar a variação de preço (R$ 2,40) com o preço inicial do produto (R$ 3,60) e não com o preço final (R$ 1,20). Na situação limite, em que o produto é oferecido gratuitamente, a variação de preço é de R$ 3,60; consequentemente, a diminuição de preço limite é de 100%. Note que se estivéssemos diante de um aumento de preço de R$ 1,20 para R$ 3,60, diríamos que o aumento foi de 200% [= 100 x (3,60 - 1,20)/1,20)].

2.6 Exercícios

1) O objetivo de um estudo da Faculdade de Medicina da USP foi avaliar a associação entre a quantidade de morfina administrada a pacientes com dores intensas provenientes de lesões medulares ou radiculares e a dosagem dessa substância em seus cabelos. Três medidas foram realizadas em cada paciente, a primeira logo após o início do tratamento e as demais após 30 e 60 dias. Detalhes podem ser obtidos no documento intitulado "morfina.doc", disponível no sítio

`http://www.ime.usp.br/~jmsinger/MorettinSinger/morfina.doc`.

A planilha morfina.xls, disponível no sítio

`http://www.ime.usp.br/~jmsinger/MorettinSinger/morfina.xls`,

foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo. Organize-a, construindo tabelas apropriadas para descrever essas características demográficas e uma planilha num formato apropriado para análise estatística.

2) A Figura 2.5 foi extraída de um relatório do Centro de Estatística Aplicada do IME/USP [ver Giampaoli et al. (2008) para detalhes]. Critique-a e reformule-a para facilitar sua leitura.

3) Utilize as sugestões para construção de planilhas apresentadas na Seção 2.2 com a finalidade de preparar os dados dos diferentes conjuntos do arquivo MorettinSingerDados.xls para análise estatística.

Sheet1

Subprefeitura	Observado	Ajustado	Nota SP
Aricanduva	0,586	0,588	0,396
Butantã	0,483	0,468	0,334
Campo Limpo	0,484	0,526	0,362
Casa Verde/Cachoeirinha	0,558	0,554	0,382
Cidade Tiradentes	0,543	0,540	0,369
Freguesia/Brasilândia	0,545	0,540	0,371
Ipiranga	0,593	0,539	0,368
Itaim Paulista	0,566	0,557	0,374
Itaquera	0,396	0,563	0,383
Jabaquara	0,533	0,533	0,364
M' Boi Mirim	0,566	0,552	0,368
São Mateus	0,523	0,511	0,354
São Miguel	0,601	0,583	0,395
Socorro	0,601	0,523	0,360
V, Prudente/Sapopemba	0,648	0,620	0,413

Figura 2.5: Tabela comparativa das notas médias da avaliação das subprefeituras no modelo padronizado para a escala [0,1]

4) A Figura 2.6 contém uma planilha encaminhada pelos investigadores responsáveis por um estudo sobre AIDS para análise estatística. Organize-a de forma a permitir sua análise por meio de um pacote computacional como o R.

5) Num estudo planejado para avaliar o consumo médio de combustível de veículos em diferentes velocidades foram utilizados 4 automóveis da marca A e 3 automóveis da marca B selecionados ao acaso das respectivas linhas de produção. O consumo (em L/km) de cada um dos 7 automóveis foi observado em 3 velocidades diferentes (40 km/h, 80 km/h e 110 km/h). Construa uma planilha apropriada para a coleta e análise estatística dos dados, rotulando-a adequadamente.

6) A planilha apresentada na Figura 2.7 contém dados de um estudo em que o limiar auditivo foi avaliado nas orelhas direita (OD) e esquerda (OE) de 13 pacientes em 3 ocasiões (Limiar, Teste 1 e Teste 2). Reformate-a segundo as recomendações da Seção 2.3.

7) Preencha a ficha de inscrição do Centro de Estatística Aplicada (`www.ime.usp.br/~cea`) com as informações de um estudo em que você está envolvido.

Grupo I	**Tempo de**			**Ganho de Peso**
registro	**Diagnóstico**	**DST**	**MAC**	**por Semana**
2847111D	pré natal	não	Pílula	11Kg em 37 semanas
3034048F	6 meses	não	pílula	?
3244701J	1 ano	não	Condon	?
2943791B	pré natal	não	não	8 Kg em 39 semanas
3000327F	4 anos	condiloma/ sífilis	não	9Kg em 39 semanas
3232893D	1 ano	não	DIU	3Kg em 39 semanas
3028772E	3 anos	não	não	3 kg em 38 semanas
3240047G	pré natal	não	pílula	9 Kg em 38 semanas
3017222G		HPV	CONDON	falta exame clínico
3015834J	2 anos	não	condon	14 Kg em 40 semanas
Grupo II	**Tempo de**			**Ganho de Peso**
registro	**Diagnóstico**	**DST**	**MAC**	**por Semana**
3173611E	3 meses	abscesso ovariano	condon	15 Kg em 40 semanas
3296159D	pré natal	não	condon	0 Kg em ? semanas
3147820D1	2 anos	não	sem dados	4 Kg em 37 semanas
3274750K	3 anos	não	condon	8 Kg em 38 semanas
3274447H	pré natal	sifílis com 3 meses	condon	
2960066D	5 anos	não	?	13 Kg em 36 semanas
3235727J	7 anos	não	Condon	(-) 2 Kg em 38 semanas
3264897E		condiloma	condon	nenhum Kg
3044120J	5 anos	HPV		3 Kg em 39 semanas 1

Figura 2.6: Planilha com dados de um estudo sobre AIDS.

Limiar	Teste1	Teste2
OD 50 / OE 55	OD/OE 50	OD/OE 80%
OD 41 /OE 40	OD 45/OE 50	OD 68% OE 80%
OD/OE 41,25	OD/OE 45	OD 64% OE 72%
OD 45/OE 43,75	OD 60/OE 50	OD 76%/OE 88%
OD51,25/ OE47,5	OD/OE 50	OD 80%/OE 88%
OD45/ OE 52,5	OD/OE 50	OD 84%/OE 96%
OD 52,5/OE 50	OD55/OE45	OD 40%/OE 28%
OD 42,15/0E48,75	OD 40/OE 50	OD80%/OE76%
OD50/ OE 48,75	OD/OE 50	OD 72%/OE 80%
OD47,5/OE46,25	OD/OE 50	OD/OE 84%
OD55/OE 56,25	OD55/OE60	OD80%/OE 84%
OD/OE 46,25	OD40/OE35	OD72%/OE 84%
OD 50/OE 47,5	OD/OE45	OD/OE 76%

Figura 2.7: Limiar auditivo de pacientes observados em 3 ocasiões.

Capítulo 3

Análise de dados de uma variável

> Você pode, certamente, ter um entendimento profundo da natureza por meio de medidas quantitativas, mas você deve saber do que está falando antes que comece a usar os números para fazer previsões.
>
> Lewis Thomas

3.1 Introdução

Neste capítulo consideraremos a análise descritiva de dados provenientes da observação de uma variável. As técnicas utilizadas podem ser empregadas tanto para dados provenientes de uma população quanto para dados oriundos de uma amostra.

A ideia de uma análise descritiva de dados é tentar responder as seguintes questões:

i) qual a frequência com que cada valor (ou intervalo de valores) aparece no conjunto de dados ou seja, qual a distribuição de frequências dos dados?
ii) quais são alguns valores típicos do conjunto de dados, como mínimo e máximo?
iii) qual seria um valor para representar a posição (ou localização) central do conjunto de dados?
iv) qual seria uma medida da variabilidade ou dispersão dos dados?
v) existem valores atípicos ou discrepantes (*outliers*) no conjunto de dados?
vi) os dados podem ser considerados simétricos?

Nesse contexto, um dos objetivos da análise descritiva é organizar e exibir os dados de maneira apropriada e para isso utilizamos

i) gráficos e tabelas;

ii) medidas para resumo de dados.

As técnicas empregadas na análise descritiva dependem do tipo de variáveis que compõem o conjunto de dados em questão. Uma possível classificação de variáveis está representada na Figura 3.1.

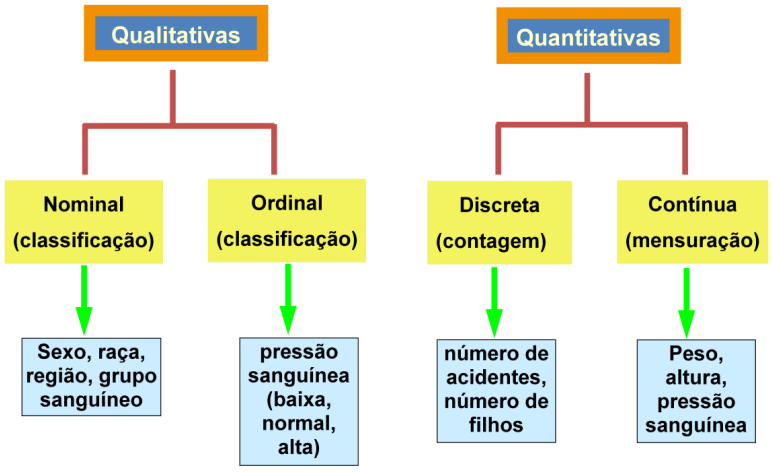


Figura 3.1: Classificação de variáveis.

Variáveis qualitativas são aquelas que indicam um atributo (não numérico) da unidade de investigação (sexo, por exemplo). Elas podem ser ordinais, quando há uma ordem nas diferentes categorias do atributo (tamanho de uma escola: pequena, média ou grande, por exemplo) ou nominais, quando não há essa ordem (região em que está localizada uma empresa: norte, sul, leste ou oeste, por exemplo).

Variáveis quantitativas são aquelas que exibem valores numéricos associados à unidade de investigação (peso, por exemplo). Elas podem ser discretas, quando assumem valores no conjunto dos números naturais (número de gestações de uma paciente) ou contínuas, quando assumem valores no conjunto dos números reais (tempo gasto por um atleta para percorrer 100 m, por exemplo). Ver Nota de Capítulo 1.

3.2 Distribuições de frequências

Exemplo 3.1: Consideremos um conjunto de dados como aquele apresentado na Tabela 3.1 obtido de um questionário respondido por 50 alunos de uma disciplina ministrado na Fundação Getúlio Vargas em São Paulo. Os dados estão disponíveis no arquivo `ceagfgv`.

Em geral, a primeira tarefa de uma análise estatística de um conjunto de dados consiste em resumi-los. As técnicas disponíveis para essa finalidade dependem do tipo de variáveis envolvidas, tema que discutiremos a seguir.

Tabela 3.1: Dados de um estudo realizado na FGV

ident	Salário (R$)	Fluência inglês	Anos de formado	Estado civil	Número de filhos	Bebida preferida
1	3500	fluente	12.0	casado	1	outra alcoólica
2	1800	nenhum	2.0	casado	3	não alcoólica
3	4000	fluente	5.0	casado	1	outra alcoólica
4	4000	fluente	7.0	casado	3	outra alcoólica
5	2500	nenhum	11.0	casado	2	não alcoólica
6	2000	fluente	1.0	solteiro	0	não alcoólica
7	4100	fluente	4.0	solteiro	0	não alcoólica
8	4250	algum	10.0	casado	2	cerveja
9	2000	algum	1.0	solteiro	2	cerveja
10	2400	algum	1.0	solteiro	0	não alcoólica
11	7000	algum	15.0	casado	1	não alcoólica
12	2500	algum	1.0	outros	2	não alcoólica
13	2800	fluente	2.0	solteiro	1	não alcoólica
14	1800	algum	1.0	solteiro	0	não alcoólica
15	3700	algum	10.0	casado	4	cerveja
16	1600	fluente	1.0	solteiro	2	cerveja
⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	1000	algum	1.0	solteiro	1	outra alcoólica
27	2000	algum	5.0	solteiro	0	outra alcoólica
28	1900	fluente	2.0	solteiro	0	outra alcoólica
29	2600	algum	1.0	solteiro	0	não alcoólica
30	3200		6.0	casado	3	cerveja
31	1800	algum	1.0	solteiro	2	outra alcoólica
32	3500		7.0	solteiro	1	cerveja
33	1600	algum	1.0	solteiro	0	não alcoólica
34	1700	algum	4.0	solteiro	0	não alcoólica
35	2000	fluente	1.0	solteiro	2	não alcoólica
36	3200	algum	3.0	solteiro	2	outra alcoólica
37	2500	fluente	2.0	solteiro	2	outra alcoólica
38	7000	fluente	10.0	solteiro	1	não alcoólica
39	2500	algum	5.0	solteiro	1	não alcoólica
40	2200	algum	0.0	casado	0	cerveja
41	1500	algum	0.0	solteiro	0	não alcoólica
42	800	algum	1.0	solteiro	0	não alcoólica
43	2000	fluente	1.0	solteiro	0	não alcoólica
44	1650	fluente	1.0	solteiro	0	não alcoólica
45		algum	1.0	solteiro	0	outra alcoólica
46	3000	algum	7.0	solteiro	0	cerveja
47	2950	fluente	5.5	outros	1	outra alcoólica
48	1200	algum	1.0	solteiro	0	não alcoólica
49	6000	algum	9.0	casado	2	outra alcoólica
50	4000	fluente	11.0	casado	3	outra alcoólica

3.2.1 Variáveis qualitativas

Uma tabela contendo as frequências (absolutas e/ou relativas) de unidades de investigação para cada categoria do atributo avaliado por uma variável qualitativa é chamada de distribuição de frequências dessa variável. As Tabelas 3.2 e 3.3, por exemplo, representam respectivamente as distribuições de frequências das variáveis "Bebida preferida" e "Fluência em inglês" para os dados do Exemplo 3.1.

Tabela 3.2: Distribuição de frequências para a variável Bebida preferida correspondente ao Exemplo 3.1

Bebida preferida	Frequência observada	Frequência relativa (%)
não alcoólica	23	46
cerveja	11	22
outra alcoólica	16	32
Total	50	100

Tabela 3.3: Distribuição de frequências para a variável Fluência em inglês correspondente ao Exemplo 3.1

Fluência em inglês	Frequência observada	Frequência relativa (%)	Frequência acumulada (%)
nenhuma	2	4	4
alguma	26	54	58
fluente	20	42	100
Total	48	100	

Obs: dois participantes não forneceram informação.

Note que para variáveis variáveis qualitativas ordinais pode-se acrescentar uma coluna com as frequências relativas acumuladas que são úteis na sua análise. Por exemplo a partir da última coluna da Tabela 3.3 pode-se afirmar que cerca de 60% dos alunos que forneceram a informação tem no máximo alguma fluência em inglês.

O resumo exibido nas tabelas com distribuições de frequências pode ser representado por meio de gráficos de barras ou gráficos do tipo pizza (ou torta). Exemplos correspondentes às variáveis "Bebida preferida" e "Fluência em inglês" são apresentados nas Figuras 3.2, 3.3 e 3.4.

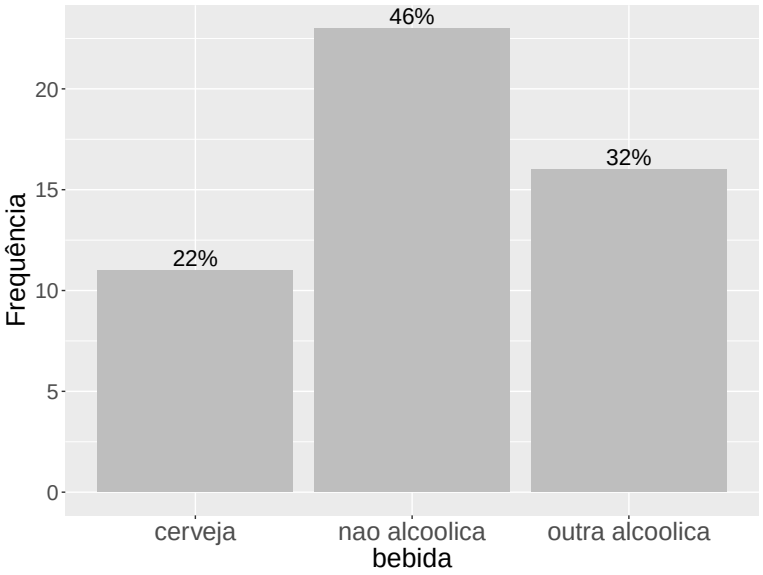


Figura 3.2: Gráfico de barras para Bebida preferida.

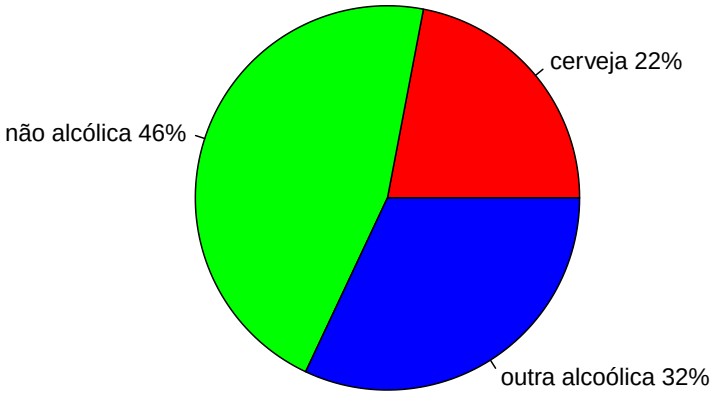


Figura 3.3: Gráfico tipo pizza para Bebida preferida.

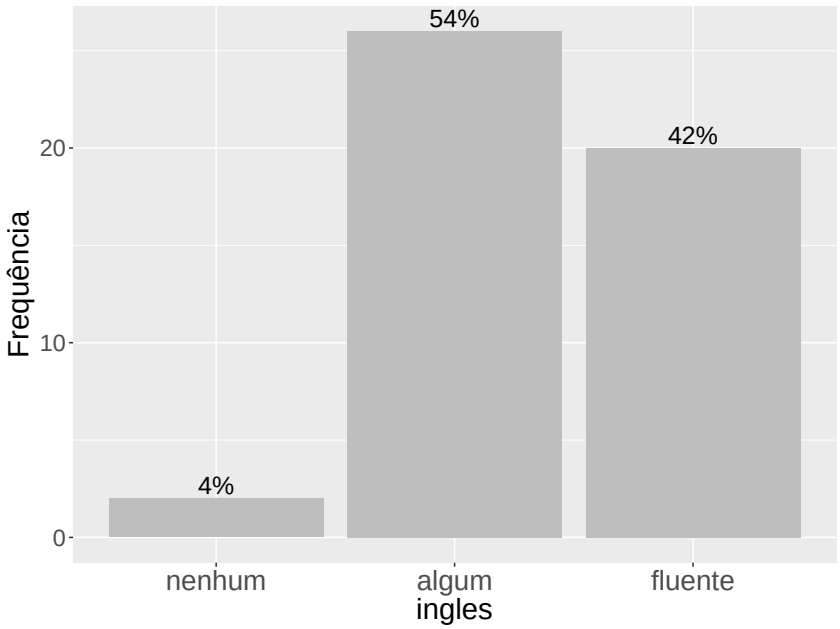


Figura 3.4: Gráfico de barras para Fluência em inglês.

Note que na Figura 3.2 as barras podem ser colocadas em posições arbitrárias; na Figura 3.4, convém colocá-las de acordo com a ordem natural das categorias.

3.2.2 Variáveis quantitativas

Se utilizássemos o mesmo critério adotado para variáveis qualitativas na construção de distribuições de frequências de variáveis quantitativas (especialmente no caso de variáveis contínuas), em geral obteríamos tabelas com frequência muito pequena (em geral 1) para as diversas categorias, deixando de atingir o objetivo de resumir os dados. Para contornar o problema, agrupam-se os valores das variáveis em classes e obtêm-se as frequências em cada classe.

Uma possível distribuição de frequências para a variável "Salário" correspondente ao Exemplo 3.1 está apresentada na Tabela 3.4.

Alternativamente a tabelas com o formato da Tabela 3.4 vários gráficos podem ser utilizados para representar a distribuição de frequências de um conjunto de dados. Os mais utilizados são apresentados a seguir.

Gráfico de dispersão unidimensional (*dotplot*)

Neste tipo de gráfico representamos os valores $x_1, \ldots, x_n$ por pontos ao longo de um segmento de reta provido de uma escala. Valores repetidos são empilhados, de modo que possamos ter uma ideia de sua distribuição. O gráfico de dispersão unidimensional para a variável Salário do Exemplo 3.1 está representado na Figura 3.5.

Tabela 3.4: Distribuição de frequências para a variável Salário correspondente ao Exemplo 3.1

Classe de salário (R%)	Frequência observada	Frequência relativa (%)	Frequência relativa acumulada (%)
$0 \vdash 1500$	6	12,2	12,2
$1500 \vdash 3000$	27	55,1	67,3
$3000 \vdash 4500$	12	24,5	91,8
$4500 \vdash 6000$	2	4,1	95,9
$6000 \vdash 7500$	2	4,1	100,0
Total	49	100,0	100,0

Obs: um dos participantes não informou o salário.

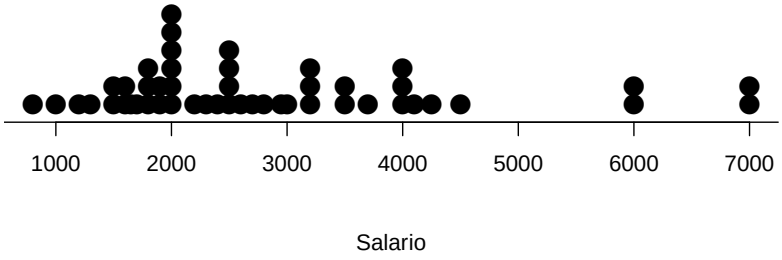


Figura 3.5: Gráfico de dispersão unidimensional para a variável Salário (Exemplo 3.1).

Gráfico Ramo e Folhas (*Stem and leaf*)

Um procedimento alternativo para reduzir um conjunto de dados sem perder muita informação sobre eles consiste na construção de um gráfico chamado ramo-e-folhas. Não há regras fixas para construí-lo, mas a ideia básica é dividir cada observação em duas partes: o **ramo**, colocado à esquerda de uma linha vertical, e a **folha**, colocada à direita.

Considere a variável "Salário" do Exemplo 3.1. Para cada observação, podemos considerar o ramo como sua parte inteira e a folha como sua parte decimal. Com esse procedimento, é fácil ver que obtemos muitos ramos, essencialmente tantos quantos são os dados e não teríamos alcançado o objetivo de resumi-los. Outra possibilidade é arredondar os dados para números inteiros e considerar o primeiro dígito como o ramo e o segundo como folha, no caso de dezenas; os dois primeiros dígitos como o ramo e o terceiro como folha, para as centenas, etc. O gráfico correspondente, apresentado na Figura 3.6 permite avaliar a forma da distribuição das observações; em particular, vemos que há quatro valores atípicos, nomeadamente, dois iguais

a R$ 6000 (correspondentes aos alunos 22 e 49) e dois iguais a R$ 7000 (correspondentes aos alunos 11 e 38) respectivamente.

```
1 | 2: representa 1200
unidade da folha: 100
n: 49

  0 | 8
  1 | 023
  1 | 55666788899
  2 | 000000234
  2 | 55556789
  3 | 0222
  3 | 557
  4 | 00012
  4 | 5
  5 |
  5 |
  6 | 00
  6 |
  7 | 00
```

Figura 3.6: Gráfico ramo-e-folhas para a variável Salário (R$).

Histograma

O histograma é um gráfico construído a partir da distribuição de frequências dos dados e é composto de retângulos contíguos cuja área total é em geral normalizada para ter valor unitário. A **área** de cada retângulo corresponde à frequência relativa associada à classe definida por sua base.

Um histograma correspondente à distribuição de frequências indicada na Tabela 3.4 está representado na Figura 3.7.

Formalmente, dados os valores $x_1, \ldots, x_n$ de uma variável quantitativa X, podemos construir uma tabela contendo

a) as frequências absolutas n_k, $k = 1, \ldots, K$, que correspondem aos números de elementos cujos valores pertencem às classes $k = 1, \ldots, K$;

b) as frequências relativas $f_k = n_k/n$, $k = 1, \ldots, K$, que são as proporções de elementos cujos valores pertencem às classes $k = 1, \ldots, K$;

c) as densidades de frequência $d_k = f_k/h_k$, $k = 1, \ldots, K$, que representam as proporções de valores pertencentes às classes $k = 1, \ldots, K$ por unidade de comprimento h_k de cada classe.

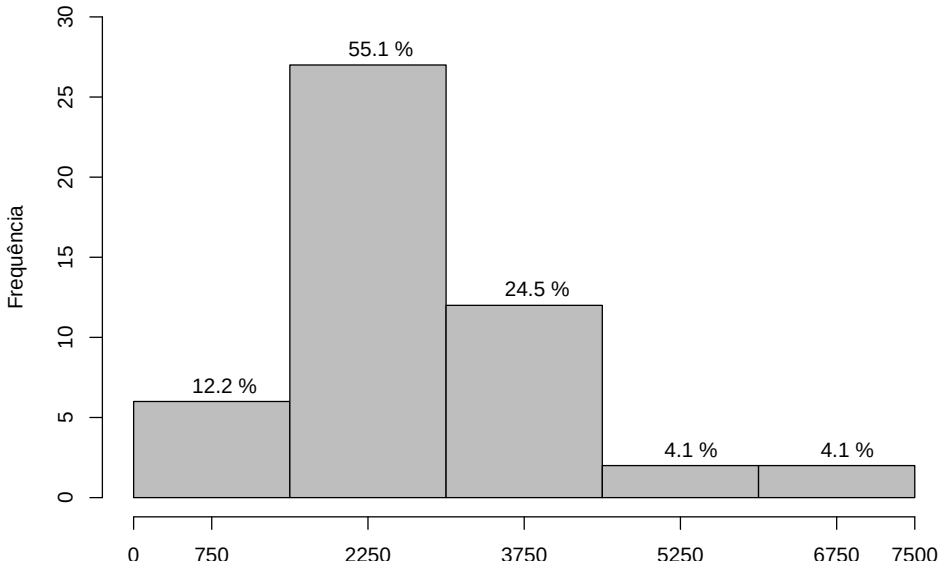


Figura 3.7: Histograma para a variável salário (R$).

Exemplo 3.2: Os dados correspondentes à população[1] (em 10000 habitantes) de 30 municípios brasileiros (IBGE, 1996) estão dispostos na Tabela 3.5. Os dados estão disponíveis no arquivo `municipios`.

Ordenemos os valores da variável X = população do município $i = 1, \ldots, 30$ do menor para o maior e consideremos a primeira classe como aquela com limite inferior igual a 40 e a última com limite superior igual a 1000; para que as classes sejam disjuntas, tomemos por exemplo, intervalos semiabertos. A Tabela 3.6 contém a distribuição de frequências para a variável X. Observemos que as duas primeiras classes têm amplitudes iguais a 100, a terceira tem amplitude 50, a penúltima tem amplitude igual 350 e a última, amplitude igual a 400. Observemos que $K = 5$, $\sum_{k=1}^{K} n_k = n = 30$ e que $\sum_{k=1}^{K} f_k = 1$. Quanto maior for a densidade de frequência de uma classe, maior será a concentração de valores nessa classe.

O valor da amplitude de classes h deve ser escolhido de modo adequado. Se h for grande, teremos poucas classes e o histograma pode não mostrar detalhes importantes; por outro lado, se h for pequeno, teremos muitas classes e algumas poderão ser vazias. A escolha do número e amplitude das classes é arbitrária. Detalhes técnicos sobre a escolha do número de classes em casos específicos podem ser encontrados na Nota de Capítulo 2. Uma definição mais técnica de histograma está apresentada na Nota de Capítulo 3.

[1] Aqui, o termo "população"se refere ao número de habitantes e é encarado como uma variável. Não deve ser confundido com população no contexto estatístico, que se refere a um conjunto (na maioria das vezes, conceitual) de valores de uma ou mais variáveis medidas. Podemos considerar, por exemplo, a população de pesos de pacotes de feijão produzidos por uma empresa.

Tabela 3.5: População de 30 municípios brasileiros (10000 habitantes)

Município	População	Município	População
São Paulo (SP)	988.8	Nova Iguaçu (RJ)	83.9
Rio de Janeiro (RJ)	556.9	São Luís (MA)	80.2
Salvador (BA)	224.6	Maceió (AL)	74.7
Belo Horizonte (MG)	210.9	Duque de Caxias (RJ)	72.7
Fortaleza (CE)	201.5	S. Bernardo do Campo (SP)	68.4
Brasília (DF)	187.7	Natal (RN)	66.8
Curitiba (PR)	151.6	Teresina (PI)	66.8
Recife (PE)	135.8	Osasco (SP)	63.7
Porto Alegre (RS)	129.8	Santo André (SP)	62.8
Manaus (AM)	119.4	Campo Grande (MS)	61.9
Belém (PA)	116.0	João Pessoa (PB)	56.2
Goiânia (GO)	102.3	Jaboatão (PE)	54.1
Guarulhos (SP)	101.8	Contagem (MG)	50.3
Campinas (SP)	92.4	S. José dos Campos (SP)	49.7
São Gonçalo (RJ)	84.7	Ribeirão Preto (SP)	46.3

Tabela 3.6: Distribuição de frequências para a X = população em dezenas de milhares de habitantes

classes	h_k	n_k	f_k	$d_k = f_k/h_k$
$00 \vdash 100$	100	17	0,567	0,00567
$100 \vdash 200$	100	8	0,267	0,00267
$200 \vdash 250$	50	3	0,100	0,00200
$250 \vdash 600$	350	1	0,033	0,00010
$600 \vdash 1000$	400	1	0,033	0,00008
Total	–	30	1,000	–

O histograma da Figura 3.8 corresponde à distribuição de frequências da variável X do Exemplo 3.2, obtido usando a função `hist` do R. O gráfico de ramo-e-folhas para os dados da Tabela 3.5 está apresentado na Figura 3.9. Pelo gráfico podemos avaliar a forma da distribuição das observações; em particular, vemos que há dois valores atípicos, $556,9$ e $988,8$, correspondentes às populações do Rio de Janeiro e São Paulo, respectivamente.

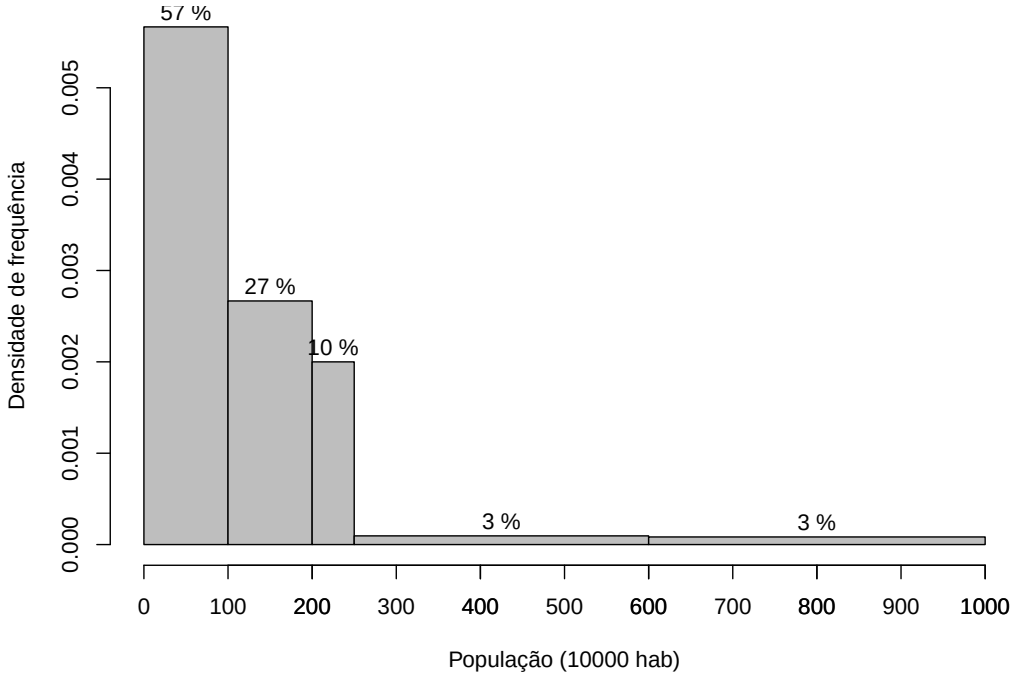


Figura 3.8: Histograma para a variável População (10000 habitantes).

```
1 | 2: representa 120
unidade da folha: 10
n: 30

    0 | 44555666666778889
    1 | 00112358
    2 | 012
    3 |
    4 |
    5 | 5
    6 |
    7 |
    8 |
    9 | 8
```

Figura 3.9: Gráfico ramo-e-folhas para a variável População (10000 habitantes).

Quando há muitas folhas num ramo, podemos considerar ramos subdivididos, como no exemplo a seguir.

Exemplo 3.3: Os dados disponíveis no arquivo `poluicao` correspondem à concentração atmosférica de poluentes ozônio O_3 e monóxido de carbono (CO) além de temperatura média e umidade na cidade de São Paulo entre 1 de janeiro e 30 de abril de 1991. O gráfico de ramo-e-folhas para a concentração de monóxido de carbono pode ser construído com dois ramos, colocando-se no primeiro folhas com dígitos de 0 a 4 e no segundo, folhas com dígitos de 5 a 9. Esse gráfico está apresentado na Figura 3.10

```
A separação decimal está em |

  4 | 77
  5 | 12
  5 | 55677789
  6 | 11111222222222333333444444
  6 | 5666677777899999999
  7 | 00122233444
  7 | 5566777778888899999999
  8 | 012334
  8 | 55678999
  9 | 0114
  9 | 557
 10 | 1333
 10 | 8
 11 | 4
 11 | 69
 12 | 0
 12 | 5
```

Figura 3.10: Gráfico ramo-e-folhas para a variável CO (ppm).

3.3 Medidas resumo

Em muitas situações deseja-se fazer um resumo mais drástico de um determinado conjunto de dados, por exemplo, por meio de um ou dois valores. A renda per capita de um país ou a porcentagem de eleitores favoráveis a um candidato são exemplos típicos. Com essa finalidade podem-se considerar as chamadas medidas de posição (localização ou de tendência central), as medidas de dispersão e medidas de assimetria, entre outras.

3.3.1 Medidas de posição

As medidas de posição mais utilizadas são a média, a mediana, a média aparada e os quantis. Para defini-las, consideremos as observações $x_1, \ldots, x_n$ de uma variável X.

A **média aritmética** (ou simplesmente média) é definida por

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.1}$$

No caso de dados agrupados numa distribuição de frequências de um conjunto com n valores, K classes e n_k valores na classe k, $k = 1, \ldots, K$, a média pode ser calculada como

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{K} n_k \tilde{x}_k = \sum_{k=1}^{K} f_k \tilde{x}_k, \tag{3.2}$$

em que $\tilde{x}_k$ é o ponto médio correspondente à classe k e $f_k = n_k/n$. Essa mesma expressão é usada para uma variável discreta, com n_k valores iguais a x_k, bastando para isso, substituir com $\tilde{x}_k$ por x_k.

A **mediana** é definida em termos das **estatísticas de ordem**, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ por

$$\mathrm{md}(x) = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ for ímpar,} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{se } n \text{ for par.} \end{cases} \tag{3.3}$$

Dado um número $0 < \alpha < 1$, a **média aparada** de ordem α, $\overline{x}(\alpha)$ é definida como a média do conjunto de dados obtido após a eliminação das $100\alpha\%$ primeiras observações ordenadas e das $100\alpha\%$ últimas observações ordenadas do conjunto original. Uma definição formal é:

$$\overline{x}(\alpha) = \begin{cases} \frac{1}{n(1-2\alpha)}\{\sum_{i=m+2}^{n-m-1} x_{(i)} + (1+m-n\alpha)[x_{(m+1)} + x_{(n-m)}]\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{se } m+2 \leq n-m+1 \\ \frac{1}{2}[x_{(m+1)} + x_{(n-m)}] \qquad\qquad \text{em caso contrário.} \end{cases} \tag{3.4}$$

em que m é o maior inteiro menor ou igual a $n\alpha$, $0 < \alpha < 0,5$. Se $\alpha = 0,5$, obtemos a mediana. Para $\alpha = 0,25$ obtemos a chamada **meia média**. Observe que se $\alpha = 0$, $\overline{x}(0) = \overline{x}$.

Exemplo 3.4: Consideremos o seguinte conjunto com $n = 10$ valores de uma variável X: $\{14, 7, 3, 18, 9, 220, 34, 23, 6, 15\}$. Então, $\overline{x} = 34,9$, $\mathrm{md}(x) = (14+15)/2 = 14,5$, e $\overline{x}(0,2) = [x_{(3)} + x_{(4)} + \ldots + x_{(8)}]/6 = 14,3$. Note que se usarmos (3.4), temos $\alpha = 0,2$ e $m = 2$ obtendo o mesmo resultado. Se $\alpha = 0,25$, então de (3.4) obtemos

$$\overline{x}(0,25) = \frac{x_{(3)} + 2x_{(4)} + 2x_{(5)} + \ldots + 2x_{(7)} + x_{(8)}}{10} = 14,2$$

Observe que a média é bastante afetada pelo valor atípico 220, ao passo que a mediana e a média aparada com $\alpha = 0,2$ não o são. Dizemos que essas duas últimas são **medidas resistentes** ou **robustas**.[2] Se substituirmos o

[2]Uma medida é dita resistente se ela muda pouco quando alterarmos um número pequeno dos valores do conjunto de dados.

valor 220 do exemplo por 2200, a média passa para $232,9$ ao passo que a mediana e a média aparada $\overline{x}(0,20)$ não se alteram.

As três medidas consideradas acima são chamadas de medidas de posição ou localização central do conjunto de dados. Para variáveis qualitativas também é comum utilizarmos outra medida de posição que indica o valor mais frequente, denominado **moda**. Quando há duas classes com a mesma frequência máxima, a variável (ou distribuição) é dita **bimodal**. A não ser que os dados de uma variável contínua sejam agrupados em classes, caso em que se pode considerar a **classe modal**, não faz sentido considerar a moda, pois em geral, cada valor da variável tem frequência unitária.

Quantis

Consideremos agora medidas de posição úteis para indicar posições não centrais dos dados. Informalmente, um quantil-p ou quantil de ordem p é **um valor da variável** (quando ela é contínua) ou **um valor interpolado entre dois valores da variável** (quando ela é discreta) que deixa $100p\%$ $(0 < p < 1)$ das observações à sua esquerda. Formalmente, definimos o quantil-p empírico (ou simplesmente quantil) como

$$Q(p) = \begin{cases} x_{(i)}, & \text{se } p = p_i = (i-0,5)/n,\ i = 1,\ldots,n \\ (1-f_i)Q(p_i) + f_i Q(p_{i+1}), & \text{se } p_i < p < p_{i+1} \\ x_{(1)}, & \text{se } 0 < p < p_1 \\ x_{(n)}, & \text{se } p_n < p < 1, \end{cases} \tag{3.5}$$

em que $f_i = (p-p_i)/(p_{i+1}-p_i)$. Ou seja, se p for da forma $p_i = (i-0,5)/n$, o quantil-p coincide com a i-ésima observação ordenada. Para um valor p entre p_i e p_{i+1}, o quantil $Q(p)$ pode ser definido como sendo a ordenada de um ponto situado no segmento de reta determinado por $[p_i, Q(p_i)]$ e $[p_{i+1}, Q(p_{i+1})]$. Escolhemos p_i como acima (e não como i/n, por exemplo) de forma que se um quantil coincidir com uma das observações, metade dela pertencerá ao conjunto de valores à esquerda de $Q(p)$ e metade ao conjunto de valores à sua direita.

Os quantis amostrais para os dez pontos do Exemplo 3.4 estão indicados na Tabela 3.7. Com essa informação, podemos calcular outros quantis; por

Tabela 3.7: Quantis amostrais para os dados do Exemplo 3.4

i	1	2	3	4	5	6	7	8	9	10
p_i	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95
$Q(p_i)$	3	6	7	9	14	15	18	23	34	220

exemplo, $Q(0,10) = [x_{(1)} + x_{(2)}]/2 = (3+6)/2 = 4,5$ com $f_1 = (0,10 - 0,05)/(0,10) = 0,5$, $Q(0,90) = [x_{(9)} + x_{(10)}]/2 = (34+220)/2 = 127$, pois $f_9 = 0,5$ e $Q(0,62) = [0,30 \times x_{(6)} + 0,70 \times x_{(7)}] = (0,3 \times 15 + 0,7 \times 18 = 17,1$ pois $f_6 = (0,62 - 0,55)/0,10 = 0,7$ Note que a definição (3.5) é compatível

com a definição de mediana apresentada anteriormente.

Os quantis $Q(0,25), Q(0,50)$ e $Q(0,75)$ são chamados **quartis** e usualmente são denotados por Q_1, Q_2 e Q_3, respectivamente. O quartil Q_2 é a mediana e a proporção dos dados entre Q_1 e Q_3 é 50%.

Outras denominações comumente empregadas são $Q(0,10)$: primeiro decil, $Q(0,20)$: segundo decil ou vigésimo percentil, $Q(0,85)$: octogésimo-quinto percentil etc.

3.3.2 Medidas de dispersão

Duas medidas de dispersão (ou de escala ou de variabilidade) bastante usadas são obtidas tomando-se a média dos desvios das observações em relação à sua média. Considere as observações $x_1, \ldots, x_n$, não necessariamente distintas. A **variância** desse conjunto de dados é definida por

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{3.6}$$

No caso de uma tabela de frequências (com K classes), a expressão para cálculo da variância é

$$\text{var}(x) = \frac{1}{n}\sum_{k=1}^{K} n_k(\tilde{x}_k - \overline{x})^2 = \sum_{k=1}^{K} f_k(\tilde{x}_k - \overline{x})^2, \tag{3.7}$$

com a notação estabelecida anteriormente. Para facilitar os cálculos, convém substituir (3.6) pela expressão equivalente

$$\text{var}(x) = n^{-1}\sum_{i=1}^{n} x_i^2 - \overline{x}^2. \tag{3.8}$$

Analogamente, podemos substituir (3.7) por

$$\text{var}(x) = n^{-1}\sum_{k=1}^{K} f_k\tilde{x}_k^2 - \overline{x}^2. \tag{3.9}$$

Como a unidade de medida da variância é o quadrado da unidade de medida da variável correspondente, convém definir outra medida de dispersão que mantenha a unidade de medida original. Uma medida com essa propriedade é a raiz quadrada positiva da variância, conhecida por **desvio padrão**, denotado $\text{dp}(x)$.

Para garantir certas propriedades estatísticas úteis para propósitos de inferência, convém modificar as definições acima. Em particular, para garantir que a variância obtida de uma amostra de dados de uma população seja um **estimador não enviesado** da variância populacional basta definir a variância como

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{3.10}$$

em substituição à definição (3.6).

Um estimador (a variância amostral S^2, por exemplo) de um determinado parâmetro (a variância populacional σ^2, por exemplo) é dito não enviesado quando seu valor esperado é o próprio parâmetro que está sendo estimado. *Grosso modo*, se um conjunto "infinito" (aqui interpretado como muito grande) de for colhido da população sob investigação e para cada uma delas for calculado o valor desse estimador não enviesado, a média desses valores será o próprio parâmetro (ou estará bem próxima dele). Dizemos que o estimador S^2 tem $n-1$ **graus de liberdade** pois "perdemos" um grau de liberdade ao estimar a média populacional μ por meio de $\overline{x}$, ou seja, dado o valor $\overline{x}$, só temos "liberdade" para escolher $n-1$ valores da variável X, pois o último valor, digamos x_n, é obtido como $x_n = n\overline{x} - \sum_{k=1}^{n-1} x_i$.

Note que se n for grande (*e.g.*, $n = 100$) (3.10) e (3.6) têm valores praticamente iguais. Para detalhes, veja Bussab e Morettin (2014) entre outros.

Em geral, S^2 é conhecida por **variância amostral**. A **variância populacional** é definida como em (3.10) com o denominador $n-1$ substituído pelo tamanho populacional N e a média amostral $\overline{x}$ substituída pela média populacional μ. As fórmulas de cálculo acima podem ser modificadas facilmente com essa definição; o desvio padrão amostral é usualmente denotado por S.

O **desvio médio** ou **desvio absoluto médio** é definido por

$$\mathrm{dm}(x) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|. \tag{3.11}$$

Outra medida de dispersão bastante utilizada é a **distância interquartis** ou **amplitude interquartis**

$$d_Q = Q_3 - Q_1. \tag{3.12}$$

A distância interquartis pode ser utilizada para estimar o desvio padrão conforme indicado na Nota de Capítulo 4.

Podemos também considerar uma medida de dispersão definida em termos de desvios em relação à mediana. Como a mediana é uma medida robusta, nada mais natural que definir o **desvio mediano absoluto** como

$$\mathrm{dma}(x) = \mathrm{md}_{1\le i\le n}|x_i - \mathrm{md}(x)|, \tag{3.13}$$

Finalmente, uma medida correspondente à média aparada é a **variância aparada**, definida por

$$S^2(\alpha) = \begin{cases} \frac{c_\alpha}{n(1-2\alpha)}\left(\sum_{i=m+2}^{n-m-1}[x_{(i)} - \overline{x}(\alpha)]^2 + A\right), & m+2 \le n-m+1 \\ \frac{1}{2}[(x_{(m+1)} - \overline{x}(\alpha))^2 + (x_{(n-m)} - \overline{x}(\alpha))^2], & \text{em caso contrário,} \end{cases} \tag{3.14}$$

em que

$$A = (1 + m - n\alpha)[(x_{(m+1)} - \overline{x}(\alpha))^2 + (x_{(n-m)} - \overline{x}(\alpha))^2],$$

m é como em (3.4) e c_α é uma constante normalizadora que torna $S^2(\alpha)$ um estimador não enviesado para σ^2. Para n grande, $c_\alpha = 1,605$. Para amostras pequenas, veja a tabela da página 173 de Johnson e Leone (1964). Em particular, para $n = 10$, $c_\alpha = 1,46$.

A menos do fator c_α, a variância aparada pode ser obtida calculando-se a variância amostral das observações restantes, após a eliminação das $100\alpha\%$ iniciais e finais (com denominador $n-l$ em que l é o número de observações desprezadas).

Considere as observações do Exemplo 3.4. Para esse conjunto de dados as medidas de dispersão apresentadas são $S^2 = 4313,9$, $S = 65,7$, $\text{dm}(x) = 37,0$, $d_Q = 23 - 7 = 16$, $S^2(0,2) = 34,3$, $S(0,20) = 5,9$, e $\text{dma}(x) = 7,0$.

Observemos que as medidas robustas são, em geral, menores do que $\overline{x}$ e S e que $d_Q/1,349 = 11,9$. Se considerarmos que esses dados constituem uma amostra de uma população com desvio padrão σ, pode-se mostrar que, $\text{dma}/0,6745$ é um estimador não enviesado para σ. A constante $0,6745$ é obtida por meio de considerações assintóticas. No exemplo, $\text{dma}/0,6745 = 10,4$. Note que esses dois estimadores do desvio padrão populacional coincidem. Por outro lado, S é muito maior, pois sofre bastante influencia do valor 220. Retirando-se esse valor do conjunto de dados, a média dos valores restantes é $14,3$ e o correspondente desvio padrão é $9,7$.

Uma outra medida de dispersão, menos utilizada na prática é a **amplitude**, definida como $max(x_1, \ldots, x_n) - min(x_1, \ldots, x_n)$.

3.3.3 Medidas de assimetria

Embora sejam menos utilizadas na prática que as medidas de posição e dispersão, as medidas de assimetria (*skewness*) são úteis para identificar modelos probabilísticos para análise inferencial.

Na Figura 3.11 estão apresentados histogramas correspondentes a dados com assimetria positiva (ou à direita) ou negativa (ou à esquerda) e simétrico. O objetivo das medidas de assimetria é quantificar sua magnitude e, em geral, são baseadas na relação entre o segundo e o terceiro **momentos centrados**, nomeadamente

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{e} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3.$$

Dentre as medidas de assimetria, as mais comuns são:

a) o coeficiente de assimetria de Fisher-Pearson: $g_1 = m_3/m_2^{3/2}$
b) o coeficiente de assimetria de Fisher-Pearson ajustado:

$$\frac{\sqrt{n(n-1)}}{n-1}\sum_{i=1}^{n}[(x_i - \overline{x})/\sqrt{m_2}]^3.$$

As principais propriedades desses coeficientes são

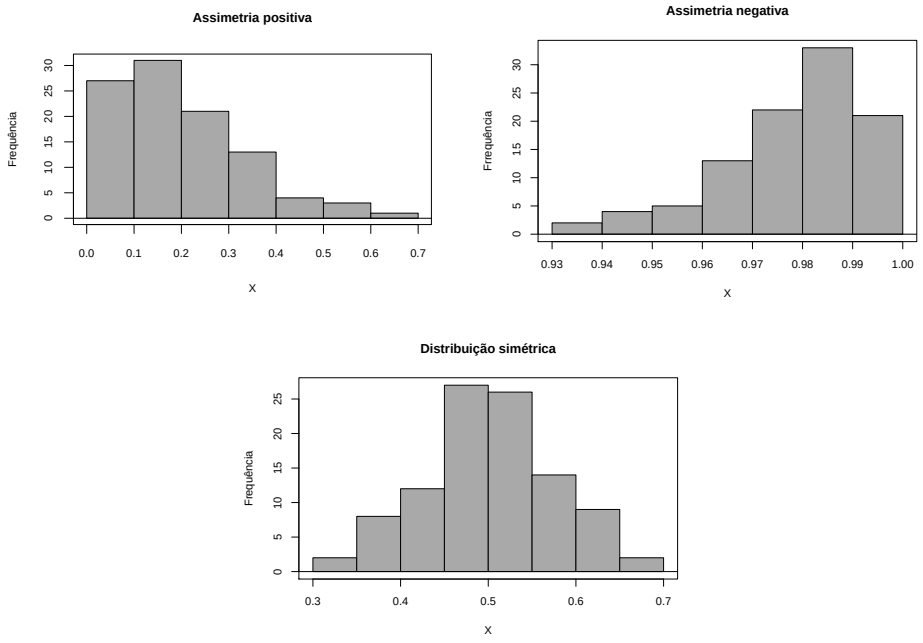


Figura 3.11: Histogramas com assimetria positiva e negativa e simétrico.

i) seu sinal reflete a direção da assimetria;
ii) comparam a assimetria dos dados com aquela da distribuição Normal;
iii) valores mais afastados do zero indicam maiores magnitudes de assimetria e consequentemente, maior afastamento da distribuição Normal;
iv) a estatística indicada em b) tem um ajuste para o tamanho amostral;
v) esse ajuste tem pequeno impacto em grandes amostras.

Outro coeficiente de assimetria mais intuitivo é o chamado Coeficiente de assimetria de Pearson 2,

$$Sk_2 = 3[\overline{x} - md(x)]/s.$$

A avaliação de assimetria também pode ser concretizada por meios gráficos. Em particular, o gráfico de $Q(p)$ *versus* p conhecido como **gráfico de quantis** é uma ferramenta importante para esse propósito.

A Figura 3.12 mostra o gráfico de quantis para os dados do Exemplo 3.2. Notamos que os pontos correspondentes a São Paulo e Rio de Janeiro são destacados. Além disso, se a distribuição dos dados for aproximadamente simétrica, a inclinação na parte superior do gráfico deve ser aproximadamente igual àquela da parte inferior, o que não acontece na figura em questão.

Os cinco valores $x_{(1)}, Q_1, Q_2, Q_3, x_{(n)}$, isto é, os extremos e os quartis, são medidas de localização importantes para avaliarmos a simetria dos dados.

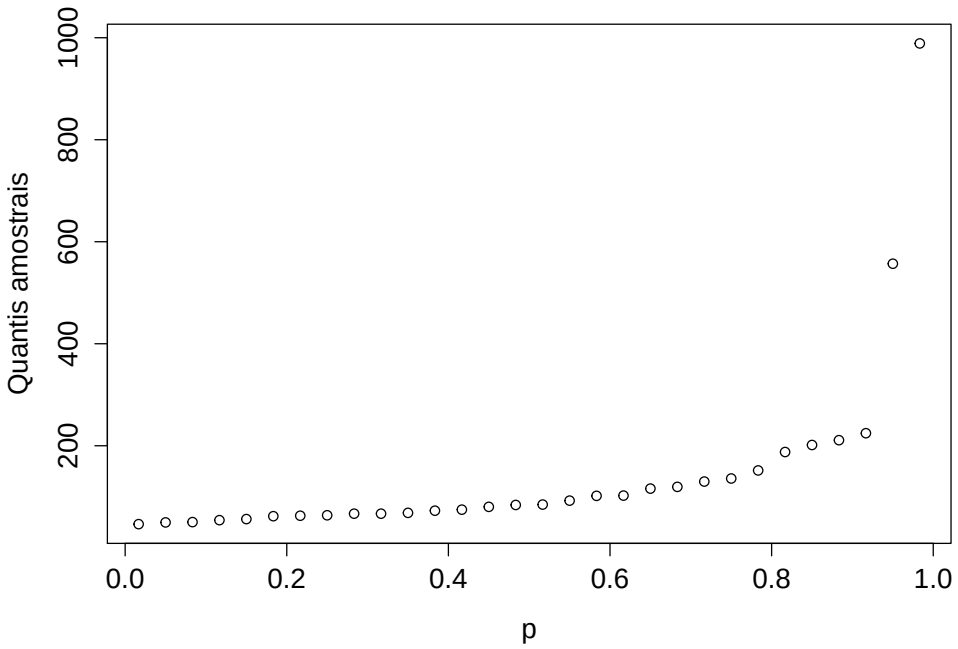


Figura 3.12: Gráfico de quantis para População (10000 habitantes).

Suponha uma distribuição simétrica (ou aproximadamente simétrica). Então,

a) $Q_2 - x_{(1)} \approx x_{(n)} - Q_2$;

b) $Q_2 - Q_1 \approx Q_3 - Q_2$;

c) $Q_1 - x_{(1)} \approx x_{(n)} - Q_3$.

Para distribuições assimétricas à direita, as diferenças entre os quantis situados à direita da mediana e a mediana são maiores que as diferenças entre a mediana e os quantis situados à esquerda da mediana. A condição (a) nos diz que a **dispersão inferior** é igual (ou aproximadamente igual) à **dispersão superior**. Notamos, também, que se uma distribuição for (aproximadamente) simétrica, vale a relação

$$Q_2 - x_{(i)} = x_{(n+1-i)} - Q_2, \quad i = 1, \ldots, [(n+1)/2], \tag{3.15}$$

em que $[x]$ representa o maior inteiro contido em x.

Chamando $u_i = Q_2 - x_{(i)},\ v_i = x_{(n+1-i)} - Q_2$, podemos considerar um **gráfico de simetria**, no qual colocamos os valores u_i como abcissas e os valores v_i como ordenadas. Se a distribuição dos dados for simétrica, os pontos (u_i, v_i) deverão estar sobre (ou próximos) da reta $u = v$.

O gráfico de simetria para os dados do Exemplo 3.2 está apresentado a Figura 3.13, na qual podemos observar que a maioria dos pontos está acima da reta $u = v$, mostrando a assimetria à direita desses dados.

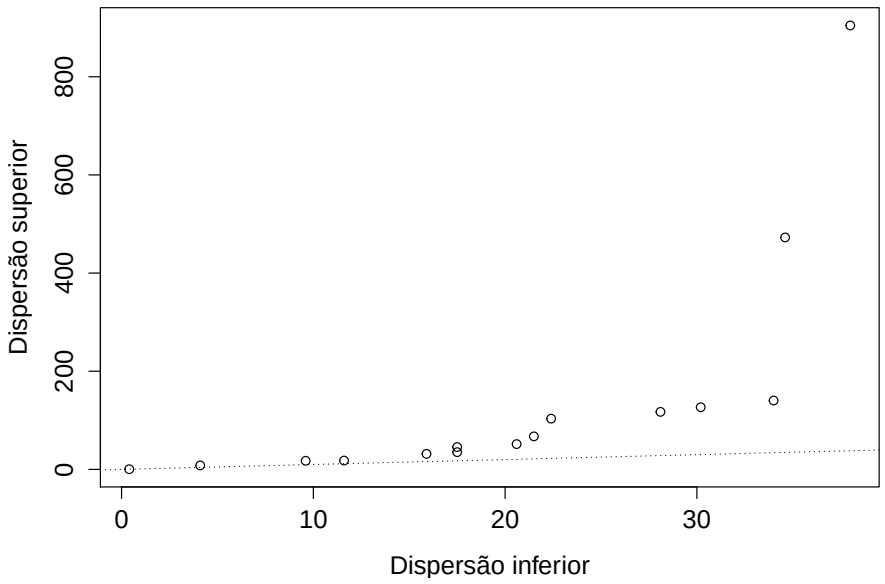
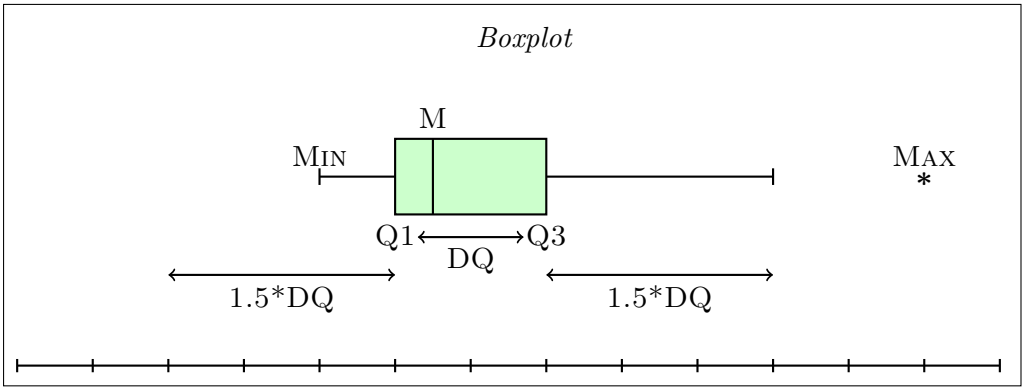


Figura 3.13: Gráfico de simetria para População (10000 habitantes).

Outra medida de interesse é a curtose, relacionada às caudas de uma distribuição. Essa medida envolve momentos de quarta ordem. Veja a Nota de Capítulo 7.

3.4 *Boxplots*

O *boxplot* é um gráfico baseado nos quantis que serve como alternativa ao histograma para resumir a distribuição dos dados. Considere um retângulo, com bases determinadas por Q_1 e Q_3, como indicado na Figura 3.14. Nesse retângulo, insira um segmento de reta correspondente à posição da mediana. Considere dois segmentos de reta denominados bigodes (*whiskers*) colocados respectivamente acima e abaixo de Q_1 e Q_3 com limites dados, respectivamente por $\min[x_{(n)}, Q_3+1,5*d_Q]$ e $\max[x_{(1)}, Q_1-1,5*d_Q]$. Pontos colocados acima do limite superior ou abaixo do limite inferior, considerados **valores atípicos** ou **discrepantes** (*outliers*) são representados por algum símbolo ($*$, por exemplo).



Q1: 1o quartil Q3: 3o quartil DQ: distância interquartis M: mediana

Figura 3.14: Detalhes para a construção de *boxplots*.

Esse gráfico permite que identifiquemos a posição dos 50% centrais dos dados (entre o primeiro e terceiro quartis), a posição da mediana, os valores atípicos, se existirem, assim como permite uma avaliação da simetria da distribuição. *Boxplots* são úteis para a comparação de vários conjuntos de dados, como veremos em capítulos posteriores.

Os *boxplots* apresentados na Figura 3.15 correspondem aos dados do Exemplo 3.2 [painel (a)] e da Temperatura do Exemplo 3.3 [painel (b)].[3] A distribuição dos dados de Temperatura tem uma natureza mais simétrica e mais dispersa do que aquela correspondente às populações de municípios. Há valores atípicos no painel (a), representando as populações do Rio de Janeiro e de São Paulo, mas não os encontramos nos dados de temperatura.

Há uma variante dos *boxplots*, denominada ***boxplot* dentado** (*notched boxplot*) que consiste em acrescentar um dente em "v" ao redor da mediana no gráfico. O intervalo determinado pelo dente, dado por

$$Q_2 \pm \frac{1,57 d_Q}{\sqrt{n}}.$$

é um intervalo de confiança para a mediana da população da qual supomos que os dados constituem uma amostra. Para detalhes, veja McGill et al. (1978) ou Chambers et al. (1983). Na Figura 3.16 apresentamos *boxplots* correspondentes àqueles da Figura 3.15 com os dentes (*notchs*) incorporados.

[3]Note que tanto a orientação horizontal (como na Figura 3.14) quanto a vertical (como na Figura 3.15) podem ser empregadas na construção dos *boxplots*.

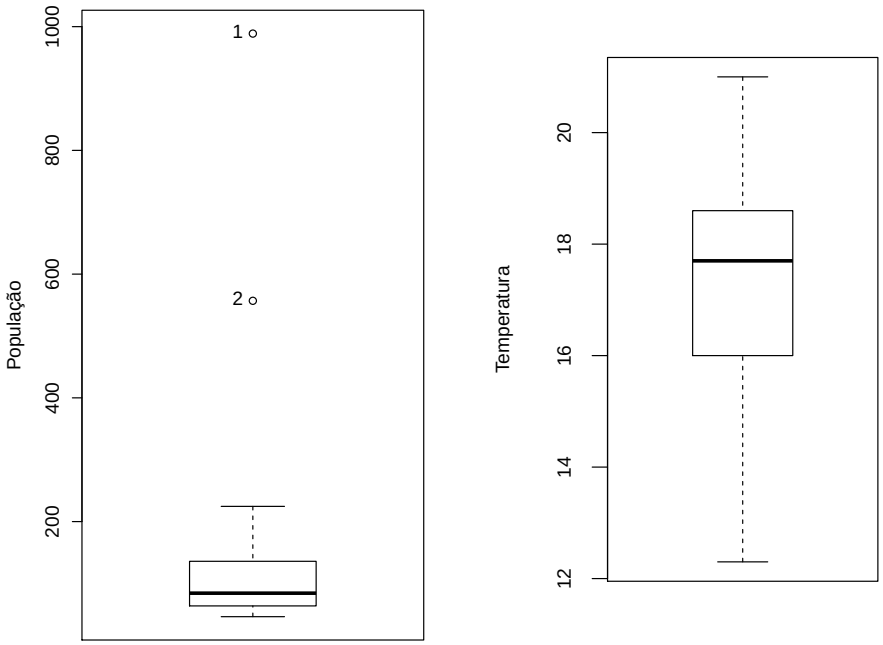


Figura 3.15: *Boxplots* para os dados dos Exemplos 3.2 (População) e 3.3 (Temperatura).

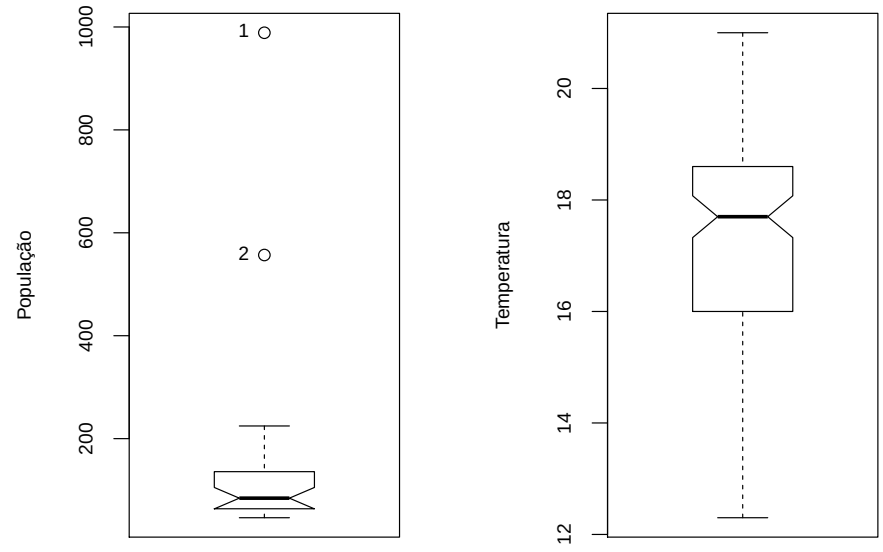


Figura 3.16: *Boxplots* dentados para os dados dos Exemplos 3.2 (População) e 3.3 (Temperatura).

3.5 Modelos probabilísticos

Um dos objetivos da Estatística é fazer inferência (ou tirar conclusões) sobre distribuições de variáveis em populações a partir de dados de uma parte dela, denominada **amostra**. A ligação entre os dados amostrais e a população depende de **modelos probabilísticos** ou seja, de modelos que representem a distribuição (desconhecida) da variável na população. Por exemplo, pode ser difícil obter informações sobre a distribuição dos salários de empregados de uma empresa com 40.000 empregados espalhados por diversos países. Nessa situação, costuma-se recorrer a uma amostra dessa população, obter as informações desejadas a partir dos valores amostrais e tentar tirar conclusões sobre toda a população a partir desses valores com base num modelo probabilístico. Esse procedimento é denominado **inferência estatística**. No exemplo acima, podemos escolher uma amostra de 400 empregados da empresa e analisar a distribuição dos salários dessa amostra.

Muitas vezes, a população para a qual se quer tirar conclusões é apenas conceitual e não pode ser efetivamente enumerada, como o conjunto de potenciais consumidores de um produto ou o conjunto de pessoas que sofrem de uma certa doença. Nesses casos, não se pode obter a correspondente distribuição de frequências de alguma característica de interesse associada a essa população e o recurso a modelos para essa distribuição faz-se necessário; esses são os chamados modelos probabilísticos e as frequências relativas correspondentes são denominadas probabilidades. Nesse sentido, o conhecido gráfico com formato de sino associado à distribuição Normal pode ser considerado como um histograma teórico. Por isso, convém chamar a média da distribuição de probabilidades (que no caso conceitual não pode ser efetivamente calculada) de **valor esperado**.

Se pudermos supor que a distribuição de probabilidades de uma variável X, definida sobre uma população possa ser representada por um determinado modelo probabilístico, nosso problema reduz-se a estimar os parâmetros que caracterizam esse modelo.

Há vários modelos probabilísticos importantes usados em situações de interesse prático. As Tabelas 3.8 e 3.9 trazem um resumo das principais distribuições discretas e contínuas, respectivamente apresentando:

a) a **função de probabilidade** (f.p.) $p(x) = P(X = x)$, no caso discreto e a **função densidade de probabilidade** (f.d.p.), $f(x)$, no caso contínuo;

b) os parâmetros que caracterizam cada distribuição;

c) a média e a variância de cada uma delas.

Detalhes podem ser encontrados em Bussab e Morettin (2017) entre outros. Para muitas dessas distribuições, as probabilidades podem ser encontradas em tabelas apropriadas ou obtidas com o uso de programas de computador.

Tabela 3.8: Modelos probabilísticos para variáveis discretas

Modelo	$P(X = x)$	Parâmetros	$E(X), \mathrm{Var}(X)$
Bernoulli	$p^x(1-p)^{1-x},\ x = 0, 1$	p	$p,\ p(1-p)$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x},\ x = 0, \ldots, n$	n, p	$np,\ np(1-p)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, \ldots$	λ	$\lambda,\ \lambda$
Geométrica	$p(1-p)^{x-1},\ x = 1, 2, \ldots$	p	$1/p,\ (1-p)/p^2$
Hipergeométrica	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, x = 0, 1, \ldots$	N, r, n	$nr/N, n\frac{r}{N}(1-\frac{r}{N})\frac{N-n}{N-1}$

Tabela 3.9: Modelos probabilísticos para variáveis contínuas

Modelo	$f(x)$	Parâmetros	$E(X), \mathrm{Var}(X)$
Uniforme	$1/(b-a),\ a < x < b$	a, b	$\frac{a+b}{2}, \frac{(b-a)^2}{12}$
Exponencial	$\alpha e^{-\alpha x},\ x > 0$	α	$1/\alpha,\ 1/\alpha^2$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} \exp\{(\frac{x-\mu}{\sigma})^2\},\ -\infty < x < \infty$	$\mu,\ \sigma$	$\mu,\ \sigma^2$
Gama	$\frac{\alpha}{\Gamma(r)}(\alpha x)^{r-1}e^{-\alpha x},\ x > 0$	$\alpha > 0, r \geq 1$	$r/\alpha,\ r/\alpha^2$
Qui-quadrado	$\frac{2^{-n/2}}{\Gamma(n/2)} x^{n/2-1}e^{-x/2},\ x > 0$	n	$n,\ 2n$
t-Student	$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}(1+\frac{x^2}{n})^{-(n+1)/2},\ -\infty < x < \infty$	n	$0,\ n/(n-2)$

3.6 Dados amostrais

Uma amostra é um subconjunto de uma população e para que possamos fazer inferências, é preciso que ela satisfaça certas condições. O caso mais comum é o de uma amostra aleatória simples (AAS). Dizemos que um conjunto de observações $x_1, \ldots, x_n$ constitui uma **amostra aleatória simples** de tamanho n de uma variável X definida sobre uma população $\mathcal{P}$ se as variáveis $X_1, \ldots, X_n$ que geraram as observações são independentes e têm a mesma distribuição de X. Como consequência, $\mathrm{E}(X_i) = \mathrm{E}(X)$ e $\mathrm{Var}(X_i) = \mathrm{Var}(X)$, $i = 1, \ldots, n$.

Nem sempre nossos dados representam uma AAS de uma população. Por exemplo, dados observados ao longo de um certo período de tempo são, em geral, correlacionados. Nesse caso, os dados constituem uma amostra de uma trajetória de um **processo estocástico** e a população correspondente pode ser considerada como o conjunto de todas as trajetórias de tal processo [detalhes podem ser encontrados em Morettin e Tolói (2018)].

Também podemos ter dados obtidos de um experimento planejado no qual uma ou mais variáveis são controladas para produzir valores de uma variável resposta. A não ser quando explicitamente indicado, para propósitos inferenciais, neste texto consideraremos os dados como provenientes de uma AAS.

Denotemos por $x_1, \ldots, x_n$ os valores efetivamente observados das variáveis $X_1, \ldots, X_n$. Denotemos por $x_{(1)}, \ldots, x_{(n)}$ esses valores observados ordenados em ordem crescente, ou seja, $x_{(1)} \leq \ldots \leq x_{(n)}$. Esses são os valores das **estatísticas de ordem** $X_{(1)}, \ldots, X_{(n)}$.

Muitas vezes não faremos distinção entre a variável e seu valor, ou seja, designaremos, indistintamente, por x a variável e um valor observado dela.

A **função distribuição acumulada** de uma variável X definida como $F(x) = P(X \leq x)$, $x \in \mathcal{R}$ pode ser estimada a partir dos dados amostrais, por meio da **função distribuição empírica** definida por

$$F_e(x) = \frac{n(x)}{n}, \quad \forall x \in \mathcal{R}, \tag{3.16}$$

em que $n(x)$ é o número de observações amostrais menores ou iguais a x.

Considere novamente as observações do Exemplo 3.4 sem o valor 220 para efeito ilustrativo. O gráfico de F_e que é essencialmente uma função em escada, com "saltos" de magnitude $1/n$ em cada $x_{(i)}$, nomeadamente

$$F_e(x_{(i)}) = \frac{i}{n}, \quad i = 1, \ldots, n$$

está disposto na Figura 3.17.

3.7 Gráficos QQ

Uma das questões fundamentais na especificação de um modelo para inferência estatística é a escolha de um modelo probabilístico para representar

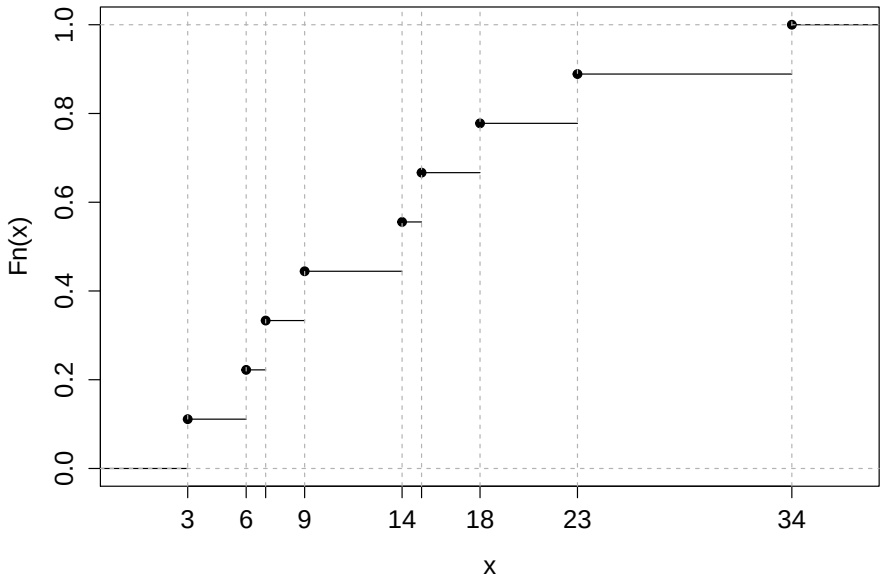


Figura 3.17: Função distribuição empírica para os dados do Exemplo 3.4 (sem o valor 220).

a distribuição (desconhecida) da variável de interesse na população. Uma possível estratégia para isso é examinar o histograma dos dados amostrais e compará-lo com histogramas teóricos associados a modelos probabilísticos candidatos. Alternativamente, os **gráficos QQ** (*QQ plots*) também podem ser utilizados com essa finalidade.

Essencialmente, gráficos QQ são gráficos cartesianos cujos pontos representam os quantis de mesma ordem obtidos das distribuições amostral (empírica) e teórica. Se os dados amostrais forem compatíveis com o modelo probabilístico proposto, esses pontos devem estar sobre uma reta (com inclinação unitária se os dados forem padronizados).

Como o modelo Normal serve de base para muitos métodos estatísticos de análise, uma primeira tentativa é construir esse tipo de gráfico baseado nos quantis dessa distribuição. Os quantis Normais padronizados $Q_N(p_i)$ são obtidos da distribuição Normal padrão $[N(0,1)]$ por meio da solução da equação

$$\int_{-\infty}^{Q_N(p_i)} \frac{1}{\sqrt{2\pi}} \exp(-x^2) = p_i,\ i = 1, \ldots n,$$

cujos resultados estão disponíveis na maioria dos pacotes computacionais destinados à análise estatística. Para facilitar a comparação, convém utilizar os quantis amostrais padronizados, $Q^\star(p_i) = [Q(p_i) - (\overline{x})]/dp(x)$ nos gráficos QQ. Ver a Nota de Capítulo 5.

Consideremos novamente os dados do Exemplo 3.4. Os quantis amostrais, quantis amostrais padronizados e Normais padronizados estão dispostos na Tabela 3.10. O correspondente gráfico QQ está representado na

Figura 3.18.

Tabela 3.10: Quantis amostrais e Normais para os dados do Exemplo 3.4

i	1	2	3	4	5	6	7	8	9	10
p_i	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0.95
$Q(p_i)$	3	6	7	9	14	15	18	23	34	220
$Q^\star(p_i)$	-0,49	-0,44	-0,42	-0,39	-0,32	-0,30	-0,26	-0,18	-0,14	2,82
$Q_N(p_i)$	-1,64	-1,04	-0,67	-0,39	-0,13	0,13	0,39	0,67	1,04	1,64

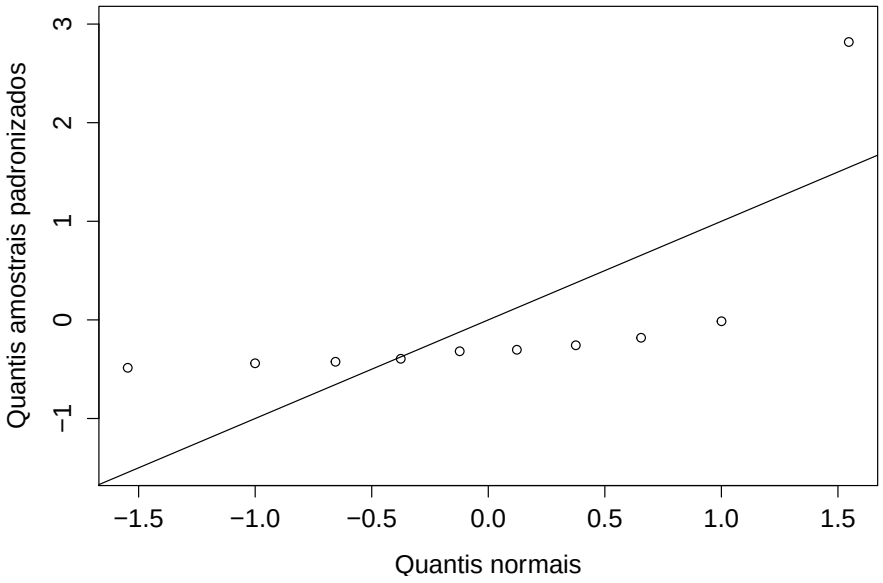


Figura 3.18: Gráfico QQ Normal para os dados do Exemplo 3.4.

Um exame da Figura 3.18 sugere que o modelo Normal não parece ser adequado para os dados do Exemplo 3.4. Uma das razões para isso, é a presença de um ponto atípico (220). Um gráfico QQ Normal para o conjunto de dados obtidos com a eliminação desse ponto está exibido na Figura 3.19 e indica que as evidências contrárias ao modelo Normal são menos aparentes.

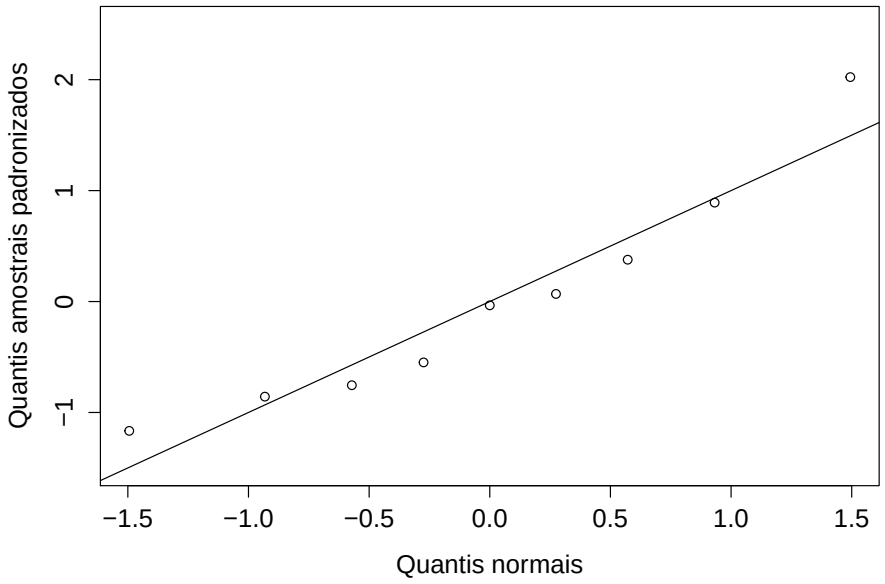


Figura 3.19: Gráfico QQ Normal para os dados do Exemplo 3.4 com a eliminação do ponto 220.

Um exemplo de gráfico QQ para uma distribuição amostral com 100 dados gerados a partir de uma distribuição Normal padrão está apresentado na Figura 3.20.

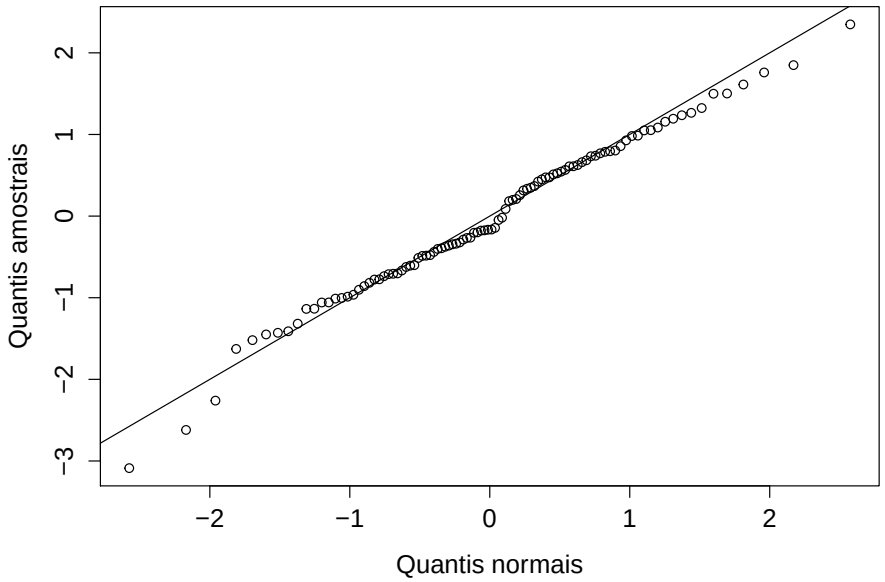


Figura 3.20: Gráfico QQ Normal para 100 dados gerados de uma distribuição Normal padrão.

Embora os dados correspondentes aos quantis amostrais da Figura 3.20 tenham sido gerados a partir de uma distribuição Normal padrão, os pontos não se situam exatamente sobre a reta com inclinação de 45 graus em função de flutuações amostrais. Em geral, a adoção de um modelo probabilístico com base num exame do gráfico QQ tem uma natureza subjetiva, mas é possível incluir bandas de confiança nesse tipo de gráfico para facilitar a decisão. Essas bandas dão uma ideia sobre a faixa de variação esperada para os pontos no gráfico. Detalhes sobre a construção dessas bandas são tratados na Nota de Capítulo 6. Um exemplo de gráfico QQ com bandas de confiança para uma distribuição amostral com 100 dados gerados a partir de uma distribuição Normal padrão está apresentado na Figura 3.21.

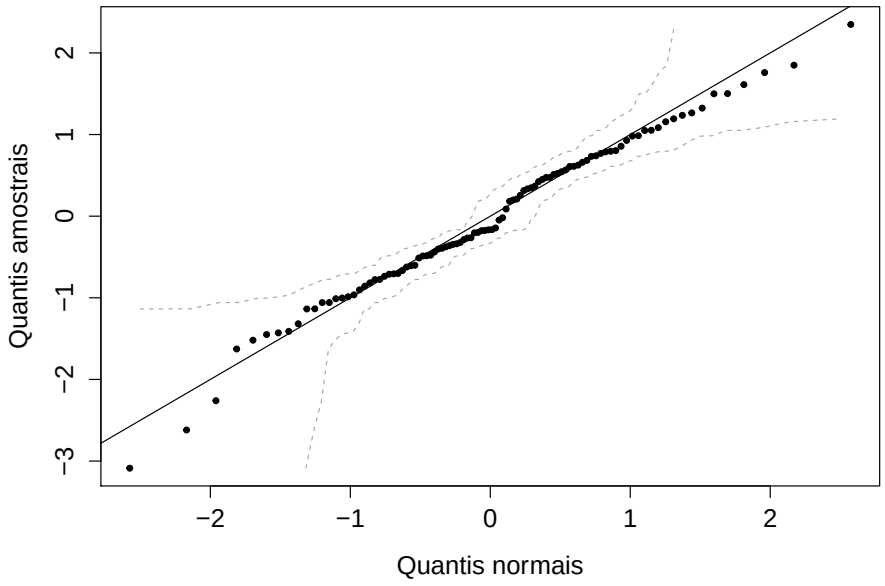


Figura 3.21: Gráfico QQ Normal para 100 dados gerados de uma distribuição Normal padrão com bandas de confiança.

Um exemplo de gráfico QQ em que as caudas da distribuição amostral (obtidas de uma amostra de 100 dados gerados a partir de uma distribuição t com 2 graus de liberdade) são mais pesadas que aquelas da distribuição Normal está apresentado na Figura 3.22.

Um exemplo de gráfico QQ Normal em que a distribuição amostral (com 100 dados gerados a partir de uma distribuição qui-quadrado com 2 graus de liberdade) é assimétrica está apresentado na Figura 3.23. O gráfico QQ correspondente, agora obtido por meio dos quantis de uma distribuição qui-quadrado com 2 graus de liberdade é apresentado na Figura 3.24.

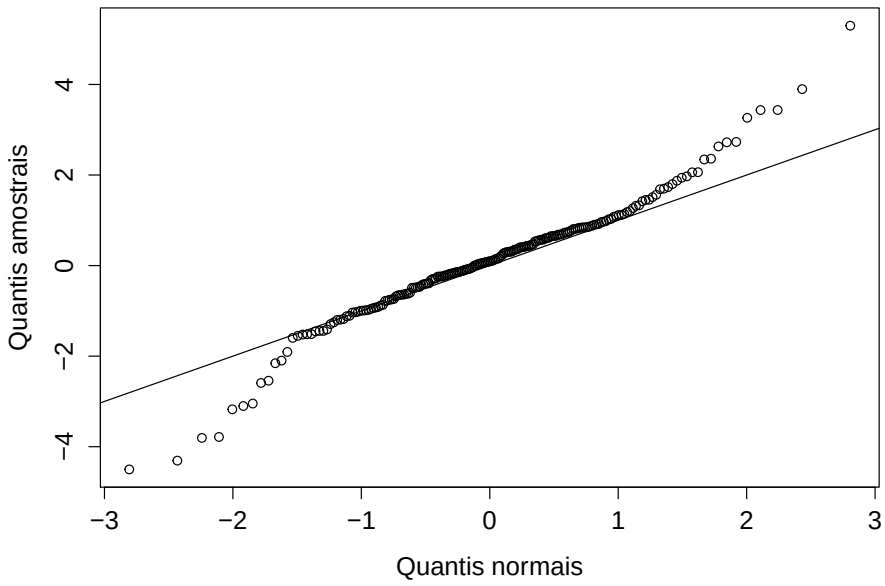


Figura 3.22: Gráfico QQ Normal para 100 dados gerados de uma distribuição t com 2 graus de liberdade.

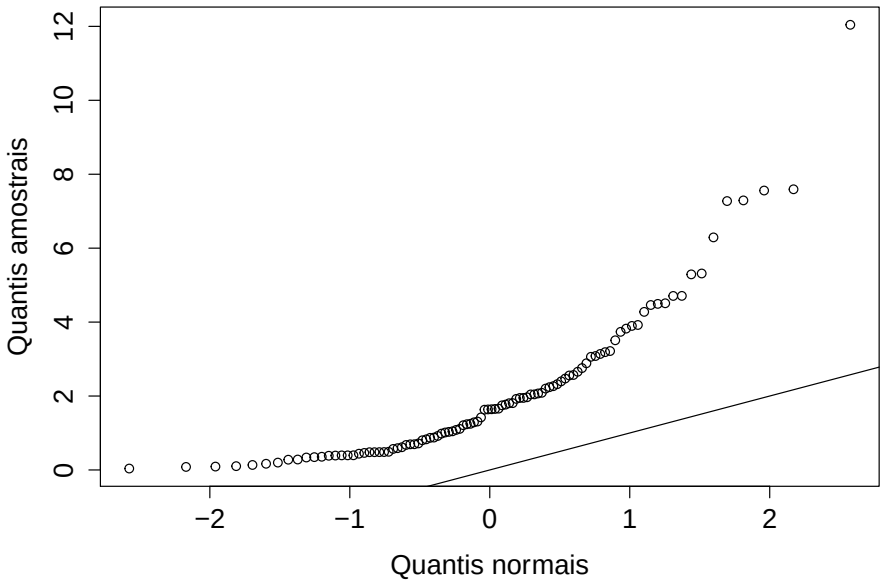


Figura 3.23: Gráfico QQ Normal para 100 dados gerados de uma distribuição qui-quadrado com 2 graus de liberdade.

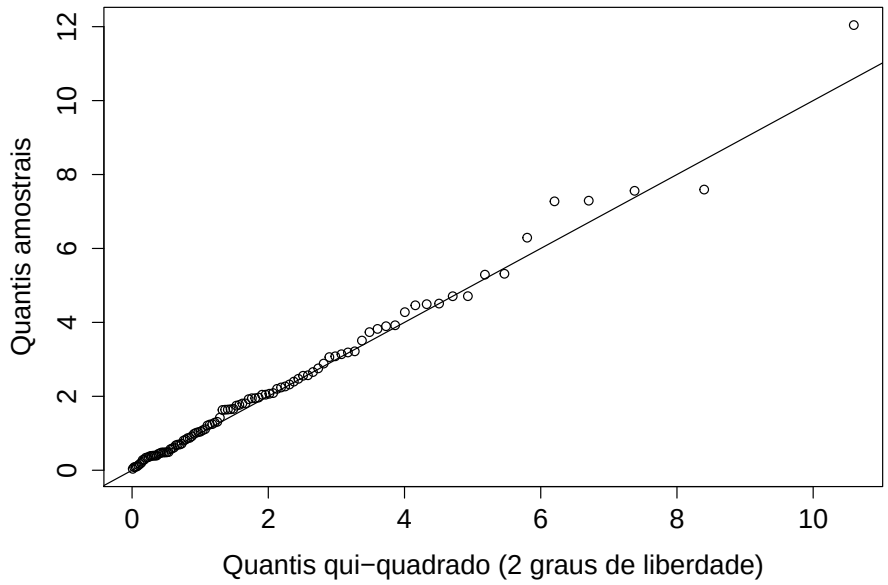


Figura 3.24: Gráfico QQ qui-quadrado para 100 dados gerados de uma distribuição qui-quadrado com 2 graus de liberdade.

3.8 Transformação de variáveis

Muitos procedimentos empregados em inferência estatística são baseados na suposição de que os valores de uma (ou mais) das variáveis de interesse provêm de uma distribuição Normal, ou seja, de que os dados associados a essa variável constituem uma amostra de uma população na qual a distribuição dessa variável é Normal. No entanto, em muitas situações de interesse prático, a distribuição dos dados na amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se quisermos utilizar os procedimentos talhados para análise de dados com distribuição Normal em situações nas quais a distribuição dos dados amostrais é sabidamente assimétrica, pode-se considerar uma transformação das observações com a finalidade de se obter uma distribuição "mais simétrica" e portanto, mais próxima da distribuição Normal. Uma transformação bastante usada com esse propósito é

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \ell n(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases} \tag{3.17}$$

Essa transformação com $0 < p < 1$ é apropriada para distribuições assimétricas à direita, pois valores grandes de x decrescem mais relativamente a valores pequenos. Para distribuições assimétricas à esquerda, basta tomar $p > 1$.

Normalmente, consideramos valores de p na sequência

$$\ldots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \ldots$$

e para cada um deles construímos gráficos apropriados (histogramas, *box-plots*) com os dados originais e transformados, com a finalidade de escolher o valor mais adequado para p. Hinkley (1977) sugere que para cada valor de p na sequência acima se calcule a média, a mediana e um estimador de escala (desvio padrão ou algum estimador robusto) e então se escolha o valor que minimiza

$$d_p = \frac{\text{média} - \text{mediana}}{\text{medida de escala}}, \tag{3.18}$$

que pode ser vista como uma medida de assimetria; numa distribuição simétrica, $d_p = 0$.

Exemplo 3.5. Consideremos a variável concentração de Fe obtidas em cascas de árvores da espécie *Tipuana tipu* disponível no arquivo `arvores`. Nas Figuras 3.25 e 3.26 apresentamos, respectivamente *boxplots* e histogramas para os valores originais da variável assim como para seus valores transformados por (3.17) com $p = 0, 1/3$ e $1/2$. Observamos que a transformação obtida com $p = 1/3$ é aquela que gera uma distribuição mais próxima da simetria.

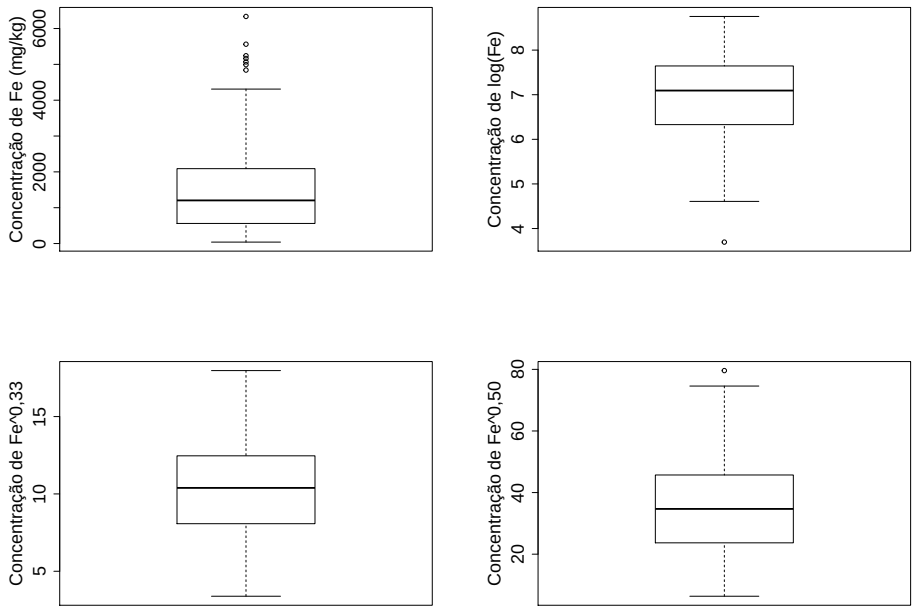


Figura 3.25: *Boxplots* com variável transformada.

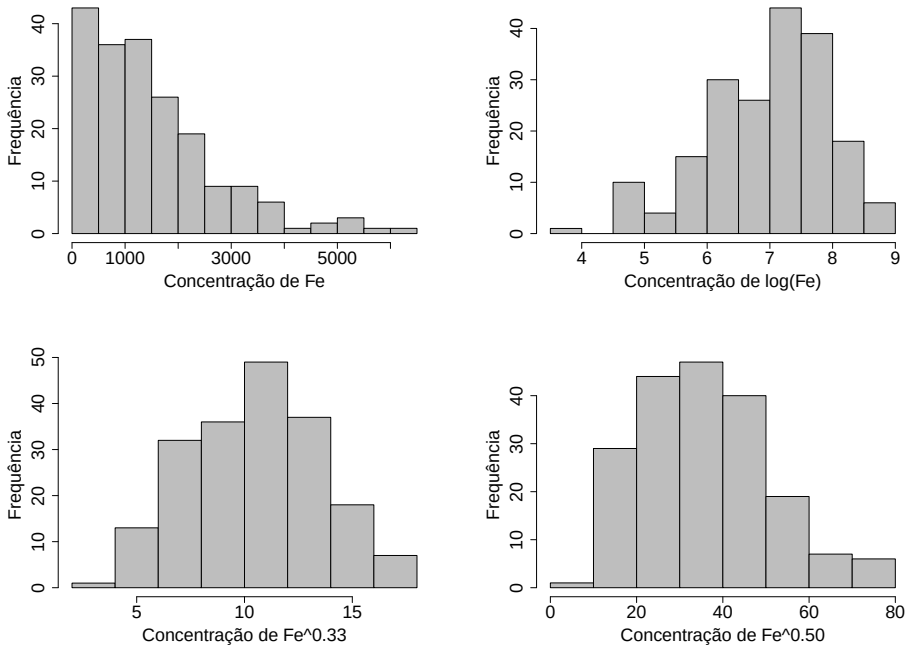


Figura 3.26: Histogramas com variável transformada.

Muitas vezes, (em **Análise de Variância**, por exemplo) é mais importante transformar os dados de modo a "estabilizar" a variância do que tornar a distribuição aproximadamente Normal. Um procedimento idealizado para essa finalidade é detalhado a seguir.

Suponhamos que X seja uma variável com $\mathrm{E}(X) = \mu$ e variância dependente da média, ou seja $\mathrm{Var}(X) = h^2(\mu)\sigma^2$, para alguma função $h(\cdot)$. Notemos que se $h(\mu) = 1$, então $\mathrm{Var}(X) = \sigma^2 =$ constante. Procuremos uma transformação $X \to g(X)$, de modo que $\mathrm{Var}[g(X)] =$ constante. Com esse propósito, consideremos uma expansão de Taylor de $g(X)$ ao redor de $g(\mu)$ até primeira ordem, ou seja

$$g(X) \approx g(\mu) + (X - \mu)g'(\mu).$$

em que g' denota a derivada de g em relação a μ. Então,

$$\mathrm{Var}[g(X)] \approx [g'(\mu)]^2 \mathrm{Var}(X) = [g'(\mu)]^2 [h(\mu)]^2 \sigma^2.$$

Para que a variância da variável transformada seja constante, devemos tomar

$$g'(\mu) = \frac{1}{h(\mu)}.$$

Por exemplo, se o desvio padrão de X for proporcional a μ, tomamos $h(\mu) = \mu$, logo $g'(\mu) = 1/\mu$ e portanto $g(\mu) = \log(\mu)$ e devemos considerar

a transformação (3.17) com $p = 0$, ou seja, $y^{(p)} = \log(x)$. Por outro lado, se a variância for proporcional à média, então usando o resultado acima, é fácil ver que a transformação adequada é $g(x) = \sqrt{x}$.

A transformação (3.19) é um caso particular das **transformações de Box-Cox** que são da forma

$$g(x) = \begin{cases} \frac{x^p - 1}{p}, & \text{se } p \neq 0 \\ \log(x), & \text{se } p = 0. \end{cases} \tag{3.19}$$

Veja Box e Cox (1964) para detalhes.

3.9 Desvio padrão e Erro padrão

Considere uma população para a qual a variável X tem média μ e variância σ^2. Imaginemos que um número grande, digamos M de amostras de tamanho n seja obtido dessa população. Denotemos por $X_{i1}, \ldots, X_{in}$ os n valores observados da variável X na i-ésima amostra, $i = 1, \ldots M$. Para cada uma das M amostras, calculemos as respectivas médias, denotadas por $\overline{X}_1, \ldots, \overline{X}_M$. Pode-se mostrar que a média e a variância da variável $\overline{X}$ (cujos valores são $\overline{X}_1, \ldots, \overline{X}_M$) são respectivamente μ e σ^2/n, *i.e.*, a média do conjunto das médias amostrais é igual à média da variável original X e a sua variância é menor (por um fator $1/n$) que a variância da variável original X. Além disso pode-se demostrar que o histograma da variável $\overline{X}$ tem o formato da distribuição Normal.

Note que a variância σ^2 é uma característica inerente à distribuição da variável original e não depende do tamanho da amostra. A variância σ^2/n da variável $\overline{X}$, depende do tamanho da amostra; quanto maior esse tamanho, mais concentrada (em torno de sua média, que é a mesma da variável original) será a sua distribuição. O desvio padrão da variável $\overline{X}$ é conhecido como **erro padrão** (da média). Detalhes podem ser obtidos em Bussab e Morettin (2017).

3.10 Intervalo de confiança

Em muitas situações, dados passíveis de análise estatística provêm de variáveis observadas em unidades de investigação (indivíduos, animais, corpos de prova, residências etc.) obtidas de uma população de interesse por meio de um processo de amostragem. Além de descrever e resumir os dados da amostra, há interesse em utilizá-los para fazer inferência sobre as distribuições populacionais dessas variáveis. Essas populações são, em geral, conceituais. Por exemplo, na avaliação de um determinado medicamento para diminuição da pressão arterial (X), a população de interesse não se resume aos pacientes (vivos) de um hospital ou de uma região; o foco é a população de indivíduos (vivos ou que ainda nascerão) que poderão utilizar essa droga. Nesse contexto, as características populacionais da diminuição da pressão arterial possivelmente provocada pela administração da droga

são desconhecidas e queremos estimá-la (ou adivinhá-las) com base nas suas características amostrais.

Não temos dúvidas sobre as características amostrais. Se a droga foi administrada a n pacientes e a redução média da pressão arterial foi de $\overline{X} = 10$ mmHg com desvio padrão $S = 3$ mmHG, não temos dúvida de que "em média" a droga reduziu a pressão arterial em 10 mmHg nos indivíduos da amostra. O problema é saber se o resultado obtido na amostra pode ser extrapolado para a população, ou seja se podemos utilizar $\overline{X}$ para estimar a média populacional (μ), que não conhecemos e que não conheceremos a não ser que seja possível fazer um censo. Obviamente, se foram tomados cuidados na seleção da amostra e se o protocolo experimental foi devidamente adotado, faz sentido supor que a redução média da pressão arterial induzida pela droga na população esteja próxima de 10 mmHg mas precisamos então especificar o que entendemos por "próxima". Isso pode ser feito por intermédio do cálculo da **margem de erro**, que, essencialmente, é uma medida de nossa incerteza na extrapolação dos resultados obtidos na amostra para a população de onde assumimos que foi obtida.

A margem de erro depende do processo amostral, do desvio padrão amostral S, do tamanho amostral n e é dada por $e = kS/\sqrt{n}$ em que k é uma constante que depende do modelo probabilístico adotado e da confiança com que pretendemos fazer a inferência. No caso de uma **amostra aleatória simples** de uma variável X obtida de uma população para a qual assumimos um modelo Normal, a margem de erro correspondente a um nível de confiança de 95% é $e = 1,96S/\sqrt{n}$. Com base nessa margem de erro, podemos construir um **intervalo de confiança** para a média populacional da variável X. Os limites inferior e superior para esse intervalo são, respectivamente, $\overline{X} - 1,96S/\sqrt{n}$ e $\overline{X} + 1,96S/\sqrt{n}$. Se considerássemos um grande número de amostras dessa população sob as mesmas condições, o intervalo construído dessa maneira conteria o verdadeiro (mas desconhecido) valor da média populacional (μ) em 95% dos casos.

Consideremos uma pesquisa eleitoral em que uma amostra de n eleitores é avaliada quanto à preferência por um determinado candidato. Podemos definir a variável resposta como $X = 1$ se o eleitor apoiar o candidato e $X = 0$ em caso contrário. A média amostral de X é a proporção amostral de eleitores favoráveis ao candidato, que representamos por $\widehat{p}$; sua variância, $p(1-p)/n$, pode ser estimada por $\widehat{p}(1-\widehat{p})/n$. Ver Exercício 32. Pode-se demonstrar que os limites inferior e superior de um intervalo de confiança com 95% de confiança para a proporção populacional p de eleitores favoráveis ao candidato são, respectivamente, $\widehat{p} - 1,96\sqrt{\widehat{p}(1-\widehat{p})/n}$ e $\widehat{p} + 1,96\sqrt{\widehat{p}(1-\widehat{p})/n}$. Se numa amostra de tamanho $n = 400$ obtivermos 120 eleitores favoráveis ao candidato, a proporção amostral será $\widehat{p} = 30\%$ e então podemos dizer que com 95% de confiança a proporção populacional p deve estar entre $25,5\%$ e $34,5\%$.

Detalhes técnicos sobre a construção de intervalos de confiança podem ser encontrados em Bussab e Morettin (2017) entre outros.

3.11 Notas de capítulo

1) **Variáveis contínuas**

Conceitualmente existem variáveis que podem assumir qualquer valor no conjunto dos números reais, como peso ou volume de certos produtos. Como na prática, todas as medidas que fazemos têm valores discretos, não é possível obter o valor π (que precisa ser expresso com infinitas casas decimais) por exemplo, para peso ou volume. No entanto, em geral, é possível aproximar as distribuições de frequências de variáveis com essa natureza por funções contínuas (como a distribuição Normal) e é essa característica que sugere sua classificação como variáveis contínuas.

2) **Amplitude de classes em histogramas**

Nos casos em que o histograma é obtido a partir dos dados de uma amostra de uma população com densidade $f(x)$, Freedman e Diaconis (1981) mostram que a escolha

$$h = 1,349\widetilde{s}\left(\frac{\log n}{n}\right)^{1/3} \tag{3.20}$$

minimiza o desvio máximo absoluto entre o histograma e a verdadeira densidade $f(x)$. Em (3.20), $\widetilde{s}$ é um estimador "robusto" do desvio padrão de X. Esse conceito será discutido adiante.

O pacote R usa como *default* o valor de h sugerido por Sturges (1926), dado por

$$h = \frac{W}{1 + 3,322\log(n)}, \tag{3.21}$$

sendo W a amplitude amostral e n o tamanho da amostra,

3) **Definição de histograma**

Consideremos um exemplo com K classes de comprimentos iguais a h. O número de classes a utilizar pode ser obtido aproximadamente como o quociente $(x_{(n)} - x_{(1)})/h$ em que $x_{(1)}$ é o valor mínimo e $x_{(n)}$, o valor máximo do conjunto de dados. Para que a área do histograma seja igual a 1, a altura do k-ésimo retângulo deve ser igual a f_k/h. Chamando $\tilde{x}_k$, $k = 1, \ldots, K$, os pontos médios dos intervalos das classes, o histograma pode ser construído a partir da seguinte função

$$\tilde{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} I(x - \tilde{x}_i; h/2), \tag{3.22}$$

em que $I(z; h)$ é a função indicadora do intervalo $[-h, h]$, ou seja,

$$I(z; h) = \begin{cases} 1, & \text{se } -h \leq z \leq h \\ 0, & \text{em caso contrário.} \end{cases}$$

Para representar essa construção, veja a Figura 3.27.

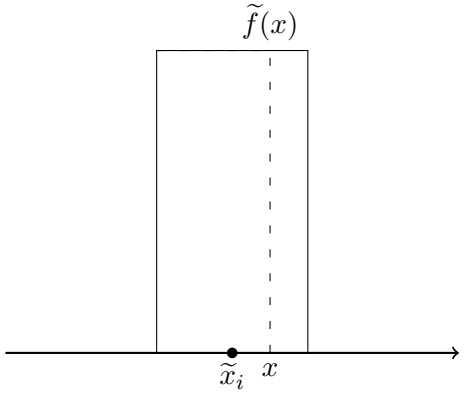


Figura 3.27: Detalhe para a construção de histogramas.

4) **Um estimador alternativo para o desvio padrão**

Pode-se verificar que, para uma distribuição Normal,

$$d_Q = 1,349\sigma$$

com σ^2 representando a variância. Logo, um estimador do desvio padrão populacional é

$$\widetilde{s} = \frac{d_Q}{1,349}.$$

Observe que substituindo $\widetilde{s}$ em (3.20), obtemos

$$h \approx d_Q \left(\frac{\log n}{n}\right)^{1/3},$$

que também pode ser utilizado para a determinação do número de classes de um histograma.

5) **Padronização de variáveis**

Para comparação de gráficos QQ, por exemplo, convém transformar variáveis com diferentes unidades de medida para deixá-las adimensionais, com a mesma média e mesma variância. Para esse efeito pode-se padronizar uma variável X com média μ e desvio padrão σ por meio da transformação $Z = (X - \mu)/\sigma$. Pode-se mostrar (ver Exercício 15) que a variável padronizada Z tem média 0 e desvio padrão 1, independentemente dos valores de μ e σ. Esse tipo de padronização também é útil em Análise de Regressão (ver Capítulo 6) quando se deseja avaliar a importância relativa de cada variável por meio dos coeficientes do modelo linear adotado.

6) **Bandas de confiança para gráficos QQ**

Seja $\{X_1, \ldots, X_n\}$ uma amostra aleatória de uma variável com função distribuição F desconhecida. A estatística de Kolmogorov-Smirnov [ver Wayne (1990, páginas 319-330), por exemplo], dada por

$$S = sup_x|F_n(x) - F_0(x)|$$

em que F_n é correspondente função distribuição empírica, serve para testar a hipótese $F = F_0$. A distribuição da estatística S é tabelada de forma que se pode obter o valor crítico s tal que $P(S \leq s) = 1 - \alpha$, $0 < \alpha < 1$. Isso implica que para qualquer valor x temos $|F_n(x) - F_0(x)| \leq s = 1 - \alpha$ ou seja, que com probabilidade $1 - \alpha$ temos $F_n(x) - s \leq F_0(x) \leq F_n(x) + s$. Consequentemente, os limites inferior e superior de um intervalo de confiança com coeficiente de confiança $1 - \alpha$ para F são respectivamente, $F_n(x) - s$ e $F_n(x) + s$. Essas bandas conterão a função distribuição Normal $N(\mu, \sigma^2)$ se

$$F_n(x) - s \leq \Phi[(x - \mu)/\sigma] \leq F_n(x) + s$$

o que equivale a ter uma reta contida entre os limites da banda definida por

$$\Phi^{-1}[F_n(x) - s] \leq (x - \mu)/\sigma \leq \Phi^{-1}[F_n(x) + s].$$

Para a construção do gráfico QQ esses valores são calculados nos pontos $X_1, \ldots, X_n$.

7) **Curtose**

Seja X uma variável aleatória qualquer, com média μ e variância σ^2. A **curtose** de X é definida por

$$K(X) = \mathrm{E}\left[\frac{(X - \mu)^4}{\sigma^4}\right]. \tag{3.23}$$

Para uma distribuição Normal, $K = 3$, razão pela qual a quantidade $e(X) = K(X) - 3$ é chamada de **excesso de curtose**. Distribuições com caudas pesadas têm curtose maior do que 3, podendo ser infinita.

Para uma amostra $\{X_1, \ldots, X_n\}$ de X, considere o r-Ésimo momento amostral

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^r,$$

em que $\widehat{\mu} = \overline{X}$. Substituindo os momentos verdadeiros de X pelos respectivos momentos amostrais, obtemos um estimador da curtose, nomeadamente

$$\widehat{K}(X) = \frac{m_4}{m_2^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\widehat{\sigma}}\right)^4, \tag{3.24}$$

em que $\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$. Então um estimador para o excesso de curtose é $\widehat{e}(X) = \widehat{K}(X) - 3$. Pode-se provar que, para uma amostra suficientemente grande de uma distribuição Normal,

$$\hat{K} \sim \mathcal{N}(3, 24/n). \tag{3.25}$$

3.12 Exercícios

1) O arquivo `rehabcardio` contém informações sobre um estudo de reabilitação de pacientes cardíacos. Elabore um relatório indicando possíveis inconsistências na matriz de dados e faça uma análise descritiva de todas as variáveis do estudo.

2) Considere um conjunto de dados $\{X_1, \ldots, X_n\}$.

   a) Calcule a média e a variância de $W_1, \ldots, W_n$ em que $W_i = X_i + k$ com k denotando uma constante.

   b) Calcule a média e a variância de $V_1, \ldots, V_n$ em que $V_i = kX_i$ com k denotando uma constante.

3) Calcule as medidas de posição e dispersão estudadas para os dados apresentados na Tabela 3.1 cujos dados estão disponíveis no arquivo `ceagfgv`.

4) Determine o valor de h dado por (3.20) para os dados do Exemplo 3.4.

5) Prove que S^2 dado por (3.10) é um estimador não enviesado da variância populacional.

6) Considere do arquivo `vento`. Observe o valor atípico $61,1$, que na realidade ocorreu devido a forte tempestade no dia 2 de dezembro. Calcule as medidas de posição e dispersão dadas na Seção 3.3. Comente os resultados.

7) Construa gráficos ramo-e-folhas e *boxplot* para os dados do Exercício 6.

8) Usando o pacote `R`, analise a variável "Temperatura" do arquivo `poluicao`.

9) Idem, para a variável "Salário de administradores", disponível no arquivo `salarioso`.

10) Construa um gráfico ramo-e-folhas e um *boxplot* para os dados de precipitação atmosférica de Fortaleza disponíveis no arquivo `precipitacao`.

11) Transforme os dados do Exercício 6 por meio de (3.17) com $p = 0, 1/4, 1/3,\ 1/2, 3/4$ e escolha a melhor alternativa de acordo com a medida d_p dada em (3.18).

12) Construa gráficos de quantis e simetria para os dados de manchas solares disponíveis no arquivo `manchas`.

13) Prove a relação (3.8). Como ficaria essa expressão para S^2?

14) Uma outra medida de assimetria é

$$A = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1},$$

que é igual a zero no caso de uma distribuição simétrica. Calcule-a para os dados do Exercício 6.

15) Considere os valores $X_1, \ldots, X_n$ de uma variável X, com média $\overline{X}$ e desvio padrão S. Mostre que a variável Z, cujos valores são $Z_i = (X_i - \overline{X})/S$, $i = 1, \ldots, n$ tem média 0 e desvio padrão 1.

16) Os dados disponíveis no arquivo `endometriose` são provenientes de um estudo sobre endometriose onde o objetivo é verificar se existe diferença entre os grupos de doentes e controle quanto a algumas características observadas.

    a) O pesquisador responsável pelo estudo tem a seguinte pergunta: as pacientes doentes apresentam mais dor na menstruação do que as pacientes não doentes? Que tipo de análise você faria para responder essa pergunta? Faça-a e tire suas conclusões.

    b) Compare as distribuições das variáveis idade e concentração de PCR durante a menstruação (PCRa) para indivíduos dos grupos controle e doente utilizando medidas resumo (mínimo, máximo, quartis, mediana, média, desvio padrão, etc), *boxplots*, histogramas, gráficos de médias e gráficos QQ. Como você considerou os valores $< 0,5$ da variável PCRa nesses cálculos? Você sugeriria uma outra maneira para considerar tais valores?

    c) Compare a distribuição da variável número de gestações para os dois grupos por intermédio de uma tabela de frequências e do teste qui-quadrado de Pearson. Utilize um método gráfico para representar essa tabela.

17) Os dados apresentados na Figura 3.28 referem-se aos instantes nos quais o centro de controle operacional de estradas rodoviárias recebeu chamados solicitando algum tipo de auxílio em duas estradas num determinado dia.

Estrada 1	12:07:00 AM	12:58:00 AM	01:24:00 AM	01:35:00 AM	02:05:00 AM
	03:14:00 AM	03:25:00 AM	03:46:00 AM	05:44:00 AM	05:56:00 AM
	06:36:00 AM	07:26:00 AM	07:48:00 AM	09:13:00 AM	12:05:00 PM
	12:48:00 PM	01:21:00 PM	02:22:00 PM	05:30:00 PM	06:00:00 PM
	07:53:00 PM	09:15:00 PM	09:49:00 PM	09:59:00 PM	10:53:00 PM
	11:27:00 PM	11:49:00 PM	11:57:00 PM		
Estrada 2	12:03:00 AM	01:18:00 AM	04:35:00 AM	06:13:00 AM	06:59:00 AM
	08:03:00 AM	10:07:00 AM	12:24:00 PM	01:45:00 PM	02:07:00 PM
	03:23:00 PM	06:34:00 PM	07:19:00 PM	09:44:00 PM	10:27:00 PM
	10:52:00 PM	11:19:00 PM	11:29:00 PM	11:44:00 PM	

Figura 3.28: Planilha com instantes de realização de chamados solicitando auxílio em estradas.

    a) Construa um histograma para a distribuição de frequências de chamados em cada uma das estradas.

b) Calcule os intervalos de tempo entre as sucessivas chamadas e descreva-os, para cada uma das estradas, utilizando medidas resumo gráficos do tipo *boxplot*. Existe alguma relação entre o tipo de estrada e o intervalo de tempo entre as chamadas?

c) Por intermédio de um gráfico do tipo QQ, verifique se a distribuição da variável "Intervalo de tempo entre as chamadas"/ em cada estrada é compatível com um modelo Normal. Faça o mesmo para um modelo exponencial. Compare as distribuições de frequências correspondentes às duas estradas.

18) As notas finais de um curso de Estatística foram: 7, 5, 4, 5, 6, 3, 8, 4, 5, 4, 6, 4, 5, 6, 4, 6, 6, 3, 8, 4, 5, 4, 5, 5 e 6.

a) Determine a mediana, os quartis e a média.

b) Separe o conjunto de dados em dois grupos denominados **aprovados**, com nota pelo menos igual a 5, e **reprovados**. Compare a variância desses dois grupos.

19) Considere o seguinte resumo descritivo da pulsação de estudantes com atividade física intensa e fraca:

Atividade	N	Média	Mediana	DP	Min	Max	Q1	Q3
Intensa	30	79,6	82	10,5	62	90	70	79
Fraca	30	73,1	70	9,6	58	92	63	77

DP: desvio padrão
Q1: primeiro quartil
Q3: terceiro quartil

Qual das seguintes afirmações está correta:

a) 5% e 50% dos estudantes com atividade física intensa e fraca, respectivamente, tiveram pulsação inferior a 70.
b) A pulsação de um estudante com fraca atividade física é provavelmente inferior a 63.
c) A atividade física não tem efeito na média da pulsação dos estudantes.
d) Quaisquer 15 estudantes com fraca atividade física têm pulsação inferior a 70.
e) Nenhuma das respostas anteriores.

20) Considere os gráficos *boxplot* da Figura 3.29. Quais deles correspondem às pulsações dos estudantes submetidos a atividade física intensa e fraca? a) A e B b) B e D c) A e C d) B e C

21) Os histogramas apresentados na Figura 3.30 mostram a distribuição das temperaturas (^{o}C) ao longo de vários dias de investigação para duas regiões (R1 e R2). Podemos dizer que:

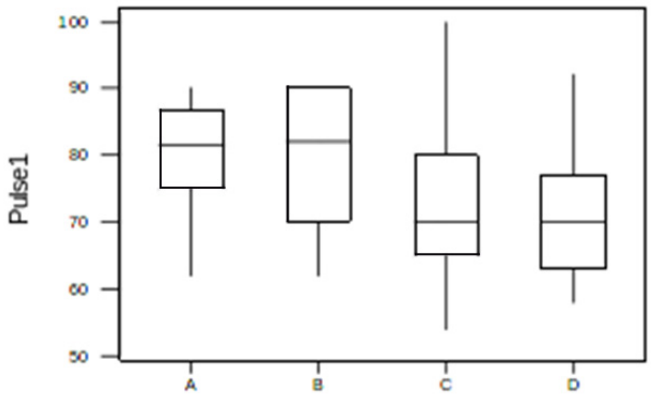


Figura 3.29: *Boxplots* para o Exercício 20.

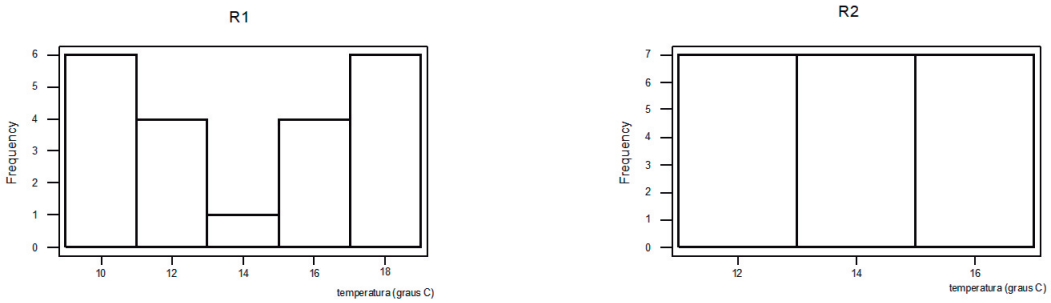


Figura 3.30: Histogramas para o Exercício 20.

a) As temperaturas das regiões R1 e R2 têm mesma média e mesma variância.

b) Não é possível compara as variâncias.

c) A temperatura média da região R2 é maior que a de R1.

d) As temperaturas das regiões R1 e R2 têm mesma média e variância diferentes.

e) Nenhuma das respostas anteriores.

22) Na companhia A, a média dos salários é 10000 unidades e o 3º quartil é 5000.

a) Se você se apresentasse como candidato a funcionário nessa firma e se o seu salário fosse escolhido ao acaso entre todos os possíveis salários, o que seria mais provável: ganhar mais ou menos que 5000 unidades?

b) Suponha que na companhia B a média dos salários seja 7000 unidades, a variância praticamente zero e o salário também seja

escolhido ao acaso. Em qual companhia você se apresentaria para procurar emprego, com base somente no salário?

23) Num conjunto de dados, o primeiro quartil é 10, a mediana é 15 e o terceiro quartil é 20. Indique quais das seguintes afirmativas são verdadeiras, justificando sua resposta:

a) A distância interquartis é 5.

b) O valor 32 seria considerado *outlier* segundo o critério utilizado na construção do *boxplot*.

c) A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior.

d) O valor mínimo é maior do que zero.

e) Nenhuma das respostas anteriores.

24) A bula de um medicamento A para dor de cabeça afirma que o tempo médio para que a droga faça efeito é de 60 seg com desvio padrão de 10 seg. A bula de um segundo medicamento B afirma que a média correspondente é de 60 seg com desvio padrão de 30 seg. Sabe-se que as distribuições são simétricas. Indique quais das seguintes afirmativas são verdadeiras, justificando sua resposta:

a) Os medicamentos são totalmente equivalentes com relação ao tempo para efeito pois as médias são iguais.

b) Com o medicamento A, a probabilidade de cura de sua dor de cabeça antes de 40 seg é maior do que com o medicamento B.

c) Com o medicamento B, a probabilidade de você ter sua dor de cabeça curada antes de 60 seg é maior que com o medicamento A.

25) Na tabela abaixo estão indicadas as durações de 335 lâmpadas.

Duração (horas)	Número de lâmpadas
$0 \vdash 100$	82
$100 \vdash 200$	71
$200 \vdash 300$	68
$300 \vdash 400$	56
$400 \vdash 500$	43
$500 \vdash 800$	15

a) Esboce o histograma correspondente.

b) Calcule os quantis de ordem p=0,1, 0,3, 0,5, 0,7 e 0,9.

26) A tabela abaixo representa a distribuição do número de dependentes por empregados da Empresa Mirante.

Dependentes	Empregados
1	40
2	50
3	30
4	20
5	10
Total	150

A mediana, média e moda são, respectivamente:

a) 50; 15; 50 b) 1; 2,1; 1 c) 50,5; 50; 50 d) 1; 1; 1

27) Com relação ao Exercício 26, qual a porcentagem de empregados da Empresa Mirante com 2 ou mais dependentes?

a) 40,1% b) 50,1% c) 60,3% d) 73,3%

28) Os dados encontrados no arquivo `esforco` são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca. As variáveis medidas durante a realização do teste foram observadas em quatro momentos distintos: repouso (REP), limiar anaeróbio (LAN), ponto de compensação respiratório (PCR) e pico (PICO). As demais variáveis são referentes às características demográficas e clínicas dos pacientes e foram registradas uma única vez.

a) Descreva a distribuição da variável consumo de oxigênio (VO2) em cada um dos quatro momentos de avaliação utilizando medidas resumo (mínimo, máximo, quartis, mediana, média, desvio padrão, etc), *boxplots* e histogramas. Você identifica algum paciente com valores de consumo de oxigênio discrepantes? Interprete os resultados.

b) Descreva a distribuição da classe funcional NYHA por meio de uma tabela de frequências. Utilize um método gráfico para representar essa tabela.

29) Num estudo na área de Oncologia, o número de vasos que alimentam o tumor está resumido na seguinte tabela.

Tabela 3.11: Distribuição de frequências do número de vasos que alimentam o tumor

Número de vasos	Frequência
$0 \vdash 5$	8 (12%)
$5 \vdash 10$	23 (35%)
$10 \vdash 15$	12 (18%)
$15 \vdash 20$	9 (14%)
$20 \vdash 25$	8 (12%)
$25 \vdash 30$	6 (9%)
Total	66 (100%)

Indique a resposta correta.

a) O primeiro quartil é 25%.

b) A mediana está entre 10 e 15.

c) O percentil de ordem 10% é 10.

d) A distância interquartis é 50.

e) Nenhuma das respostas anteriores.

30) Utilizando o mesmo enunciado da questão anterior, indique a resposta correta:

a) Não é possível estimar nem a média nem a variância com esses dados.

b) A variância é menor que 30.

c) A média estimada é 12.8.

d) Em apenas 35% dos casos, o número de vasos é maior que 10.

e) Nenhuma das anteriores.

31) Em dois estudos realizados com o objetivo de estimar o nível médio de colesterol total para uma população de indivíduos saudáveis observaram-se os dados indicados na tabela seguinte:

Tabela 3.12: Medidas descritivas dos estudos A e B

Estudo	n	Média	Desvio padrão
A	100	160 mg/dL	60 mg/dL
B	49	150 mg/dL	35 mg/dL

- Indique a resposta correta:

a) Não é possível estimar o nível médio de colesterol populacional só com esses dados.

b) Se os dois estudos foram realizados com amostras da mesma população não deveria haver diferença entre os desvios padrões amostrais.

c) Com os dados do estudo B, o colesterol médio populacional pode ser estimado com mais precisão do que com os dados do estudo A.

d) Ambos os estudos sugerem que a distribuição do colesterol na população é simétrica.

e) Nenhuma das respostas anteriores.

32) Considere uma amostra aleatória simples $X_1, \ldots, X_n$ de uma variável X que assume o valor 1 com probabilidade $0 < p < 1$ e o valor 0 com probabilidade $1 - p$. Seja $\widehat{p} = n^{-1} \sum_{i=1}^{n} X_i$. Mostre que

i) $\mathrm{E}(X_i) = p$ e $\mathrm{Var}(X_i) = p(1-p)$.

ii) $\mathrm{E}(\widehat{p}) = p$ e $\mathrm{Var}(\widehat{p}) = p(1-p)/n$.

iii) $0 < \mathrm{Var}(X_i) < 0,25$.

Com base nesses resultados, utilize o Teorema Limite Central [ver Sen et al. (2009), por exemplo] para construir um intervalo de confiança aproximado conservador (*i.e.* com a maior amplitude possível) para p. Utilize o Teorema de Sverdrup [ver Sen et al. (2009), por exemplo] para construir um intervalo de confiança aproximado para p com amplitude menor que a do intervalo mencionado acima.

33) Com a finalidade de entender a diferença entre "desvio padrão" e "erro padrão",

a) Simule 10000 dados de uma distribuição normal com média 12 e desvio padrão 4. Construa o histograma correspondente, calcule a média e o desvio padrão amostrais e compare os valores obtidos com aqueles utilizados na geração dos dados.

b) Simule 500 amostras de tamanho $n = 4$ dessa população. Calcule a média amostral de cada amostra, construa o histograma dessas médias e estime o correspondente desvio padrão (que é o erro padrão da média).

c) Repita os passos a) e b) com amostras de tamanhos $n = 9$ e $n = 100$. Comente os resultados.

d) Repita os passos a) - c) simulando amostras de uma distribuição qui-quadrado com 3 graus de liberdade.

Capítulo 4

Análise de dados de duas variáveis

> Não é a linha reta que me atrai. Dura, inflexível, criada pelo homem. O que me atrai é a curva livre e natural.
>
> Oscar Niemeyer

4.1 Introdução

Neste capítulo trataremos da análise descritiva da **associação** entre duas variáveis. *Grosso modo*, dizemos que existe associação entre duas variáveis, se o conhecimento do valor de uma delas nos dá alguma informação sobre a distribuição da outra. Podemos estar interessados, por exemplo, na associação entre o grau de instrução e o salário de um conjunto de indivíduos. Nesse caso, esperamos que quanto maior seja o nível educacional de um indivíduo, maior deve ser o seu salário. Como na análise de uma única variável, também discutiremos o emprego de tabelas e gráficos para representar a distribuição conjunta das variáveis de interesse além de medidas resumo para avaliar o tipo e a magnitude da associação. Podemos destacar três casos:

i) as duas variáveis são qualitativas;
ii) as duas variáveis são quantitativas;
iii) uma variável é qualitativa e a outra é quantitativa.

As técnicas para analisar dados nos três casos acima são distintas. No primeiro caso, a análise é baseada no número de unidades de investigação (amostrais, por exemplo) em cada cela de uma tabela de dupla entrada. No segundo caso, as observações são obtidas por mensurações, e técnicas envolvendo gráficos de dispersão ou de quantis são apropriadas. Na terceira situação, podemos comparar as distribuições da variável quantitativa para cada categoria da variável qualitativa.

Aqui, é importante considerar a classificação das variáveis segundo outra característica, intimamente ligada à forma de coleta dos dados. **Variáveis**

explicativas são aquelas cujas categorias ou valores são fixos, seja por planejamento ou seja por condicionamento. **Variáveis respostas** são aquelas cujas categorias ou valores são aleatórios.

Num estudo em que se deseja avaliar o efeito da quantidade de aditivo adicionado ao combustível no consumo de automóveis, cada um de 3 conjuntos de 5 automóveis (de mesmo modelo) foi observado sob o tratamento com uma de 4 quantidades de aditivo. O consumo (em km/L) foi avaliado após um determinado período de tempo. Nesse contexto, a variável qualitativa "Quantidade de aditivo" (com 4 categorias) é considerada como explicativa e a variável quantitativa "Consumo de combustível" é classificada como resposta.

Num outro cenário, em que se deseja estudar a relação entre o nível sérico de colesterol (mg/dL) e o nível de obstrução coronariana (em %), cada paciente de um conjunto de 30 selecionados de um determinado hospital foi submetido a exames de sangue e tomográfico. Nesse caso, tanto a variável "Nível sérico de colesterol" quanto a variável "Nível de obstrução coronariana" devem ser encaradas como respostas. Mesmo assim, sob um **enfoque condicional**, em que se deseja avaliar o "Nível de obstrução coronariana" para pacientes com um determinado "Nível sérico de colesterol" a primeira é encarada como variável resposta e a segunda, como explicativa.

4.2 Duas variáveis qualitativas

Nessa situação, as classes das duas variáveis podem ser organizadas numa tabela de dupla entrada, em que as linhas correspondem aos níveis de uma das variáveis e as colunas, aos níveis da outra.

Exemplo 4.1. Os dados disponíveis no arquivo `coronarias` contém dados do projeto "Fatores de risco na doença aterosclerótica coronariana", coordenado pela Dra. Valéria Bezerra de Carvalho (INTERCOR). O arquivo contém informações sobre cerca de 70 variáveis observadas em 1500 indivíduos.

Para fins ilustrativos, consideramos apenas duas variáveis qualitativas nominais, a saber, hipertensão arterial (X) e insuficiência cardíaca (Y), ambas codificadas com os atributos 0=não tem e 1=tem observadas em 50 pacientes. Nesse contexto, as duas variáveis são classificadas como respostas. A Tabela 4.1 contém a **distribuição de frequências conjunta** das duas variáveis. Essa distribuição indica, por exemplo, que 12 indivíduos têm hipertensão arterial **e** insuficiência cardíaca, ao passo que 4 indivíduos não têm hipertensão **e** têm insuficiência cardíaca. Para efeito de comparação com outros estudos envolvendo as mesmas variáveis mas com número de pacientes diferentes, convém expressar os resultados na forma de porcentagens. Com esse objetivo, podemos considerar porcentagens em relação ao total da tabela, em relação ao total de suas linhas ou em relação ao total de suas colunas. Na Tabela 4.2 apresentamos as porcentagens correspondentes à Tabela 4.1 calculadas em relação ao seu total. Os dados da Tabela 4.2

Tabela 4.1: Distribuição conjunta das variáveis $X=$ hipertensão arterial e $Y=$ insuficiência cardíaca

Insuficiência	Hipertensão arterial		Total
cardíaca	Tem	Não tem	
Tem	12	4	16
Não tem	20	14	34
Total	32	18	50

Tabela 4.2: Porcentagens para os dados da Tabela 4.1 em relação ao seu total

Insuficiência	Hipertensão		Total
cardíaca	Tem	Não tem	
Tem	24%	8%	32%
Não tem	40%	28%	68%
Total	64%	36%	100%

permitem-nos concluir que 24% dos indivíduos avaliados têm hipertensão **e** insuficiência cardíaca, ao passo que 36% dos indivíduos avaliados não sofrem de hipertensão.

Também podemos considerar porcentagens calculadas em relação ao total das colunas como indicado na Tabela 4.3. Com base nessa tabela, po-

Tabela 4.3: Porcentagens com totais nas colunas

Insuficiência	Hipertensão		Total
cardíaca	Tem	Não tem	
Tem	$37,5\%$	$22,2\%$	32%
Não tem	$62,5\%$	$77,8\%$	68%
Total	$100,0\%$	$100,0\%$	$100,0\%$

demos dizer que independentemente do *status* desses indivíduos quanto à presença de hipertensão, 32% têm insuficiência cardíaca. Esse cálculo de porcentagens é mais apropriado quando uma das variáveis é considerada explicativa e a outra, considerada resposta.

No exemplo, apesar de o planejamento do estudo indicar que as duas variáveis são respostas (a frequência de cada uma delas não foi fixada *a priori*), para efeito da análise, uma delas (Hipertensão arterial) será considerada explicativa. Isso significa que não temos interesse na distribuição de frequências de hipertensos ou não dentre os 50 pacientes avaliados apesar de ainda querermos avaliar a associação entre as duas variáveis. Nesse caso, dizemos que a variável "Hipertensão arterial" é considerada explicativa **por condicionamento**. Se houvéssemos fixado *a priori* um certo número de hipertensos e outro de não hipertensos e então observado quantos dentre cada um desses dois grupos tinham ou não insuficiência cardíaca, diríamos

que a variável "Hipertensão arterial" seria considerada explicativa **por planejamento**. Nesse caso, apenas as porcentagens calculadas como na Tabela 4.3 fariam sentido. Uma enfoque análogo poderia ser adotado se fixássemos as frequências de "Insuficiência cardíaca" e considerássemos "Hipertensão arterial" como variável resposta. Nesse caso, as porcentagens deveriam ser calculadas em relação ao total das linhas da tabela.

Tabelas com a natureza daquelas descritas acima são chamadas de **tabelas de contingência** ou **tabelas de dupla entrada**. Essas tabelas são classificadas como tabelas $r \times c$ em que r é o número de linhas e c é o número de colunas. As tabelas apresentadas acima são, portanto, tabelas 2×2. Se a variável X tiver 3 categorias e a variável Y, 4 categorias, a tabela de contingência correspondente será uma tabela 3×4.

Suponha, agora, que queiramos verificar se as variáveis X e Y são associadas. No caso da Tabela 4.2 (em que as duas variáveis são consideradas respostas), dizer que as variáveis não são associadas corresponde a dizer que essas variáveis são (estatisticamente) independentes. No caso da Tabela 4.3 (em que uma variáveis é explicativa, quer por condicionamento, quer por planejamento e a outra é considerada resposta), dizer que as variáveis não são associadas corresponde a dizer que as distribuições de frequências da variável resposta ("Insuficiência cardíaca") para indivíduos classificados em cada categoria da variável explicativa ("Hipertensão arterial") são homogêneas.

Nas Tabelas 4.2 ou 4.3, por exemplo, há diferenças, que parecem não ser "muito grandes", o que nos leva a conjecturar que **para a população de onde esses indivíduos foram extraídos**, as duas variáveis não são associadas. Para avaliar essa conjectura, pode-se construir um **teste formal** para essa **hipótese de inexistência de associação** (independência ou homogeneidade), nomeadamente

$$H \ : \ X \text{ e } Y \ \text{ são não associadas.}$$

Convém sempre lembrar que a hipótese H refere-se à associação entre as variáveis X e Y na população (geralmente conceitual) de onde foi extraída uma amostra cujos dados estão dispostos na tabela. Não há dúvidas de que na tabela, as distribuições de frequências correspondentes às colunas rotuladas por "Tem" e "Não tem" hipertensão são diferentes.

Se as duas variáveis não fossem associadas, deveríamos ter porcentagens iguais nas colunas da Tabela 4.3 rotuladas "Tem" e "Não tem". Podemos então calcular as frequências esperadas nas celas da tabela admitindo que a hipótese H seja verdadeira. Por exemplo, o valor $10,2$ corresponde a 32% de 32, ou ainda, $10,2 = (32 \times 16)/50$. Observe que os valores foram arredondados segundo a regra usual e que as somas de linhas e colunas são as mesmas da Tabela 4.1.

Tabela 4.4: Valores esperados das frequências na Tabela 4.3 sob H

Insuficiência Cardíaca	Hipertensão		Total
	Tem	Não Tem	
Tem	10,2	5,8	16
Não Tem	21,8	12,2	34
Total	32	18	50

Chamando os valores observados por o_i e os esperados por e_i, $i = 1, 2, 3, 4$, podemos calcular os **resíduos** $r_i = o_i - e_i$ e verificar que $\sum_i r_i = 0$. Uma medida da discrepância entre o valores observados e aqueles esperados sob a hipótese H é a chamada estatística ou **qui-quadrado** de Pearson, dada por

$$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i}. \tag{4.1}$$

No nosso exemplo, $\chi^2 = 1,3$. Quanto maior esse valor, maior a **evidência** de que a hipótese H não é verdadeira, ou seja de que as variáveis X e Y são associadas (na população de onde foi extraída a amostra que serviu de base para os cálculos). Resta saber se o valor observado é suficientemente grande para concluirmos que H não é verdadeira. Com essa finalidade, teríamos que fazer um teste formal, o que não será tratado nesse texto. Pode-se mostrar que sob a hipótese H, a estatística (4.1) segue uma distribuição qui-quadrado com número de graus de liberdade igual a $(r-1)(c-1)$ de forma que a decisão de rejeitar ou não a hipótese pode ser baseada nessa distribuição. Veja Bussab e Morettin (2017), entre outros, para detalhes.

A própria estatística de Pearson poderia servir como medida da intensidade da associação mas o seu valor aumenta com o tamanho da amostra; uma alternativa para corrigir esse problema é o **coeficiente de contingência de Pearson**, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{4.2}$$

Para o Exemplo 4.1, temos que $C = \sqrt{1,3/(1,3+50)} = 0,16$, que é um valor pequeno. Esse coeficiente tem interpretação semelhante à do **coeficiente de correlação**, a ser tratado na próxima seção. Mas enquanto esse último varia entre -1 e $+1$, o coeficiente C, como definido acima, não varia entre 0 e 1 (em módulo). O valor máximo de C depende do número de linhas, r, e do número de colunas, c, da tabela de contingência. Uma modificação de C é o **coeficiente de Tschuprov**,

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(c-1)}}}, \tag{4.3}$$

que atinge o valor máximo igual a 1 quando $r = c$. No Exemplo 4.1, $T = 0,16$.

Em estudos que envolvem a mesma característica observada sob duas condições diferentes (gerando duas variáveis, X e Y, cada uma correspondendo à observação da característica sob uma das condições), sabe-se *a priori* que elas são associadas e o interesse recai sobre a avaliação da **concordância** dos resultados em ambas as condições. Nesse contexto, consideremos um exemplo em que as redações de 445 alunos são classificadas por cada um de dois professores (A e B) como "ruim", "média" ou "boa" com os resultados resumidos na Tabela 4.5.

Tabela 4.5: Frequências de redações classificadas por dois professores

	Professor B		
Professor A	ruim	média	boa
ruim	192	1	5
média	2	146	5
boa	11	12	71

Se todas as frequências estivessem dispostas ao longo da diagonal principal da tabela, diríamos que a haveria completa concordância entre os dois professores com relação ao critério de avaliação das redações. Como em geral isso não acontece, é conveniente construir um índice para avaliar a magnitude da concordância. O índice

$$\kappa = \frac{\sum_{i=1}^{3} p_{ii} - \sum_{i=1}^{3} p_{i+}p_{+i}}{1 - \sum_{i=1}^{3} p_{i+}p_{+i}},$$

denominado $\boldsymbol{\kappa}$ de Cohen (1960) é o mais utilizado com esse propósito. Nessa expressão, p_{ij} representa frequência relativa associada à cela correspondente à linha i e coluna j da tabela e p_{i+} e p_{+j} representam a soma das frequências relativas associadas à linha i e coluna j, respectivamente. O numerador corresponde à diferença entre a soma das frequências relativas correspondentes à diagonal principal da tabela e a soma das frequências relativas que seriam esperadas se as avaliações dos dois professores fossem independentes. Portanto, quando há concordância completa, $\sum_{i=1}^{3} p_{ii} = 1$, o numerador é igual ao denominador e o valor do índice de Cohen é $\kappa = 1$. Quando os dois professores não concordam em nenhuma das avaliações, $\kappa < 0$. Para os dados da Tabela 4.5 temos $\kappa = 0.87$ sugerindo uma "boa" concordância entre as avaliações dos dois professores. Embora o nível de concordância medido pelo índice κ seja subjetivo e dependa da área em que se realiza o estudo gerador dos dados, há autores que sugerem modelos de classificação, como aquele proposto por Viera and Garrett (2005) e reproduzido na Tabela 4.6

Tabela 4.6: Níveis de concordância segundo o índice κ de Cohen

κ de Cohen	Nível de concordância
< 0	Menor do que por acaso
0,01–0,20	Leve
0,21–0,40	Razoável
0,41–0,60	Moderado
0,61–0,80	Substancial
0,81–0,99	Quase perfeito

Para salientar discordâncias mais extremas como no exemplo, um professor classifica a redação como "ruim" e o outro como "boa", pode-se considerar o índice κ ponderado, definido como

$$\kappa_p = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}p_{ij} - \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}p_{i+}p_{+j}},$$

em que $w_{ij}, i, j = 1, 2, 3$ é um conjunto de pesos convenientes. Por exemplo, $w_{ii} = 1$, $w_{ij} = 1 - (i-j)/(I-1)$ em que I é o número de categorias em que a característica de interesse é classificada. Para o exemplo, $w_{12} = w_{21} = w_{23} = w_{32} = 1 - 1/2 = 1/2$, $w_{13} = w_{31} = 1 - 2/2 = 0$.

Exemplo 4.2. A Tabela 4.7 contém dados sobre o tipo de escola cursada por alunos aprovados no vestibular da USP em 2018.

Tabela 4.7: Frequências de alunos aprovados no vestibular de 2018 na USP

Tipo de escola	Área do conhecimento			
frequentada	Biológicas	Exatas	Humanas	Total
Pública	341	596	731	1668
Privada	1327	1957	2165	5449
Principalmente pública	100	158	178	436
Principalmente privada	118	194	196	508
Total	1886	2905	3270	8061

O valor da estatística de Pearson (4.1) correspondente aos dados da Tabela 4.7 é $\chi^2 = 15$; com base na distribuição χ^2 com $6 = (4-1)(3-1)$ graus de liberdade obtemos $p = 0,02$ o que sugere uma associação entre as duas variáveis (tipo de escola e área do conhecimento). No entanto, essa conclusão não tem significância prática, pois a estatística de Pearson terá um valor tanto maior quanto maior for o total da tabela, mesmo que a associação entre as variáveis seja muito tênue. Nesse contexto, convém avaliar essa associação por intermédio dos coeficientes de contingência de

Pearson (4.2) ou de Tschuprov (4.3), entre outros. Para o exemplo, seus valores são, respectivamente, $0,043$ e $0,027$, sugerindo uma associação de pequena intensidade.

Para comparar as preferências de formação profissional entre alunos que frequentaram diferentes tipos de escola, consideramos as frequências relativas tomando como base os totais das linhas; os resultados estão dispostos na Tabela 4.8.

Tabela 4.8: Frequências relativas de preferências por área de conhecimento (por tipo de escola)

Tipo de escola	Área do conhecimento			
frequentada	Biológicas	Exatas	Humanas	Total
Pública	20,5%	35,7%	43,8%	100,0%
Privada	24,4%	35,9%	39,7%	100,0%
Principalmente pública	23,0%	36,2%	40,8%	100,0%
Principalmente privada	23,2%	38,2%	38,6%	100,0%
Total	23,4%	36,0%	40,6%	100,0%

Grosso modo podemos dizer que cerca de 40% dos alunos que frequentaram escolas publicas ou privadas, mesmo que parcialmente, matricularam-se em cursos de Ciências Humanas, cerca de 36% de alunos com as mesmas características matricularam-se em cursos de Ciências Exatas e os demais 24% em cursos de Ciências Biológicas. Note que foi necessário um ajuste em algumas frequências relativas (por exemplo, o valor correspondente à cela Escola Pública/Ciências Biológicas deveria ser 20,4% e não 20,5%) para que o total somasse 100% mantendo a aparência da tabela com apenas uma casa decimal.

Se, por outro lado, o objetivo for avaliar o tipo de escola frequentado por alunos matriculados em cada área do conhecimento, devemos calcular as frequências relativas tomando como base o total da colunas; os resultados estão dispostos na Tabela 4.9 e sugerem que dentre os alunos que optaram por qualquer das três áreas, cerca de 21% são oriundos de escolas públicas, cerca de 68% de escolas privadas com os demais 11% tendo cursado escolas públicas ou privadas parcialmente.

Tabela 4.9: Frequências relativas tipo de escola cursada (por área do conhecimento)

Tipo de escola	Área do conhecimento			
frequentada	Biológicas	Exatas	Humanas	Total
Pública	18,1%	20,5%	22,4%	20,7%
Privada	70,3%	67,4%	66,2%	67,6%
Principalmente pública	5,3%	5,4%	5,4%	5,4%
Principalmente privada	6,3%	6,7%	6,0%	6,3%
Total	100,0%	100,0%	100,0%	100,0%

Risco atribuível, risco relativo e razão de chances

Em muitas áreas do conhecimento há interesse em avaliar a associação entre um ou mais **fatores de risco** e uma variável resposta. Num estudo epidemiológico, por exemplo, pode haver interesse em avaliar a associação entre o hábito tabagista (fator de risco) e a ocorrência de algum tipo de câncer pulmonar (variável resposta). Um exemplo na área de Seguros pode envolver a avaliação da associação entre estado civil e sexo (considerados como fatores de risco) e o envolvimento em acidente automobilístico (variável resposta).

No primeiro caso, os dados obtidos de uma amostra de 50 fumantes e 100 não fumantes, por exemplo, para os quais se observa a ocorrência de câncer pulmonar após um determinado período podem ser dispostos no formato da Tabela 4.10. Esse tipo de estudo em que se fixam os níveis do fator de risco (hábito tabagista) e se observa a ocorrência do evento de interesse (câncer pulmonar) após um determinado tempo é conhecido como **estudo prospectivo**.

Tabela 4.10: Frequências de doentes observados num estudo prospectivo

Hábito	Câncer pulmonar		Total
tabagista	sem	com	
não fumante	80	20	100
fumante	35	15	50

Para a população da qual essa amostra é considerada oriunda (e para a qual se quer fazer inferência), a tabela correspondente pode ser expressa como na Tabela 4.11.

Tabela 4.11: Probabilidades de ocorrência de doença

Hábito	Câncer pulmonar		Total
tabagista	sem	com	
não fumante	$1-\pi_0$	π_0	1
fumante	$1-\pi_1$	π_1	1

O parâmetro π_0 corresponde à proporção (ou probabilidade) de indivíduos que contraem câncer pulmonar dentre os que **sabemos** ser não fumantes; analogamente, π_1 corresponde à proporção (ou probabilidade) de indivíduos que contraem câncer pulmonar dentre os que **sabemos** ser fumantes.

Nesse contexto podemos definir algumas medidas de associação (entre o fator de risco e a variável resposta).

i) **Risco atribuível**: $d = \pi_1 - \pi_0$ que corresponde à diferença entre as probabilidades de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.

ii) **Risco relativo**: $r = \pi_1/\pi_0$ que corresponde ao quociente entre as probabilidades de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.

iii) **Razão de chances** (*odds ratio*)[1]: $\omega = [\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$ que corresponde ao quociente entre as chances de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.

No exemplo da Tabela 4.10 essas medidas de associação podem ser estimadas como

i) Risco atribuível: $d = 0,30 - 0,20 = 0,10$ (o risco de ocorrência de câncer pulmonar aumenta de 10% para fumantes relativamente aos não fumantes).

ii) Risco relativo: $r = 0,30/0,20 = 1,50$ (o risco de ocorrência de câncer pulmonar para fumantes é $1,5$ vezes o risco correspondente para não fumantes).

iii) Chances: a chance de ocorrência de câncer pulmonar para fumantes é $0,429 = 0,30/0,70$; a chance de ocorrência de câncer pulmonar para não fumantes é $0,250 = 0,20/0,80$.

iiv) Razão de chances: $\omega = 0,429/0,250 = 1,72$ (a chance de ocorrência de câncer pulmonar para fumantes é $1,71$ vezes a chance correspondente para não fumantes).

Em geral, a medida de associação de maior interesse prático pela facilidade de interpretação, seja o risco relativo, a razão de chances talvez seja a mais utilizada na prática. Primeiramente, observemos que

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = r\frac{1-\pi_0}{1-\pi_1} \longrightarrow r, \text{ quando } \pi_0 \text{ e } \pi_1 \longrightarrow 0$$

ou seja, para eventos raros [cujas probabilidade π_1 ou π_0 são muito pequenas], a razão de chances serve como uma boa aproximação do risco relativo.

Em geral, estudos prospectivos com a natureza daquele que motivou a discussão acima não são praticamente viáveis em função do tempo decorrido até o diagnóstico da doença. Uma alternativa é a condução de **estudos retrospectivos** em que, por exemplo, são selecionados 35 pacientes com e 115 pacientes sem câncer pulmonar e se determinam quais dentre eles eram fumantes e não fumantes. Nesse caso, os papéis das variáveis explicativa e resposta se invertem, sendo o *status* relativo à presença da moléstia encarado

[1]Lembremos que **probabilidade** é uma medida de frequência de ocorrência de um evento (quanto maior a probabilidade de um evento, maior a frequência com que ele ocorre) cujos valores variam entre 0 e 1 (ou entre 0% e 100%). Uma medida de frequência equivalente mas com valores entre 0 e ∞ é conhecida como **chance** (*odds*). Por exemplo, se um evento ocorre com probabilidade 0.8 (80%), a chance de ocorrência é 4 (= 80% / 20%) ou mais comumente de 4 para 1, indicando que em cinco casos, o evento ocorre em 4 e não ocorre em 1.

como variável explicativa e o hábito tabagista, como variável resposta. A Tabela 4.12 contém dados hipotéticos de um estudo retrospectivo planejado com o mesmo intuito do estudo prospectivo descrito acima, ou seja, avaliar a associação entre tabagismo e ocorrência de câncer de pulmão.

Tabela 4.12: Frequências de fumantes observados num estudo retrospectivo

Hábito	Câncer pulmonar	
tabagista	sem	com
não fumante	80	20
fumante	35	15
Total	115	35

A Tabela 4.13 representa as probabilidades pertinentes.

Tabela 4.13: Probabilidades de hábito tabagista

Hábito	Câncer pulmonar	
tabagista	sem	com
não fumante	$1 - p_0$	$1 - p_1$
fumante	p_0	p_1
Total	1	1

O parâmetro p_0 corresponde à proporção (ou probabilidade) de fumantes **dentre** os indivíduos que **sabemos** não ter câncer pulmonar; analogamente, p_1 corresponde à proporção (ou probabilidade) de não fumantes **dentre** os indivíduos que **sabemos** ter câncer pulmonar. Nesse caso, não é possível calcular nem o risco atribuível nem o risco relativo, pois não se conseguem estimar as probabilidades de ocorrência de câncer pulmonar, π_1 ou π_0. No entanto, pode-se demonstrar (ver Nota de Capítulo 1) que a razão de chances obtida por meio de um estudo retrospectivo é igual àquela que seria obtida por intermédio de um estudo prospectivo correspondente ou seja

$$\omega = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}.$$

Num estudo retrospectivo, pode-se afirmar que a chance de ocorrência do evento de interesse (câncer pulmonar, por exemplo) para indivíduos expostos ao fator de risco é ω vezes a chance correspondente para indivíduos não expostos, embora não se possa estimar quanto valem essas chances.

Detalhes sobre estimativas e intervalos de confiança para o risco relativo e razão de chances são apresentados na Nota de Capítulo 7.

A partir das frequências da Tabela 4.12 podemos estimar a chance de um indivíduo ser fumante dado que tem câncer pulmonar como $0,751 = 0,429/0,571$ e a chance de um indivíduo ser fumante dado que não tem câncer pulmonar como $0,437 = 0,304/0,696$; a razão de chances correspondente é $\omega = 0,751/0,437 = 1,72$. Essas chances não são aquelas de

interesse pois gostaríamos de conhecer as chances de ter câncer pulmonar para indivíduos fumantes e não fumantes. No entanto a razão de chances tem o mesmo valor que aquela calculada por meio de um estudo prospectivo, ou seja, a partir da análise dos dados da Tabela 4.12, não é possível calcular a chance de ocorrência de câncer pulmonar nem para fumantes nem para não fumantes mas podemos concluir que a primeira é $1,72$ vezes a segunda.

Avaliação de testes diagnósticos

Dados provenientes de estudos planejados com o objetivo de avaliar a capacidade de testes laboratoriais ou exames médicos para diagnóstico de alguma doença envolvem a classificação de indivíduos segundo duas variáveis; a primeira corresponde ao verdadeiro *status* relativamente à presença da moléstia (doente ou não doente) e a segunda ao resultado do teste (positivo ou negativo). Dados correspondentes aos resultados de um determinado teste aplicado a n indivíduos podem ser dispostos no formato da Tabela 4.14.

Tabela 4.14: Frequência de pacientes submetidos a um teste diagnóstico

Verdadeiro	Resultado do teste		Total
status	positivo (T+)	negativo (T-)	
doente (D)	n_{11}	n_{12}	n_{1+}
não doente (ND)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Aqui, n_{ij} corresponde à frequência de indivíduos com o i-ésimo *status* relativo à doença ($i = 1$ para doentes e $i = 2$ para não doentes) e j-ésimo *status* relativo ao resultado do teste ($j = 1$ para resultado positivo e $j = 2$ para resultado negativo). Além disso, $n_{i+} = n_{i1} + n_{i2}$ e $n_{+j} = n_{1j} + n_{2j}$, $i, j = 1, 2$. As seguintes características associadas aos testes diagnóstico são bastante utilizadas na prática.

i) **Sensibilidade**: corresponde à probabilidade de resultado positivo para pacientes doentes $[S = P(T + |D)]$ e pode ser estimada por $s = n_{11}/n_{1+}$;

ii) **Especificidade**: corresponde à probabilidade de resultado negativo para pacientes não doentes $[E = P(T - |ND)]$ e pode ser estimada por $e = n_{22}/n_{2+}$;

iii) **Falso positivo**: corresponde à probabilidade de resultado positivo para pacientes não doentes $[FP = P(ND|T+)]$ e pode ser estimada por $fp = n_{21}/n_{+1}$;

iv) **Falso negativo**: corresponde à probabilidade de resultado negativo para pacientes doentes $[FN = P(D|T-)]$ e pode ser estimada por $fn = n_{12}/n_{+2}$;

v) **Valor preditivo positivo**: corresponde à probabilidade de que o paciente seja doente dado que o resultado do teste é positivo $[VPP =$

$P(D|T+)]$ e pode ser estimada por $vpp = n_{11}/n_{+1}$;

vi) **Valor preditivo negativo**: corresponde à probabilidade de que o paciente não seja doente dado que o resultado do teste é negativo $[VPN = P(ND|T-)]$ e pode ser estimada por $vpn = n_{22}/n_{+2}$;

vii) **Acurácia**: corresponde à probabilidade de resultados corretos $[AC = P\{(D \cap T+) \cup (ND \cap T-)\}]$ e pode ser estimada por $ac = (n_{11}+n_{22})/n$.

A sensibilidade de um teste corresponde à proporção de doentes identificados por seu intermédio, ou seja, é um indicativo da capacidade de o teste detectar a doença. Por outro lado, a especificidade de um teste corresponde à sua capacidade de identificar indivíduos que não têm a doença.

Quanto maior a sensibilidade de um teste, menor é a possibilidade de que indique falsos positivos. Um teste com sensibilidade de 95%, por exemplo, consegue identificar um grande número de pacientes que realmente têm a doença e por esse motivo testes com alta sensibilidade são utilizados em triagens. Quanto maior a especificidade de um teste, maior é a probabilidade de apresentar um resultado negativo para pacientes que não têm a doença. Se, por exemplo, a especificidade de um teste for de 99% dificilmente um paciente que não tem a doença terá um resultado positivo. Um bom teste é aquele que apresenta alta sensibilidade e alta especificidade, mas nem sempre isso é possível.

O valor preditivo positivo indica a probabilidade de um indivíduo ter a doença dado que o resultado do teste é positivo e valor preditivo negativo indica a probabilidade de um indivíduo não ter a doença dado um resultado negativo no teste.

Sensibilidade e especificidade são características do teste, mas tanto o valor preditivo positivo quanto o valor preditivo negativo dependem da **prevalência** (porcentagem de indivíduos doentes) da doença. Consideremos um exemplo em que o mesmo teste diagnóstico é aplicado em duas comunidades com diferentes prevalências de uma determinada doença. A Tabela 4.15 contém os dados da comunidade em que a doença é menos prevalente e a Tabela 4.16 contém os dados da comunidade em que a doença é mais prevalente.

Tabela 4.15: Frequência de pacientes submetidos a um teste diagnóstico (prevalência da doença = 15%)

Verdadeiro status	Resultado do teste		Total
	positivo (T+)	negativo (T-)	
doente (D)	20	10	30
não doente (ND)	80	90	170
Total	100	100	200

Tabela 4.16: Frequência de pacientes submetidos a um teste diagnóstico (prevalência da doença = 30%)

Verdadeiro	Resultado do teste		Total
status	positivo (T+)	negativo (T-)	
doente (D)	40	20	60
não doente (ND)	66	74	140
Total	106	94	200

Os valores estimados para a sensibilidade, especificidade, valores preditivo positivo e negativo além da acurácia estão dispostos na Tabela 4.17

Tabela 4.17: Características do teste aplicado aos dados das Tabelas 4.15 e 4.16

	População com doença	
Característica	menos prevalente	mais prevalente
Sensibilidade	67%	67%
Especifidade	53%	53%
VPP	20%	38%
VPN	90%	79%
Acurácia	55%	57%

4.3 Duas variáveis quantitativas

Uma das principais ferramentas para avaliar a associação entre duas variáveis quantitativas é o **gráfico de dispersão**. Consideremos um conjunto de n pares de valores (x_i, y_i) de duas variáveis X e Y; o gráfico de dispersão correspondente é um gráfico cartesiano em que os valores de uma das variáveis são colocados no eixo das abscissas e os da outra, no eixo das ordenadas.

Exemplo 4.3 Os dados contidos na Tabela 4.18, disponíveis no arquivo `figado`, correspondem a um estudo cujo objetivo principal era avaliar a associação entre o volume (cm^3) do lobo direito de fígados humanos medido ultrassonograficamente e o seu peso (g). Um objetivo secundário era avaliar a concordância de medidas ultrassonográficas do volume (Volume1 e Volume2) realizadas por dois observadores. O volume foi obtido por meio da média das duas medidas ultrassonográficas. Detalhes podem ser obtidos em Zan (2005).

O gráfico de dispersão correspondente às variáveis Volume e Peso está apresentado na Figura 4.1. Nesse gráfico pode-se notar que a valores menores do volume correspondem valores menores do peso e a valores maiores do volume correspondem valores maiores do peso, sugerindo uma associação positiva e possivelmente linear entre as duas variáveis. Além disso, o gráfico permite identificar um possível ponto discrepante (*outlier*) correspondente à

Tabela 4.18: Peso e volume do lobo direito de enxertos de fígado

Volume1 (cm^3)	Volume2 (cm^3)	Volume (cm^3)	Peso (g)
672,3	640,4	656,3	630
686,6	697,8	692,2	745
583,1	592,4	587,7	690
850,1	747,1	798,6	890
729,2	803,0	766,1	825
776,3	823,3	799,8	960
715,1	671,1	693,1	835
634,5	570,2	602,3	570
773,8	701,0	737,4	705
928,3	913,6	920,9	955
916,1	929,5	922,8	990
983,2	906,2	944,7	725
750,5	881,7	816,1	840
571,3	596,9	584,1	640
646,8	637,4	642,1	740
1021,6	917,5	969,6	945

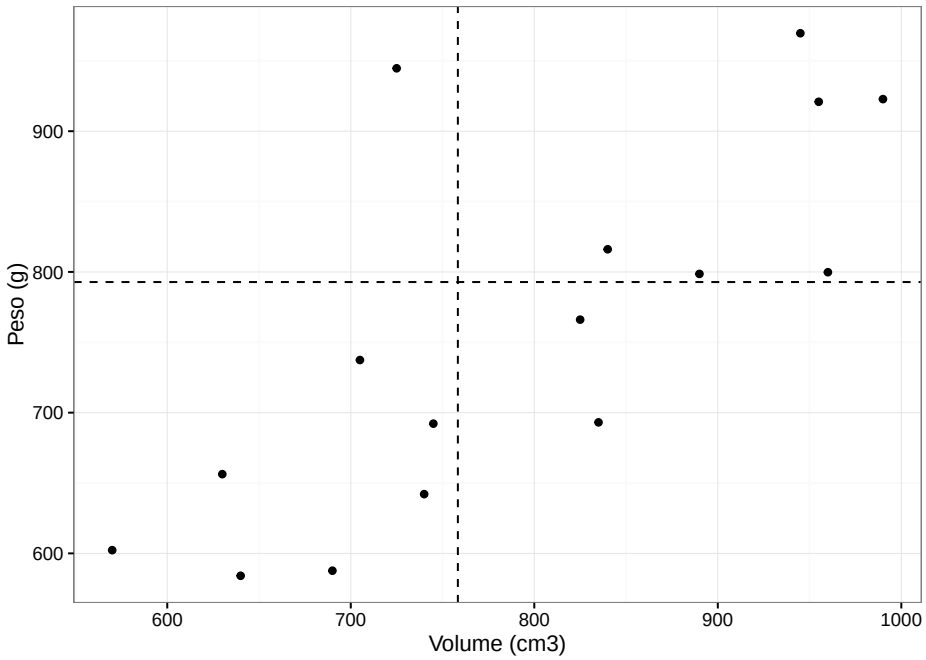


Figura 4.1: Gráfico de dispersão entre peso e volume do lobo direito de enxertos de fígado.

unidade amostral em que o volume é $725cm^3$ e o peso é $944,7g$. A utilização dessas constatações para a construção de um modelo que permita estimar o peso como função do volume é o objeto da técnica conhecida como **Análise de Regressão** que será considerada no Capítulo 6.

Dado um conjunto de n pares (x_i, y_i), a associação (linear) entre as variáveis quantitativas X e Y pode ser quantificada por meio do **coeficiente de correlação (linear)** de Pearson, definido por

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{[\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2]^{1/2}}. \tag{4.4}$$

Pode-se mostrar que $-1 \leq r_P \leq 1$ e, na prática, se o valor r_P estiver próximo de -1 ou $+1$, pode-se dizer que as variáveis são fortemente associadas ou (linearmente) correlacionadas; por outro lado, se o valor de r_P estiver próximo de zero, dizemos que as variáveis são não correlacionadas. Quanto mais próximos de uma reta estiverem os pontos (x_i, y_i), maior será a intensidade da correlação (linear) entre elas.

Não é difícil mostrar que

$$r_P = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\left[\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)\right]^{1/2}}. \tag{4.5}$$

Essa expressão é mais conveniente que (4.4), pois basta calcular: (a) as médias amostrais $\overline{x}$ e $\overline{y}$; (b) a soma dos produtos $x_i y_i$ e (c) a soma dos quadrados dos x_i e a soma dos quadrados dos y_i.

Para os dados do Exemplo 4.3, o coeficiente de correlação de Pearson é $0,76$. Se excluirmos o dado discrepante identificado no gráfico de dispersão, o valor do coeficiente de correlação de Pearson é $0,89$, evidenciando a falta de robustez desse coeficiente relativamente a observações com essa natureza. Nesse contexto, uma medida de associação mais robusta é o coeficiente de correlação de Spearman, cuja expressão é similar à (4.4) com os valores das variáveis X e Y substituídos pelos respectivos postos.[2] Mais especificamente, o coeficiente de correlação de Spearman é

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{[\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2]^{1/2}}, \tag{4.6}$$

em que R_i corresponde ao posto da i-ésima observação da variável X entre seus valores e $\overline{R}$ à média desses postos e S_i e $\overline{S}$ têm interpretação similar para a variável Y. Para efeito de cálculo pode-se mostrar que a expressão (4.6) é equivalente a

$$r_S = 1 - 6\sum_{i=1}^{n}(R_i - S_i)^2/[n(n^2-1)]. \tag{4.7}$$

[2] O posto de uma observação x_i é o índice correspondente à sua posição no conjunto ordenado $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Por exemplo, dado o conjunto de observações $x_1 = 4$, $x_2 = 7$, $x_3 = 5$, $x_4 = 13$, $x_5 = 6$, $x_6 = 5$, o posto correspondente à x_5 é 4. Quando há observações com o mesmo valor, o posto correspondente a cada uma delas é definido como a média dos postos correspondentes. No exemplo, os postos das observações x_3 e x_6 são iguais a $2,5 = (2+3)/2$.

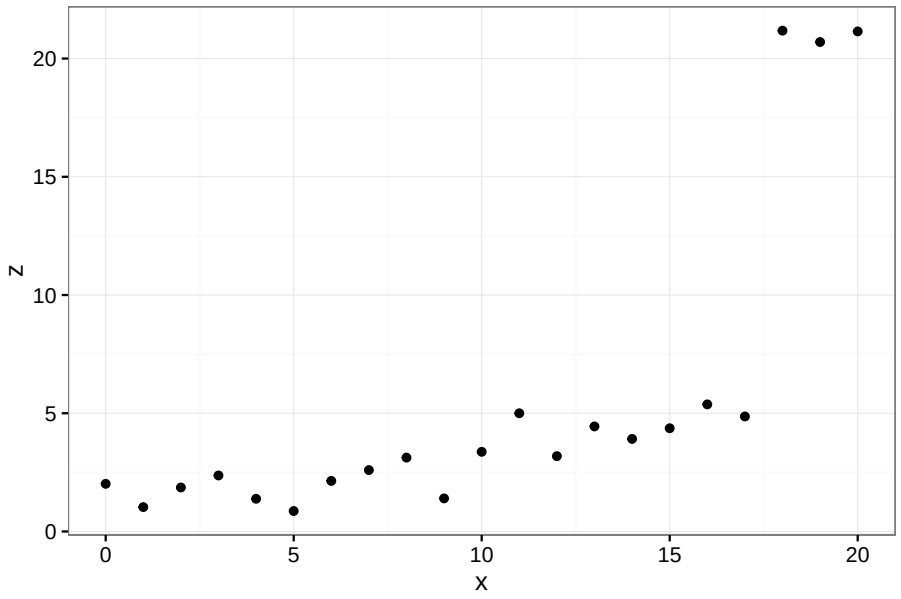


Figura 4.2: Gráfico de dispersão entre valores de duas variáveis X e Z.

Os dados correspondentes à Figura 4.2 foram gerados a partir da expressão $z_i = 1 + 0,25x_i + e_i$ com e_i simulado a partir de uma distribuição Normal padrão e com as três últimas observações acrescidas de 15. Para esses dados obtemos $r_P = 0.73$ e $r_S = 0.90$. Eliminando as três observações com valores discrepantes, os coeficientes de correlação correspondentes são $r_P = 0.85$ e $r_S = 0.84$, indicando que o primeiro é mais afetado do que o segundo.

Além disso, o coeficiente de correlação de Spearman também é mais apropriado para avaliar associações não lineares, desde que sejam monotônicas, *i.e.*, em que os valores de uma das variáveis só aumentam ou só diminuem conforme a segunda variável aumenta (ou diminui). Os dados representados na Figura 4.3 foram gerados a partir da expressão $y_i = \exp(0.4x_i)$, $i = 1, \ldots, 20$.

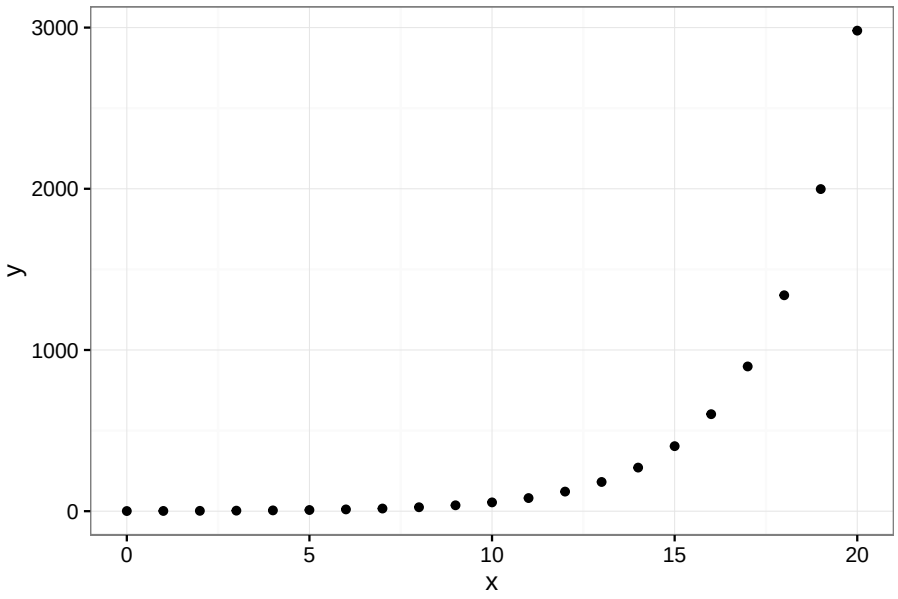


Figura 4.3: Gráfico de dispersão entre valores de duas variáveis X e Y.

Nesse caso, os valores dos coeficientes de correlação de Pearson e de Spearman são, respectivamente, $r_P = 0.75$ e $r_S = 1$ indicando que apenas este último é capaz de realçar a associação perfeita entre as duas variáveis.

Gráficos de perfis individuais

Para dados longitudinais, *i.e.*, aqueles em que a mesma variável resposta é observada em cada unidade amostral mais do que uma vez ao longo do tempo (ou de outra escala ordenada, como distância de uma fonte poluidora, por exemplo), uma das ferramentas descritivas mais importantes são os chamados **gráficos de perfis individuais**. Eles são essencialmente gráficos de dispersão (com o tempo na abscissa e a resposta na ordenada) em que os pontos associados a uma mesma unidade amostral são unidos por segmentos de reta. Em geral, os perfis médios são sobrepostos a eles. Esse tipo de gráfico pode ser utilizado para sugerir modelos de regressão (ver Capítulo 6) construídos para modelar o comportamento temporal da resposta esperada e também para identificar possíveis unidades ou observações discrepantes.

Exemplo 4.4. Os dados do arquivo `lactato` foram obtidos de um estudo realizado na Escola de Educação Física da Universiade de São Paulo com o objetivo de comparar a evolução da concentração sérica de lactato de sódio (mmol/L) como função da velocidade de dois grupos de atletas: 14 fundistas e 12 triatletas. A concentração sérica de lactato de sódio tem sido utilizada como um indicador da condição física de atletas. Nesse estudo, cada atleta correu durante certos períodos com velocidades pré-estabelecidas e a concentração de lactato de sódio foi registrada logo após cada corrida. A observação repetida da resposta em cada atleta caracteriza a natureza

longitudinal dos dados. Por meio dos comandos

```
> library(gdata)
> library(ggplot2)
> library(reshape2)
> library(dplyr)
>
> lactato <-read.xls("/home/jmsinger/Desktop/lactato.xls",
                     sheet='dados', method="tab")
> fundistas <- lactato[which(lactato$group == 0), ]
> fundistas1 <- fundistas[-1]
> fundistas2 <- melt(fundistas1, id.vars = "ident")
> fundistaslong <- group_by(fundistas2, ident)
>
> g1 <- ggplot(fundistaslong) +
+     geom_line(aes(variable, value, group = ident))
> g2 <- g1 + theme_bw() + annotate("text", x = 5, y = 5,
                           label = "atleta 9")
> g3 <- g2 + labs(x="velocidade",
                  y="Concentração de lactato de sódio")
> g4 <- g3 + theme(text=element_text(size=18))
> g4
```

obtemos o gráfico de perfis individuais para os fundistas que está representado na Figura 4.4 e sugere que i) a relação entre a concentração esperada de lactato de sódio pode ser representada por uma curva quadrática no intervalo de velocidades considerado e ii) o atleta 9 é um possível *outlier*. Na realidade, verificou-se que esse atleta era velocista e não fundista.

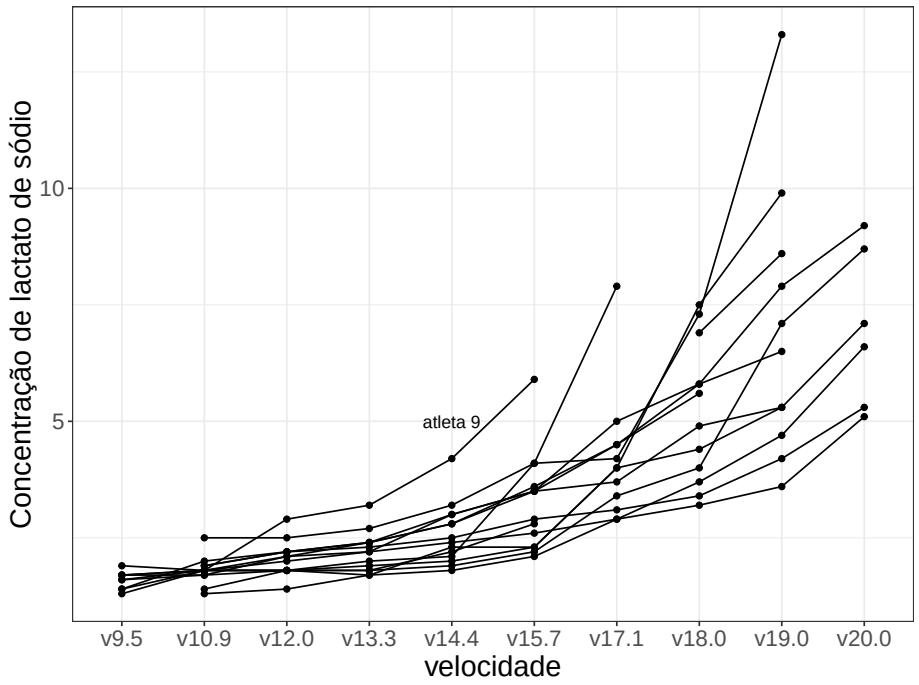


Figura 4.4: Gráfico de perfis individuais para os dados do Exemplo 4.4 (atletas fundistas).

Gráficos QQ para comparação de duas distribuições amostrais

Uma ferramenta adequada para comparar as distribuições de uma característica observada sob duas condições diferentes é o gráfico QQ utilizado na Seção 3.7 para a comparação de uma distribuição empírica com uma distribuição teórica. Um exemplo típico é aquele referente ao objetivo secundário mencionado na descrição do Exemplo 4.3, em que se pretende avaliar a concordância entre as duas medidas ultrassonográficas do volume do lobo direito do fígado.

Denotando por X uma das medidas e por Y, a outra, sejam $Q_X(p)$ e $Q_Y(p)$ os quantis de ordem p das duas distribuições que pretendemos comparar. O gráfico QQ é um gráfico cartesiano de $Q_X(p)$ em função de $Q_Y(p)$ (ou vice-versa) para diferentes valores de p. Se as distribuições de X e Y forem iguais, os pontos nesse gráfico devem estar sobre a reta $x = y$. Se uma das variáveis for uma função linear da outra, os pontos também serão dispostos sobre uma reta, porém com intercepto possivelmente diferente de zero e com inclinação possivelmente diferente de 1.

Quando os números de observações das duas variáveis for igual, o gráfico QQ é essencialmente um gráfico dos dados ordenados de X, ou seja $x_{(1)} \leq \ldots \leq x_{(n)}$, *versus* os dados ordenados de Y, nomeadamente, $y_{(1)} \leq \ldots \leq y_{(n)}$.

Quando os números de observações das duas variáveis forem diferentes,

digamos $m > n$, calculam-se os quantis amostrais referentes àquela variável com menos observações utilizando $p_i = (i-0,5)/n$, $i = 1, \ldots, n$ e obtêm-se os quantis correspondentes à segunda variável por meio de interpolações como aquelas indicadas em (3.5). Consideremos, por exemplo os conjuntos de valores $x_{(1)} \leq \ldots \leq x_{(n)}$ e $y_{(1)} \leq \ldots \leq y_{(m)}$. Primeiramente, determinemos $p_i = (i - 0,5)/n$, $i = 1, \ldots, n$ para obter os quantis $Q_X(p_i)$; em seguida, devemos obter índices j tais que

$$\frac{j - 0,5}{m} = \frac{i - 0,5}{n} \text{ ou seja } j = \frac{m}{n}(i - 0,5) + 0,5.$$

Se j obtido dessa forma for inteiro, o ponto a ser disposto no gráfico QQ será $(x_{(i)}, y_{(j)})$; em caso contrário, teremos $j = [j] + f_j$ em que $[j]$ é o maior inteiro contido em j e $0 < f_j < 1$ é a correspondente parte fracionária ($f_j = j - [j]$). O quantil correspondente para a variável Y será:

$$Q_Y(p_i) = (1 - f_j)y_{([j])} + f_j y_{([j]+1)}.$$

Por exemplo, sejam $m = 45$ e $n = 30$; então, para $i = 1, \ldots, 30$ temos

$$p_i = (i - 0,5)/30 \text{ e } Q_X(p_i) = x_{(i)}$$

logo $j = 45/30(i - 0,5) + 0,5 = 1,5i - 0,25$ e $[j] = [1,5i - 0,25]$. Consequentemente, no gráfico QQ, o quantil $Q_X(p_i)$ deve ser pareado com o quantil $Q_Y(p_i)$ conforme o seguinte esquema

i	p_i	j	$[j]$	$j - [j]$	$Q_X(p_i)$	$Q_Y(p_i)$
1	0,017	1,25	1	0,25	$x_{(1)}$	$0,75y_{(1)} + 0,25y_{(2)}$
2	0,050	2,75	2	0,75	$x_{(2)}$	$0,25y_{(2)} + 0,75y_{(3)}$
3	0,083	4,25	4	0,25	$x_{(3)}$	$0,75y_{(4)} + 0,25y_{(5)}$
4	0,117	5,75	5	0,75	$x_{(4)}$	$0,25y_{(5)} + 0,75y_{(6)}$
5	0,150	7,25	7	0,25	$x_{(5)}$	$0,75y_{(7)} + 0,25y_{(8)}$
6	0,183	8,75	8	0,75	$x_{(6)}$	$0,25y_{(8)} + 0,75y_{(9)}$
7	0,216	10,25	10	0,25	$x_{(7)}$	$0,75y_{(10)} + 0,25y_{(11)}$
8	0,250	11,75	11	0,75	$x_{(8)}$	$0,25y_{(11)} + 0,25y_{(12)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	0,983	44,75	44	0,75	$x_{(30)}$	$0,25y_{(44)} + 0,75y_{(45)}$

Suponha, por exemplo, que duas variáveis, X e Y, sejam tais que $Y = aX + b$, indicando que suas distribuições são iguais, exceto por uma transformação linear. Então,

$$p = P(X \leq Q_X(p)) = P(aX + b \leq aQ_X(p) + b) = P(Y \leq Q_Y(p)),$$

ou seja, $Q_Y(p) = aQ_X(p) + b$, indicando que o gráfico QQ correspondente mostrará uma reta com inclinação a e intercepto b.

Para a comparação das distribuições do volume ultrassonográfico do lobo direito do fígado medidas pelos dois observadores mencionados no Exemplo 4.3, o gráfico QQ está disposto na Figura 4.5.

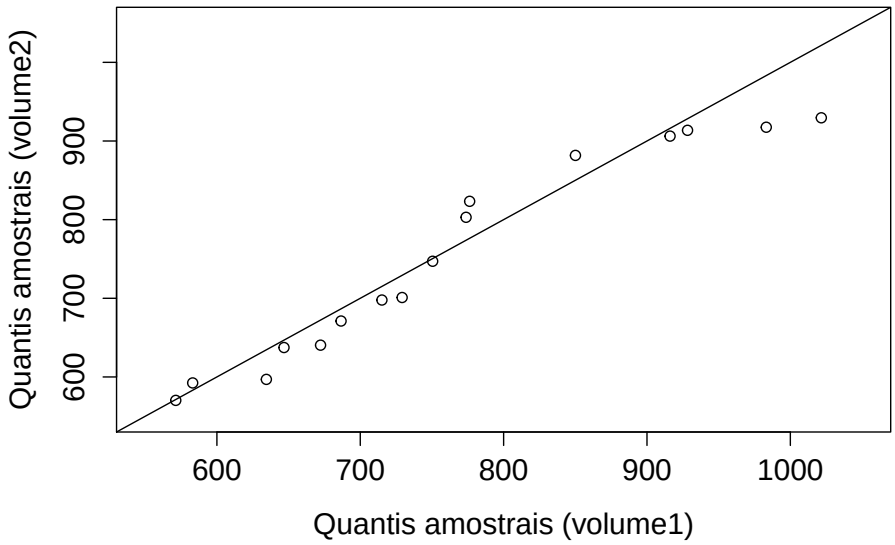


Figura 4.5: Gráfico QQ para avaliação da concordância de duas medidas ultrassonográficas do lobo direito do fígado.

Os pontos distribuem-se em torno da reta $x = y$ sugerindo que as medidas realizadas pelos dois observadores tendem a ser similares. Em geral os gráficos QQ são mais sensíveis a diferenças nas caudas das distribuições, se estas forem aproximadamente simétricas e com a aparência de uma distribuição Normal. Enquanto os diagramas de dispersão mostram uma relação sistemática global entre X e Y, os gráficos QQ relacionam valores pequenos de X com valores pequenos de Y, valores medianos de X com valores medianos de Y e valores grandes de X com valores grandes de Y.

Uma ferramenta geralmente utilizada para avaliar concordância entre as distribuições de duas variáveis contínuas com o mesmo espírito da estatística κ é o **gráfico de médias/diferenças** originalmente proposto por Tukey e popularizado como **gráfico de Bland-Altman**. Essencialmente, essa ferramenta consiste num gráfico das diferenças entre as duas observações pareadas $(X_{2i} - X_{1i})$ em função das médias correspondentes $[(X_{1i} + X_{2i})/2]$, $i = 1, \ldots, n$. Esse procedimento transforma a reta com coeficiente angular igual 1 apresentada no gráfico QQ numa reta horizontal passando pelo ponto zero no gráfico de médias/diferenças de Tukey e facilita a percepção das diferenças entre as duas medidas da mesma característica.

Note que enquanto gráficos QQ são construídos a partir do quantis amostrais, gráficos de Bland-Altman baseiam-se no próprios valores das variáveis em questão. Por esse motivo, para a construção de gráficos de Bland-Altman as observações devem ser pareadas ao passo que gráficos QQ podem ser construídos a partir de conjuntos de dados desbalanceados (com número diferentes de observações para cada variável).

O gráfico de médias/diferenças de Tukey (Bland-Altman) correspondente aos volumes medidos pelos dois observadores e indicados na Tabela 4.18 está

apresentado na Figura 4.6.

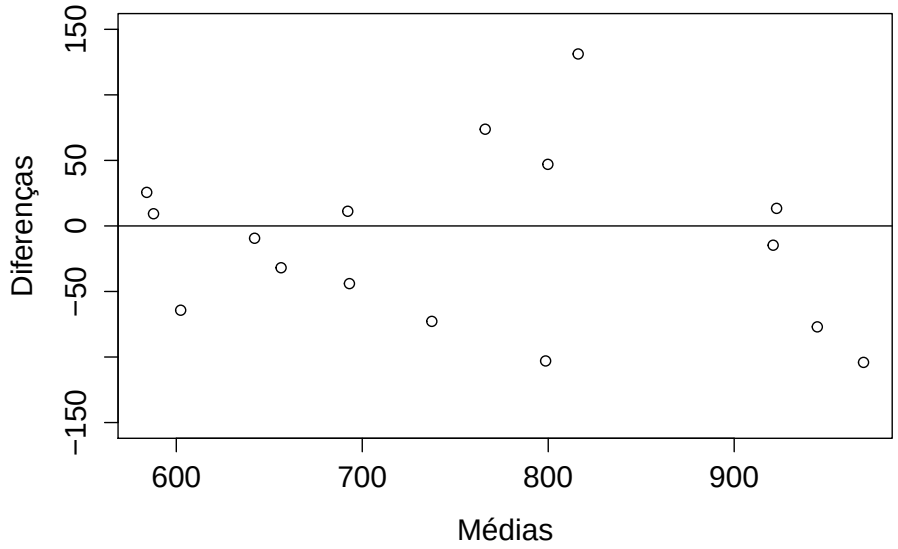


Figura 4.6: Gráfico de médias/diferenças de Tukey (Bland-Altman) para avaliação da concordância de duas medidas ultrassonográficas do lobo direito do fígado.

Os pontos no gráfico de médias/diferenças de Tukey distribuem-se de forma não regular em torno do valor zero e não sugerem evidências de diferenças entre as distribuições correspondentes. Por essa razão, para diminuir a variabilidade, decidiu-se adotar a média das medidas obtidas pelos dois observadores como volume do lobo direito do fígado para avaliar sua associação com o peso correspondente.

Exemplo 4.5 Os dados contidos na Tabela 4.19 foram extraídos de um estudo para avaliação de insuficiência cardíaca e correspondem à frequência cardíaca em repouso e no limiar anaeróbio de um exercício em esteira para 20 pacientes. O conjunto de dados completos está disponível no arquivo `esforco`.

Tabela 4.19: Frequência cardíaca em repouso (fcrep) e no limiar anaeróbio (fclan) de um exercício em esteira

paciente	fcrep	fclan	paciente	fcrep	fclan
1	89	110	11	106	157
2	69	100	12	83	127
3	82	112	13	90	104
4	89	104	14	75	82
5	82	120	15	100	117
6	75	112	16	97	122
7	89	101	17	76	140
8	91	135	18	77	97
9	101	131	19	85	101
10	120	129	20	113	150

Os gráficos QQ e de médias/diferenças de Tukey correspondentes aos dados da Tabela 4.19 estão apresentados nas Figuras 4.7 e 4.8.

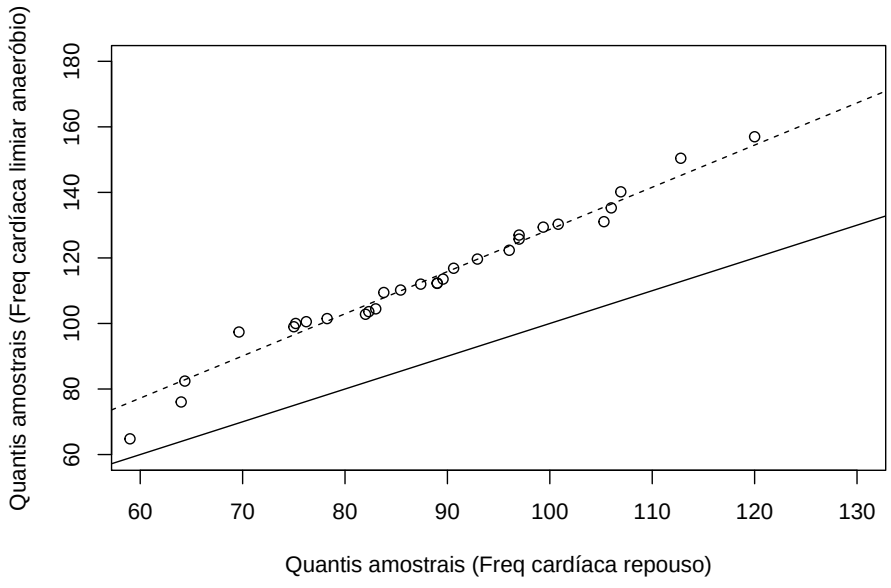


Figura 4.7: Gráfico QQ para comparação das distribuições de frequência cardíaca em repouso e no limiar anaeróbio.

Na Figura 4.7, a curva pontilhada corresponde à reta $Q_Y(p) = 1.29Q_X(p)$ sugerindo que a frequência cardíaca no limiar anaeróbio (Y) tende a ser cerca de 30% maior que aquela em repouso (X) em toda faixa de variação. Isso também pode ser observado, embora com menos evidência, no gráfico de Bland-Altman da Figura 4.8.

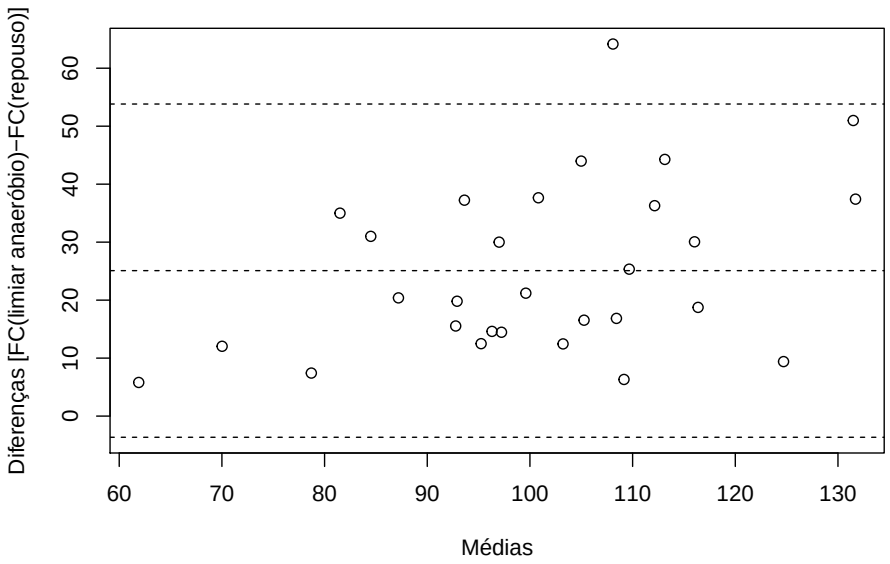


Figura 4.8: Gráfico de médias/diferenças de Tukey (Bland-Altman) para comparação das distribuições de frequência cardíaca em repouso e no limiar anaeróbio.

Exemplo 4.6. Considere o arquivo `temperaturas`, contendo dados de temperatura para Ubatuba e Cananéia. O gráfico QQ correspondente está apresentado na Figura 4.9. Observamos que a maioria dos pontos está acima da reta $y = x$, mostrando que as temperaturas de Ubatuba são em geral maiores do que as de Cananeia para valores maiores do que 17 graus.

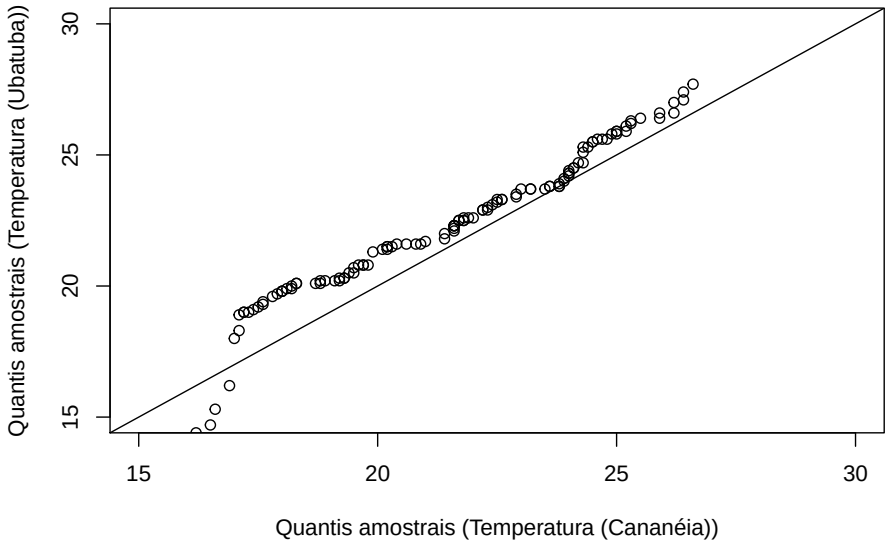


Figura 4.9: Gráfico QQ para comparação das distribuições de temperaturas de Ubatuba e Cananéia.

O gráfico de Bland-Altman correspondente, apresentado na Figura 4.10, sugere que acima de 17 graus, em média Ubatuba tende a ser 1 grau mais quente que Cananeia.

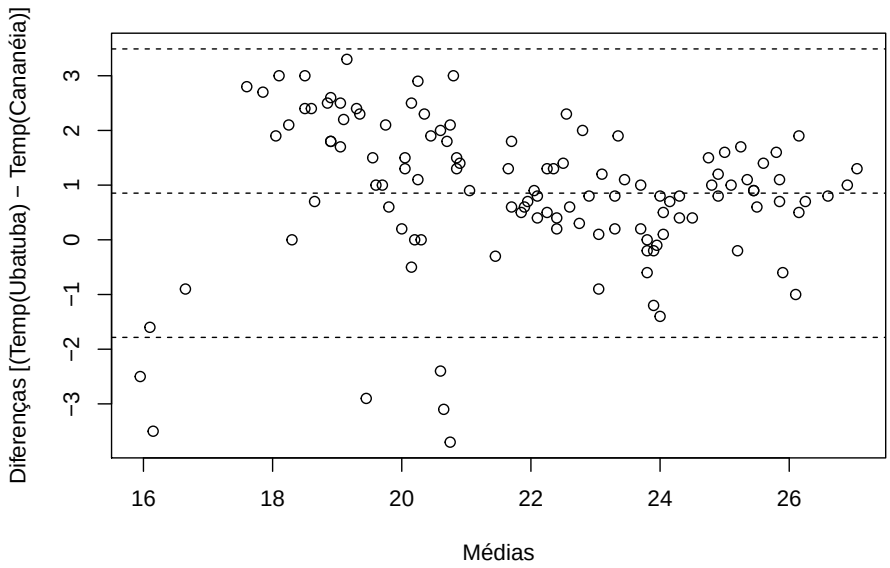


Figura 4.10: Gráfico de médias/diferenças de Tukey (Bland-Altman) para comparação das distribuições de temperaturas de Ubatuba e Cananéia.

4.4 Uma variável qualitativa e outra quantitativa

Um estudo da associação entre uma variável quantitativa e uma qualitativa consiste essencialmente na comparação das distribuições da primeira nos diversos níveis da segunda. Essa análise pode ser conduzida por meio de medidas resumo, histogramas, *boxplots* etc.

Exemplo 4.7. Num estudo coordenado pelo Laboratório de Poluição Atmosférica Experimental da USP, foram colhidos dados de concentração de vários elementos captados nas cascas de árvores em diversos pontos do centro expandido do município de São Paulo com o intuito de avaliar sua associação com a poluição atmosférica oriunda do tráfego. Os dados disponíveis no arquivo `arvores` foram extraídos desse estudo e contêm a concentração de Zn (ppm) em 497 árvores classificadas segundo a espécie (*alfeneiro*, *sibipiruna* e *tipuana*) e a localização em termos da proximidade do tipo de via (arterial, coletora, local I, local II, em ordem decrescente da intensidade de tráfego). Para efeito didático, consideramos primeiramente as 193 *tipuanas*. Medidas resumo para a concentração de Zn segundo os níveis de espécie e tipo de via estão indicadas na Tabela 4.20.

Tabela 4.20: Medidas resumo para a concentração de Zn (ppm) em cascas de *tipuanas*

Tipo de via	Média	Desvio padrão	Min	Q1	Mediana	Q3	Max	n
Arterial	199,4	110,9	29,2	122,1	187,1	232,8	595,8	59
Coletora	139,7	90,7	35,2	74,4	127,4	164,7	385,5	52
Local I	100,6	73,4	20,1	41,9	73,0	139,4	297,7	48
Local II	59,1	42,1	11,0	31,7	45,7	79,0	206,4	34

Min: mínimo Max: máximo
Q1: primeiro quartil Q3: terceiro quartil

Os resultados indicados na Tabela 4.20 mostram que tanto a concentração média e mediana de Zn quanto o correspondente desvio padrão decrescem à medida que a intensidade de tráfego diminui, sugerindo que essa variável pode ser utilizada como um indicador da poluição produzida por veículos automotores. Os *boxplots* apresentados na Figura 4.11 confirmam essas conclusões e também indicam que as distribuições apresentam uma leve assimetria, especialmente para as vias coletoras e locais I além de alguns pontos discrepantes.

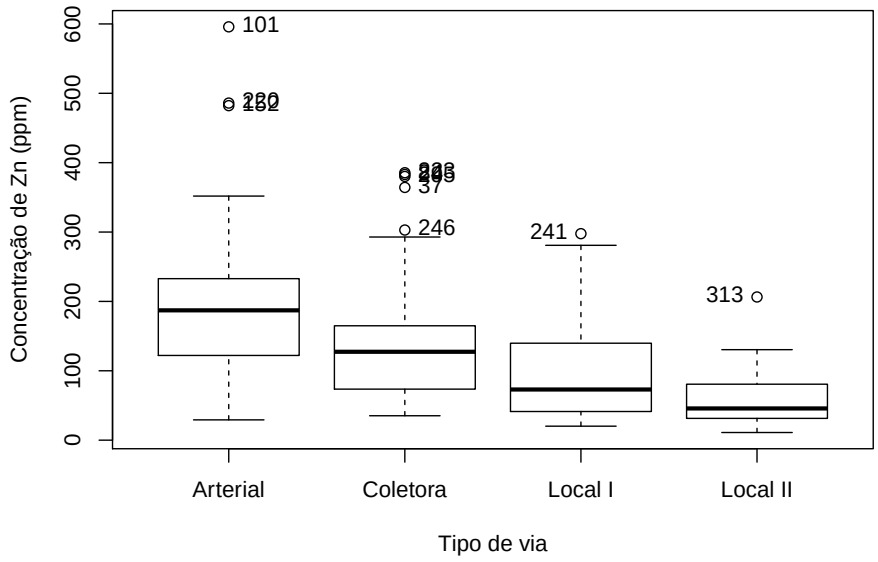


Figura 4.11: *Boxplots* para comparação das distribuições da concentração de Zn nas cascas de *tipuanas*.

Outro tipo de gráfico útil para avaliar a associação entre a variável quantitativa (concentração de Zn, no exemplo) e a variável qualitativa (tipo de via, no exemplo) especialmente quando esta tem níveis ordinais (como no exemplo) é o **gráfico de perfis médios**. Nesse gráfico cartesiano as médias

(e barras representando desvios padrões, erros padrões ou intervalos de confiança)[3] da variável quantitativa são representadas no eixo das ordenadas e os níveis da variável quantitativa, no eixo das abscissas. O gráfico de perfis médios para a concentração de Zn medida nas cascas de *Tipuana*s está apresentado na Figura 4.12 e reflete as mesmas conclusões obtidas com as análises anteriores.

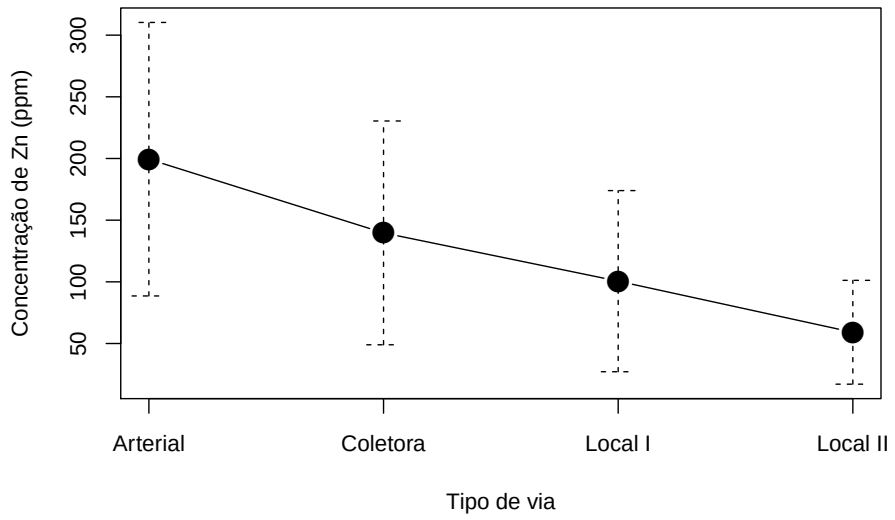


Figura 4.12: Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de *tipuanas*.

No título do gráfico, deve-se sempre indicar o que representam as barras; desvios padrões são úteis para avaliar como a dispersão dos dados em torno da média correspondente varia com os níveis da variável quantitativa (e não dependem do número de observações utilizadas para o cálculo da média); erros padrões são indicados para avaliação da precisão das médias (e dependem do número de observações utilizadas para o cálculo delas); intervalos de confiança servem para comparação das médias populacionais correspondentes e dependem de suposições sobre a distribuição da variável quantitativa.

Uma análise similar para os 76 *alfeneiros* está resumida na Tabela 4.21, e Figuras 4.13 e 4.14.

[3]Ver Nota de Capítulo 6

Tabela 4.21: Medidas resumo para a concentração de Zn (ppm) em cascas de *alfeneiros*

Tipo de via	Média	Desvio padrão	Min	Q1	Mediana	Q3	Max	n
Arterial	244,2	102,4	58,5	187,4	244,5	283,5	526,0	19
Coletora	234,8	102,7	15,6	172,4	231,6	311,0	468,6	31
Local I	256,3	142,4	60,0	154,9	187,0	403,7	485,3	19
Local II	184,4	96,4	45,8	131,1	180,8	247,6	306,6	7

Min: mínimo Max: máximo
Q1: primeiro quartil Q3: terceiro quartil

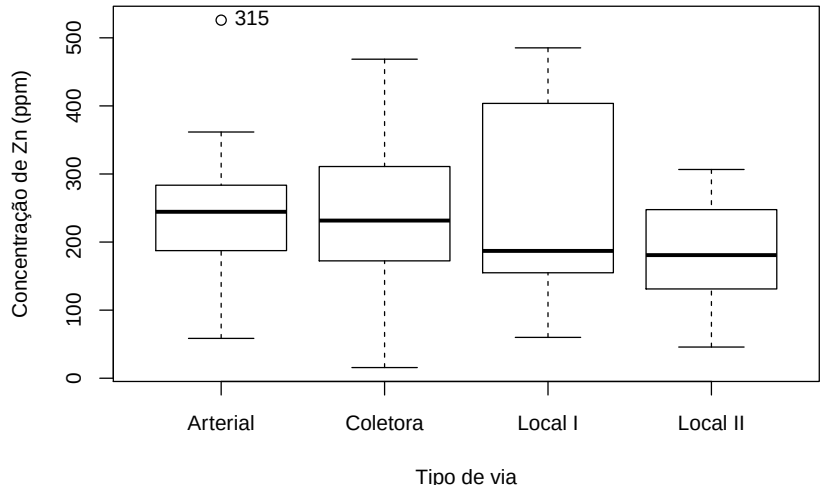


Figura 4.13: *Boxplots* para comparação das distribuições da concentração de Zn nas cascas de *alfeneiros*.

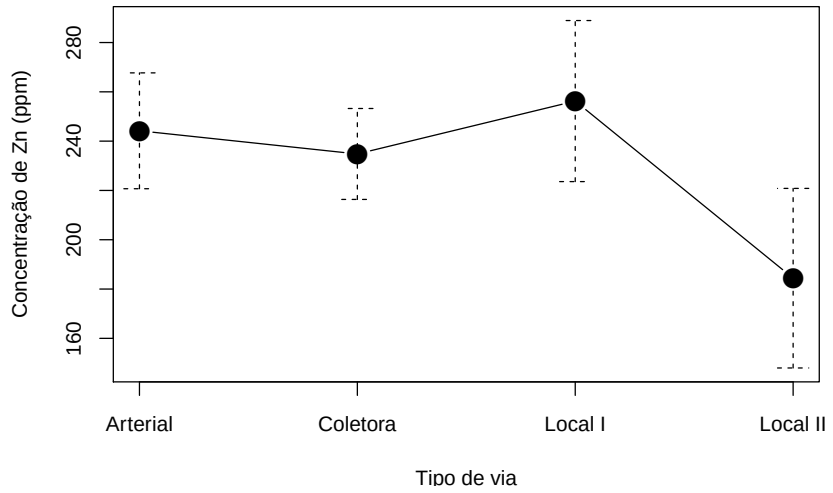


Figura 4.14: Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de *alfeneiros*.

Os valores dispostos na Tabela 4.21 e as Figuras 4.13 e 4.14 indicam que as concentrações de Zn em *alfeneiros* tendem a ser maiores do que aquelas encontradas em *tipuanas* porém são menos sensíveis a variações na intensidade de tráfego com exceção de vias locais II; no entanto, convém lembrar que apenas 7 *alfeneiros* foram avaliados nas proximidades desse tipo de via.

Exemplo 4.8. Consideremos os dados do arquivo `empresa`, referentes à informações sobre 36 funcionários de uma certa empresa. Nosso objetivo é avaliar a associação entre as variáveis "Salário" (S) expressa em número de salários mínimos e "Grau de instrução" (GI), com a classificação "fundamental", "médio" ou "superior".

Medidas resumo para "Salário" em função dos níveis de "Grau de instrução" são apresentadas na Tabela 4.22.

Tabela 4.22: Medidas resumo para a variável "Salário" (número de salários mínimos)

Grau de instrução	n	Média $\overline{S}$	Variância var(S)	Min	Q1	Q2	Q3	Max
Fundam	12	7,84	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	20,46	4,00	7,55	10,17	14,06	23,30

Min: mínimo Max: máximo
Q1: primeiro quartil Q2: mediana Q3: terceiro quartil

A leitura desses resultados sugere associação entre salários e grau de instrução: o salário médio tende a aumentar conforme aumenta o grau de instrução. O salário médio dos 36 funcionários é 11,12 salários mínimos; para funcionários com curso superior, o salário médio é de 16,48 salários mínimos, enquanto que funcionários com primeiro grau completo recebem, em média, 7,82 salários mínimos.

Embora nos dois exemplos apresentados a variável qualitativa seja ordinal, o mesmo tipo de análise pode ser empregado no caso de variáveis qualitativas nominais, tendo o devido cuidado na interpretação, pois não se poderá afirmar que a média da varável quantitativa aumenta com o aumento dos níveis da variável quantitativa.

Como nos casos anteriores, é conveniente poder contar com uma medida que quantifique o grau de associação entre as duas variáveis. Com esse intuito, convém observar que as variâncias podem ser usadas como insumos para construir essa medida. A variância da variável quantitativa ("Salário") para todos os dados, *i.e.*, calculada sem usar a informação da variável qualitativa ("Grau de instrução"), mede a dispersão dos dados em torno da média global (média salarial de todos os funcionários). Se as variâncias da variável "Salário" calculadas dentro de cada categoria da variável qualita-

tiva forem pequenas (comparativamente à variância global), essa variável pode ser usada para melhorar o conhecimento da distribuição da variável quantitativa, sugerindo a existência de uma associação entre ambas.

Na Tabela 4.22 pode-se observar que as variâncias do salário dentro das três categorias são menores do que a variância global e além disso, que aumentam com o grau de instrução. Uma medida resumo da variância **entre** as categorias da variável qualitativa é a média das variâncias ponderada pelo número de observações em cada categoria, ou seja,

$$\overline{\text{Var}(S)} = \frac{\sum_{i=1}^{k} n_i \text{Var}_i(S)}{\sum_{i=1}^{k} n_i}, \tag{4.8}$$

em que k é o número de categorias ($k = 3$ no exemplo) e $\text{Var}_i(S)$ denota a variância de S dentro da categoria i, $i = 1, \ldots, k$. Pode-se mostrar que $\overline{\text{Var}(S)} \leq \text{Var}(S)$, de modo que podemos definir o grau de associação entre as duas variáveis como o ganho relativo na variância obtido pela introdução da variável qualitativa. Explicitamente,

$$R^2 = \frac{\text{Var}(S) - \overline{\text{Var}(S)}}{\text{Var}(S)} = 1 - \frac{\overline{\text{Var}(S)}}{\text{Var}(S)}. \tag{4.9}$$

Além disso, pode-se mostrar que $0 \leq R^2 \leq 1$.

Quando as médias da variável resposta (salário, no exemplo) nas diferentes categorias da variável explicativa forem iguais, $\overline{\text{Var}(S)} = \text{Var}(S)$ e $R^2 = 0$, indicando a inexistência de associação entre as duas variáveis relativamente às suas médias. Esse é o princípio que norteia a técnica conhecida por Análise de Variância, cuja finalidade é comparar médias (populacionais) de distribuições Normais independentes com mesma variância. O leitor deve consultar Kutner et al. (2004), por exemplo, para detalhes sobre essa técnica. A estatística R^2 também é utilizada para avaliar a qualidade do ajuste de modelos de regressão, o tópicos abordado no Capítulo 6.

Para os dados do Exemplo 4.8, temos

$$\overline{\text{Var}(S)} = \frac{12 \times 7,77 + 18 \times 13,10 + 6 \times 16,89}{12 + 18 + 6} = 11,96.$$

Como $\text{Var}(S) = 20,46$, obtemos $R^2 = 1 - (11,96/20,46) = 0,415$, sugerindo que $41,5\%$ da variação total do salário é **explicada** pelo grau de instrução.

4.5 Notas de capítulo

1) **Probabilidade condicional e razões de chances**

Considere a seguinte tabela 2x2

	Doente (D)	Não doentes ($\overline{D}$)	Total
Exposto (E)	n_{11}	n_{12}	n_{1+}
Não exposto ($\overline{E}$)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

correspondente a um estudo em que o interesse é avaliar a associação entre a exposição de indivíduos a um certo fator de risco e a ocorrência de uma determinada moléstia. Em **estudos prospectivos** (*prospective, follow-up, cohort*) o planejamento envolve a escolha de amostras de tamanhos n_{1+} e n_{2+} de indivíduos respectivamente expostos e não expostos ao fator de risco e a observação da ocorrência ou não da moléstia após um certo intervalo de tempo. A razão de chances (de doença entre indivíduos expostos e não expostos) é definida como:

$$\omega_1 = \frac{P(D|E)P(\overline{D}|\overline{E})}{P(\overline{D}|E)P(D|\overline{E})}.$$

Em **estudos retrospectivos** ou **caso-controle**, o planejamento envolve a escolha de amostras de tamanhos $n_{\bullet 1}$ e $n_{\bullet 2}$ de indivíduos não doentes (controles) e doentes (casos), respectivamente e a observação retrospectiva de sua exposição ou não ao fator de risco. Nesse caso, a razão de chances é definida por:

$$\omega_2 = \frac{P(E|D)P(\overline{E}|\overline{D})}{P(\overline{E}|D)P(E|\overline{D})}.$$

Utilizando a definição de probabilidade condicional [ver Bussab e Morettin (2017), por exemplo], temos

$$\begin{aligned}\omega_1 &= \frac{[P(D\cap E)/P(E)][P(\overline{D}\cap\overline{E})/P(\overline{E})]}{[P(\overline{D}\cap E)/P(E)][P(D\cap\overline{E})/P(\overline{E})} = \frac{P(D\cap E)P(\overline{D}\cap\overline{E})}{P(\overline{D}\cap E)P(D\cap\overline{E})} \\ &= \frac{[P(E|D)/P(D)][P(\overline{E}|\overline{D})/P(\overline{D})]}{[P(E|\overline{D})/P(\overline{D})][P(\overline{E}|D)/P(D)]} = \omega_2\end{aligned}$$

Embora não se possa calcular o risco relativo de doença em estudos retrospectivos, a razão de chances obtida por meio desse tipo de estudo é igual àquela que seria obtida por intermédio de um estudo prospectivo, que em muitas situações práticas não pode ser realizado devido ao custo.

2) **Medidas de dependência entre duas variáveis**

Dizemos que X e Y são **comonotônicas** se Y (ou X) for uma função estritamente crescente de X (ou Y) e são **contramonotônicas** se a função for estritamente decrescente.

Consideremos duas variáveis X e Y e seja $\delta(X,Y)$ uma medida de dependência entre elas. As seguintes propriedades são desejáveis para δ (Embrechts et al., 2003):

(i) $\delta(X,Y) = \delta(Y,X)$;

(ii) $-1 \leq \delta(X,Y) \leq 1$;

(iii) $\delta(X,Y) = 1$ se X e Y são comonotônicas e $\delta(X,Y) = -1$ se X e Y são contramonotônicas;

(iv) Se T for uma transformação monótona,

$$\delta(T(X),Y) = \begin{cases} \delta(X,Y), & \text{se T for crescente,} \\ -\delta(X,Y), & \text{se T for decrescente.} \end{cases}$$

(v) $\delta(X,Y) = 0$ se e somente se X e Y são independentes.

O **coeficiente de correlação (linear)** entre X e Y é definido por

$$\rho = \frac{\text{Cov}(X,Y)}{DP(X)DP(Y)} \tag{4.10}$$

com $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$, $DP(X) = \text{E}\{[X - \text{E}(X)]^2\}$ e $DP(Y) = \text{E}\{[Y - \text{E}(Y)]^2\}$. Pode-se provar que $-1 \leq \rho \leq 1$ e que satisfaz as propriedades (i)-(ii). Além disso, ρ requer que as variâncias de X e Y sejam finitas e $\rho = 0$ não implica independência entre X e Y, a não ser que (X,Y) tenha uma distribuição Normal bivariada. Também, mostra-se que ρ não é invariante sob transformações não lineares estritamente crescentes.

3) **Dependência linear entre duas variáveis**

Convém reafirmar que $\rho(X,Y)$ mede dependência linear entre X e Y e não outro tipo de dependência. De fato, suponha que uma das variáveis possa ser expressa linearmente em termos da outra, por exemplo $X = aY + b$, e seja $d = E(|X - aY - b|^2)$. Então, pode-se provar (veja Exercício 13) que o mínimo de d ocorre quando

$$a = \frac{\sigma_X}{\sigma_Y}\rho(X,Y), \qquad b = E(X) - aE(Y), \tag{4.11}$$

e é dado por

$$\min d = \sigma_X^2[1 - \rho(X,Y)^2]. \tag{4.12}$$

Portanto, quanto maior o valor absoluto do coeficiente de correlação entre X e Y, melhor a acurácia com que uma das variáveis pode ser representada como uma combinação linear da outra. Obviamente, este mínimo se anula se e somente se $\rho = 1$ ou $\rho = -1$. Então de (4.12) temos

$$\rho(X,Y) = \frac{\sigma_X^2 - \min_{a,b} E(|X - aY - b|^2)}{\sigma_X^2}, \tag{4.13}$$

ou seja, $\rho(X,Y)$ mede a redução relativa na variância de X por meio de uma regressão linear de X sobre Y.

4) **Medidas de dependência robustas**

O coeficiente de correlação não é uma medida robusta. Uma alternativa robusta para a associação entre duas variáveis quantitativas pode ser construída como indicamos na sequência. Considere as variáveis padronizadas

$$\tilde{x}_k = \frac{x_k}{S_x(\alpha)}, \quad \tilde{y}_k = \frac{y_k}{S_y(\alpha)}, \ \ k = 1, \ldots, n,$$

em que $S_x^2(\alpha)$ e $S_y^2(\alpha)$ são as variâncias α-aparadas para os dados x_i e y_i, $i = 1, \ldots, n$, respectivamente. Um coeficiente de correlação robusto é definido por

$$r(\alpha) = \frac{S^2_{\tilde{x}+\tilde{y}}(\alpha) - S^2_{\tilde{x}-\tilde{y}}(\alpha)}{S^2_{\tilde{x}+\tilde{y}}(\alpha) + S^2_{\tilde{x}-\tilde{y}}(\alpha)}, \tag{4.14}$$

em que, por exemplo, $S^2_{\tilde{x}+\tilde{y}}(\alpha)$ é a variância α-aparada da soma dos valores padronizados de x_i e y_i, $i = 1, \ldots, n$. Pode-se mostrar que $r(\alpha) = r_P$ se $\alpha = 0$. Esse método é denominado de **método de somas e diferenças padronizadas**.

Exemplo 4.9. Consideremos os dados (x_i, y_i), $i = 1, \ldots, n$ apresentados na Tabela 4.23 e dispostos num diagrama de dispersão na Figura 4.15.

Tabela 4.23: Valores hipotéticos de duas variáveis X e Y

x	y	x	y
20,2	24,0	19,3	18,5
50,8	38,8	19,3	18,5
12,0	11,5	19,3	18,5
25,6	25,8	10,2	11,1
20,2	24,0	12,0	12,9
7,2	7,2	7,2	7,2
7,2	7,2	13,5	12,9
7,2	7,2		

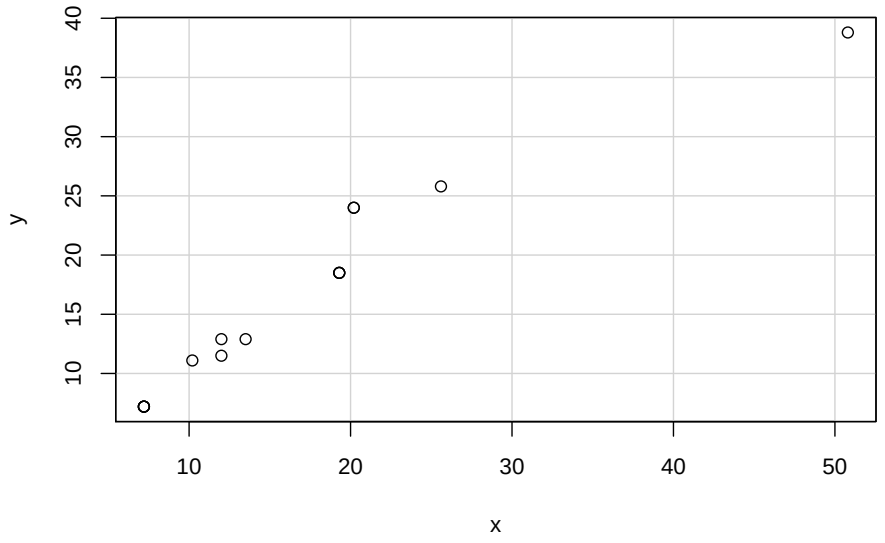


Figura 4.15: Gráfico de dispersão para os dados do Exemplo 4.9.

Para $\alpha = 0,05$, obtemos:

$$\overline{x}(\alpha) = 14,86, \quad \overline{y}(\alpha) = 15,33, \quad S_x(\alpha) = 5,87, \quad S_y(\alpha) = 6,40,$$

$$(\overline{\tilde{x} + \tilde{y}})(\alpha) = 4,93, \quad (\overline{\tilde{x} - \tilde{y}})(\alpha) = 0,14,$$

$$S^2_{\tilde{x}+\tilde{y}}(\alpha) = 3,93, \quad S^2_{\tilde{x}-\tilde{y}}(\alpha) = 0,054.$$

Então de (4.14) obtemos $r(\alpha) = 0,973$, o que indica uma alta correlação entre as duas variáveis.

5) **Gráficos PP**

Na Figura 4.16, observe que $p_x(q) = P(X \leq q) = F_X(q)$ e que $p_y(q) = P(Y \leq q) = F_Y(q)$. O gráfico cartesiano com os pares $[p_x(q), p_y(q)]$, para qualquer q real, é chamado de gráfico de probabilidades ou **gráfico PP**. O gráfico cartesiano com os pares $(Q_X(p), Q_Y(p))$, para $0 < p < 1$, é o gráfico de quantis *versus* quantis (gráfico QQ).

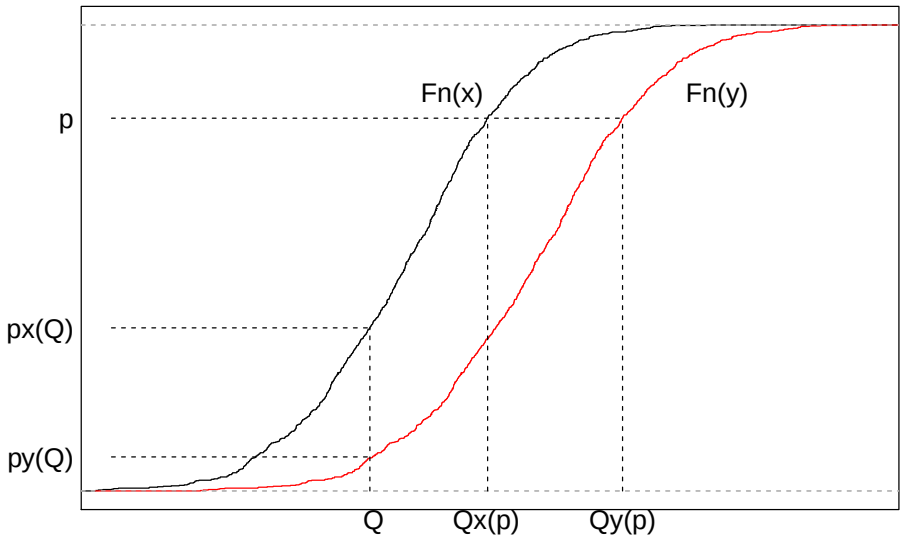


Figura 4.16: Quantis e probabilidades associados a duas distribuições.

Se as distribuições de X e Y forem iguais, então $F_X = F_Y$ e os pontos dos gráficos PP e QQ se situam sobre a reta $x = y$. Em geral os gráficos QQ são mais sensíveis a diferenças nas caudas das distribuições se estas forem aproximadamente simétricas e com a aparência de uma distribuição Normal. Suponha que $Y = aX + b$, ou seja, que as

distribuições de X e Y são as mesmas, exceto por uma transformação linear. Então,

$$p = P(X \leq Q_X(p)) = P(aX + b \leq aQ_X(p) + b) = P(Y \leq Q_Y(p)),$$

ou seja,

$$Q_Y(p) = aQ_X(p) + b.$$

O gráfico QQ correspondente será representado por uma reta com inclinação a e intercepto b. Essa propriedade não vale para gráficos PP.

Gráficos PP e QQ para a distribuição da concentração de Zn em cascas de árvores da espécie *Tipuana*, disponíveis no arquivo `arvores` estão dispostos na Figura 4.17 e salientam a maior capacidade dos últimos para detectar assimetrias em distribuições de frequência.

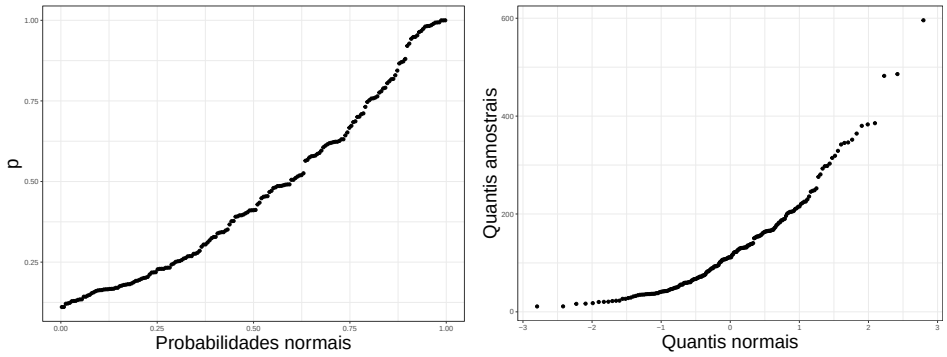


Figura 4.17: Gráficos PP e QQ para concentração de Zn em cascas de árvores da espécie *Tipuana*.

6) **Diferenças significativas**

Consideremos a distribuição de uma variável X (pressão arterial, por exemplo) em duas populações, A e B e admitamos que as médias de X sejam μ_A e μ_B (desconhecidas), respectivamente. Além disso, admitamos que ambas as distribuições tenham desvios padrões iguais a $\sigma = 10$ (conhecido). Nosso objetivo é saber se existem evidências de que $\mu_A = \mu_B$ com base em amostras aleatórias $X_{A1}, \ldots, X_{An}$ da população A e $X_{B1}, \ldots, X_{Bn}$ da população B. Admitamos que $n = 100$ e que as correspondentes médias amostrais sejam $\overline{X}_A = 13$ e $\overline{X}_B = 10$, respectivamente. Nesse caso, dizemos que a diferença $|\overline{X}_A - \overline{X}_B| = 3$ é **significativa** com $p < 0,05$, concluindo que há evidências de que para acreditar que $\mu_A \neq \mu_B$. Consideremos agora uma amostra de tamanho $n = 25$ de cada população, com médias $\overline{X}_A = 15$ e $\overline{X}_B = 10$. Nesse caso, dizemos que a diferença $|\overline{X}_A - \overline{X}_B| = 5$ **não é significativa** com $p > 0,05$, concluindo que não há razão para acreditar que $\mu_A \neq \mu_B$, embora a diferença entre as médias amostrais $\overline{X}_A$ e

$\overline{X}_B$ seja maior que no primeiro caso. Essencialmente, queremos saber qual é a interpretação da expressão "a diferença entre as médias é significativa".

O cerne do problema é que não queremos tirar conclusões sobre as médias amostrais, $\overline{X}_A$ e $\overline{X}_B$ (cuja diferença é evidente, pois a conhecemos) e sim sobre as médias populacionais μ_A e μ_B, que desconhecemos. Para associar as amostras às populações, precisamos de um modelo probabilístico. No caso do exemplo, um modelo simples supõe que as distribuições de frequências da variável X nas populações A e B são Normais, independentes com médias μ_A e μ_B, respectivamente e desvio padrão comum $\sigma = 10$.

No primeiro caso ($n = 100$), admitindo que as duas distribuições têm médias iguais ($\mu_A = \mu_B$), a probabilidade de que a diferença (em valor absoluto) entre as médias amostrais seja maior ou igual a 3 é

$$P(|\overline{X}_A - \overline{X}_B| \geq 3) = P(|Z| > 3/(\sqrt{2}\sigma/\sqrt{100}) = P(|Z| \geq 2,82) < 0,05$$

em que Z representa uma distribuição Normal padrão, *i.e.*, com média zero e variância 1. Em outras palavras, se as médias populacionais forem iguais, a probabilidade de se obter uma diferença de magnitude 3 entre as médias de amostras de tamanho $n = 100$ é menor que 5% e então dizemos que a diferença (entre as médias amostrais) é significativa ($p < 0,05$), indicando que a evidência de igualdade entre as médias populacionais μ_A e μ_B é pequena.

No segundo caso ($n = 25$), temos

$$P(|\overline{X}_A - \overline{X}_B| \geq 5) = P(|Z| > 5/(\sigma/\sqrt{25}) = P(|Z| > 1,76) > 0,05,$$

e então dizemos que a diferença (entre as médias amostrais) não é significativa ($p > 0,05$), indicando que não há evidências fortes o suficiente para acreditarmos que as médias populacionais μ_A e μ_B sejam diferentes.

Apesar de que no segundo caso, a diferença amostral é maior do que aquela do primeiro caso, concluímos que a evidência de diferença entre as médias populacionais é menor. Isso ocorre porque o tamanho amostral desempenha um papel importante nesse processo; quanto maior o tamanho amostral, mais fácil será detectar diferenças entre as médias populacionais em questão.

Grosso modo, afirmar que uma diferença entre duas médias amostrais é significativa é dizer que as médias das populações de onde as amostras forma extraídas não devem ser iguais; por outro lado, dizer que a diferença entre as médias amostrais não é significativa é dizer que não há razões para acreditar que exista diferença entre as médias populacionais correspondentes.

7) **Intervalos de confiança para o risco relativo e razão de chances**

Consideremos a seguinte tabela 2 × 2

Tabela 4.24: Frequência de pacientes

Fator	Status do paciente		
de risco	doente	são	Total
presente	n_{11}	n_{12}	n_{1+}
ausente	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Estimativas dos riscos de doença para pacientes expostos e não expostos ao fator de risco são, respectivamente, $p_1 = n_{11}/n_{1+}$ e $p_2 = n_{21}/n_{2+}$. Sob a suposição de que as distribuições de n_{11} e n_{21} são binomiais, as variâncias de p_1 e p_2 são respectivamente estimadas por $\text{Var}(p_1) = p_1(1-p_1)/n_{1+}$ e $\text{Var}(p_2) = p_2(1-p_2)/n_{2+}$.

Em vez de calcular a variância associada à estimativa do risco relativo, $rr = p_1/p_2$, é mais conveniente calcular a variância de $\log(rr)$. Com essa finalidade, recorremos ao **método Delta**[4], obtendo

$$\begin{aligned}
\text{Var}[\log(rr)] &= \text{Var}[\log(p_1) - \log(p_2)] = \text{Var}[\log(p_1)] + \text{Var}[\log(p_2)] \\
&= \frac{p_1(1-p_1)}{p_1^2 n_{1+}} + \frac{p_2(1-p_2)}{p_2^2 n_{2+}} = \frac{1-p_1}{p_1 n_{1+}} + \frac{1-p_2}{p_2 n_{2+}} \\
&= \frac{1-p_1}{n_{11}} + \frac{1-p_2}{n_{21}} \\
&= \frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}
\end{aligned}$$

Os limites inferior e superior de um itervalo de confiança com coeficiente de confiança aproximado de 95% para o logaritmo do risco relativo (populacional) RR são obtidos de

$$\log(p_1/p_2) \pm 1.96\sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}}. \tag{4.15}$$

Os limites do intervalo de confiança correspondente para o risco relativo podem ser obtidos exponenciando-se os limites indicados em 4.15.

A razão de chances RC de doença entre indivíduos expostos e não expostos ao fator de risco é estimada por $rc = p_1(1-p_2)/p_2(1-p_1)$.

[4]O método Delta é utilizado para calcular a variância de funções de variáveis aleatórias. Essencialmente, se x é tal que $\text{E}(X) = \mu$ e $\text{Var}(X) = \sigma^2$, então sob certas condições de regularidade (em geral satisfeitas nos casos mais comuns), $\text{Var}[g(X)] = [g'(\mu)]^2\sigma^2$ em que g é uma função com derivada $g'(z)$ no ponto z. Para detalhes, o leitor poderá consultar Sen et al. (2009).

Como no caso do risco relativo é mais conveniente estimar a variância de $\log(rc)$, que é

$$\text{Var}[\log(rc)] = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}.$$

Os limites inferior e superior de um itervalo de confiança com coeficiente de confiança aproximado de 95% para o logaritmo do razão de chances (populacional) RC são obtidos de

$$\log[p_1(1-p_2)/p_2(1-p_1)] \pm 1.96\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \quad (4.16)$$

Assim como no caso do risco relativo, os limites do intervalo de confiança correspondente para a razão de chances pode ser obtido por meio da exponenciação dos limites indicados em (4.16).

4.6 Exercícios

1) Considere o conjunto de dados disponível no arquivo `empresa`. Compare as distribuições de frequências de cada variável para indivíduos residentes na capital, interior e outros.

2) Considere o conjunto de dados disponível no arquivo `regioes`. Avalie a relação entre as variáveis "Região" e "Densidade populacional".

3) Considere o conjunto de dados disponível no arquivo `salarios`.

   a) Construa um gráfico QQ para as variáveis "Salário de professor secundário" e "Salário de administrador".

   b) Calcule o coeficiente de correlação de Pearson, r_P e o coeficiente de correlação robusto, $r(\alpha)$ com $\alpha = 0,10$ entre essas duas variáveis.

4) Considere o conjunto de dados disponível no arquivo `coronarias`.

   a) Construa gráficos QQ para comparar as distribuições das variáveis imc e idade de pacientes masculinos (=1) e femininos (=0) e discuta os resultados.

   b) Calcule o coeficiente de correlação de Pearson e o coeficiente de correlação de Spearman entre as variáveis "Altura" e "Peso".

   c) Construa uma tabela de contingência para avaliar a distribuição conjunta das variáveis "Tabagismo" (com 6 níveis) e "Arteriopatia" (com 4 níveis) e calcule a intensidade de associação entre elas utilizando a estatística de Pearson, o coeficiente de contingência de Pearson e o coeficiente de "Tschuprov".

5) Considere o conjunto de dados disponível no arquivo `esforco`.

a) Compare as distribuições de frequências da variável "VO2" em repouso e no pico do exercício para pacientes classificados em cada um dos níveis da variável "Etiologia" por meio de gráficos QQ e de medidas resumo. Comente os resultados.

b) Repita o item a) utilizando gráficos de Bland-Altman.

c) Utilize *boxplots* e gráficos de perfis médios para comparar as distribuições da variável "Frequência cardíaca" correspondentes a pacientes nos diferentes níveis da variável "NYHA". Comente os resultados.

6) Para os dados do arquivo `salarios`, considere a variável "Região", com as classes "América do Norte", "América Latina", "Europa"e "Outros"e a variável "Salário de professor secundário". Analise as duas variáveis.

7) Considere os dados do arquivo `figadodiag`. Calcule a sensibilidade, especificidade, taxas de falsos positivos e falsos negativos, valores preditivos positivos e negativos e acurácia das técnicas radiológicas para detecção de alterações anatômicas tendo os resultados intraoperatórios como padrão áureo.

8) Analise a variável "Preço de veículos" segundo as categorias N (nacional) e I (importado) para o conjunto de dados disponíveis no arquivo `veiculos`.

9) Utilizando a definição da Nota de Capítulo 4, prove que se $\alpha = 0$, então $r(\alpha) = r$.

10) Prove que (4.1) pode ser escrita como

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*},$$

em que n_{ij} é a frequência absoluta observada na linha i e coluna j e n_{ij}^* é a respectiva frequência esperada.

11) Prove que (4.1) pode ser escrita em termos de frequências relativas como

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*},$$

com notação similar à do problema anterior.

12) Prove que (4.4) e (4.5) são equivalentes.

13) Prove as relações (4.11)-(4.13).

14) Os dados da Tabela 4.25 são provenientes de um estudo em que um dos objetivos era avaliar o efeito da dose de radiação gama (em centigrays) na formação de múltiplos micronúcleos em células de indivíduos normais. Analise os dados descritivamente, calculando o risco relativo de ocorrência de micronúcleos para cada dose tomando como base a dose nula. Repita a análise calculando as razões de chances correspondentes. Quais as conclusões de suas análises?

Tabela 4.25: Número de células

Dose de radiação gama (cGy)	Frequência de células com múltiplos micronúcleos	Total de células examinadas
0	1	2373
20	6	2662
50	25	1991
100	47	2047
200	82	2611
300	207	2442
400	254	2398
500	285	1746

15) De uma tabela construída para avaliar a associação entre tratamento (ativo e placebo) e cura (sim ou não) de uma certa moléstia obteve-se uma razão de chances igual a 2,0. Explique por que não se pode concluir daí que a probabilidade de cura para pacientes submetidos ao tratamento ativo é 2 vezes a probabilidade de cura para pacientes submetidos ao placebo.

16) Um criminologista desejava estudar a relação entre: X (densidade populacional = número de pessoas por unidade de área) e Y (índice de assaltos = número de assaltos por 100000 pessoas) em grandes cidades. Para isto sorteou 10 cidades observando em cada uma delas os valores de X e Y. Os resultados obtidos estão dispostos na Tabela 4.26

Tabela 4.26: Densidade populacional e índice de assaltos em grandes cidades

Cidade	1	2	3	4	5	6	7	8	9	10
X	59	49	75	65	89	70	54	78	56	60
Y	190	180	195	186	200	204	192	215	197	208

a) Classifique as variáveis envolvidas.

b) Calcule a média, mediana, desvio-padrão e a distância interquartis para cada variável.

c) Construa o diagrama de dispersão entre Y e X e faça comentários sobre a relação entre as duas variáveis.

17) Considere a seguinte tabela.

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

O que se pode dizer sobre a relação entre as variáveis X e Y?

a) Não há associação entre X e Y.
b) Há relação linear positiva.
c) Há relação linear negativa.
d) Há relação quadrática.

18) Em um teste de esforço cardiopulmonar aplicado a 55 mulheres e 104 homens, foram medidas entre outras, as seguintes variáveis:

- Grupo: Normais, Cardiopatas ou DPOC (portadores de doença pulmonar obstrutiva crônica).
- VO2MAX: consumo máximo de O2 (ml/min).
- VCO2MAX: consumo máximo de CO2 (ml/min).

Algumas medidas descritivas e gráficos são apresentados abaixo nas Tabelas 4.27 e 4.28 e Figura 4.18

Tabela 4.27: VO2MAX

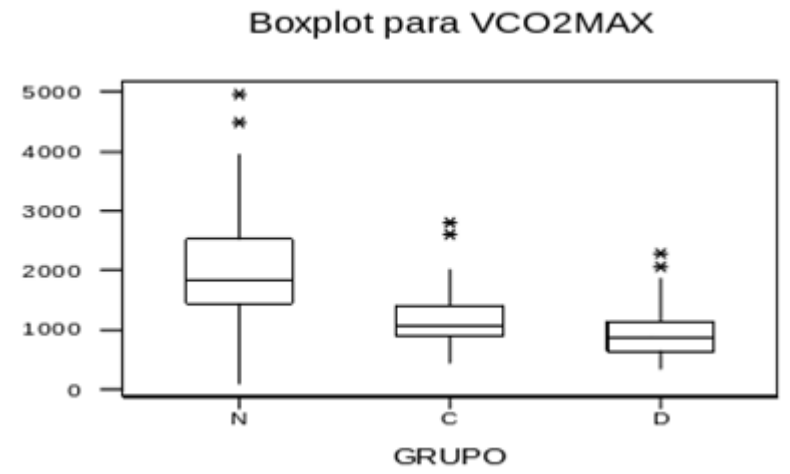
Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	1845	1707	795
Cardiopatas	57	1065	984	434
DPOC	46	889	820	381

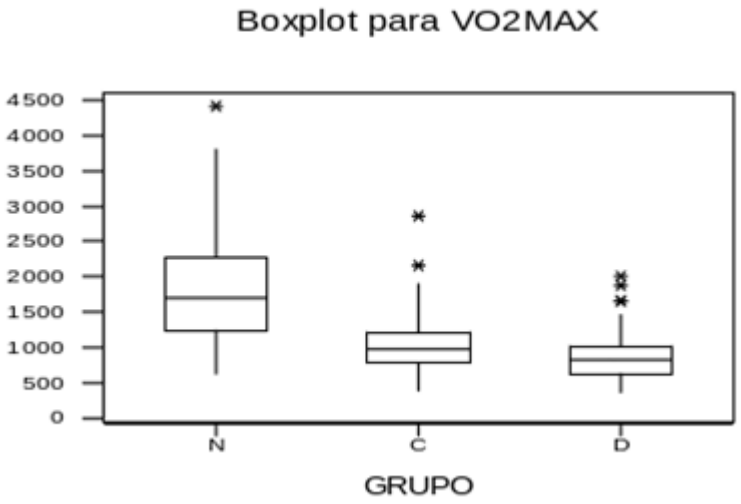
Tabela 4.28: VCO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	2020	1847	918
Cardiopatas	57	1206	1081	479
DPOC	46	934	860	430

Coeficiente de correlação entre VO2MAX e VCO2MAX = 0,92.

a) Que grupo tem a maior variabilidade?
b) Compare as médias e as medianas dos 3 grupos.
c) Compare as distâncias interquartis dos 3 grupos para cada variável. Você acha razoável usar a distribuição normal para esse conjunto de dados?





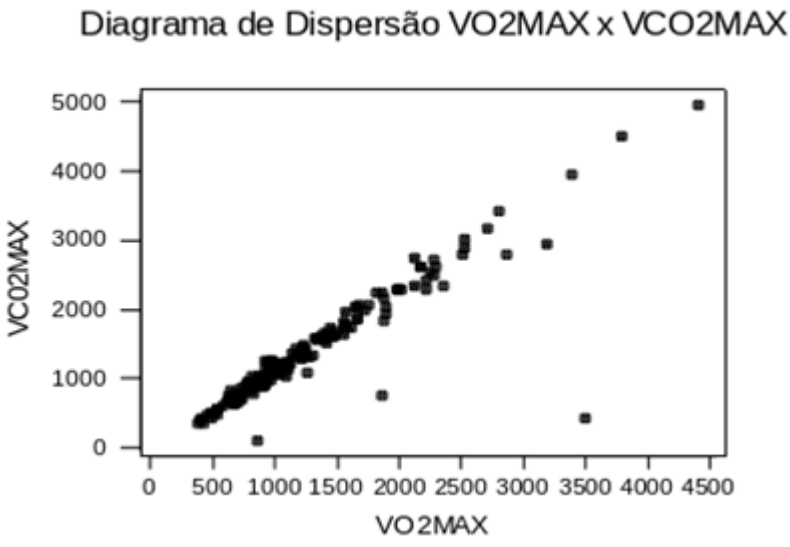


Figura 4.18: Gráficos para o Exercício 18.

d) O que representam os asteriscos nos *boxplots*?

e) Que tipo de função você ajustaria para modelar a relação entre o consumo máximo de CO2 e o consumo máximo de O2? Por quê?

f) Há informações que necessitam verificação quanto a possíveis erros? Quais?

19) Para avaliar a associação entre a persistência do canal arterial (PCA) em recém-nascidos pré-termo (RNPT) e óbito ou hemorragia intracraniana, um pesquisador obteve os dados dispostos na seguinte tabela

Frequências absolutas e relativas de óbitos e hemorragia intracraniana em recém-nascidos

PCA	Óbito			Hemorragia intracraniana		
	Sim	Não	Total	Sim	Não	Total
Presente	8	13	21	7	14	21
Ausente	1	39	40	7	33	40
Total	9	52	61	14	44	61

Um resumo das análises para óbitos e hemorragia intracraniana está disposto na tabela seguinte

		Razão de chances e Intervalo de confiança (95%)		
Variável	valor p	Estimativa	Lim inf	Lim sup
Óbito	0,001	24,0	2,7	210,5
Hemorragia intracraniana	0,162	2,4	0,7	8,0

a) Interprete as estimativas das razões de chances, indicando claramente a que pacientes elas se referem.

b) Analogamente, interprete os intervalos de confiança correspondentes, indicando claramente a que pacientes eles se referem.

c) Com base nos resultados anteriores, o que você pode concluir sobre a associação entre persistência do canal arterial e óbito para RNPT em geral? E sobre a associação entre a persistência do canal arterial e a ocorrência de hemorragia interna? Justifique suas respostas.

d) Qual a hipótese nula testada em cada caso?

e) Qual a interpretação dos níveis descritivos (p-value) em cada caso?

Detalhes podem ser obtidos em Afiune (2000).

20) Em um estudo comparativo de duas drogas para hipertensão os resultados indicados nas Tabelas 4.29, 4.30 e 4.31 e Figura 4.19 foram usados para descrever a eficácia e a tolerabilidade das drogas ao longo de 5 meses de tratamento.

Tabela 4.29: Frequências absoluta e relativa do efeito colateral para as duas drogas

	Droga 1		Droga 2	
Efeito Colateral	n	%	n	%
não	131	61,22	144	65,45
sim	83	38,79	76	34,54

Tabela 4.30: Distribuição de frequências para as drogas 1 e 2

Variação	Droga 1		Droga 2	
Pressão	n	%	n	%
$0 \vdash 5$	9	4,20561	5	2,27273
$5 \vdash 10$	35	16,3551	29	13,1818
$10 \vdash 20$	115	53,7383	125	56,8181
$20 \vdash 30$	54	25,2336	56	25,4545
$30 \vdash 40$	1	0,46729	5	2,27273

Tabela 4.31: Medidas resumo das drogas 1 e 2

Droga	Média	DP	Mediana
1	15,58	6,09	15,49
2	16,82	6,37	17,43

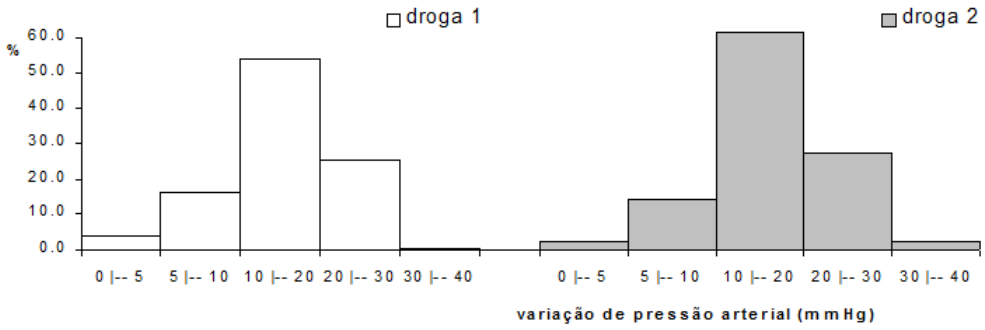


Figura 4.19: Histogramas para a variação de pressão arterial.

a) Com a finalidade de melhorar a apresentação dos resultados, faça as alterações que você julgar necessárias em cada uma das tabelas

e figura.

b) Calcule a média, o desvio padrão e a mediana da variação de pressão arterial para cada uma das duas drogas por meio do histograma.

c) Compare os resultados obtidos no item b) com aqueles obtidos diretamente dos dados da amostra (Tabela 4.31).

21) Numa cidade A em que não foi veiculada propaganda, a porcentagem de clientes que desistem do plano de TV a cabo depois de um ano é 14%. Numa cidade B, em que houve uma campanha publicitária, essa porcentagem é de 6%. Então, considerando uma aproximação de 2 asas decimais, podemos dizer a razão de chances (rc) de desistência entre as cidades A e B é

a) rc = 2,33 b) rc = 2,55 c) rc = 8,00 d) rc = 1,75 e) Nenhuma das respostas anteriores está correta.

22) Em um estudo realizado para avaliar o efeito do tabagismo nos padrões de sono foram consideradas amostras de tamanhos 12 e 15 de duas populações: Fumantes e Não Fumantes, respectivamente. A variável observada foi o tempo, em minutos, que se leva para dormir. Os correspondentes *boxplots* e gráficos de probabilidade Normal são apresentados nas Figuras 4.20 e 4.21.

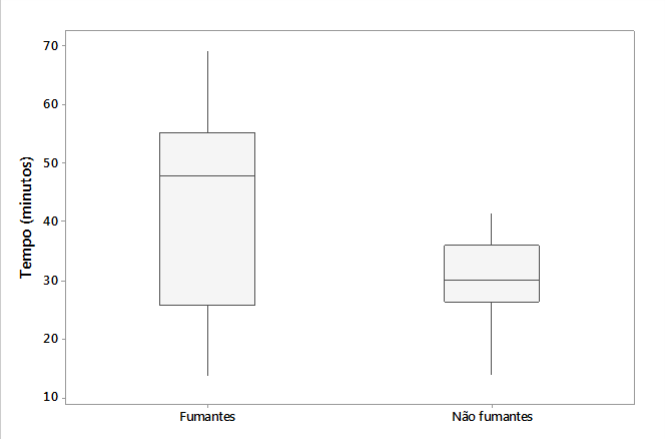


Figura 4.20: *Boxplots* do tempo até dormir nas populações Fumantes e Não Fumantes.

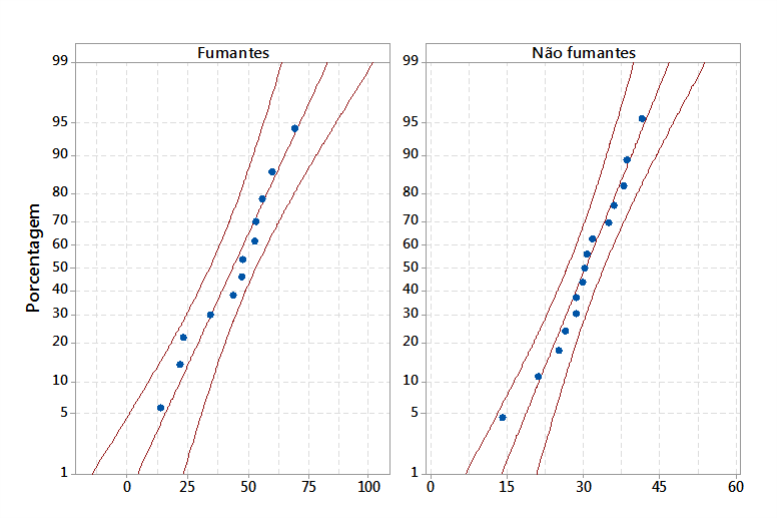


Figura 4.21: Gráfico QQ para as populações Fumantes e Não Fumantes.

Esses gráficos sugerem que:

a) a variabilidade do tempo é a mesma nas duas populações estudadas;

b) as suposições para a aplicação do teste t-Student para comparar as médias dos tempos nas duas populações estão válidas;

c) os fumantes tendem a apresentar um tempo maior para dormir do que os não fumantes;

d) as informações fornecidas permitem concluir que o estudo foi bem planejado;

e) nenhuma das respostas anteriores está correta.

23) Considere os dados do arquivo `entrevista`. Calcule estatísticas κ sem e com ponderação para quantificar a concordância entre as duas observadoras (G e P) para as variáveis "Impacto"e "Independência" e comente os resultados.

24) Considere os dados do arquivo `endometriose`. Construa um gráfico QQ para comparar as distribuições da variável "Idade" de pacientes do grupos Controle e Doente.

25) Considere duas amostras de uma variável X com n unidades amostrais cada. Utilize a definição (4.8) para mostrar que $\overline{\text{Var}(X)} = \text{Var}(X)$ quando as médias das duas amostras são iguais.

26) Utilize o método Delta para calcular uma estimativa da variância da razão de chances (ver Nota de Capítulo 7).

27) Considere os dados do arquivo `neonatos` contendo pesos de recém nascidos medidos por via ultrassonográfica (antes do parto) e ao nascer. Construa gráficos QQ e gráficos Bland-Altman para avaliar a concordância entre as duas distribuições. Comente os resultados.

Capítulo 5

Análise de dados de várias variáveis

> Nothing would be done at all if a man waited til he could do it so well that no one could find fault with it.
>
> John Henry Newman

5.1 Introdução

Em várias situações práticas, os valores de mais de duas variáveis são observados em cada unidade amostral (ou populacional). Por exemplo, o conjunto de dados disponível no arquivo `veiculos` corresponde a uma amostra de 30 veículos fabricados no Brasil ou importados em cada qual foram observadas 4 variáveis: "Preço" (preco), "Comprimento" (comp), "Potência do motor" (motor) e "Procedência" (proc). As três primeiras são variáveis quantitativas contínuas e a quarta é uma variável qualitativa nominal. O conjunto de dados disponível no arquivo `poluicao` contém 4 variáveis quantitativas contínuas, nomeadamente, concentrações atmosféricas de CO (monóxido de carbono) e O3 (ozônio), além de temperatura (temp) e umidade do ar (umid) observadas ao longo de 120 dias.

Modelos probabilísticos para esse tipo de dados envolvem distribuições conjuntas para as p variáveis, digamos $X_1, \ldots, X_p$, sob investigação. No caso discreto, eles envolvem funções de probabilidade $P(X_1 = x_1, \ldots, X_p = x_p)$, e no caso contínuo, funções densidade de probabilidade, $f(x_1, \ldots, x_p)$. Medidas resumo (média, variância, correlação etc.) são extensões daquelas estudadas no Capítulo 3 e são abordadas na Seção 5.4.

Quando todas as p variáveis são observadas em cada uma de n unidades amostrais, podemos dispo-las em uma matriz com dimensão $n \times p$, chamada **matriz de dados**. No exemplo dos veículos, essa matriz tem dimensão 30×4 e nem todos os seus elementos são numéricos. No conjunto de dados de poluição, a matriz de dados correspondente tem dimensão 120×4.

A análise de dados com essa estrutura deve levar em consideração a provável correlação (intraunidades amostrais) entre as variáveis estudadas.

Em geral, as observações realizadas em duas unidades amostrais diferentes não são correlacionadas embora haja exceções. No exemplo de dados de poluição, as unidades amostrais são os n diferentes dias e o conjunto das n observações de cada variável corresponde a uma série temporal. Nesse contexto, também se esperam correlações entre as observações realizadas entre unidades amostrais diferentes.

Embora seja possível considerar cada variável separadamente e aplicar as técnicas do Capítulo 3, a análise da relação entre elas precisa ser avaliada de forma conjunta. Muitas análises desse tipo de dados consistem na redução de sua dimensionalidade, considerando algum tipo de transformação que reduza o número de variáveis mas conserve a maior parte da informação do conjunto original. Com essa finalidade, uma técnica de análise de dados multivariados (dados de várias variáveis) bastante utilizada é a **Análise de Componentes Principais**, também conhecida por Análise de Funções Empíricas Ortogonais em muitas ciências físicas. Esse tópico será discutido no Capítulo 12.

Recursos gráficos para representar as relações entre as variáveis são mais complicados quando temos mais de duas variáveis. Neste livro trataremos apenas de alguns casos, com ênfase em três variáveis. Mais opções e detalhes podem ser encontrados em Chambers et al. (1983).

5.2 Gráficos para três variáveis

Gráfico do desenhista (*Draftsman's display)*

Esse tipo de gráfico consiste de uma matriz (ou dos componentes situados abaixo ou acima da diagonal principal) cujos elementos são gráficos de dispersão para cada par de variáveis. Muitas vezes incluem-se coeficientes de correlação de Pearson entre os diferentes pares de variáveis nos painéis situados acima ou abaixo da diagonal.

Exemplo 5.1. O gráfico do desenhista para as variáveis Preço, Comprimento e Potência do motor do arquivo `veiculos` está apresentado na Figura 5.1. Observam-se associações positivas tanto entre Potência do motor e Comprimento quanto entre potência do motor e preço: maiores potências do motor estão associadas tanto com maiores comprimentos quanto com preços maiores. Esse tipo de relação não está tão aparente quanto consideramos as variáveis Preço e Comprimento: para comprimentos entre 3,5 m e 4 m, os preços situam-se entre 5,000 e 15,000 enquanto os preços para veículos com comprimentos entre 4 e 5 metros distribuem-se entre 5,000 e 40,000. A Figura 5.2 contém o mesmo tipo de gráfico para as variáveis CO, O3 e temp do arquivo `poluicao` e não mostra evidências de associação entre cada par dessas variáveis.

Gráfico de dispersão simbólico

Gráficos de dispersão simbólicos ou estéticos (*aesthetic*) são essen-

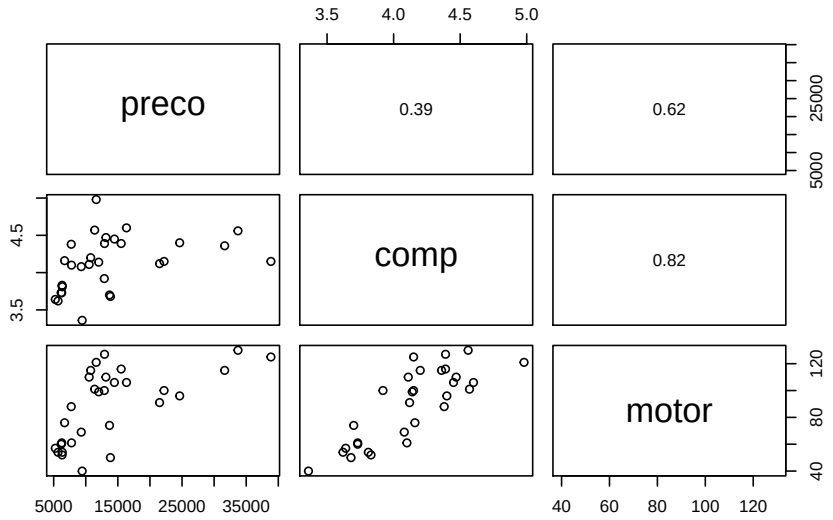


Figura 5.1: Gráfico do desenhista para os dados do arquivo `veiculos`.

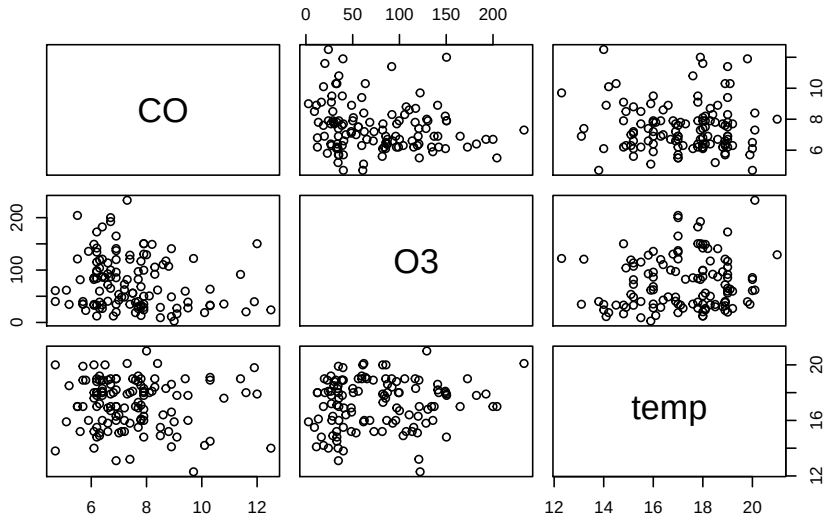


Figura 5.2: Gráfico do desenhista para os dados do arquivo `poluicao`.

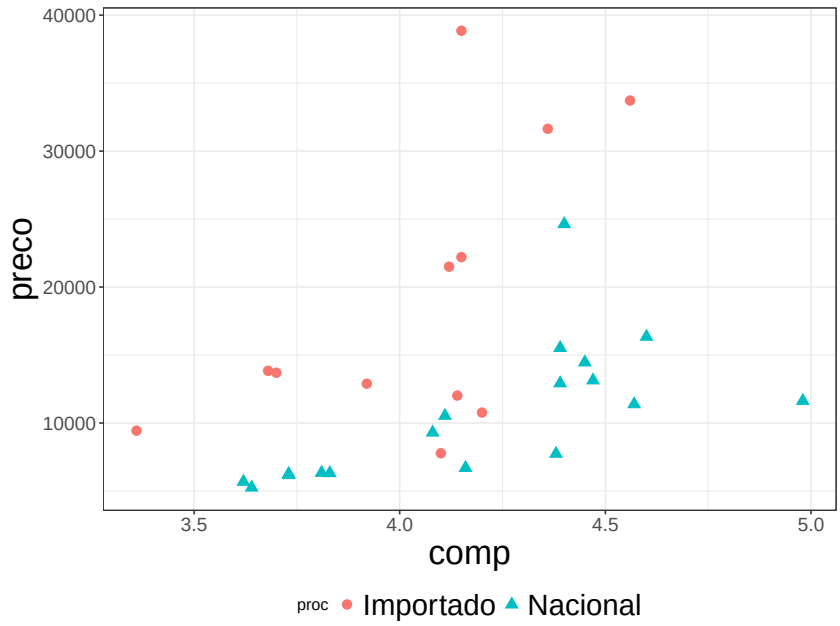


Figura 5.3: Gráfico de dispersão simbólico para os dados do arquivo `veiculos`.

cialmente gráficos de dispersão em que mais do que duas variáveis são representadas. Para distingui-las usam-se diferentes símbolos, cores ou forma dos pontos.

Exemplo 5.2. Consideremos novamente os dados do arquivo `veiculos`, concentrando a atenção em duas variáveis quantitativas [Preço (y) e Comprimento (x)] e em uma terceira variável qualitativa [Procedência (nacional ou importado)]. Para cada par (x, y), usamos um símbolo, como $\triangle$, para representar a categoria "nacional" e $\circ$, para indicar a categoria "importado". Essa escolha permite-nos construir o gráfico de dispersão disposto na Figura 5.3, em que se pode notar que os preços maiores correspondem, de modo geral, a carros importados, o que era esperado. Os carros nacionais pequenos têm os menores preços. Os comandos do pacote `ggplot2` do `R` utilizados para a construção do gráfico disposto na Figura 5.3 são

```
library(ggplot2)
library{gdata}

veiculos<-read.xls("/dados/veiculos.xls",sep=",",dec=",",h=T)
g1 <- ggplot(veiculos,aes(comp,preco))
      + geom_point(aes(shape=proc, colour=proc), size=3)
      + theme_bw()
g2 <- g1 + theme(axis.title = element_text(size=23))
g3 <- g2 + theme(legend.position="bottom",
                 legend.direction="horizontal",
```

```
                  legend.text=element_text(size=20))
g4 <- g3 + theme(axis.text.x = element_text(face="plain", size=13),
                  axis.text.y = element_text(face="plain", size=13))
```

Outra alternativa para a representação gráfica das associações entre três variáveis quantitativas desse conjunto de dados consiste em adotar um símbolo com diferentes tamanhos para representar cada uma delas. Por exemplo, na Figura 5.4, apresentamos o gráfico de dispersão de Preço *versus* Comprimento, com a variável Potência do Motor representada por círculos com tamanhos variando conforme a potência: círculos menores para potências entre 40 e 70, círculos médios para potências entre 70 e 100 e círculos maiores para potências entre 100 e 130. O gráfico permite evidenciar que carros com maior potência do motor são em geral mais caros e têm maior comprimento.

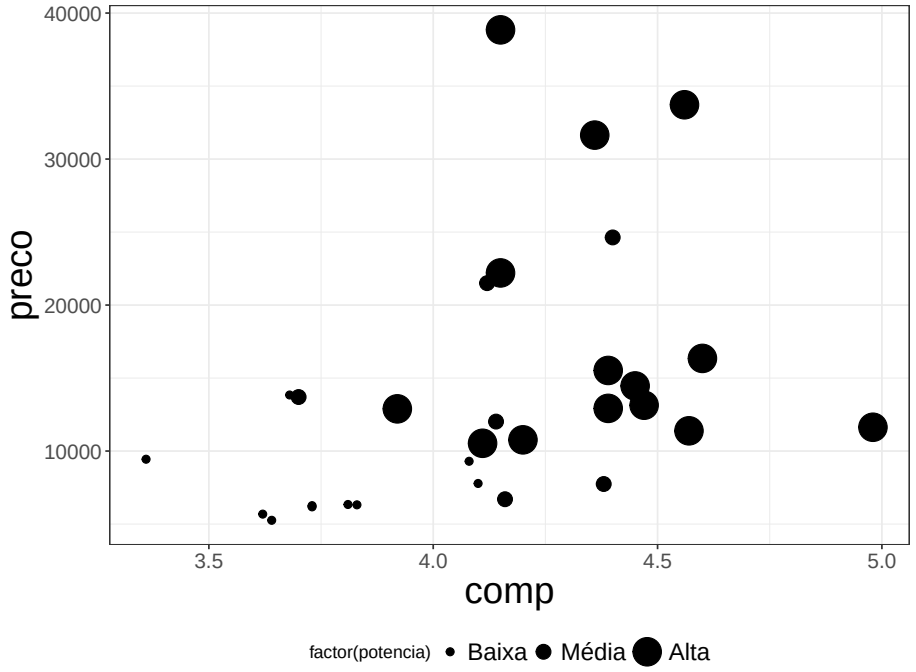


Figura 5.4: Gráfico de dispersão simbólico para os dados do CD-veiculos com potência do motor.

Os comandos do pacote `ggplot2` do `R` utilizados para a construção do gráfico disposto na Figura 5.4 são

```
library(ggplot2)
library(gdata)

veiculos<-read.table("/dados/veiculos.xls", sep=",",dec=",",h=T)
categ_motor=rep(NA,length(motor))
categ_motor[motor>=40 & motor<70]="Baixa Potencia"
```

```
categ_motor[motor>=70 & motor<100]="Media Potencia"
categ_motor[motor>=100 & motor<=130]="Alta Potencia"
categ_motor=factor(categ_motor)
potencia = 2*c(categ_motor == "Baixa Potencia")+
  4*c(categ_motor == "Media Potencia")+
  8*c(categ_motor== "Alta Potencia")

ggplot(veiculos, aes(comp,preco))
       + geom_point(aes(size = factor(potencia)))

g1 <- ggplot(veiculos,aes(comp,preco))
      + geom_point(aes(size = factor(potencia))) + theme_bw()
g2 <- g1 + theme(axis.title = element_text(size=23))
g3 <- g2 + theme(legend.position="bottom",
 legend.direction="horizontal",
 legend.text=element_text(size=15))
g4 <- g3 + theme(axis.text.x = element_text(face="plain",
           size=13), axis.text.y = element_text(face="plain",
           size=13))
g5 <- g4 + scale_size_manual(labels = c("Baixa", "Media",
                             "Alta"), values = c(2, 4, 8))
```

Exemplo 5.3. No pacote `ggplot2` encontramos o conjunto de dados `mpg`, que consiste de observações de 38 modelos de carros nos EUA, com várias variáveis, dentre as quais destacamos: `displ` = potência do motor, `hwy` = eficiência do carro em termos de gasto de combustível, `class` = tipo do carro (duas portas, compacto, SUV etc.) e `drv` = tipo de tração (4 rodas, rodas dianteiras e rodas traseiras).

Consideremos o comando

```
ggplot(data=mpg)
          + geom_point(mapping=aes(x=displ,y=hwy,color=drv))
          + geom_smooth(mapping=aes(x=displ,y=hwy,color=drv))
```

que usa a opção `geom_smooth` para ajustar curvas suaves aos dados (usando o procedimento de suavização **lowess**) de cada conjunto de pontos da variável `drv` (*drive*), ou seja uma curva para os pontos com o valor 4 (*four-wheel drive*), outra para os pontos com o valor f (*front-wheel drive*) e uma curva para os pontos com valor r (*rear-wheel drive*). O resultado está apresentado na Figura 5.5. As curvas lowess são úteis para identificar possíveis modelos de regressão que serão discutidos no Capítulo 6. Detalhes sobre curvas lowess podem ser obtidos na Nota de Capítulo 2.

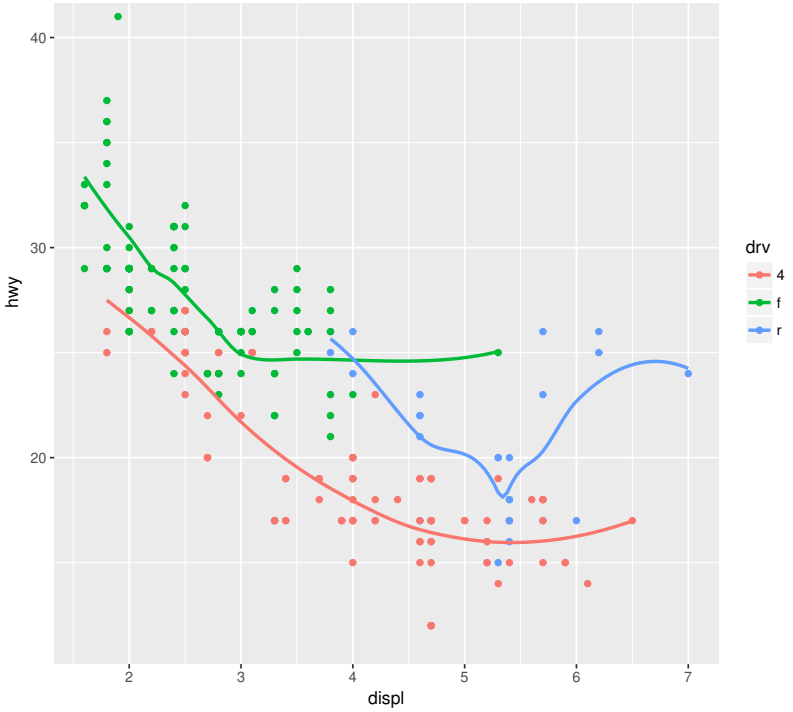


Figura 5.5: Gráfico de dispersão simbólico das variáveis `hwy` versus `displ`, categorizado pela variável `drv` com pontos e curvas.

Partição e Janelamento

Uma abordagem alternativa aos gráficos de dispersão simbólicos consiste em dividir as n observações disponíveis em subconjuntos de acordo com os valores de uma das variáveis e construir um gráfico de dispersão envolvendo as outras duas variáveis para cada subconjunto.

Por exemplo, para os dados do arquivo `veiculos`, podemos construir gráficos de dispersão para as variáveis Preço e Comprimento de acordo com a Procedência (nacional ou importado), como na Figura 5.3 ou construir gráficos de dispersão para Preço e Comprimento segundo as classes em que foram divididas as unidades amostrais com base na variável quantitativa Potência de motor. Esse gráfico está apresentado na Figura 5.6.

Figura 5.6: Janelamento para as variáveis Preço versus Comprimento, categorizado pela variável Potência do motor.

Gráfico de perfis médios

Os gráficos de perfis médios considerados no Capítulo 4 para duas variáveis podem ser facilmente estendidos para acomodar situações com duas variáveis explicativas categorizadas, usualmente denominadas **fatores** e uma variável

resposta. Como ilustração, consideremos o conjunto de dados `arvores` com o objetivo de comparar as concentrações médias de Mg obtidas nas cascas de três espécies de árvores localizadas nas proximidades de vias com diferentes intensidades de tráfego. Nesse contexto, estamos diante de um problema com dois fatores, nomeadamente, "Espécie de árvores" e "Tipo de via" e uma variável resposta contínua, "Concentração de Mg". O gráfico de perfis médios correspondente, apresentado na Figura 5.7 pode ser obtido por intermédio dos seguintes comandos

```
> library(gdata)
> library(ggplot2)
> library(plyr)
>
> arvores <-read.xls("/home/jmsinger/Desktop/arvores.xls",
                      sheet='dados', method="tab")
> resumo <- ddply(arvores, c("especie", "tipovia"), summarise,
+ N    = sum(!is.na(Mg)),
+ mean = mean(Mg, na.rm=TRUE),
+ sd   = sd(Mg, na.rm=TRUE),
+ se   = sd / sqrt(N)
+ )
>
> pd <- position_dodge(0.1)
> ggplot(resumo, aes(x=tipovia, y=mean, colour=especie)) +
+   geom_errorbar(aes(ymin=mean-se, ymax=mean+se),  width=.1,
                  position=pd) +
+   geom_line(aes(group = especie)) + geom_point(position=pd) +
+   theme_bw() + labs(x="Tipo de via", y="Concentração de Mg") +
+   theme(text=element_text(size=18))
```

O gráfico permite concluir que as concentrações médias Mg nas Tipuanas são mais elevadas que aquelas obtidas em Alfeneiros, cujas concentrações médias de Mg são mais elevadas que aquelas obtidas em Sibipirunas. Além disso, nota-se que a variação das concentrações médias de Mg são similares para Tipuanas e Alfeneiros para os quatro tipos de vias considerados. As concentrações médias de Mg em Sibipirunas, por outro lado, seguem um padrão diferente.

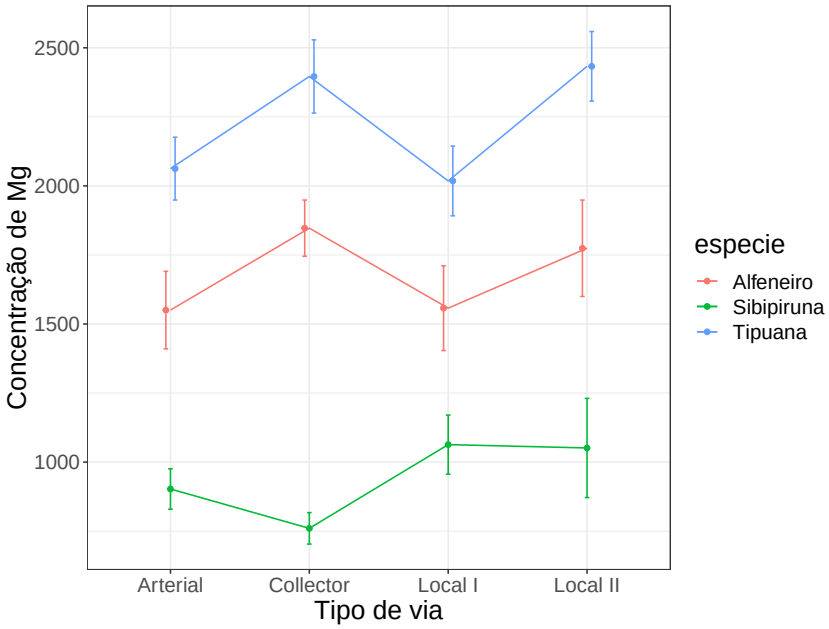


Figura 5.7: Gráfico de perfis médios para a concentração de Mg em cascas de árvores (as barras correspondem a erros padrões).

Nesse tipo de estudo, o objetivo inferencial é avaliar o "**efeito**" de cada fator e de sua "**interação**" na distribuição de uma variável resposta quantitativa contínua. Os termos "efeito" e "interação"estão entre aspas porque precisam ser definidos. Quando os as observações são independentes e a distribuição (populacional) da variável resposta é Normal com a mesma variância para todas as combinações dos níveis dos fatores, as comparações de interesse restringem-se aos correspondentes valores esperados. Esse é o típico figurino dos problemas analisados por meio da técnica conhecida como **Análise de Variância**, comumente cognominada ANOVA (do inglês, *ANalysis Of VAriance*).

Com o objetivo de definir os "efeitos" dos fatores e sua "interação", consideremos um exemplo simples em que cada um dos dois fatores tem dois níveis. Um dos fatores, que representamos por A, por exemplo, pode ser o tipo de droga (com níveis ativa e placebo) e o outro, digamos B, pode ser faixa etária (com níveis < 60 anos e ≥ 60 anos) e a variável resposta poderia ser pressão diastólica.

De uma forma geral, admitamos que m unidades amostrais tenham sido observadas sob cada tratamento, *i.e.*, para cada combinação dos a níveis do fator A e dos b níveis do fator B e que a variável resposta seja denotada por y. A estrutura de dados coletados sob esse esquema está apresentada na Figura 5.1.

Tabela 5.1: Estrutura de dados para ANOVA com dois fatores

Droga	Idade	Paciente	PDiast	Droga	Idade	Paciente	PDiast
Ativa	< 60	1	y_{111}	Placebo	< 60	1	y_{211}
Ativa	< 60	2	y_{112}	Placebo	< 60	2	y_{212}
Ativa	< 60	3	y_{113}	Placebo	< 60	3	y_{213}
Ativa	≥ 60	1	y_{121}	Placebo	≥ 60	1	y_{221}
Ativa	≥ 60	2	y_{122}	Placebo	≥ 60	2	y_{222}
Ativa	≥ 60	3	y_{123}	Placebo	≥ 60	3	y_{223}

O "efeito" de cada um dos fatores e da "interação" entre eles podem ser definidos em termos dos valores esperados das distribuições da resposta sob os diferentes tratamentos (combinações dos níveis dos dois fatores). Um modelo comumente considerado para análise inferencial de dados com essa estrutura é

$$y_{ijk} = \mu_{ij} + e_{ijk}, \tag{5.1}$$

$i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, m$, em que $\mathrm{E}(e_{ijk}) = 0$, $\mathrm{Var}(e_{ijk}) = \sigma^2$ e $\mathrm{E}(e_{ijk}e_{i'j'k'}) = 0$, $i \neq i'$ ou $j \neq j'$ ou $k \neq k'$, ou seja, os e_{ijk} são erros não correlacionados. Aqui, y_{ijk} denota a resposta observada para a k-ésima unidade amostral submetida ao tratamento definido pela combinação do nível i do fator A e nível j do fator B.

Esta é a **parametrização** conhecida como de **parametrização de médias de celas** pois o **parâmetro de localização** μ_{ij} corresponde ao valor esperado (médio) da resposta de unidades amostrais submetidas ao tratamento correspondente à combinação do nível i do fator A e nível j do fator B. Outra parametrização bastante utilizada está discutida na Nota de Capítulo 3.

Fazendo $a = b = 2$ para facilidade de exposição, o **efeito** do fator A (droga) para unidades amostrais no nível j do fator B (faixa etária) pode ser definido como a diferença $\mu_{1j} - \mu_{2j}$, que, por exemplo, corresponde à diferença entre o valor esperado da pressão diastólica de unidades amostrais na faixa etária j submetidas à droga 1 (ativa) e o valor esperado da pressão diastólica de unidades amostrais na mesma faixa etária submetidas à droga 2 (placebo). Analogamente, o **efeito** do fator B (faixa etária) para unidades amostrais no nível i do fator A (droga) pode ser definido como a diferença $\mu_{i1} - \mu_{i2}$.

A **interação** entre os fatores A e B pode ser definida como a diferença entre o efeito do fator A para unidades amostrais no nível 1 do fator B e o efeito do fator A para unidades amostrais no nível 2 do fator B, nomeadamente, $(\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$. Outras definições equivalentes, como $(\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$, também podem ser utilizadas. A escolha entre as alternativas deve ser feita em função dos detalhes do problema; por exemplo, se a droga 1 for uma droga padrão e a faixa etária 1 corresponder a indivíduos mais jovens, esta última proposta pode ser mais conveniente.

Quando a interação é nula, o efeito do fator A é o mesmo para unidades amostrais submetidas a qualquer um dos níveis do fator B e pode-se definir o **efeito principal** do fator A como $(\mu_{11}+\mu_{12})/2-(\mu_{21}+\mu_{22})/2$, que corresponde à diferença entre o valor esperado da resposta para unidades amostrais submetidas ao nível 1 do fator A e o valor esperado da resposta para unidades amostrais submetidas ao nível 2 do fator A (**independentemente** do nível do fator B). Similarmente, o efeito principal do fator B pode ser definido como $(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$.

Em muitos casos, essas definições de efeitos principais podem ser consideradas mesmo na presença de interação, desde que ela seja **não essencial**. A interação entre os fatores A e B é não essencial quando as diferenças $\mu_{11} - \mu_{21}$ e $\mu_{12} - \mu_{22}$ têm o mesmo sinal, mas magnitudes diferentes. Por exemplo, se $\mu_{11} - \mu_{21} = K_1 > 0$ e $\mu_{12} - \mu_{22} = K_2 > 0$ com $K_1 \neq K_2$, a resposta esperada sob o nível 1 do fator A é maior que a resposta esperada sob o nível 2 do fator A tanto no nível 1 quanto no nível 2 do fator B, embora as magnitudes das diferenças não sejam iguais. Se essas magnitudes tiverem sinais diferentes, a interação é essencial. Por outro lado, se $K_1 = K_2$, não há interação. O leitor pode consultar Kutner et al. (2004) para uma discussão sobre a consideração de efeitos principais em situações com interação não essencial. Na Figura 5.8 apresentamos gráficos de perfis médios (populacionais) com interações essencial e não essencial entre dois fatores, A e B, cada um com dois níveis.

Na prática, tnato a interação entre os fatores bem como seus efeitos (que são parâmetros populacionais são estimados pelas correspondentes funções das médias amostrais $\overline{y}_{ij} = m^{-1}\sum_{k=1}^{m} y_{ijk}$. Os correspondentes gráficos de perfis médios são construídos com essas médias amostrais e desvios padrões (ou erros padrões) associados e servem para sugerir uma possível interação entre os fatores envolvidos ou os seus efeitos.

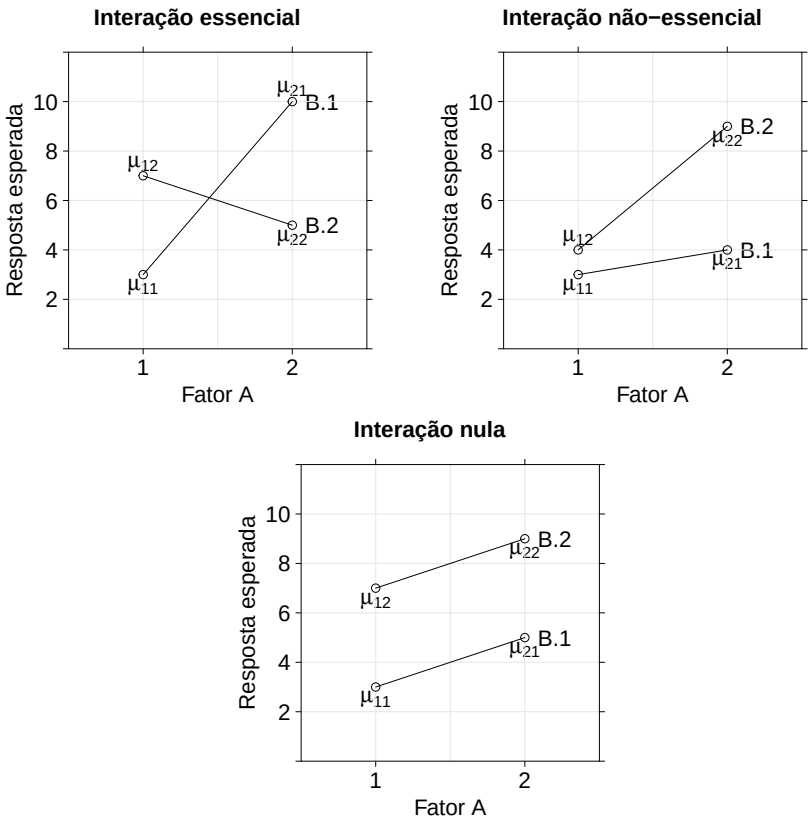


Figura 5.8: Gráfico de perfis médios (populacionais) com diferentes tipos de interação.

Exemplo 5.4. Consideremos um estudo cujo objetivo é avaliar o efeito de dois fatores, a saber, tipo de adesivo odontológico e instante em que foi aplicada uma carga cíclica na resistência à tração de corpos de prova odontológicos (variável resposta). O fator **Adesivo** tem três níveis (CB, RX e RXQ) e o fator **Instante** tem três níveis (início, após 15 minutos e após 2 horas) para os adesivos CB e RXQ e quatro níveis (após fotoativação além de início, após 15 minutos e após 2 horas) para o cimento RX. Os dados, disponíveis no arquivo `adesivo` estão dispostos na Tabela 5.2 e contêm omissões causadas pela quebra acidental dos corpos de prova. Detalhes sobre o estudo podem ser encontrados em Witzel et al. (2000).

Tabela 5.2: Resistência à tração de corpos de prova de um estudo sobre cimentos odontológicos

Adesivo	Instante carga	Repet	Resist	Adesivo	Instante carga	Repet	Resist
CB	inic	1	8,56	RX	2h	1	16,76
CB	inic	2	5,01	RX	2h	2	16,80
CB	inic	3	2,12	RX	2h	3	13,07
CB	inic	4	1,70	RX	2h	4	11,47
CB	inic	5	4,78	RX	2h	5	15,86
CB	15min	1	5,67	RX	fativ	1	13,42
CB	15min	2	4,07	RX	fativ	2	13,82
CB	15min	3	5,99	RX	fativ	3	19,63
CB	15min	4	5,52	RX	fativ	4	15,60
CB	15min	5		RX	fativ	5	17,87
CB	2h	1	8,57	RXQ	inic	1	3,95
CB	2h	2	6,94	RXQ	inic	2	6,49
CB	2h	3		RXQ	inic	3	4,60
CB	2h	4		RXQ	inic	4	6,59
CB	2h	5		RXQ	inic	5	4,78
RX	inic	1	20,81	RXQ	15min	1	8,14
RX	inic	2	12,14	RXQ	15min	2	3,70
RX	inic	3	9,96	RXQ	15min	3	
RX	inic	4	15,95	RXQ	15min	4	
RX	inic	5	19,27	RXQ	15min	5	
RX	15min	1	14,25	RXQ	2h	1	4,39
RX	15min	2	14,21	RXQ	2h	2	6,76
RX	15min	3	13,60	RXQ	2h	3	4,81
RX	15min	4	11,04	RXQ	2h	4	10,68
RX	15min	5	21,08	RXQ	2h	5	

Médias e desvios padrões da resistência à tração para as observações realizadas sob cada tratamento (correspondentes ao cruzamento dos níveis de cada fator) estão apresentados na Tabela 5.3.

O gráfico de perfis médios correspondente está apresentado na Figura 5.9.

Tabela 5.3: Estatísticas descritivas para os dados da Tabela 5.2

Adesivo	Instante	n	Média	Desvio Padrão
CB	0 min	5	4,43	2,75
	15 min	4	5,31	0,85
	120 min	2	7,76	1,15
RXQ	0 min	5	5,28	1,19
	15 min	2	5,92	3,14
	120 min	4	6,66	2,87
RX	0 min	5	15,63	4,60
	15 min	5	14,84	3,73
	120 min	5	14,79	2,40
	fativ	5	16,07	2,66

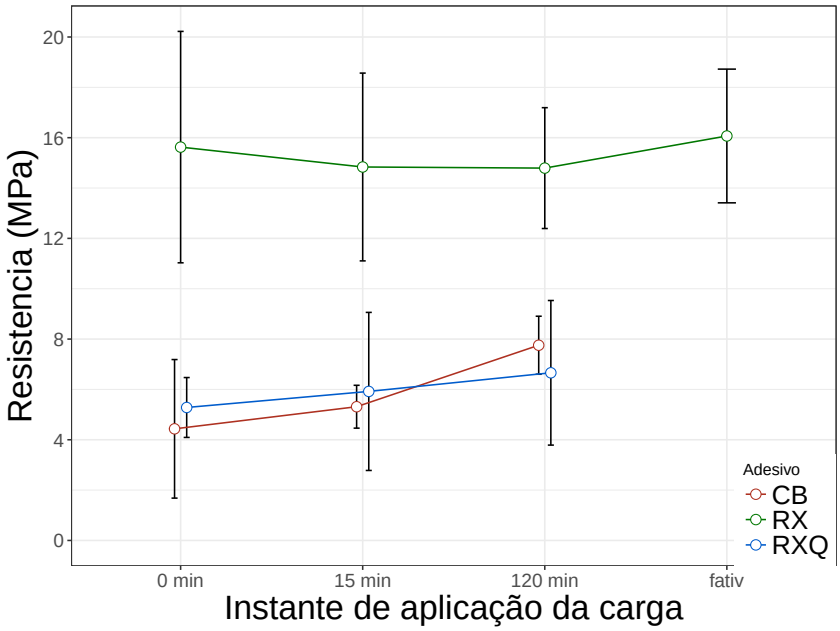


Figura 5.9: Gráfico de perfis de médias (com barras de desvios padrões) para os dados da Tabela 5.2.

Esse gráfico sugere que não existe interação entre os dois fatores (pois os perfis são "paralelos" (lembremos que os perfis apresentados são amostrais e que servem apenas para sugerir o comportamento dos perfis populacionais correspondentes). Além disso, a variabilidade dos dados (aqui representada pelas barras de desvios padrões) deve ser levada em conta para avaliar as possíveis diferenças. Nesse contexto, podemos esperar um efeito principal do fator Adesivo, segundo o qual, os adesivos CB e RXQ têm valores esperados iguais, mas menores que valor esperado do adesivo RX. Também é razoável

esperar que não exista um efeito principal de Instante de aplicação, dado que os três perfis são "paralelos" ao eixo das abscissas. Finalmente, convém reafirmar que as conclusões acima são apenas exploratórias precisam ser confirmadas por técnicas de ANOVA para efeitos inferenciais. Os seguintes comandos `R` geram a tabela ANOVA apresentada em seguida.

```
library(gdata)
adesivo<-read.xls("/dados/adesivo.xls", sheet='dados',
                  method="tab")
adesivo.anova <- aov(resistencia ~ adesivo + instante +
adesivo*instante, data=adesivo)
summary(adesivo.anova)
```

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
adesivo           2  987.8   493.9  59.526 1.65e-11 ***
instante          3    9.3     3.1   0.373    0.773
adesivo:instante  4   16.5     4.1   0.498    0.737
Residuals        32  265.5     8.3
```

O resultado não sugere evidências nem de interação entre Adesivo e Instante de aplicação ($p = 0,737$) nem de efeito principal de Instante de aplicação ($p = 0,773$), mas sugere forte evidência de efeito de Adesivo ($p < 0,001$).

Comparações múltiplas entre os níveis de Adesivo realizadas por meio da técnica de Tukey a partir do comando

```
TukeyHSD(adesivo.anova, which = "adesivo")
```

corroboram a sugestão de que os efeitos dos adesivos CB e RXQ são iguais ($p < 0,899$), porém diferentes do efeito do adesivo RXQ ($p < 0,001$).

```
             diff        lwr       upr     p adj
RX-CB   9.9732273   7.316138 12.630317 0.0000000
RXQ-CB  0.5418182  -2.476433  3.560069 0.8986306
RXQ-RX -9.4314091 -12.088499 -6.774319 0.0000000
```

5.3 Gráficos para quatro ou mais variáveis

Os mesmos tipos de gráficos examinados na seção anterior podem ser considerados para a análise conjunta de quatro ou mais variáveis. Como ilustração, consideremos dados de concentração de elementos químicos observados em cascas de diferentes espécies de árvores na cidade de São Paulo, utilizados para avaliar os níveis de poluição. Os dados estão disponíveis no arquivo `arvores`.

Exemplo 5.5. Na Figura 5.10 apresentamos um gráfico do desenhista com $\binom{4}{2} = 6$ painéis correspondentes aos elementos Mn, Fe, Cu e Zn observados em árvores da espécie *Tipuana* localizadas junto a vias coletoras. Aqui também observam-se evidências de correlações moderadas entre as variáveis.

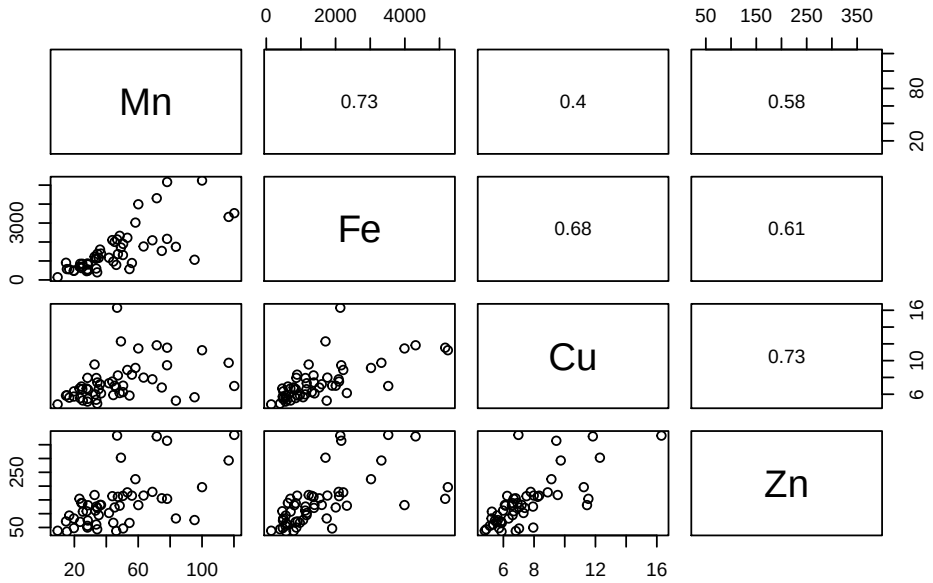


Figura 5.10: Gráfico do desenhista para os dados da concentração de elementos químicos em cascas de árvores.

Outros tipos de gráficos podem ser encontrados em Cleveland (1979) e Chambers et al. (1983), entre outros.

5.4 Medidas resumo multivariadas

Consideremos valores de p variáveis $X_1, \ldots, X_p$, medidas em n unidades amostrais dispostos na forma de uma matriz de dados $\mathbf{X}$, de ordem $n \times p$, *i.e.*,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1v} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2v} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{ii} & x_{i2} & \cdots & x_{iv} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nv} & \cdots & x_{np} \end{bmatrix}. \tag{5.2}$$

Para cada variável X_i podemos considerar as medidas resumo já estudadas no Capítulo 3 (média, mediana, quantis, variância etc.). Para cada par de variáveis, X_i e X_j, também podemos considerar as medidas de correlação (linear) já estudadas no Capítulo 4, a saber, covariância e coeficiente de correlação. O vetor de dimensão $p \times 1$ contendo as p médias é chamado de **vetor de médias**. Similarmente, a matriz simétrica com dimensão $p \times p$

contendo as variâncias ao longo da diagonal principal e as covariâncias dispostas acima e abaixo dessa diagonal é chamada de **matriz de covariâncias** de $X_1, \ldots, X_p$ ou, equivalentemente, do vetor $\mathbf{X} = (X_1, \ldots, X_p)^\top$. Tanto o vetor de médias quanto a matriz de covariâncias (ou de correlações) correspondentes ao vetor de variáveis podem ser facilmente calculados por meio de operações matriciais como detalhado na Nota de Capítulo 1.

Exemplo 5.6 Consideremos as variáveis CO, O3, Temp e Umid do arquivo `poluicao`. A matriz de covariâncias correspondente é

	CO	O3	Temp	Umid
CO	2.38	-14.01	-0.14	1.46
O3	-14.01	2511.79	9.43	-239.02
Temp	-0.14	9.43	3.10	0.14
Umid	1.46	-239.02	0.14	153.63

Note que Cov(CO,O3)= -14,01 = Cov(O3,CO) etc. Para obter a correspondente **matriz de correlações**, basta usar a definição (4.10) para cada par de variáveis, obtendo-se a matriz

	CO	O3	Temp	Umid
CO	1.00	-0.18	-0.05	0.08
O3	-0.18	1.00	0.11	-0.38
Temp	-0.05	0.11	1.00	0.01
Umid	0.08	-0.38	0.01	1.00

As correlações entre as variáveis são irrelevantes, exceto para O3 e Umid.

Exemplo 5.7 Consideremos agora os dados do arquivo `iris`. A matriz de correlações correspondente é

	i1	i2	i3	i4
i1	1.00	-0.12	0.87	0.82
i2	-0.12	1.00	-0.43	-0.37
i3	0.87	-0.43	1.00	0.96
i4	0.82	-0.37	0.96	1.00

em que i1=Sepal.Length, i2=Sepal.Width, i3=Petal.Length e i4=Petal.Width.

Na Figura 5.11 dispomos o gráfico do desenhista para as quatro variáveis, com o acréscimo dos coeficientes de correlação de Pearson na parte superior, confirmando a relevância das associação positiva observada entre os pares de variáveis (i1,i3), (i1,i4) e (i3,i4).

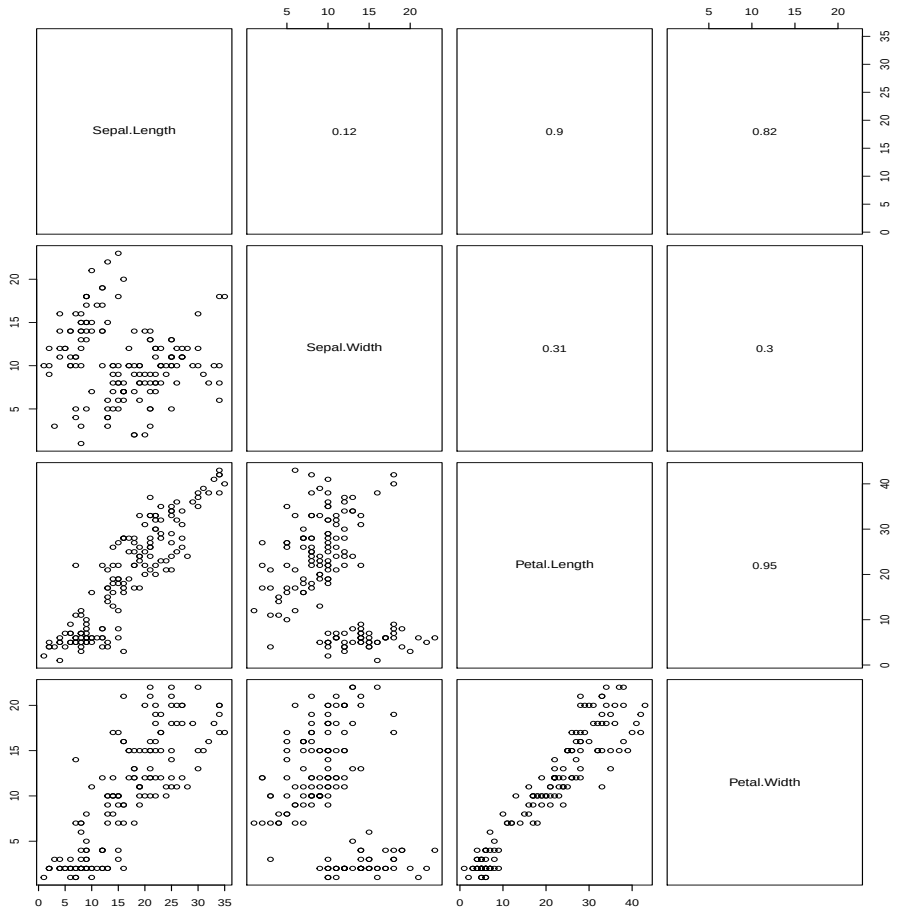


Figura 5.11: Gráfico do desenhista para os dados do CD-iris.

5.5 Tabelas de contingência de múltiplas entradas

A análise de dados de três ou mais variáveis qualitativas (ou quantitativas categorizadas) pode ser realizada nos moldes daquela abordada na Seção 4.2 para duas variáveis. A distribuição de frequências conjunta correspondente pode ser representada por meio de tabelas de contingência de múltiplas entradas. Nesse contexto, as frequências de um conjunto de dados com três variáveis qualitativas com 3, 3 e 2 níveis, respectivamente, são representadas numa tabela $3 \times 3 \times 2$. Como ilustração, consideremos o seguinte exemplo.

Exemplo 5.8 Consideremos as variáveis `dismenorreia`, `esterelidade` e `endometriose` do arquivo `endometrise2`. A tabela de frequências conjunta pode ser obtida por meio dos seguintes comandos

```
> endomet1 <-read.xls("/home/jmsinger/Desktop/endometriose2.xls",
             sheet='dados', method="tab")
> endomet1$dismenorreia <- reorder(endomet1$dismenorreia,
```

```
                          new.order=c("nao", "leve", "moderada",
                          "intensa", "incapacitante"))
> attach(endomet1)
> tab <- ftable(dismenorreia, esterelidade, endometriose)
> tab
                          endometriose nao sim
dismenorreia  esterelidade
nao           nao                     482  36
              sim                     100  27
leve          nao                     259  31
              sim                      77  14
moderada      nao                      84  71
              sim                      31  45
intensa       nao                     160 134
              sim                      52  67
incapacitante nao                     106  43
              sim                      28  24
```

Quando o objetivo é estudar as relações de dependência entre as três variáveis encaradas como resposta, as frequências relativas calculadas em relação ao total de pacientes é obtida por meio do comendo

```
> tabprop <- prop.table(tab)
> tabprop <- round(tabprop, 2)
> tabprop
                          endometriose  nao  sim
dismenorreia  esterelidade
nao           nao                     0.26 0.02
              sim                     0.05 0.01
leve          nao                     0.14 0.02
              sim                     0.04 0.01
moderada      nao                     0.04 0.04
              sim                     0.02 0.02
intensa       nao                     0.09 0.07
              sim                     0.03 0.04
incapacitante nao                     0.06 0.02
              sim                     0.01 0.01
```

Nesse caso, as análises de interesse geralmente envolvem hipóteses de independência conjunta, independência marginal e independência condicional e são estudadas com técnicas de anaálise de dados categorizados, por meio de modelos log-lineares.

Alternativamente, o interesse pode recair na avaliação do efeito de duas das variáveis (encaradas como fatores) e de sua interação na distribuição da outra variável, encarada como variável resposta, como o mesmo espírito daquele envolvendo problemas de ANOVA. As frequências relativas correspondentes devem ser calculadas em relação ao total das linhas da tabela. Com essa finalidade, consideremos os comandos

```
> tabprop12 <- prop.table(tab,1)
> tabprop12 <- round(tabprop12,2)
> tabprop12
                           endometriose  nao  sim
dismenorreia  esterelidade
nao           nao                        0.93 0.07
              sim                        0.79 0.21
leve          nao                        0.89 0.11
              sim                        0.85 0.15
moderada      nao                        0.54 0.46
              sim                        0.41 0.59
intensa       nao                        0.54 0.46
              sim                        0.44 0.56
incapacitante nao                        0.71 0.29
              sim                        0.54 0.46
```

Medidas de associação entre `esterelidade` e `endometriose` podem ser obtidas para cada nível de `dismenorreia` por meio das tabelas marginais, com os seguintes comandos do pacote `vcd`

```
> nao <- subset(endomet1, dismenorreia =="nao", na.rm=TRUE)
> attach(nao)
> tab1 <- ftable(esterelidade, endometriose)
> tab1
             endometriose nao sim
esterelidade
nao                        482  36
sim                        100  27
> assocstats(tab1)
                    X^2 df   P(> X^2)
Likelihood Ratio 19.889  1 8.2064e-06
Pearson          23.698  1 1.1270e-06

Phi-Coefficient   : 0.192
Contingency Coeff.: 0.188
Cramer's V        : 0.192
```

Razões de chances (e intervalos de confiança) correspondentes às variáveis `esterelidade` e `endometriose` podem ser obtidas para cada nível de `dismenorreia` são obtidas com os seguintes comandos do pacote `epiDisplay`

```
> endomet1 %$% mhor(esterelidade, endometriose, dismenorreia,
                    graph = F)

Stratified analysis by  dismenorreia
                              OR lower lim. upper lim.  P value
dismenorreia nao            3.61      2.008       6.42 7.73e-06
dismenorreia leve           1.52      0.708       3.12 2.63e-01
```

```
dismenorreia moderada       1.71      0.950       3.12 6.86e-02
dismenorreia intensa        1.54      0.980       2.42 5.11e-02
dismenorreia incapacitante  2.10      1.042       4.25 2.69e-02
M-H combined                1.91      1.496       2.45 1.36e-07

M-H Chi2(1) = 27.77 , P value = 0
Homogeneity test, chi-squared 4 d.f. = 6.9 , P value = 0.141
```

O resultado obtido por meio da razão de chances combinada pelo **método de Mantel-Haenszel** sugerem que a chance de endometriose para pacientes com sintomas de esterilidade é 1,91 (IC95%: 1,5 - 2,45) vezes a chance de endometriose para pacientes sem esses sintomas, independentemente da intensidade da dismenorreia. Detalhes sobre a técnica de Mantel-Haenszel são apresentados na Nota de Capítulo 4.

5.6 Notas de capítulo

1) **Notação matricial para variáveis multivariadas**

Nesta seção iremos formalizar a notação matricial usualmente empregada para representar medidas resumo multivariadas.

Denotemos cada coluna da matriz de dados $\mathbf{X}$ por $\mathbf{x}_j,\ j = 1, \ldots, p$, com elementos $x_{ij},\ i = 1, \ldots, n$. Então, definindo $\overline{x}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$, o vetor de médias é expresso como $\overline{\mathbf{x}} = (\overline{x}_1, \ldots, \overline{x}_p)^\top$.

Se denotarmos por $\mathbf{1}_n$ o vetor coluna de ordem $n \times 1$ contendo todos os elementos iguais a um, podemos escrever o vetor de médias como

$$\overline{\mathbf{x}}^\top = \frac{1}{n} \mathbf{1}_n^\top \mathbf{X} = [\overline{x}_1, \ldots, \overline{x}_p]. \tag{5.3}$$

A matriz de desvios de cada observação em relação à média correspondente é

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}^\top \tag{5.4}$$

de forma que a matriz de covariâncias pode ser expressa como

$$\mathbf{S} = \frac{1}{n-1} \mathbf{Y}^\top \mathbf{Y}. \tag{5.5}$$

Na diagonal principal de $\mathbf{S}$ constam as variâncias amostrais $s_{jj},\ j = 1, \ldots, p$ e nas demais diagonais temos as covariâncias amostrais

$$s_{uv} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{iu} - \overline{x}_u)(x_{iv} - \overline{x}_v), \quad u, v = 1, \ldots, p,$$

em que $s_{uv} = s_{vu}$, para todo u, v. Ou seja, podemos escrever

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}.$$

O desvio padrão amostral da j-ésima componente é $s_j = (s_{jj})^{1/2},\ j = 1, \ldots, p$. Denotando por $\mathbf{D}$ a matriz diagonal de ordem $p \times p$ com o j-ésimo elemento da diagonal igual a s_j, a **matriz de correlações** é definida por

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1} = [r_{uv}]. \tag{5.6}$$

em que $r_{vv} = r_v = 1,\ v = 1, \ldots, p$ e $r_v \geq r_{uv}$ para todo $u \neq v$.

O coeficiente de correlação amostral entre as variáveis X_u e X_v é dado por

$$r_{uv} = \frac{s_{uv}}{\sqrt{s_u s_v}}, \tag{5.7}$$

com $-1 \leq r_{uv} \leq 1$ e $r_{uv} = r_{vu}$ para todo u, v.

Em muitas situações também são de interesse as somas de quadrados de desvios, nomeadamente

$$W_{vv} = \sum_{i=1}^{n}(x_{iv} - \overline{x}_v)^2, \quad v = 1, \ldots, p \tag{5.8}$$

e as somas dos produtos de desvios, a saber,

$$W_{uv} = \sum_{i=1}^{n}(x_{iu} - \overline{x}_u)(x_{iv} - \overline{x}_v), \quad u, v = 1, \ldots, p. \tag{5.9}$$

Exemplo 5.9. Os dados dispostos na Tabela 5.4 correspondem a cinco agentes de seguros para os quais foram observados os valores das variáveis X_1 = número de anos de serviço e X_2 = número de clientes.

Tabela 5.4: Número de anos de serviço e número de clientes para cinco agentes de seguros

Agente	X_1	X_2
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72

Para os dados da Tabela 5.4, a matriz de dados é

$$\mathbf{X}^\top = \begin{bmatrix} 2 & 4 & 5 & 6 & 8 \\ 48 & 56 & 64 & 60 & 72 \end{bmatrix},$$

de modo que

$$\overline{x}_1 = \frac{1}{5}\sum_{i=1}^{5} x_{i1} = \frac{1}{5}(2 + 4 + 5 + 6 + 8) = 5,$$

$$\overline{x}_2 = \frac{1}{5}\sum_{i=1}^{5} x_{i2} = \frac{1}{5}(48 + 56 + 64 + 60 + 72) = 60$$

e o vetor de médias é

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 60 \end{bmatrix}.$$

A matriz de desvios em relação às médias é

$$\mathbf{Y} = \begin{bmatrix} 2 & 48 \\ 4 & 56 \\ 5 & 64 \\ 6 & 60 \\ 8 & 72 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [5 \quad 60] = \begin{bmatrix} 2\text{-}5 & 48\text{-}60 \\ 4\text{-}5 & 56\text{-}60 \\ 5\text{-}5 & 64\text{-}60 \\ 6\text{-}5 & 60\text{-}60 \\ 8\text{-}5 & 72\text{-}60 \end{bmatrix} = \begin{bmatrix} \text{-}3 & \text{-}12 \\ \text{-}1 & \text{-}4 \\ 0 & 4 \\ 1 & 0 \\ 3 & 12 \end{bmatrix}$$

e as correspondentes matrizes de covariâncias e correlações são, respectivamente,

$$\mathbf{S} = \frac{1}{5-1}\mathbf{Y}^\top\mathbf{Y} = \begin{bmatrix} 5 & 19 \\ 19 & 80 \end{bmatrix} \quad \text{e} \quad \mathbf{R} = \begin{bmatrix} 1 & 0,95 \\ 0,95 & 1 \end{bmatrix}.$$

As variâncias e covariâncias amostrais são respectivamente,

$$s_{11} = \frac{1}{4}\sum_{i=1}^{5}(x_{i1} - \overline{x}_1)^2 = 5, \quad s_{22} = \frac{1}{4}\sum_{i=1}^{5}(x_{i2} - \overline{x}_2)^2 = 80,$$

$$s_{12} == s_{21} = \frac{1}{4}\sum_{i=1}^{5}(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) = 19$$

ao passo que as correlações amostrais são dadas por

$$r_{11} = r_{22} = 1 \;\text{ e }\; r_{12} = r_{21} = \frac{s_{12}}{\sqrt{s_{11}s_{22}}} = \frac{19}{\sqrt{5 \times 80}} = 0,95.$$

2) **Lowess**

Muitas vezes, gráficos de dispersão (simbólicos ou não) são utilizados para a identificação de curvas que possam representar a relação entre as variáveis sob avaliação. Por exemplo, pode haver interesse em saber se uma variável é uma função linear ou quadrática da outra. O ajuste de uma curva suave aos dados pode ser realizado or meio da técnica conhecida como **lowess** (*locally weighted regression scatterplot smoothing*). Essa técnica de **suavização** é realizada por meio de sucessivos ajustes de retas por mínimos quadrados ponderados (ver Capítulo 6) a subconjuntos dos dados.

Consideremos, as coordenadas (x_j, y_j), $j = 1, \ldots, n$ de um conjunto de dados, por exemplo, correspondentes aos pontos associados aos veículos `drv`=4 na Figura 5.5. O ajuste de uma curva suave a esses pontos por meio da técnica lowess é baseado na substituição da coordenada y_j por um valor suavizado $\widehat{y}_j$ obtido segundo os seguintes passos:

i) Escolha uma faixa vertical centrada em (x_j, y_j) contendo q pontos conforme ilustrado na Figura 5.12 (em que $q = 9$). Em geral, escolhemos $q = [n \times p]$ em que $0 < p < 1$, tendo em conta que quanto maior for p, maior será o grau de suavização.

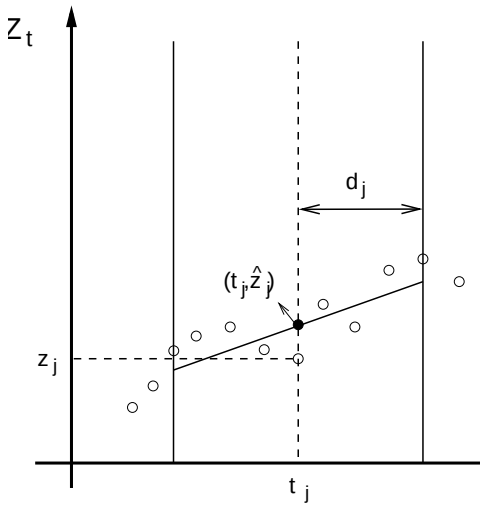


Figura 5.12: Faixa centrada em (x_j, y_j) para suavização por lowess.

ii) Use uma função simétrica em torno de x_j para atribuir pesos aos pontos na vizinhança de (x_j, y_j). Essa função é escolhida de forma que o maior peso seja atribuído a (x_j, y_j) e que os demais pesos diminuam à medida que x se afasta de x_j. Com essa finalidade, utiliza-se a **função tricúbica**

$$h(u) = \begin{cases} (1 - |u|^3)^3, & \text{se } |u| < 1 \\ 0, & \text{em caso contrário.} \end{cases}$$

O peso atribuído a (x_k, y_k) é $h[(x_j - x_k)/d_j]$ em que d_j é a distância entre x_j e seu vizinho mais afastado dentro da faixa selecionada em i) conforme ilustrado na Figura 5.13.

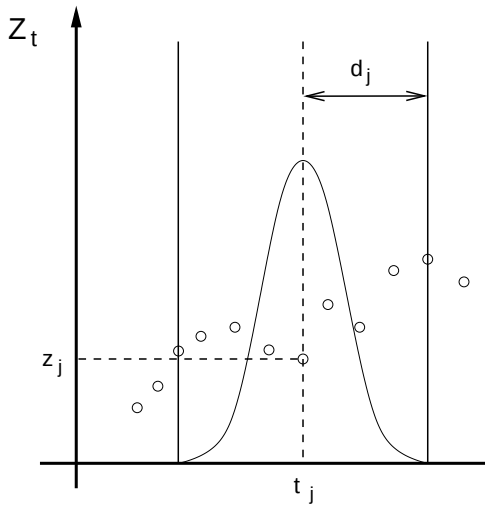


Figura 5.13: Atribuição de pesos para suavização por lowess.

iii) Ajuste uma reta $y = \alpha + \beta x + e$ aos q pontos da faixa centrada em x_j, por meio da minimização de

$$\sum_{k=1}^{q} h_j(x_k)(y_k - \alpha - \beta x_k)^2,$$

obtendo os estimadores $\widehat{\alpha}$ e $\widehat{\beta}$. O valor suavizado de y_k é $\widehat{y}_k = \widehat{\alpha} + \widehat{\beta} x_k$, $k = 1, \ldots, q$.

iv) Calcule os resíduos $\widehat{e}_k = y_k - \widehat{y}_k$, $k = 1, \ldots, q$ e por meio de um gráfico de dispersão, por exemplo, identifique possíveis pontos discrepantes (*outliers*). Quando existirem, refaça os cálculos, atribuindo pesos menores aos maiores resíduos, por meio da **função biquadrática**

$$g(u) = \begin{cases} (1 - |u|^2)^2, & \text{se } |u| < 1 \\ 0, & \text{em caso contrário.} \end{cases}$$

O peso atribuído ao ponto (x_k, y_k) é $g(x_k) = g(\widehat{e}_k/6m)$ em que m é a mediana dos valores absolutos dos resíduos ($|\widehat{e}_k|$). Se o resíduo $\widehat{e}_k$ for muito menor do que $6m$, o peso a ele atribuído será próximo de 1; em caso contrário, será próximo de zero. A razão pela qual utilizamos o denominador $6m$ é que se os resíduos tiverem uma distribuição Normal com variância σ^2, então $m \approx 2/3$ e $6m \approx 4\sigma$. Isso implica que para resíduos Normais, raramente teremos pesos pequenos.

v) Finalmente, ajuste uma nova reta aos pontos (x_k, y_k) com pesos $h(x_k)g(x_k)$. Se (x_k, y_k) corresponder a um ponto discrepante, o resíduo $\widehat{e}_k$ será grande, mas o peso atribuído a ele será pequeno.

vi) Repita o procedimento duas ou mais vezes, observando que a presença de pontos discrepantes exige um maior número de iterações.

Gráficos das funções tricúbica $[h(u)]$ e biquadrática $[g(u)]$ estão exibidos na Figura 5.14.

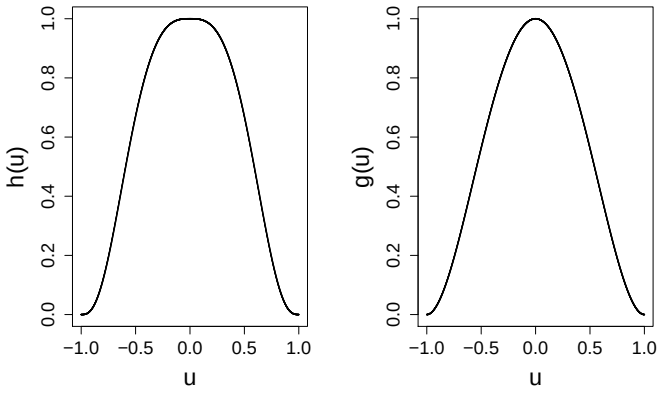


Figura 5.14: Gráficos das funções tricúbica $[h(u)]$ e biquadrática $[g(u)]$.

Para mais detalhes sobre o método lowess bem como sobre outros métodos de suavização o leitor poderá consultar Morettin e Tolói (2018), por exemplo.

Por meio dos comandos

```
> rotarod<-read.xls("/home/jmsinger/Desktop/rotarod.xls",
          sheet = 'dados', method = "tab")
> par(mar=c(5.1,5.1,4.1,2.1))
> plot(rotarod$tempo, rotarod$rotarod, type='p',
       xlab = "Tempo", ylab ="Rotarod",
       cex.axis = 1.3, cex.lab = 1.6)
> lines(lowess(rotarod$rotarod ~ rotarod$tempo, f=0.1),
         col=1, lty=2, lwd =2)
> lines(lowess(rotarod$rotarod ~ rotarod$tempo, f=0.4),
         col=2, lty=1, lwd =2)
```

podemos obter o gráfico disposto na Figura 5.15, em que apresentamos curvas lowess (com dois níveis de suavização) ajustadas aos pontos de um conjunto de dados obtidos de um estudo cujo objetivo era propor um modelo para avaliar a evolução de uma variável ao longo do tempo. O gráfico sugere um modelo de **regressão segmentada**, *i.e.* em que a resposta média assume um valor constante até um ponto de mudança, a partir do qual a uma curva quadrática pode representar a sua variação temporal.

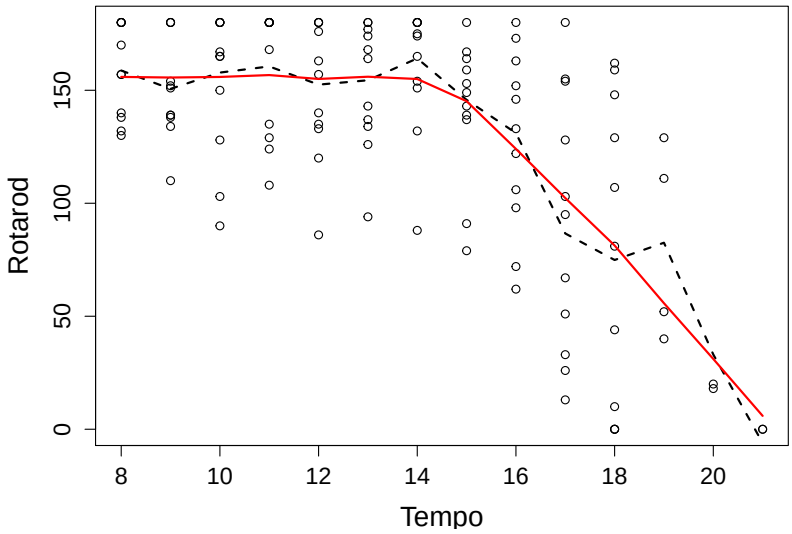


Figura 5.15: Curvas lowess com diferentes parâmetros de suavização ajustadas a um conjunto de dados.

3) **Parametrização de desvios médios**

Com a finalidade de explicitar efeitos principais e interação no modelo, é comum considerar-se a reparametrização $\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$, que implica o modelo

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}, \tag{5.10}$$

$i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, m$. Muitos autores, como Nelder et al. (1988), interpretam erroneamente os parâmetros μ, α_i, β_j, $\alpha\beta_{ij}$, respectivamente, como "média geral", "efeito principal do nível i do fator A", "efeito principal do nível j do fator B" e "interação entre os níveis i do fator A e j do fator B". Esse modelo também é **inidentificável**[1] e seus parâmetros são não estimáveis e as restrições de identificabilidade mais frequentemente utilizadas e correspondentes às parametrizações de desvios de médias e cela de referência são, respectivamente,

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \alpha\beta_{ij} = \sum_{j=1}^{b} \alpha\beta_{ij} = 0 \tag{5.11}$$

[1]Um modelo $F(\theta)$, dependendo do parâmetro $\theta \subset \Theta$, é identificável se para todo $\theta_1, \theta_2 \subset \Theta$, $\theta_1 \neq \theta_2$ temos $F(\theta_1) \neq F(\theta_2)$. Em caso contrário, o modelo é dito inidentificável. Por exemplo, consideremos o modelo $y_i \sim N(\mu + \alpha_i, \sigma^2)$, $i = 1, 2$ em que y_1 e y_2 são independentes. Tomando $\boldsymbol{\theta} = (\mu, \alpha_1, \alpha_2)^\top$ como parâmetro, o modelo é inidentificável, pois tanto para $\boldsymbol{\theta}_1 = (5, 1, 0)^\top$ quanto para $\boldsymbol{\theta}_2 = (4, 2, 1)^\top \neq \boldsymbol{\theta}_1$, a distribuição conjunta de (y_1, y_2) é $N_2[(6, 6)^\top, \sigma^2\mathbf{I}_2]$. O leitor poderá consultar Bickel e Doksum (2001), entre outros, para detalhes.

e

$$\alpha_1 = \beta_1 = \alpha\beta_{11} = \ldots = \alpha\beta_{1b} = \alpha\beta_{21} = \ldots = \alpha\beta_{a1} = 0 \qquad (5.12)$$

Sob as restrições (5.11), pode-se mostrar que

$$\mu = (ab)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}, \;\; \alpha_i = b^{-1}\sum_{j=1}^{b}\mu_{ij} - \mu, \;\; \beta_j = a^{-1}\sum_{i=1}^{a}\mu_{ij} - \mu$$

e que

$$\alpha\beta_{ij} = \mu_{ij} - b^{-1}\sum_{j=1}^{b}\mu_{ij} - a^{-1}\sum_{i=1}^{a}\mu_{ij}.$$

Sob as restrições (5.12), temos

$$\mu = \mu_{11}, \;\; \alpha_i = \mu_{ij} - \mu_{1j}, \;\; i = 2, \ldots, a, \;\; \beta_j = \mu_{ij} - \mu_{i1}, \;\; j = 2, \ldots, b,$$

e que

$$\alpha\beta_{ij} = \mu_{ij} - (\mu_{11} + \alpha_i + \beta_j), \;\; i = 2, \ldots, a, \; j = 2, \ldots, b,$$

de forma que os parâmetros α_i, $i = 2, \ldots, a$ podem ser interpretados como efeitos diferenciais entre as respostas esperadas das unidades amostrais submetidas ao nível i do fator A relativamente àquelas obtidas por unidades amostrais submetidas ao tratamento associado ao nível 1 do fator A, mantido fixo o nível correspondente ao fator B. Analogamente, os parâmetros β_j, $j = 2, \ldots, b$ podem ser interpretados como efeitos diferenciais entre as respostas esperadas das unidades amostrais submetidas ao nível j do fator B relativamente àquelas obtidas por unidades amostrais submetidas ao tratamento associado ao nível 1 do fator B, mantido fixo o nível correspondente do fator A. Os parâmetros $\alpha\beta_{ij}$, $i = 2, \ldots, a$, $j = 2, \ldots b$ podem ser interpretados como diferenças entre as respostas esperadas das unidades amostrais submetidas ao tratamento correspondente à cela (i, j) e aquela esperada sob um modelo sem interação.

4) **A estatística de Mantel-Haenszel**

A estatística de Mantel-Haenszel é utilizada para avaliar a associação em conjuntos de tabelas 2×2 obtidas de forma estratificada segundo o paradigma indicado na Tabela 5.5, com apenas dois estratos, para simplificação.

Tabela 5.5: Frequência de pacientes

Estrato	Fator de risco	Status do paciente		Total
		doente	são	
1	presente	n_{111}	n_{112}	n_{11+}
	ausente	n_{121}	n_{122}	n_{12+}
	Total	n_{1+1}	n_{1+2}	n_{1++}
2	presente	n_{211}	n_{212}	n_{21+}
	ausente	n_{221}	n_{222}	n_{22+}
	Total	n_{2+1}	n_{2+2}	n_{2++}

Uma estimativa da razão de chances para o estrato h é

$$rc_h = \frac{n_{h11}nh22}{n_{h12}n_{h21}}.$$

A estimativa da razão de chances comum proposta por Mantel e Haenszel (1959) é uma média ponderada das razões de chances de cada um dos H estratos com pesos

$$w_h = \frac{n_{h12}n_{h21}}{n_{h++}} / \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}},$$

ou seja

$$\begin{aligned} rc_{MH} &= \sum_{h=1}^{H} w_h rc_h = \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}} \times \frac{n_{h11}n_{h22}}{n_{h12}n_{h21}} / \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}} \\ &= \sum_{h=1}^{H} \frac{n_{h11}n_{h22}}{n_{h++}} / \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}} \end{aligned}$$

Consideremos, por exemplo, os dados dispostos na Tabela 5.6 provenientes de um estudo cujo objetivo é avaliar a associação entre um fator de risco e a ocorrência de uma determinada moléstia com dados obtidos em três clínicas diferentes.

Tabela 5.6: Frequências de pacientes em um estudo com três estratos

Clínica	Fator de risco	Doença		Total	Razão de chances
		sim	não		
A	presente	5	7	12	2,86
	ausente	2	8	10	
B	presente	3	9	12	2,00
	ausente	1	6	7	
C	presente	3	4	7	2,63
	ausente	2	7	9	

A estimativa da razão de chances proposta por Mantel-Haenszel é

$$rc_{MH} = \frac{(5 \times 8)/22 + (3 \times 6)/19 + (3 \times 7)/16}{(7 \times 2)/22 + (9 \times 1)/19 + (4 \times 2)/16} = 2,53.$$

Uma das vantagens da razão de chances de Mantel-Haenszel é que ela permite calcular a razão de chances comum mesmo quando há frequências nulas. Vamos admitir que uma das frequências da Tabela 5.6, fosse nula, como indicado na Tabela 5.7

Tabela 5.7: Tabela com frequência nula

	Fator	Doença			Razão
Clínica	de risco	sim	não	Total	de chances
A	presente	5	7	12	∞
	ausente	0	10	10	
B	presente	3	9	12	2,00
	ausente	1	6	7	
C	presente	3	4	7	2,63
	ausente	2	7	9	

Embora a razão de chances para o estrato A seja "infinita", a razão de chances de Mantel-Haenszel pode ser calculada,

$$rc_{MH} = \frac{(5 \times 10)/22 + (3 \times 6)/19 + (3 \times 8)/16}{(7 \times 0)/22 + (9 \times 1)/19 + (4 \times 2)/16} = 6,56.$$

Outra vantagem da estatística de Mantel-Haenszel é que ela não é afetada pelo **Paradoxo de Simpson**, que ilustramos por meio de um exemplo em que a opinião sobre um determinado projeto foi avaliada com moradores de duas regiões, obtendo-se os dados apresentados na Tabela 5.8.

Tabela 5.8: Tabela com frequências relacionadas com a prefrência por dois projetos

		Opinião			Razão
Região	Projeto	favorável	desfavorável	Total	de chances
1	A	50	950	1000	0,47
	B	1000	9000	10000	
	Total	1050	9950	10000	
2	A	5000	5000	1000	0,05
	B	95	5	100	
	Total	5095	5005	10100	

Segundo a Tabela 5.8, em ambas as regiões, há uma preferência pelo Projeto B ou seja, a chance de preferência pelo projeto B é pelo menos o dobro daquela de preferência pelo Projeto A. Se agruparmos os dados somando os resultados de ambas as regiões, obteremos as frequências dispostas na Tabela 5.9.

Tabela 5.9: Frequências agrupadas correspondentes à Tabela 5.8

	Opinião			Razão
Projeto	favorável	desfavorável	Total	de chances
A	5050	5950	11000	6,98
B	1095	9005	10100	
Total	6145	9950	21100	

A razão de chances obtida com os dados agrupados indicam que a chance de preferência pelo Projeto A é cerca de 7 vezes aquela de preferência pelo Projeto B. Essa aparente incongruência é conhecida como o Paradoxo de Simpson e pode ser explicado por uma forte associação (com $rc = 0,001$) entre a variável Região e Projeto como indicado na Tabela 5.10.

Tabela 5.10: Frequências de pacientes favoráveis a cada projeto

	Região			Razão
Projeto	1	2	Total	de chances
A	1000	10000	11000	0,001
B	10000	100	10100	
Total	11000	10100	21100	

A estatística de Mantel-Haenszel correspondente é

$$rc_{MH} = \frac{(50 \times 9000)/11000 + (5000 \times 5)/10100}{(950 \times 1000)/11000 + (5000 \times 95)/10100} = 0,33$$

preservando a associação entre as duas variáveis de interesse. Detalhes sobre o Paradoxo de Simpson podem ser encontrados em Paulino e Singer (2006).

5.7 Exercícios

1) Um laboratório de pesquisa desenvolveu uma nova droga para febre tifóide com a mistura de duas substâncias químicas (A e B). Foi realizado um ensaio clínico com o objetivo de estabelecer as dosagens

adequadas (baixa ou alta, para a substância A, e baixa, média ou alta, para a substância B) na fabricação da droga. Vinte e quatro voluntários foram aleatoriamente distribuídos em 6 grupos de 4 indivíduos e cada grupo foi submetido a um dos 6 tratamentos. A resposta observada foi o tempo para o desaparecimento dos sintomas (em dias). Os resultados obtidos estão dispostos na Tabela 5.11

Tabela 5.11: Tempo para o desaparecimento dos sintomas (dias)

Dose da	Dose da substância B		
substância A	baixa	média	alta
baixa	10,4	8,9	4,8
baixa	12,8	9,1	4,5
baixa	14,6	8,5	4,4
baixa	10,5	9,0	4,6
alta	5,8	8,9	9,1
alta	5,2	9,1	9,3
alta	5,5	8,7	8,7
alta	5,3	9,0	9,4

a) Faça uma análise descritiva dos dados com o objetivo de avaliar qual a combinação de dosagens das substâncias faz com que os sintomas desapareçam em menos tempo.

b) Especifique o modelo para a comparação dos 6 tratamentos quanto ao tempo para o desaparecimento dos sintomas. Identifique os fatores e seus níveis.

c) Construa o gráfico dos perfis médios e interprete-o. Com base nesse gráfico, você acha que existe interação entre os fatores? Justifique sua resposta.

d) Confirme suas conclusões do item c) por meio de uma ANOVA com dois fatores.

2) Um experimento foi realizado em dois laboratórios de modo independente com o objetivo de verificar o efeito de três tratamentos (A1, A2 e A3) na concentração de uma substância no sangue de animais (dados hipotéticos). As concentrações observadas nos dois laboratórios são apresentas na Tabela 5.12.

Tabela 5.12: Concentração de uma substância no sangue de animais

	Laboratório 1				Laboratório 2		
	A1	A2	A3		A1	A2	A3
	8	4	3		4	6	5
	3	8	2		5	7	4
	1	10	8		3	7	6
	4	6	7		5	8	5
Total	16	28	20	Total	16	28	20

a) O que você pode comentar sobre as médias dos três tratamentos nos dois laboratórios?

b) Sem nenhum cálculo, apenas olhando os dados, em qual dos dois laboratórios será observado o maior valor da estatística F na análise de variância?

3) Um estudo foi realizado com o objetivo de avaliar a influência da exposição ao material particulado fino (MP2,5) na capacidade vital forçada (% do predito) em indivíduos que trabalham em ambiente externo. Deseja-se verificar se o efeito da exposição depende da ocorrência de hipertensão ou diabetes. Os 101 trabalhadores na amostra foram classificados quanto à exposição e presença de diabetes ou hipertensão. As médias da capacidade vital forçada em cada combinação das categorias de diabetes ou hipertensão e exposição estão representadas na Figura 5.16.

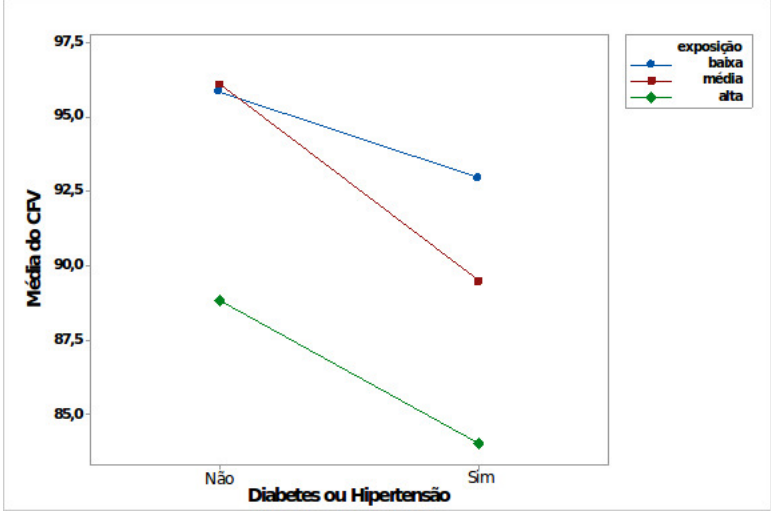


Figura 5.16: Capacidade vital forçada (% do predito).

a) Comente descritivamente os resultados obtidos, discutindo a interação entre diabetes e exposição ao material particulado.

b) Que comparações você faria para explicar a interação?

4) Considere os dados do arquivo `esforco`.

   a) Para cada etiologia, construa gráficos do desenhista (*draftman's plots*) para representar a relação entre carga na esteira e VO2 nos quatro momentos de avaliação, indicando os coeficientes de correlação de Pearson e de Spearman correspondentes.

   b) Construa gráficos de perfis médios da frequência cardíaca para as diferentes combinações dos níveis de etiologia e gravidade da doença avaliada pelo critério NYHA. Avalie descritivamente as evidências de efeitos dos fatores Etiologia e Gravidade da doença e de sua interação.

   c) Utilize ANOVA para avaliar se as conclusões descritivas podem ser extrapoladas para a populção de onde a amostra foi obtida.

5) Considere os dados do arquivo `arvores`. Obtenha os vetores de médias e matrizes de covariâncias e correlações entre as concentrações dos elementos Mn, Fe, Cu, Zn, Sr, Ba, Mg, Al, P, S, Cl e Ca para cada combinação dos níveis de espécie e tipo de via.

6) Um novo tipo de bateria está sendo desenvolvido. Sabe-se que o tipo de material da placa e a temperatura podem afetar o tempo de vida da bateria. Há três materiais possíveis a testar em três temperaturas escolhidas de forma a serem consistentes com o ambiente de uso do produto: -9 ^{o}C, 21 ^{o}C e 50 ^{o}C. Quatro baterias foram testadas em cada combinação de material e temperatura em ordem aleatória. As médias observadas do tempo de vida (h) e intervalos de confiança de 95% para as médias populacionais em cada combinação de temperatura e material estão representados no gráfico da Figura 5.17 .

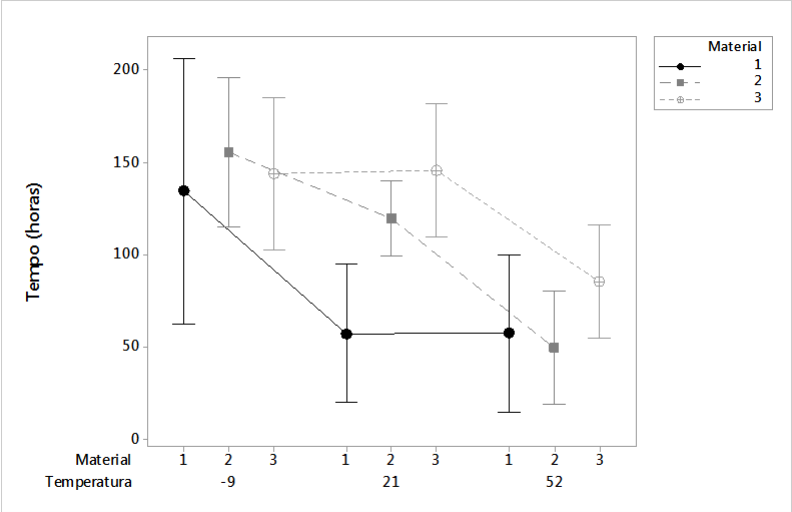


Figura 5.17: Gráfico das médias observadas do tempo de vida (h) e intervalos de confiança de 95% para as médias populacionais em cada combinação de temperatura e material.

Com base nesse gráfico pode-se conjecturar que:

a) a escolha do material com o qual é obtida a maior média do tempo de vida independe da temperatura;
b) as menores médias de tempo de vida foram observadas quando foi utilizado o material 1;
c) a temperatura em que foram observadas as maiores médias do tempo de vida é a de 21 °C;
d) há interação entre Temperatura e Tempo de vida;
e) nenhuma das alternativas acima é correta.

7) O gráfico apresentado na Figura 5.18 considera a associação entre as variáveis pressão sistólica e idade de imigrantes com menos de dez anos (Migra1) e com mais de dez anos (Migra2) desde a migração.

A dispersão dos pontos indica que:

a) existem muitos pontos aberrantes.
b) existe correlação linear positiva entre as variáveis para o grupo Migra2.
c) independentemente do tempo desde a migração as variáveis são altamente correlacionadas.
d) existe correlação linear positiva entre as variáveis para o grupo Migra1.

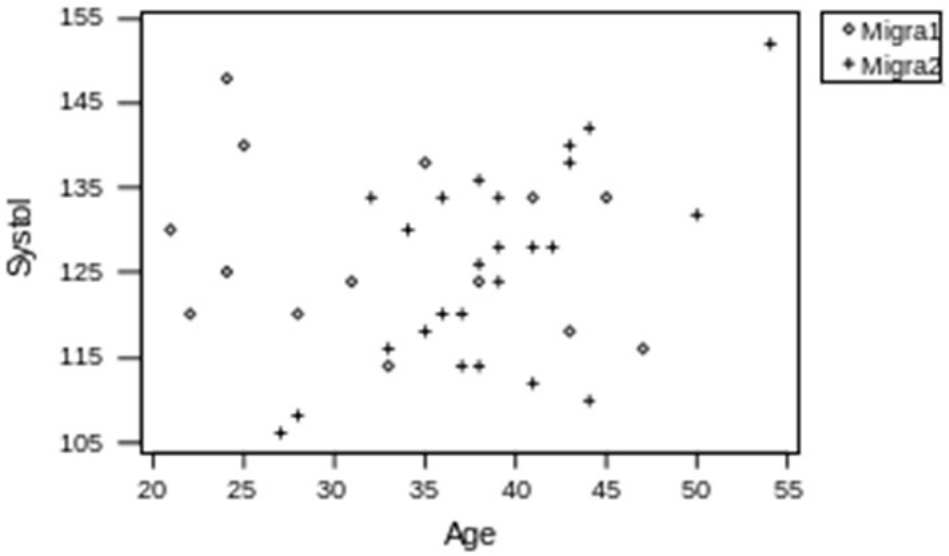


Figura 5.18: Gráfico de Pressão sistólica *versus* Idade para imigrantes.

8) Os dados do arquivo `palato` provêm de um estudo realizado no Laboratório Experimental de Poluição Atmosférica da Faculdade de Medicina da Universidade de São Paulo para avaliar os efeitos de agentes oxidantes no sistema respiratório. Espera-se que a exposição a maiores concentrações de agentes oxidantes possa causar danos crescentes às células ciliares e excretoras de muco, que constituem a principal defesa do sistema respiratório contra agentes externos. Cinquenta e seis palatos de sapos foram equitativamente e aleatoriamente alocados a um de seis grupos; cada grupo de 8 palatos foi imerso por 35 minutos numa solução de peróxido de hidrogênio numa concentração especificada, nomeadamente 0, 1, 8, 16, 32 ou 64 μM. A variável resposta de interesse é a velocidade de transporte mucociliar relativa (mm/s), definida como o quociente entre a velocidade de transporte mucociliar num determinado instante e aquela obtida antes da intervenção experimental. Essa variável foi observada a cada cinco minutos após a imersão.

   a) Obtenha os vetores de médias e matrizes de covariâncias/correlações para os dados correspondentes aos diferentes níveis do fator interunidades amostrais (concentração de peróxido de hidrogênio).

   b) Construa gráficos de perfis individuais com perfis médios e curvas lowess sobrepostas para os diferentes níveis da concentração de peróxido de hidrogênio.

   c) Compare os resultados obtidos com os diferentes níveis do fator interunidades amostrais.

9) Os dados abaixo reportam-se a uma avaliação do desempenho de um

conjunto de 203 estudantes universitários em uma disciplina introdutória de Álgebra e Cálculo. Os estudantes, agrupados segundo os quatro cursos em que estavam matriculados, foram ainda aleatoriamente divididos em dois grupos por curso, a cada um dos quais foi atribuído um de dois professores que lecionaram a mesma matéria. O desempenho de cada aluno foi avaliado por meio da mesma prova.

Frequências de aprovação/reprovação de estudantes.

		Desempenho	
Curso	Professor	Aprovado	Reprovado
Ciências Químicas	A	8	11
	B	11	13
Ciências Farmacêuticas	A	10	14
	B	13	9
Ciências Biológicas	A	19	25
	B	20	18
Bioquímica	A	14	2
	B	12	4

a) Para valiar a associação entre Professor e Desempenho, calcule a razão de chances em cada estrato.

b) Calcule a razão de chances de Mantel-Haenszel correspondente.

c) Expresse suas conclusões de forma não técnica.

10) Com base nos dados do arquivo `coronarias`, construa uma tabela de contingência $2\times2\times2\times2$ envolvendo of fatores sexo (`SEXO`), idade (`IDA55`) e hipertensão arterial (`HA`) e a variável resposta lesão obstrutiva coronariana $\geq 50\%$ (`LO3`). Obtenha as razões de chances entre cada fator e a variável resposta por meio das correspondentes distribuições marginais. Comente os resultados, indicando possíveis problemas com essa estratégia.

Capítulo 6

Análise de Regressão

> Models are, for the most part, caricatures of reality, but if they are good, like good caricatures, they portray, though perhaps in a disturbed manner, some features of the real world.
>
> Mark Kač

6.1 Introdução

Neste capítulo estaremos interessados em avaliar, de modo exploratório, um dos modelos estatísticos mais utilizados na prática, conhecido como **modelo de regressão**. O exemplo mais simples serve para a análise de dados pareados $(x_1, y_1), \ldots, (x_n, y_n)$ de duas variáveis contínuas X e Y num contexto em que sabemos a priori que a distribuição de frequências de Y pode depender de X, ou seja, na linguagem introduzida no Capítulo 4, em que X é a variável explicativa e Y é a variável resposta.

Exemplo 6.1: Para efeito de ilustração, considere os dados apresentados na Tabela 6.1 (disponíveis no arquivo `distancia`), oriundos de um estudo cujo objetivo é avaliar como a distância com que motoristas conseguem distinguir um determinado objeto (doravante indicada simplesmente como distância) varia com a idade. Aqui, a variável resposta é a distância e a variável explicativa é a idade. O gráfico de dispersão correspondente está apresentado na Figura 6.1 e mostra uma tendência decrescente da distância com a idade. O objetivo da análise de regressão é quantificar essa tendência. Como a resposta para motoristas com a mesma idade (ou com idades bem próximas) varia, o foco da análise é a estimação de uma tendência média (representada pela reta sobreposta aos dados na Figura 6.1).

No caso geral em que temos n pares de dados, o modelo de regressão utilizado para essa quantificação é

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n, \tag{6.1}$$

em que α e β são coeficientes (usualmente chamados de **parâmetros**) desconhecidos (e que se pretende estimar com base nos dados) e e_i são erros

Tabela 6.1: Distância com que motoristas conseguem distinguir certo objeto

Ident	Idade (anos)	Distância (m)	Ident	Idade (anos)	Distância (m)
1	18	170	16	55	140
2	20	197	17	63	117
3	22	187	18	65	140
4	23	170	19	66	100
5	23	153	20	67	137
6	25	163	21	68	100
7	27	187	22	70	130
8	28	170	23	71	107
9	29	153	24	72	123
10	32	137	25	73	93
11	37	140	26	74	140
12	41	153	27	75	153
13	46	150	28	77	120
14	49	127	29	79	103
15	53	153	30	82	120

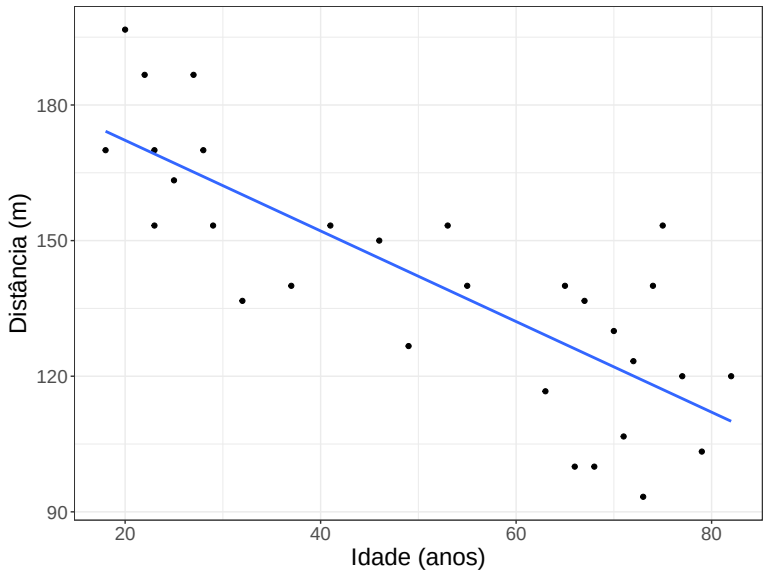


Figura 6.1: Gráfico de dispersão para os dados da Tabela 6.1.

(aleatórios) que representam desvios entre as observações y_i e a reta $\alpha + \beta x$ que corresponde à tendência esperada.[1] Em geral, supõe-se que a média (ou valor esperado) dos erros é nula, o que significa, *grosso modo*, que existe uma compensação entre erros positivos e negativos e que, consequentemente, o objetivo da análise é modelar a resposta média. Consequentemente,

$$\mathrm{E}(y_i) = \alpha + \beta x_i.$$

Nesse contexto, podemos interpretar o parâmetro α como a distância esperada com que um recém-nascido, *i.e.*, um motorista com idade $x = 0$, consegue distinguir o determinado objeto e o parâmetro β como a diminuição esperada nessa distância para cada aumento de um ano na idade. Como a interpretação de α não faz muito sentido nesse caso, um modelo mais adequado é

$$y_i = \alpha + \beta(x_i - 18) + e_i, \quad i = 1, \ldots, n. \tag{6.2}$$

Para esse modelo, parâmetro α corresponde à distância esperada com que um motorista com idade $x = 18$ anos consegue distinguir o determinado objeto e o parâmetro β tem a mesma interpretação apresentada para o modelo (6.1).

O modelo (6.1) é chamado de **regressão linear simples** e o adjetivo **linear** refere-se ao fato que os parâmetros α e β são incluídos de forma linear. Nesse sentido, o modelo

$$y_i = \alpha + \exp(\beta x_i) + e_i, \quad i = 1, \ldots, n \tag{6.3}$$

seria um **modelo não linear**. Por outro lado, o modelo

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + e_i, \quad i = 1, \ldots, n, \tag{6.4}$$

é também um modelo linear, pois embora a variável explicativa x esteja elevada ao quadrado, os parâmetros α, β e γ aparecem de forma linear. Modelos como esse, que envolvem funções polinomiais da variável explicativa, são conhecidos como **modelos de regressão polinomial** e serão analisados na Seção 6.3.

Nosso principal objetivo não é discutir em detalhes o problema da estimação dos parâmetros desses modelos, mas considerar métodos gráficos que permitam avaliar se eles são ou não adequados para descrever conjuntos de dados com a estrutura descrita. Infelizmente não poderemos prescindir de apresentar alguns detalhes técnicos. Um tratamento mais aprofundado sobre o ajuste de modelos lineares e não lineares pode ser encontrado em inúmeros textos, dentre o quais destacamos Kutner et al. (2004) para uma primeira abordagem.

Vários pacotes computacionais dispõem de códigos que permitem ajustar esses modelos. Em particular, mencionamos a função `lm()` do `R`. Na Seção

[1] Uma notação mais elucidativa para (6.1) é $y_i|x_i = \alpha + \beta x_i + e_i$, cuja leitura como "valor observado y_i da variável resposta Y para um dado valor da variável x_i da variável explicativa X" deixa claro que o interesse da análise está centrado na distribuição de Y e não naquela de X.

6.2, discutiremos, com algum pormenor, o ajuste de modelos da forma (6.1) e depois indicaremos como o caso geral de uma regressão linear múltipla (com mais de duas variáveis) pode ser abordado.

6.2 Regressão Linear Simples

Consideramos o modelo (6.1), supondo que os erros e_i são não correlacionados, tenham média 0 e variância σ^2. Nosso primeiro objetivo é estimar os parâmetros α e β. Um possível método para obtenção dos estimadores consite em determinar $\widehat{\alpha}$ e $\widehat{\beta}$ que minimizem a distância entre cada observação a reta definida por $\mathrm{E}(y_i) = \alpha + \beta x_i$. Com esse objetivo, consideremos a soma dos quadrados dos erros e_i,

$$Q(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{6.5}$$

Os **estimadores de mínimos quadrados** são obtidos minimizando-se (6.5) com relação a α e β. Com essa finalidade, derivamos $Q(\alpha, \beta)$ em relação a esses parâmetros e obtemos as **equações de estimação** igualando as expressões resultantes a zero. A solução dessas equações são os estimadores de mínimos quadrados,

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{6.6}$$

e

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \tag{6.7}$$

em que $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i$ e $\overline{y} = n^{-1}\sum_{i=1}^{n} .y_i$ Um estimador não enviesado de σ^2 é

$$S^2 = \frac{1}{n-2} Q(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{e}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2, \tag{6.8}$$

em que onde $Q(\widehat{\alpha}, \widehat{\beta})$ é a **soma dos quadrados dos resíduos**, abreviadamente, *SQRes*. Note que no denominador de (6.8) temos $n-2$, pois perdemos dois graus de liberdade em função da estimação de dois parâmetros (α e β). Alguns resultados referentes à inferência baseada nesse tipo de modelos são apresentados na Nota de Capítulo 1.

Os valores ajustados, $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}$, são utilizados para obtenção dos **resíduos**

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i),\ i = 1, \ldots, n.$$

Num contexto inferencial, ou seja, em que os dados correspondem a uma amostra de uma população (geralmente conceitual), os valores dos parâmetros α, β e σ^2 não podem ser conhecidos, a menos que toda a população seja avaliada. Consequentemente, os erros e_i não são conhecidos, mas os resíduos $\widehat{e}_i$ podem ser calculados e correspondem a estimativas desses erros.

A proposta de um modelo de regressão linear simples pode ser baseada em argumentos teóricos, como no caso em que dados são coletados para a avaliação do espaço percorrido num movimento uniforme ($s = s_0 + vt$) ou num gráfico de dispersão entre a variável resposta e a variável explicativa como aquele da Figura 6.1 em que parece razoável representar a variação esperada da distância com a idade por meio de uma reta.

Uma vez ajustado o modelo, convém avaliar a qualidade do ajuste e um dos indicadores mais utilizados para essa finalidade é o **coeficiente de determinação** definido como

$$R^2 = \frac{SQTot - SQRes}{SQTot} = \frac{SQReg}{SQTot} = 1 - \frac{SQRes}{SQTot}$$

em que a soma de quadrados total é $SQTot = \sum_{i=1}^{n}(y_i - \overline{y})^2$, a soma de quadrados dodos resíduos é $SQRes = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ e a soma de quadrados da regressão é $SQReg = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$. Para mais detalhes, ver a Nota de Capítulo 3. Em essência, esse coeficiente mede a porcentagem da variação total dos dados (em relação à sua média) explicada pelo modelo de regressão.

O coeficiente de determinação deve ser acompanhado de outras ferramentas para a avaliação do ajuste, pois não está direcionado para identificar se todas as suposições do modelo são compatíveis com os dados sob investigação. Em particular, mencionamos os gráficos de resíduos, gráficos de Cook e gráficos de influência local. Tratamos dos dois primeiros na sequência e remetemos os últimos para as Notas de Capítulo 4 e 5.

Resultados do ajuste do modelo de regressão linear simples $distancia_i = \alpha + \beta(idade_i - 18) + e_i,\ i = 1, \ldots, n$ aos dados da Tabela 6.1 por meio da função `lm()` do pacote `MASS` estão apresentados abaixo. Note que a variável preditora está especificada como `id= idade - 18`.

```
> lm(formula = distancia ~ id, data = distancia)
Residuals:
    Min      1Q  Median      3Q     Max
-26.041 -13.529   2.388  11.478  35.994
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 174.2296     5.5686  31.288  < 2e-16 ***
id           -1.0039     0.1416  -7.092 1.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.6 on 28 degrees of freedom
Multiple R-squared:  0.6424,Adjusted R-squared:  0.6296
F-statistic: 50.29 on 1 and 28 DF,  p-value: 1.026e-07
```

As estimativas dos parâmetros α (distância esperada para motoristas com 18 anos) e β (diminuição esperada da distância para cada ano adicional na idade) com erros padrões entre parênteses são, respectivamente, $\widehat{\alpha} = 174,2\ (5,6)$ e $\widehat{\beta} = -1,004\ (0,14)$.

A estimativa do desvio padrão dos erros (σ) é $S = 16,6$, com 30-2=28 graus de liberdade e o coeficiente de determinação é $R^2 = 0,63$. Se usássemos o modelo (6.1), a estimativa de α seria $192,3$ $(7,8)$ e a de β seria a mesma.

Uma das ferramentas mais úteis para a avaliação da qualidade do ajuste de modelos de regressão é o **gráfico de resíduos** em que os resíduos ($\widehat{e}_i$) são dispostos no eixo das ordenadas e os correspondentes valores da variável explicativa (x_i), no eixo das abscissas.

O gráfico de resíduos correspondente ao modelo ajustado aos dados da Tabela 6.1 está apresentado na Figura 6.2.

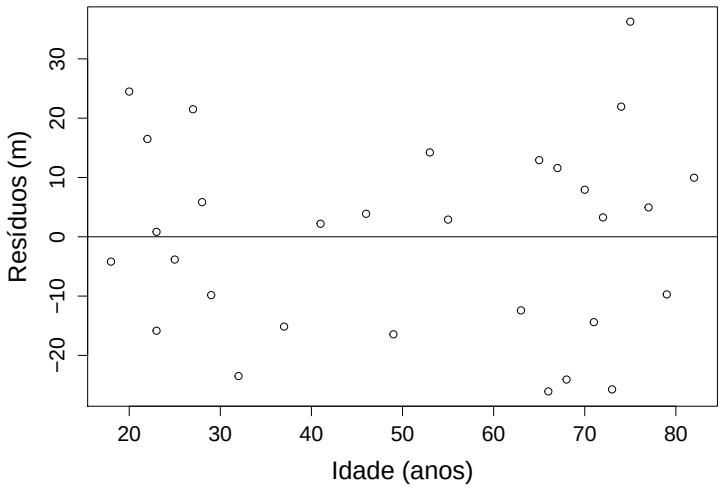


Figura 6.2: Gráfico de resíduos para o ajuste do modelo de regressão linear simples aos dados da Tabela 6.1.

Para facilitar a visualização em relação à dispersão dos resíduos e para efeito de comparação entre ajustes de modelos em que as variáveis respostas têm unidades de medida diferentes, convém padronizá-los, *i.e.*, dividi-los pelo respectivo desvio padrão para que tenham variância igual a 1. Como os resíduos (ao contrário dos erros) são correlacionados, pode-se mostrar que

$$DP(\widehat{e}_i) = \sigma\sqrt{1-h_{ii}} \;\text{ com }\; h_{ii} = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2},$$

de forma que os **resíduos padronizados**, também chamados de **resíduos studentizados** são definidos por

$$\widehat{e}_i^{\,*} = \widehat{e}_i/(S\sqrt{1-h_{ii}}). \tag{6.9}$$

Os resíduos padronizados são adimensionais e têm variância igual a 1, independentemente da variância da variável resposta (σ^2). Além disso, para erros com distribuição Normal, cerca de 99% dos resíduos padronizados têm valor entre -3 e +3.

O gráfico de resíduos padronizados correspondente àquele da Figura 6.2 está apresentado na Figura 6.3.

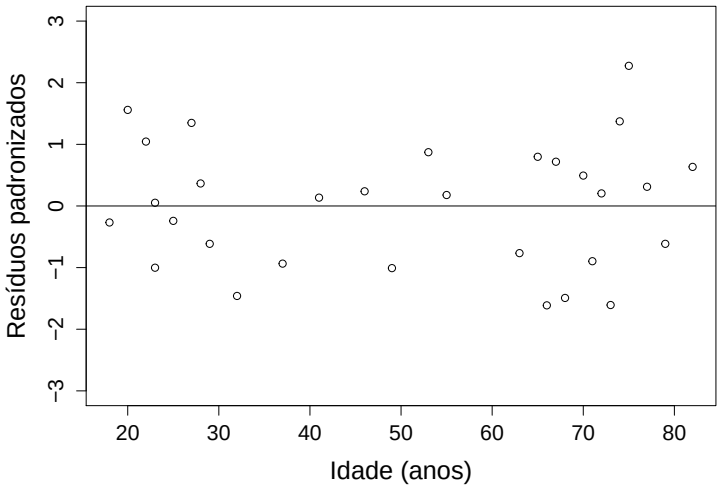


Figura 6.3: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados da Tabela 6.1.

Na Figura 6.3, nota-se que resíduos positivos e negativos estão distribuídos sem algum padrão sistemático e que sua variabilidade é razoavelmente uniforme ao longo dos diferentes valores da variável explicativa, sugerindo que relativamente à suposição de **homocedasticidade** (variância constante) o modelo adotado é (pelo menos, aproximadamente) adequado.

Exemplo 6.2: Os gráficos de dispersão e de resíduos padronizados correspondentes ao ajuste do modelo $CO_i = \alpha + \beta\ tempo_i + e_i, \quad i = 1, \ldots, n$ em que CO representa a concentração atmosférica de monóxido de carbono no dia i contado a partir de 1 de janeiro de 1991 (arquivo `poluicao`) estão apresentados nas Figuras 6.4 e 6.5. Ambos sugerem uma deficiência no ajuste: no primeiro, observa-se uma curvatura não compatível com o ajuste de uma reta; no segundo, nota-se um padrão na distribuição dos resíduos, que são positivos nos primeiros dias, negativos em seguida e espalhados ao final das observações diárias. Além disso, a dispersão dos resíduos varia com o tempo.

A saída da aplicação da função `lm()` está abaixo. Na fórmula, t indica o dia, $t = 1, \ldots, 120$.

```
> lm(formula = co ~ t, data = poluicao)
Residuals:
    Min      1Q  Median      3Q     Max
-3.7655 -0.9157 -0.1788  0.6613  4.5104
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.264608   0.254847  24.582  < 2e-16 ***
t           0.019827   0.003656   5.424 3.15e-07 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.387 on 118 degrees of freedom
Multiple R-squared:  0.1996,Adjusted R-squared:  0.1928
F-statistic: 29.42 on 1 and 118 DF,  p-value: 3.148e-07
```

O coeficiente de determinação correspondente é $0,19$, sugerindo que o modelo de regressão explica apenas uma pequena parcela da variabilidade dos dados. Um modelo (linear) de regressão polinomial alternativo em que termos quadrático e cúbico são incluídos, *i.e.*, $CO_i = \alpha + \beta\ tempo_i + \gamma\ tempo_i^2 + \delta\ tempo_i^3 + e_i, \quad i = 1, \ldots, n$ tem um melhor ajuste, como se pode notar tanto pelo acréscimo no coeficiente de determinação cujo valor é 0,35 para o modelo alternativo quanto pelo gráfico de resíduos padronizados disposto na Figura 6.6. Detalhes sobre o ajuste de modelos de regressão polinomial como esse, serão apresentados na Seção 6.3.

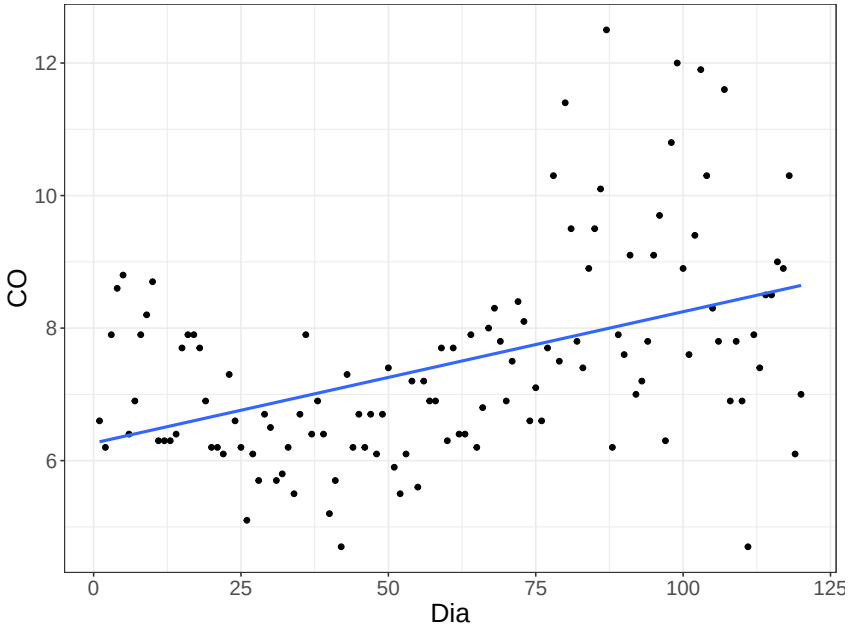


Figura 6.4: Gráfico de dispersão para os dados de monóxido de carbono.

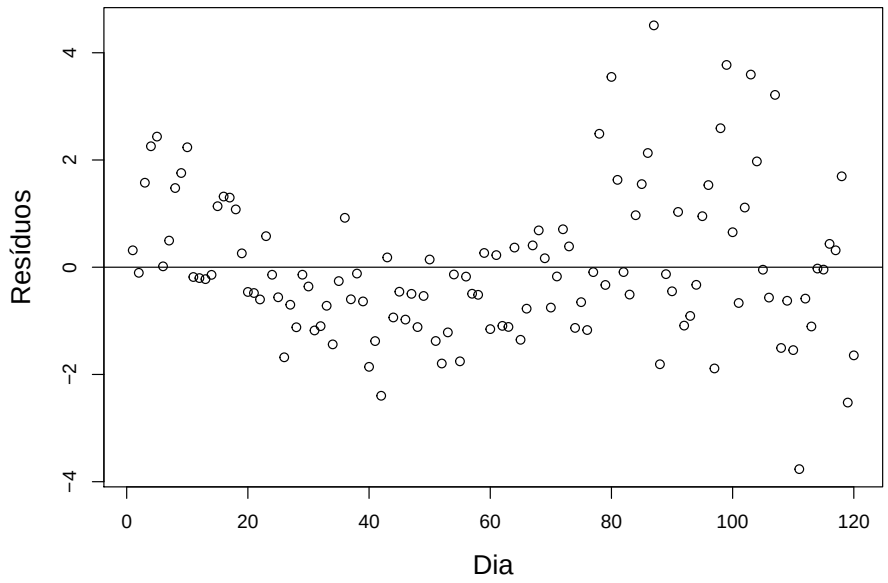


Figura 6.5: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados da concentração de CO.

Ainda assim, esse modelo polinomial não é o mais adequado em virtude da presença de **heterocedasticidade**, ou seja, de variâncias que não são constantes ao longo do tempo. Há modelos que incorporam heterogeneidade de variâncias, mas estão fora do objetivo deste texto. Para detalhes, pode-se consultar Kutner et al. (2004), por exemplo.

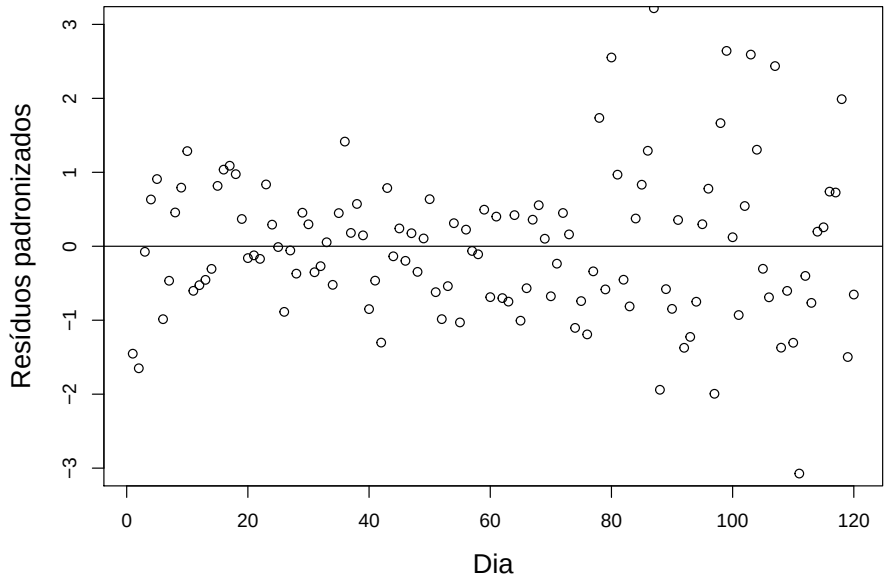


Figura 6.6: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear polinomial aos dados da concentração de CO.

Exemplo 6.3: Os dados da Tabela 6.2 são provenientes da mensuração da velocidade do vento no aeroporto de Philadelphia (EUA), sempre à uma hora da manhã, para os primeiros 15 dias de dezembro de 1974 (Graedel e Kleiner, 1985). Esses dados estão disponíveis no arquivo `vento`.

Tabela 6.2: Velocidade do vento no aeroporto de Philadelphia (v_t), com os resíduos obtidos do ajuste de modelos resistentes

t	v_t	$\widehat{r}_t^{(0)}$	$\widehat{r}_t^{(1)}$
1	22,2	1,56	-1,11
2	61,1	41,38	38,90
3	13,0	-5,80	-8,10
4	27,8	9,92	7,81
5	22,2	5,24	3,31
6	7,4	-8,64	-10,39
7	7,4	-7,72	-9,28
8	7,4	-6,80	-8,18
9	20,4	7,12	5,93
10	20,4	8,04	7,03
11	20,4	8,96	8,13
12	11,1	0,58	-0,06
13	13,0	3,40	2,94
14	7,4	-1,28	-1,55
15	14,8	7,04	6,95

O diagrama de dispersão dos dados no qual está indicada a reta obtida pelo ajuste de um modelo linear simples, nomeadamente,

$$\widehat{v}_t = 30,034 - 1,454t, \quad t = 1, \ldots, 15$$

e o correspondente gráfico de resíduos padronizados estão apresentados nas Figuras 6.7 e 6.8. Nesses gráficos pode-se notar que tanto a observação associada ao segundo dia ($t = 2$, com $v_t = 61,1$) quanto o resíduo correspondente destoam dos demais, gerando estimativas dos coeficientes da reta diferentes daqueles que se espera. Essa é uma **observação discrepante** (*outlier*). Na Nota de Capítulo 8, apresentamos um modelo alternativo com a finalidade de obter estimativas "resistentes" (também chamadas de "robustas") a pontos desse tipo.

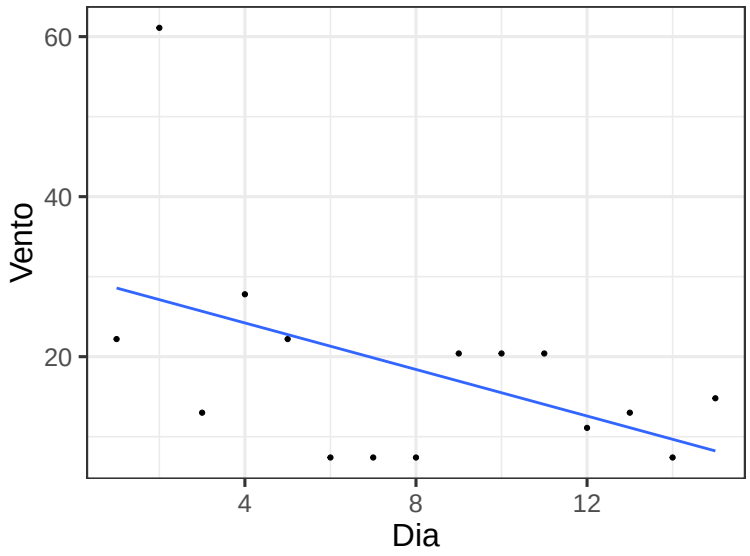


Figura 6.7: Gráfico de dispersão para os dados da Tabela 6.2 com reta de mínimos quadrados sobreposta.

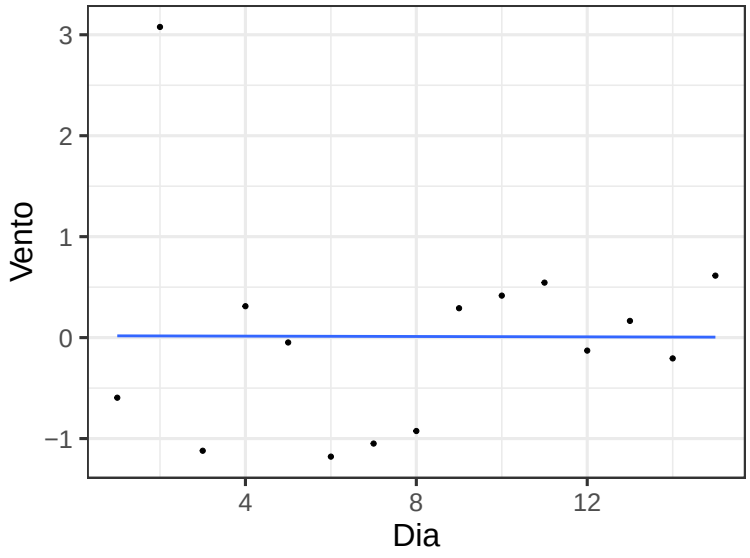


Figura 6.8: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados da Tabela 6.2.

Exemplo 6.4: Consideremos agora os dados (hipotéticos) dispostos na Tabela 6.3 aos quais ajustamos um modelo de regressão linear simples.

O gráfico de dispersão (com os dados representados por círculos e com a reta de regressão representada pela linha sólida) e o correspondente gráfico de resíduos padronizados estão apresentados nas Figuras 6.9 e 6.10.

Tabela 6.3: Dados hipotéticos

X	10	8	13	9	11	14	6	4	12	7	5	18
Y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68	6,31

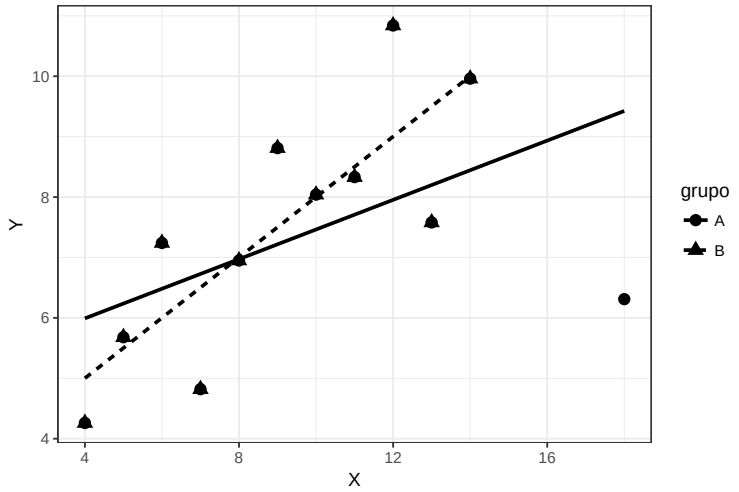


Figura 6.9: Gráfico de dispersão (com retas de regressão sobrepostas) para os dados da Tabela 6.3; curva sólida para dados completos e curva interrompida para dados com ponto influente eliminado.

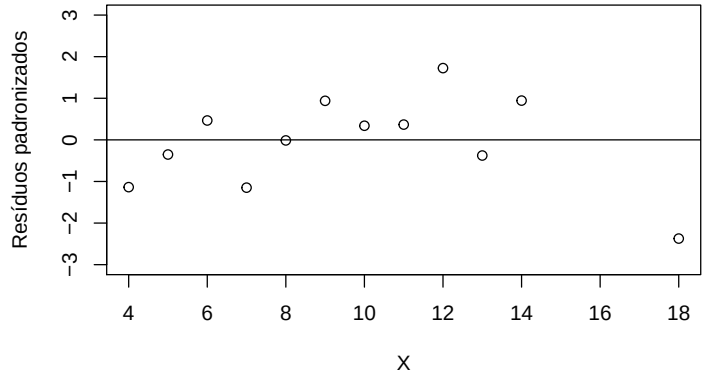


Figura 6.10: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados da Tabela 6.3.

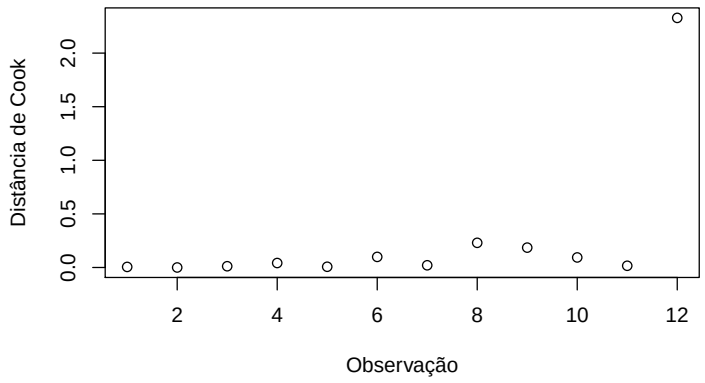


Figura 6.11: Gráfico de Cook correspondente ao ajuste do modelo de regressão linear aos dados da Tabela 6.3.

Os dois gráficos contêm indicações de que o ponto associado aos valores $(X = 18, Y = 6.31)$ pode ser um ponto discrepante. Isso fica mais evidente quando consideramos outra ferramenta diagnóstica conhecida como **gráfico de Cook** apresentado na Figura 6.11. Esse gráfico é baseado na chamada **distância de Cook** (ver Nota de Capítulo 4) que serve para indicar as observações que têm grande influência em alguma característica do ajuste do modelo. Em particular, salienta os pontos [chamados de **pontos influentes** ou **pontos alavanca** (*high leverage points*)] que podem alterar de forma relevante as estimativas dos parâmetros. Em geral, como no caso estudado aqui, esses pontos apresentam valores das respetivas abscissas afastadas daquelas dos demais pontos do conjunto de dados. Neste exemplo, a eliminação do ponto mencionado altera as estimativas do intercepto [de 5,01 (1,37) para 3,00 (1,12)] e da inclinação [de 0,25 (0,13) para 0,50 (0,12)] da reta ajustada. A reta correspondente ao ajuste quando o ponto $(X = 18, Y = 6.31)$ é eliminado do conjunto de dados está representada na Figura 6.9 pela curva interrompida.

Nos casos em que se supõe que os erros têm distribuição Normal, pode-se utilizar gráficos QQ com o objetivo de avaliar se os dados são compatíveis com essa suposição. É importante lembrar que esses gráficos QQ devem ser construídos com os quantis amostrais baseados nos resíduos e não com as observações da variável resposta, pois apesar de suas distribuições também serem normais, suas médias variam com os valores associados da variável explicativa, ou seja, a média da variável resposta correspondente a y_i é $\alpha+\beta x_i$. Convém observar que sob normalidade dos erros, os resíduos padronizados seguem uma distribuição t com $n-2$ graus de liberdade e é dessa distribuição que se devem obter os quantis teóricos para a construção do gráfico QQ. Também deve-se lembrar que para valores de n maiores que 20 ou 30, os quantis da distribuição t se aproximam daqueles da distribuição Normal, tornando-as intercambiáveis para a construção do correspondente gráfico

QQ.

Gráficos QQ (com bandas de confiança) correspondentes aos ajustes de modelos de regressão linear simples aos dados das Tabelas 6.1 e 6.3 (com e sem a eliminação da observação influente) estão respectivamente apresentados nas Figuras 6.12, 6.13 e 6.14.

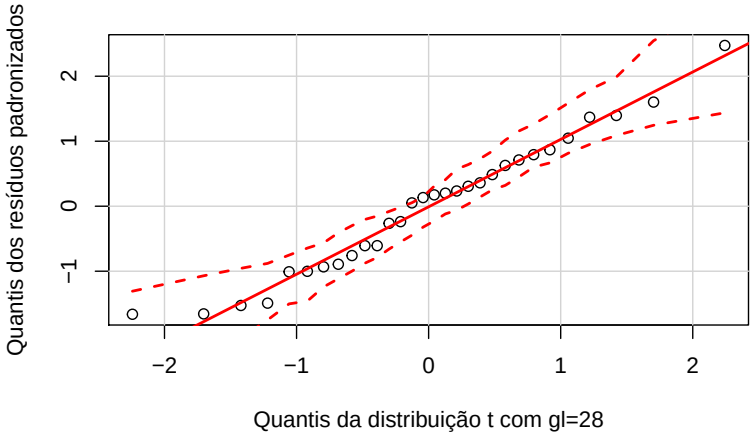


Figura 6.12: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 6.1.

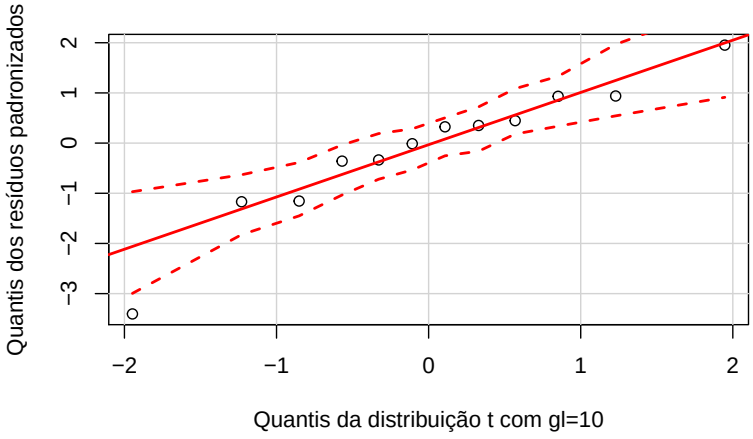


Figura 6.13: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 6.3 (com todas as observações).

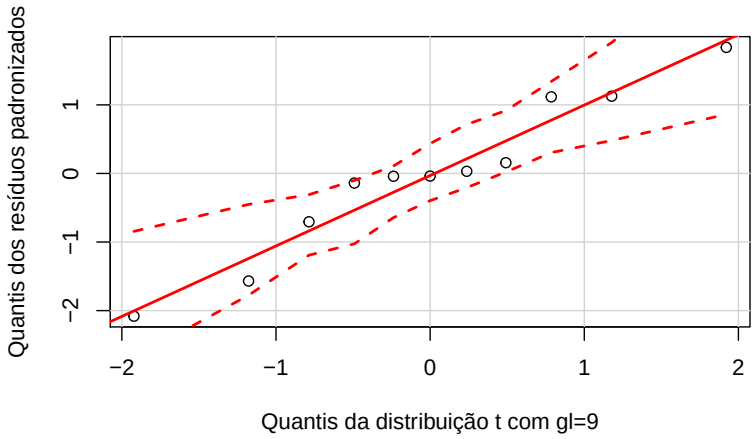


Figura 6.14: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 6.3 (sem a observação influente).

Nos três casos, não há evidências fortes contra a suposição de normalidade dos erros (apesar do ponto fora da banda de confiança salientado na Figura 6.13). Em geral, especialmente com poucos dados, é difícil observar casos em que essa suposição não parece razoável.

Convém lembrar que se o objetivo for avaliar a inclinação da reta de regressão, ou seja, avaliar a taxa com que a resposta esperada muda por unidade de variação da variável explicativa, essa suposição tem efeito marginal no estimador de mínimos quadrados desse parâmetro., Pode-se mostrar que esse estimador segue uma distribuição aproximadamente Normal quando o tamanho da amostra é suficientemente grande, por exemplo, 30 ou mais, mesmo quando a suposição de normalidade para a variável resposta não for verdadeira. Mais detalhes e uma referência estão apresentados na Nota de Capítulo 1.

Em geral, a suposição de que os erros do modelo linear são não correlacionados deve ser questionada com base no procedimento de coleta de dados. Como ilustração, consideramos dois exemplos nos quais essa característica justifica a dúvida. O primeiro exemplo é um caso simples dos problemas abordados pelas técnicas de análise de séries cronológicas; o segundo exemplo é o caso típico de análise de **dados longitudinais** e será apresentado na Seção 6.4. Ambos são apresentados aqui com a finalidade de mostrar como as técnicas de análise de regressão podem ser empregadas para analisar modelos mais gerais do que aqueles governados pelo paradigma de Gauss-Markov (ver Nota de Capítulo 1).

Exemplo 6.5: Na Tabela 6.4 apresentamos valores do peso de um bezerro observado a cada duas semanas após o nascimento com o objetivo de avaliar seu crescimento nesse período. O gráfico de dispersão correspondente está disposto na Figura 6.15.

Tabela 6.4: Peso (kg) de um bezerro nas primeiras 26 semanas após o nascimento

Semana	Peso	Semana	Peso
0	32,0	14	81,1
2	35,5	16	84,6
4	39,2	18	89,8
6	43,7	20	97,4
8	51,8	22	111,0
10	63,4	24	120,2
12	76,1	26	134,2

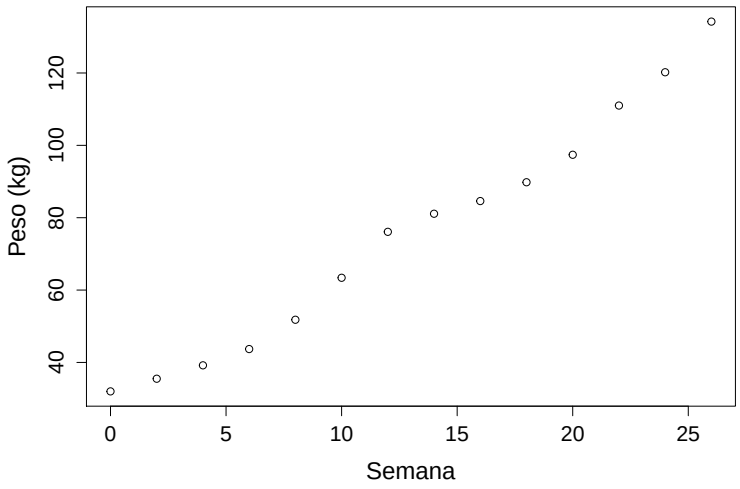


Figura 6.15: Gráfico de dispersão para os dados da Tabela 6.4

Tendo em vista o gráfico de dispersão apresentado na Figura 6.15, um possível modelo seria

$$y_t = \alpha + \beta t + \gamma t^2 + e_t, \tag{6.10}$$

$i = 1, \ldots, 14$ em que y_t representa o peso do bezerro no instante t, α denota o valor esperado de seu peso ao nascer, β e γ representam os coeficientes dos termos linear e quadrático da curva que rege a variação temporal do peso esperado no intervalo de tempo estudado e e_t denota um erro aleatório com média 0 e variância σ^2. Utilizamos t como índice para salientar que as observações são colhidas sequencialmente ao longo do tempo.

O coeficiente de determinação ajustado, $R^2_{aj} = 0,987$ indica que o ajuste (por mínimos quadrados) do modelo com $\widehat{\alpha} = 29,9$ $(2,6)$, $\widehat{\beta} = 2,7$ $(2,5)$ e $\widehat{\gamma} = 0,05$ $(0,02)$ é excelente (sob essa ótica, obviamente). Por outro lado, o gráfico de resíduos apresentado na Figura 6.16 mostra sequências de resíduos positivos seguidas de sequências de resíduos negativos, sugerindo uma possível correlação positiva entre eles (autocorrelação).

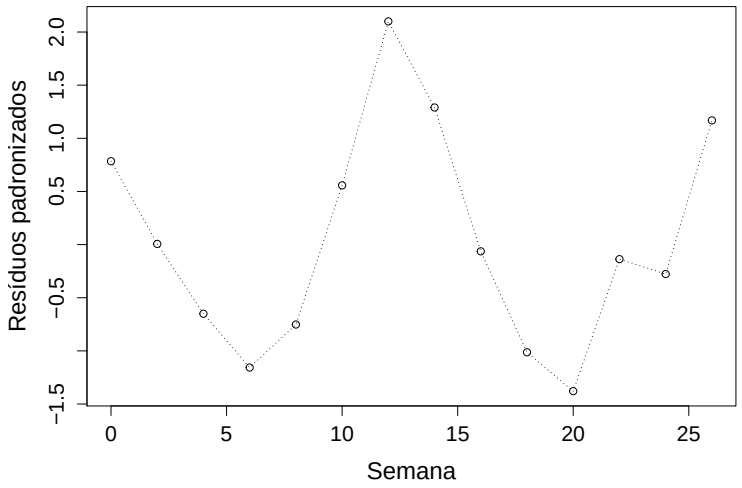


Figura 6.16: Resíduos studentizados obtidos do ajuste do modelo (6.10)

Uma maneira de contornar esse problema, é modificar os componentes aleatórios do modelo para incorporar essa possível autocorrelação nos erros. Nesse contexto, podemos considerar o modelo (6.10) com

$$e_t = \rho e_{t-1} + u_t, \quad t = 1, \ldots, n \tag{6.11}$$

em que $u_t \sim N(0, \sigma^2)$, $t = 1, \ldots, n$, independentes e e_0 é uma constante (geralmente igual a zero). Essas suposições implicam que $\text{Var}(e_t) = \sigma^2/(1-\rho^2)$ e que $\text{Cov}(e_t, e_{t-s}) = \rho^s[\sigma^2/(1-\rho^2)]$.

Para testar a hipótese de que os erros são não correlacionados pode-se utilizar a **estatística de Durbin-Watson**:

$$D = \sum_{t=2}^{n} (\widehat{e}_t - \widehat{e}_{t-1})^2 / \sum_{t=1}^{n} \widehat{e}_t^2, \tag{6.12}$$

em que $\widehat{e}_t$, $t = 1, \ldots, n$ são os resíduos obtidos do ajuste do modelo (6.10) por mínimos quadrados. Expandindo (6.12) obtemos

$$\begin{aligned} D &= \frac{\sum_{t=2}^{n} \widehat{e}_t^2}{\sum_{t=1}^{n} \widehat{e}_t^2} + \frac{\sum_{t=2}^{n} \widehat{e}_{t-1}^2}{\sum_{t=1}^{n} \widehat{e}_t^2} - 2\frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=1}^{n} \widehat{e}_t^2} \\ &\approx 2 - 2\frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=1}^{n} \widehat{e}_t^2}, \end{aligned} \tag{6.13}$$

Se os resíduos não forem correlacionados, então $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx 0$ e consequentemente, $D \approx 2$; se, por outro lado, os resíduos forem altamente correlacionados, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx \sum_{t=2}^{n} \widehat{e}_t^2$ e então $D \approx 0$; finalmente, se os resíduos tiverem uma grande correlação negativa, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx -\sum_{t=2}^{n} \widehat{e}_t^2$ e nesse caso, $D \approx 4$. Durbin and Watson

(1950, 1951, 1971) produziram tabelas da distribuição da estatística D que podem ser utilizados para avaliar a suposição de que os erros não são correlacionados.

O valor da estatística de Durbin-Watson para os dados do Exemplo 6.5 sob o modelo (6.10) é $D = 0,91$ ($p < 0,0001$), sugerindo um alto grau de autocorrelação dos resíduos. Uma estimativa do coeficiente de autocorrelação ρ é 0,50. Nesse caso, o modelo (6.10) - (6.11) poderá ser ajustado pelo textbfmétodo de mínimos quadrados generalizados ou por métodos de **Séries Temporais**.

Exemplo 6.6: Os dados dispostos na Tabela 6.5 são extraídos de um estudo conduzido na Faculdade de Odontologia da Universidade de São Paulo e correspondem a medidas de um índice de placa bacteriana obtidas de 26 crianças em idade pré-escolar, antes e depois do uso de uma escova de dentes experimental (Hugger) e de uma escova convencional (dados disponíveis no arquivo `placa`). O objetivo do estudo era comparar os dois tipos de escovas com respeito à eficácia na remoção da placa bacteriana. Os dados do estudo foram analisados por Singer and Andrade (1997) e são apresentados aqui apenas com intuito didático para mostrar a flexibilidade dos modelos de regressão. Analisamos somente os dados referentes à escova experimental e não incluímos a variável sexo dado que a análise completa não indicou diferenças entre meninas e meninos com relação à remoção da placa bacteriana.

Tabela 6.5: Índices de placa bacteriana antes e depois da escovação com uma escova de dentes experimental

ident	antes	depois	ident	antes	depois
1	2,18	0,43	14	1,40	0,24
2	2,05	0,08	15	0,90	0,15
3	1,05	0,18	16	0,58	0,10
4	1,95	0,78	17	2,50	0,33
5	0,28	0,03	18	2,25	0,33
6	2,63	0,23	19	1,53	0,53
7	1,50	0,20	20	1,43	0,43
8	0,45	0,00	21	3,48	0,65
9	0,70	0,05	22	1,80	0,20
10	1,30	0,30	23	1,50	0,25
11	1,25	0,33	24	2,55	0,15
12	0,18	0,00	25	1,30	0,05
13	3,30	0,90	26	2,65	0,25

Embora as duas variáveis, índices de placa bacteriana antes e depois da escovação correspondam essencialmente a variáveis respostas, é possível considerar uma **análise condicional**, tomando o índice pré escovação como variável explicativa (x_i) e o índice pós escovação como variável resposta (y_i). Nesse contexto, a pergunta que se deseja responder é "qual é o valor

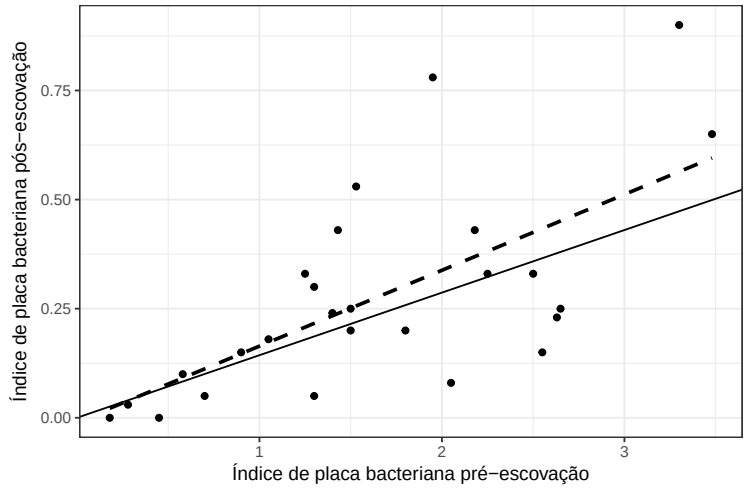


Figura 6.17: Gráfico de dispersão para os dados da Tabela 6.5; curva sólida para o modelo de regressão linear sem intercepto (6.20) e curva interrompida para o modelo (6.19).

esperado do índice pós escovação **dado** um determinado valor do índice pré escovação?".

O gráfico de dispersão dos dados da Tabela 6.5 está apresentado na Figura 6.17 (a linha tracejada corresponde ao modelo de regressão linear simples ajustado aos dados originais) em que se pode notar um aumento da dispersão do índice de placa observado pós escovação com o aumento do índice pré escovação. Isso invalida a adoção de um modelo como (6.1) cujo ajuste exige homocedasticidade (variância constante).

Singer and Andrade (1997) analisaram os dados do estudo completo por meio de um modelo não linear da forma

$$y_i = \beta x_i^{\gamma} e_i, \quad i = 1, \ldots, 26, \tag{6.14}$$

em que $\beta > 0$ e γ são parâmetros desconhecidos e e_i são erros (multiplicativos) positivos, justificando-o por meio das seguintes constatações:

i) os índices de placa bacteriana são positivos ou nulos;

ii) a relação entre X e Y deve ser modelada por uma função que passa pela origem (uma medida nula de X implica uma medida nula de Y);

iii) espera-se que a variabilidade de Y seja menor para valores menores de X, pois o índice de placa pós escovação deve ser menor ou igual ao índice pré escovação.

Note que y/x denota a taxa de redução do índice de placa e $E(y)/x$ denota a taxa esperada de redução do índice de placa. Por (6.14), temos

$$\frac{\mathrm{E}(y_i)}{x_i} = \frac{\beta x_i^{\gamma} \mathrm{E}(e_i)}{x_i} = \beta x_i^{\gamma-1} \mathrm{E}(e_i),$$

lembrando que $\mathrm{E}(e_i) > 0$. Logo, se $\gamma = 1$, essa taxa de redução esperada é constante; se $\gamma > 1$ a taxa de redução esperada aumenta e se $\gamma < 1$, ela diminui com o aumento do índice de placa x_i. Por outro lado, quanto menor for β (por exemplo, se $0 < \beta < 1$), maior será a redução do índice de placa.

Na Figura 6.18 apresentamos o histograma de X e o respectivo *boxplot*, mostrando que a distribuição do índice de placa pré escovação é moderadamente assimétrica à direita. Embora não faça sentido construir o histograma e o *boxplot* correspondente ao índice de placa bacteriana pós escovação Y (pois sob o modelo, sua média depende de X), é razoável supor que a distribuição condicional de Y dado X também seja assimétrica.

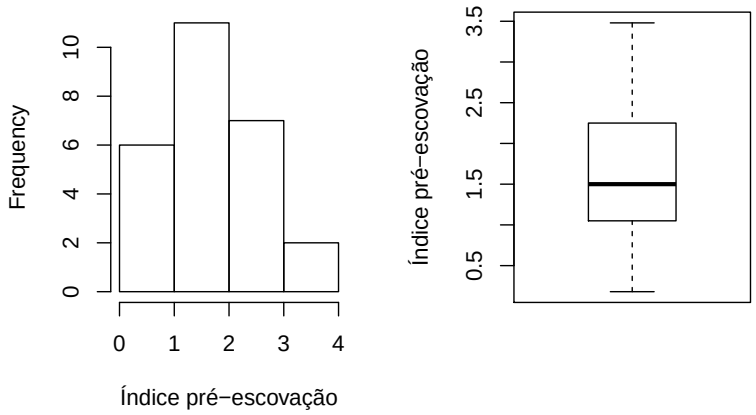


Figura 6.18: Histograma e *boxplot* para o índice de placa bacteriana pré escovação.

Esses resultados sugerem que uma transformação da forma z^θ, com $0 \leq \theta < 1$ pode ser adequada para tornar os dados mais simétricos e estabilizar a variância (consulte a Seção 3.8 para detalhes sobre transformações de variáveis). Podemos considerar, por exemplo, os casos $\theta = 0$, ou seja, a transformação logarítmica, $\theta = 1/3$ (raiz cúbica) ou $\theta = 1/2$ (raiz quadrada). A transformação logarítmica é mais conveniente, pois permite a **linerização** do modelo, deixando-o no formato de um modelo de regressão linear simples para o qual dispomos de técnicas padrão de ajuste. Esse modelo, no entanto, exige que eliminemos os dois pares para os quais $Y = 0$, reduzindo para 24 o número de observações. O modelo resultante obtido com a tranformação logarítmica é

$$y_i^* = \beta^* + \gamma x_i^* + e_i^*, \quad i = 1, \ldots, 24, \tag{6.15}$$

em que $y_i^* = \log(y_i)$, $x_i^* = \log(x_i)$, parâmetros γ, $\beta^* = \log \beta$ e erros $e_i^* = \log(e_i)$, que supomos ter média 0 e variância σ^2. Se, adicionalmente, supusermos que e_i^* tem distribuição Normal, os erros originais, e_i terão disdribuição Log-normal, defnida apenas para valores positivos, o que é compatível com as suposições adotas para o modelo (6.14).

Na Figura 6.19 apresentamos o diagrama de dispersão entre $\log x_i$ e $\log y_i$ sugerindo que a transformação induz uma menor dispersão dos dados, embora ainda haja um maior acúmulo de pontos para valores "grandes" de X.

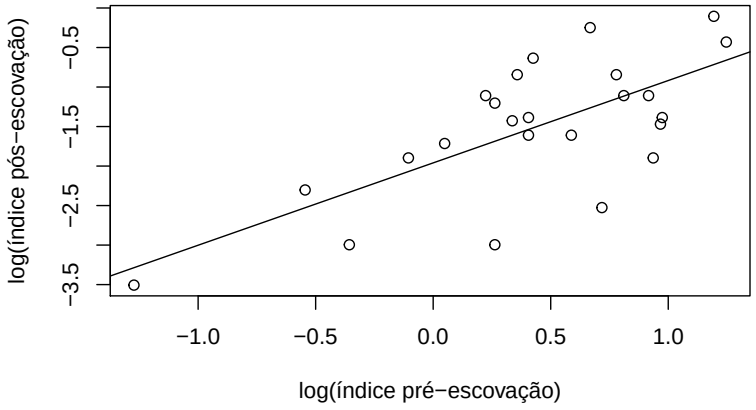


Figura 6.19: Gráfico de dispersão para os dados da Tabela 6.5 sob transformação logarítimica.

Usando o método de mínimos quadrados, a reta ajustada é

$$\widehat{y}_i^* = -1,960 + 1,042x_i^*. \tag{6.16}$$

que corresponde a

$$\widehat{y}_i = 0,141x_i^{1,042}, \tag{6.17}$$

na concepção original, já que $\widehat{\beta} = \exp(\widehat{\beta}^*) = \exp(-1,960) = 1,042$. Note que $\widehat{\beta} < 1$ e $\widehat{\gamma}$ tem valor muito próximo de 1. Podemos testar a hipótese $\gamma = 1$, para avaliar se esse resultado traz evidência suficiente para concluir que a taxa de redução do índice de placa bacteriana na população para a qual se deseja fazer inferência é constante. Para testar $H_0 : \gamma = 1$ contra a alternativa $H_A : \gamma > 1$ usamos a estatística

$$t = \frac{\widehat{\gamma} - 1}{S/\sqrt{\sum(x_i^* - \overline{x}^*)^2}}$$

cuja distribuição sob H_0 é t com $n-2$ graus de liberdade (veja a Seção 5.2 e Bussab e Morettin, 2017). O valor-p correspondente ao valor observado da estatística t é

$$p = P(\widehat{\gamma} > 1,042) = P[t_{22} > \frac{1,042 - 1}{S}\sqrt{\sum(x_i^* - \overline{x^*})^2}].$$

Como $\sum_i x_i^* = 10,246$, $\overline{x^*} = 0,427$, $\sum_i y_i^* = -36,361$, $\overline{y^*} = -1,515$, $\sum_i(x_i^* - \overline{x^*})^2 = \sum_i(x_i^*)^2 - (24)(\overline{x^*})^2 = 12,149 - (24)(0,182) = 7,773$, obtemos $S^2 = 7,773/22 = 0,353$ e $S = 0,594$. Então

$$p = P(t_{22} > (0,42)(2,788)/(0,594)) = P(t_{22} > 1,971) \approx 0,06$$

indicando que não há evidências fortes para rejeitar H_0. Como consequência, podemos dizer que a taxa esperada de redução da placa é

$$\mathrm{E}(y_i)/x_i = \beta \mathrm{E}(e_i),$$

que é constante. Como concluímos que $\gamma = 1$, o modelo linear nos logaritmos das variáveis pode ser reescrito como

$$y_i^* = \beta^* + x_i^* + e_i^*, \tag{6.18}$$

e para estimar β^* basta considerar a soma de quadrados dos erros

$$Q(\beta^*) = \sum_i (y_i^* - \beta^* - x_i^*)^2,$$

derivá-la em relação a β^* e obter o estimador de mínimos quadrados de β^* como

$$\widehat{\beta}^* = \overline{y}^* - \overline{x}^*.$$

No nosso caso, $\widehat{\beta}^* = -1,515 - 0,427 = -1,942$, e o modelo ajustado é

$$\widehat{y_i^*} = -1,942 + x_i^*. \tag{6.19}$$

O modelo original ajustado aos dados corresponde à reta

$$\widehat{y}_i = 0,1434 x_i, \quad i = 1, \ldots, 24, \tag{6.20}$$

que passa pela origem e está representada por meio de uma linha sólida na Figura 6.17.

Os resíduos dos modelos (6.14) e (6.15) estão representados nas Figuras 6.20 e 6.21, respectivamente. Esses gráficos sugerem que os resíduos dos dois modelos são aproximadamente simétricos, mas não são totalmente compatíveis com a distribuição adotada, pois a variabilidade desses resíduos ainda é grande. Para uma análise do conjunto de dados do qual este exemplo foi extraído e em que a suposição de hetrocedasticidade é levada em conta, o leitor deve consultar Singer and Andrade (1997).

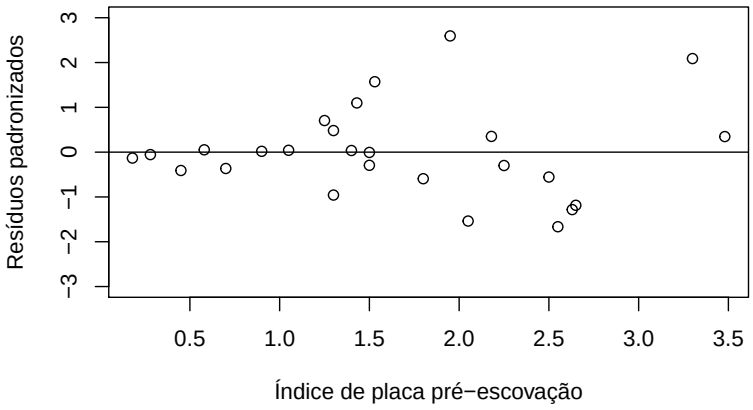


Figura 6.20: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados da Tabela 6.5.

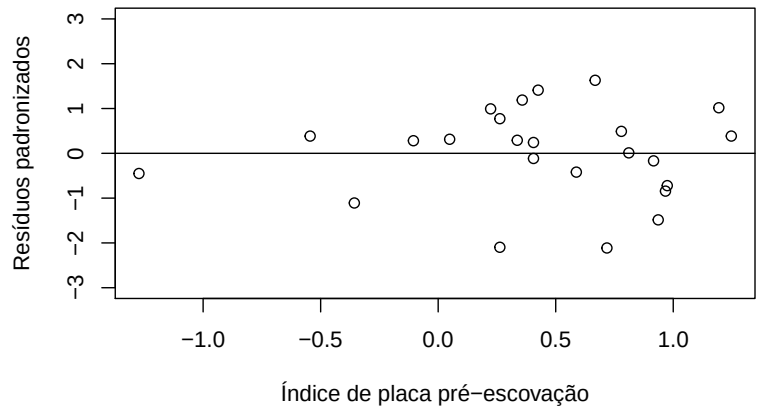


Figura 6.21: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados logaritmizados da Tabela 6.5.

6.3 Regressão Linear Múltipla

Com p variáveis explicativas $X_1, \ldots, X_p$ e uma variável resposta Y, o **modelo de regressão linear múltipla** é expresso como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i, \quad i = 1, \ldots, n. \tag{6.21}$$

O coeficiente β_0 é o chamado **intercepto** e a variável explicativa associada a ele, x_{i0}, tem valor constante igual a 1. para completar a especificação do modelo, supõe-se que os erros e_i são não correlacionados, tenham média zero e variância comum (desconhecida) σ^2.

Se quisermos testar hipóteses a respeito dos coeficientes do modelo ou construir intervalos de confiança para eles por meio de estatísticas com distribuições exatas, a suposição de que a distribuição de frequências dos erros é Normal deve ser adicionada. O modelo (6.21) tem $p + 2$ parâmetros desconhecidos, a saber, $\beta_0, \beta_1 \ldots, \beta_p$ e σ^2, que precisam que ser estimados com base nos dados observados.

Definindo $x_{i0} = 1$, $i = 1, \ldots, n$, podemos escrever (6.21) na forma

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i, \quad i = 1, \ldots, n.$$

Minimizando a soma dos quadrados do erros e_i, *i.e.*,

$$Q(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{p} \beta_j x_{ij}]^2,$$

em relação a $\beta_0, \ldots, \beta_p$ obtemos os **estimadores de mínimos quadrados**

(EMQ) $\widehat{\beta}_j,\ j = 1, \ldots, p$, de modo que

$$\widehat{y}_i = \sum_{j=0}^{p} \widehat{\beta}_j x_{ij}, \quad i = 1, \ldots, n$$

são os **valores estimados** (sob o modelo). Os termos

$$\widehat{e}_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n \tag{6.22}$$

são os **resíduos**, cuja análise é fundamental para avaliar se modelos da forma (6.21) se ajustam bem aos dados.

Para efeitos computacionais os dados correspondentes a problemas de regressão linear múltipla devem ser dispostos como indicado na Tabela 6.6.

Tabela 6.6: Matriz de dados

Y	X_1	X_2	$\cdots$	X_p
y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Em geral, a variável correspondente ao intercepto (que é constante e igual a um) não precisa ser incluída na matriz de dados; os pacotes computacionais incluem-na naturalmente no modelo a não ser que se indique o contrário.

Para facilitar o desenvolvimento metodológico, convém expressar o modelo na forma matricial

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{6.23}$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$ é o vetor cujos elementos são os valores da variável resposta Y, $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ é a matriz cujos elementos são os valores das variáveis explicativas, com $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$ contendo os valores da variável X_j, $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_p)^\top$ contém os respectivos coeficientes e $\mathbf{e} = (e_1, \ldots, e_n)^\top$ é o vetor de **erros aleatórios**.

Exemplo 6.7 : Os dados da Tabela 6.7 (disponíveis no arquivo `esteira` são provenientes de um estudo cujo objetivo é avaliar o efeito do índice de massa corpórea (IMC) e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO2) numa determinada fase do exercício. Para associar a distribuição do consumo de oxigênio (Y) com as informações sobre carga na esteira ergométrica (X_1) e IMC (X_2), consideramos o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \tag{6.24}$$

$i = 1, \ldots, 28$ com as suposições usuais sobre os erros (média zero, variância constante σ^2 e não correlacionados). Aqui, o parâmetro β_1 representa a

Tabela 6.7: VO2, IMC e carga na esteira ergométrica para 28 indivíduos

ident	VO2 (mL/kg/min)	IMC (kg/m^2)	carga (W)	ident	VO2 (mL/kg/min)	IMC (kg/m^2)	carga (W)
1	14,1	24,32	71	15	22,0	22,45	142
2	16,3	27,68	91	16	13,2	30,86	62
3	9,9	23,93	37	17	16,2	25,79	86
4	9,5	17,50	32	18	13,4	33,56	86
5	16,8	24,46	95	19	11,3	22,79	40
6	20,4	26,41	115	20	18,7	25,65	105
7	11,8	24,04	56	21	20,1	24,24	105
8	29,0	20,95	104	22	24,6	21,36	123
9	20,3	19,03	115	23	20,5	24,48	136
10	14,3	27,12	110	24	29,4	23,67	189
11	18,0	22,71	105	25	22,9	21,60	135
12	18,7	20,33	113	26	26,3	25,80	189
13	9,5	25,34	69	27	20,3	23,92	95
14	17,5	29,93	145	28	31,0	24,24	151

variação no VO2 esperada por unidade carga para indivíduos com o mesmo IMC. O parâmetro β_2 tem interpretação semelhante com a substituição de carga na esteira por IMC e IMC por carga na esteira. Como não temos dados para indivíduos com IMC menor que 17,50 e carga menor que 32, o parâmetro β_0 deve ser interpretado como um fator de ajuste do plano que aproxima a verdadeira função que relaciona o valor esperado da variável resposta com as variáveis explicativas na região em que há dados disponíveis. Se substituíssemos X_1 por $X_1 - 32$ e X_2 por $X_2 - 17.5$, o termo β_0 corresponderia ao VO2 esperado para um indivíduo com IMC = 17,50 submetido a uma carga igual a 32 na esteira ergométrica.

O modelo (6.24) pode ser expresso na forma matricial (6.23) com

$$\mathbf{y} = \begin{bmatrix} 14,1 \\ 16,3 \\ \vdots \\ 31,0 \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & 24,32 & 71 \\ 1 & 27,68 & 91 \\ \vdots & \vdots & \vdots \\ 1 & 24,34 & 151 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \ \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{28} \end{bmatrix}.$$

Para problemas com diferentes tamanhos de amostra (n) e diferentes números de variáveis explicativas (p), basta alterar o número de elementos do vetor de respostas $\mathbf{y}$ e do vetor de coeficientes $\boldsymbol{\beta}$ e modificar a matriz com os valores das variáveis explicativas, alterando o número de linhas e colunas convenientemente. Note que o modelo de regressão linear simples também pode ser expresso em notação matricial; nesse caso, a matriz $\mathbf{X}$ terá 2 colunas e o vetor $\boldsymbol{\beta}$, dois elementos (α e β). Uma das vantagens da expressão do modelo de regressão linear múltipla em notação matricial é que o método de mínimos quadrados utilizado para estimar o vetor de parâmetros $\boldsymbol{\beta}$ no modelo (6.23) pode ser desenvolvido de maneira universal

e corresponde à minimização da forma quadrática

$$Q(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2. \tag{6.25}$$

Por meio da utilização de operações matriciais, obtém-se a seguinte expressão para os estimadores de mínimos quadrados

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{6.26}$$

Sob a suposição de que $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ e $\mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$, em que $\mathbf{I}_n$ denota a matriz identidade de dimensão n, temos

i) $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,

ii) $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

Além disso, se adicionarmos a suposição de que os erros têm distribuição Normal, pode-se mostrar que o estimador (6.26) tem uma distribuição Normal multivariada, o que permite a construção de intervalos de confiança para ou testes de hipóteses sobre os elementos (ou combinações lineares deles) de $\boldsymbol{\beta}$ por meio de estatísticas com distribuições exatas. Mesmo sem a suposição de normalidade para os erros, um recurso ao **Teorema Limite Central** (ver Nota de Capítulo 1) permite mostrar que a distribuição aproximada do estimador (6.26) é Normal, com média a $\boldsymbol{\beta}$ e matriz de covariâncias $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

Um estimador não enviesado de σ^2 é

$$\begin{aligned} s^2 &= [n - (p+1)]^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= [n - (p+1)]^{-1} \mathbf{y}^\top [\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top] \mathbf{y}. \end{aligned}$$

Com duas variáveis explicativas, o gráfico de dispersão precisa ser construído num espaço tridimensional, que ainda pode ser representado em duas dimensões; para mais que 2 variáveis explicativas, o gráfico de dispersão requer um espaço com mais do que três dimensões que não pode ser representado no plano. Por isso, uma alternativa é construir gráficos de dispersão entre a variável resposta e cada uma das variáveis explicativas.

Para os dados da Tabela 6.7, o gráfico de dispersão com três dimensões incluindo o plano correspondente ao modelo de regressão múltipla ajustado está disposto na Figura 6.22. Os gráficos de dispersão correspondentes a cada uma das duas variáveis explicativas estão dispostos na Figura 6.23 e indicam que a distribuição do VO2 varia positivamente com a carga na esteira e negativamente com o IMC.

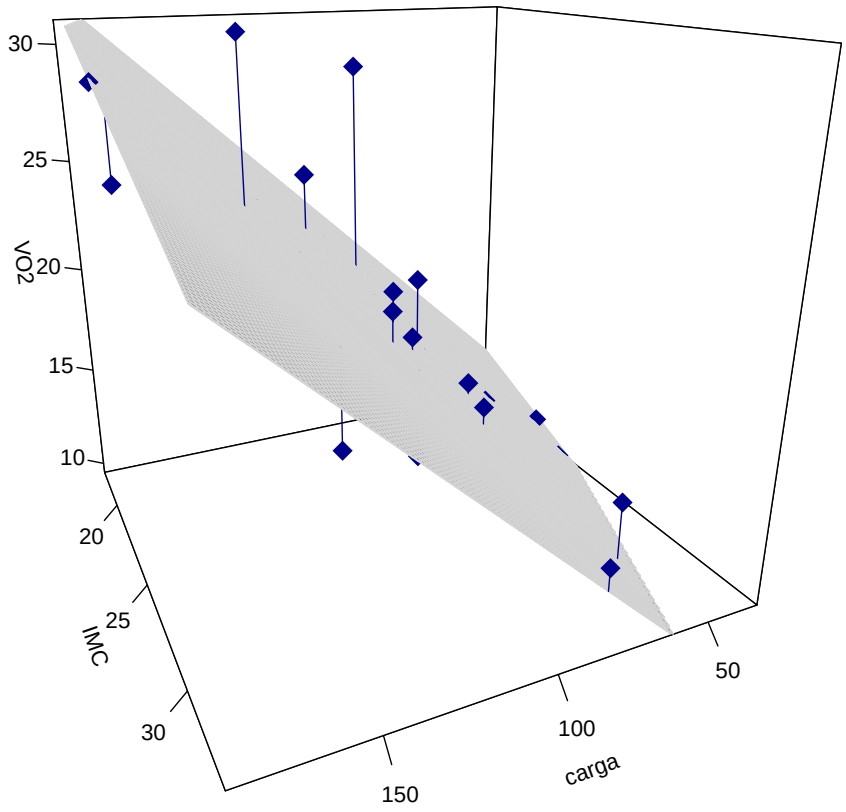


Figura 6.22: Gráficos de dispersão tridimensional para os dados da Tabela 6.7.

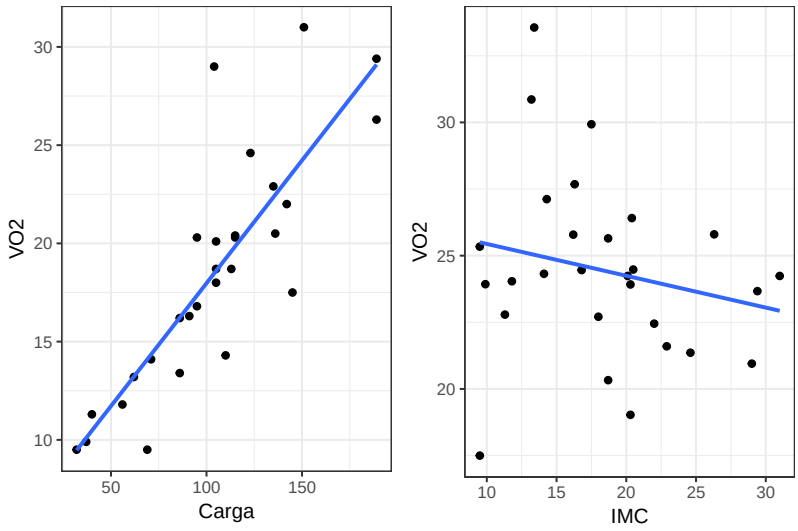


Figura 6.23: Gráficos de dispersão para os dados da Tabela 6.7.

O uso da função lm() conduz aos seguintes resultados.

```
Call:
lm(formula = VO2 ~ IMC + carga, data = esteira)
Residuals:
    Min      1Q  Median      3Q     Max
-4.1834 -2.0162 -0.2929  1.0646  9.0868
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.44726    4.45431   3.468  0.00191 **
IMC         -0.41317    0.17177  -2.405  0.02389 *
carga        0.12617    0.01465   8.614 5.95e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.057 on 25 degrees of freedom
Multiple R-squared:  0.759,Adjusted R-squared:  0.7397
F-statistic: 39.36 on 2 and 25 DF,  p-value: 1.887e-08
```

Essa saída nos diz que os coeficientes (erro padrão) correspondentes ao ajuste do modelo (6.24) aos dados da Tabela 6.7 são $\widehat{\beta}_0 = 15,45$ $(4,45)$, $\widehat{\beta}_1 = 0,13$ $(0,01)$ e $\widehat{\beta}_2 = -0,41$ $(0,17)$. Então, segundo o modelo, o valor esperado do VO2 para um indivíduo (IMC fixado) aumenta de $0,13$ unidades para cada aumento de uma unidade da carga na esteira; similarmente, o valor esperado do VO2 para indivíduos submetidos à mesma carga na esteira diminui de $0,41$ unidades com o aumento de uma unidade no IMC.

Embora o coeficiente de determinação $R^2 = 0,74$ sugira a adequação do modelo, convém avaliá-la por meio de outras ferramentas diagnósticas. No caso de regressão linear múltipla, gráficos de resíduos podem ter cada uma

das variáveis explicativas ou os valores ajustados no eixo das abscissas. Para o exemplo, esses gráficos estão dispostos na Figura 6.24 juntamente com o gráfico contendo as distâncias de Cook.

Os gráficos de resíduos padronizados não indicam um comprometimento da hipótese de homoscedasticidade embora seja possível suspeitar de dois ou três pontos discrepantes (correspondentes aos indivíduos com identificação 4, 8 e 28) que também são salientados no gráfico das distâncias de Cook. A identificação desses pontos está baseada num critério bastante utilizado (não sem controvérsias) na literatura em que resíduos associados a distâncias de Cook maiores que $4/n$ [ou $4/(n-p)$] são considerados "influentes". Em todo o caso, convém lembrar que o propósito dessas ferramentas é essencialmente exploratório e que as decisões sobre a exclusão de pontos discrepantes ou a escolha do modelo dependem de outras considerações. Esses pontos também fazem com que a suposição de normalidade possa ser posta em causa como se observa pelo painel esquerdo da Figura 6.25.

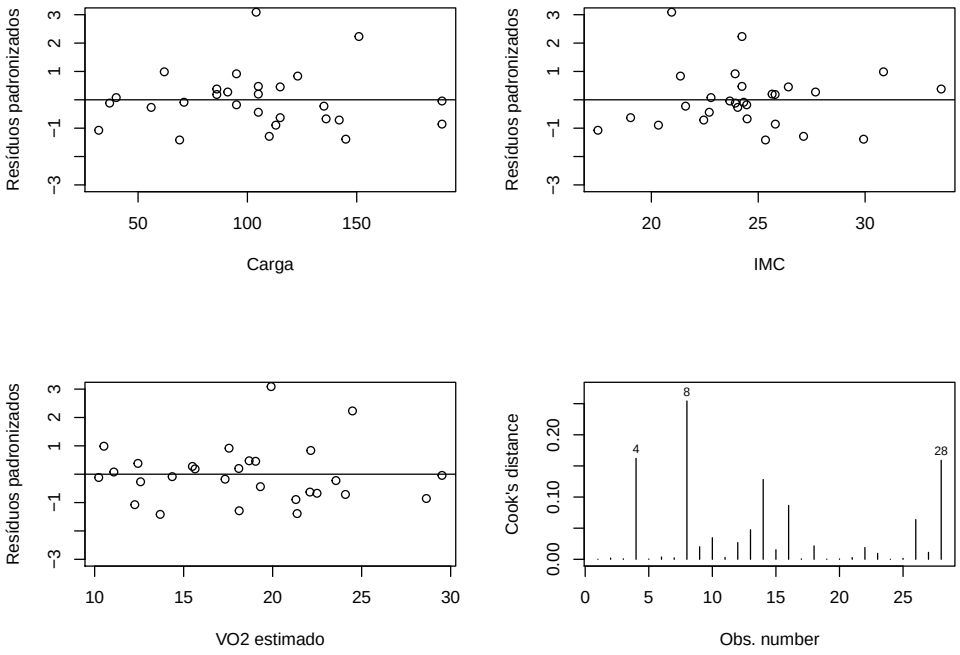


Figura 6.24: Gráficos de resíduos padronizados e distâncias de Cook para o ajuste do modelo (6.24) aos dados da Tabela 6.7.

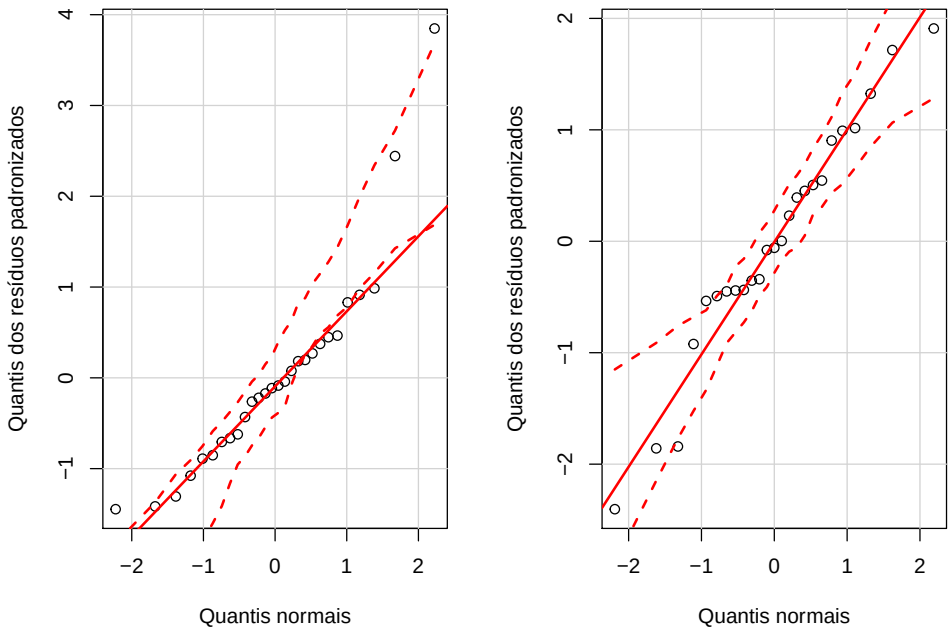


Figura 6.25: Gráficos QQ correspondentes ao ajuste do modelo (6.24) aos dados da Tabela 6.7 com (painel esquerdo) e sem (painel direito) os pontos com identificação 4, 8 e 28.

Os coeficientes do modelo ajustado ao conjunto de 25 dados obtidos com a exclusão dos pontos com identificação 4, 8 e 28 são $\widehat{\beta}_0 = 14,89$ $(3,47)$, $\widehat{\beta}_1 = 0,11$ $(0,01)$ e $\widehat{\beta}_2 = -0,36$ $(0,13)$. O coeficiente de determinação correspondente é $R^2 = 0,74$. Os gráficos de dispersão, resíduos padronizados e de distâncias de Cook correspondentes estão dispostos na Figura 6.26 e também sugerem um ajuste melhor.

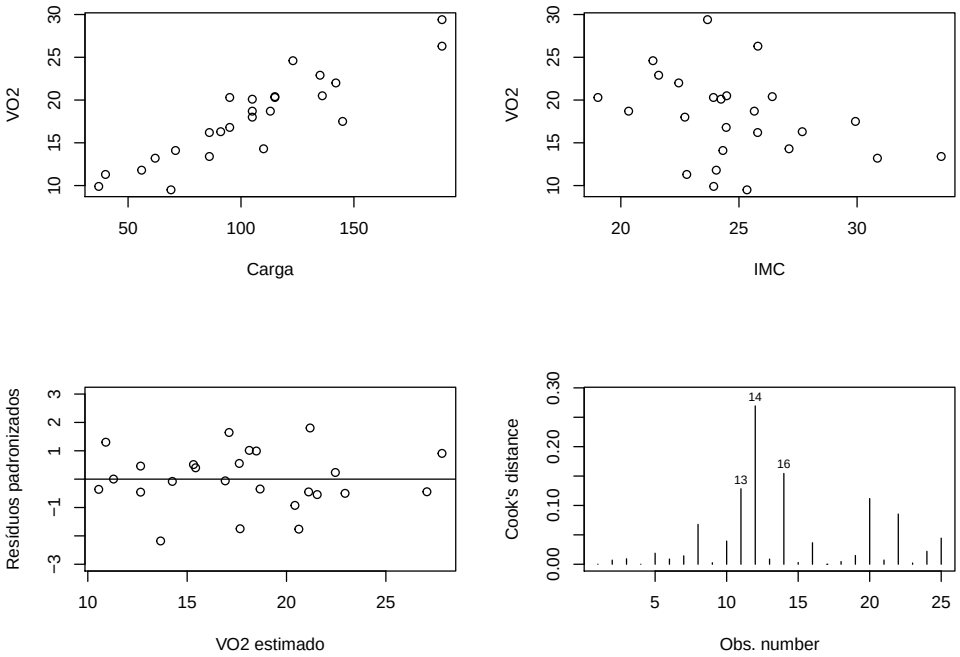


Figura 6.26: Gráficos de dispersão, resíduos padronizados e de distância de Cook correspondentes ao ajuste do modelo (6.24) aos dados da Tabela 6.7 sem os pontos com identificação 4, 8 e 28.

Exemplo 6.8: Os dados dispostos na Tabela 6.8 (disponíveis no arquivo `producao`) contêm informações sobre a produção (ton), potência instalada (1000 kW) e área construída (m^2) de 10 empresas de uma certa indústria. O objetivo é avaliar como a produção média varia em função da potência instalada e área construída. Os gráficos de dispersão entre a variável resposta (produção) e cada uma das variáveis explicativas estão dispostos na Figura 6.27 e sugerem que essas duas variáveis são linearmente associadas com a produção.

Tabela 6.8: Produção (ton), potência instalada (1000 kW) e área construída (100 m^2) de empresas de uma certa indústria

Produção	4,5	5,9	4,2	5,2	8,1	9,7	10,7	11,9	12,0	12,3
Potência	0,9	2,5	1,3	1,8	3,1	4,6	6,1	6,0	5,9	6,1
Área	7,1	10,4	7,2	8,2	8,5	11,9	12,1	12,5	12,0	11,3

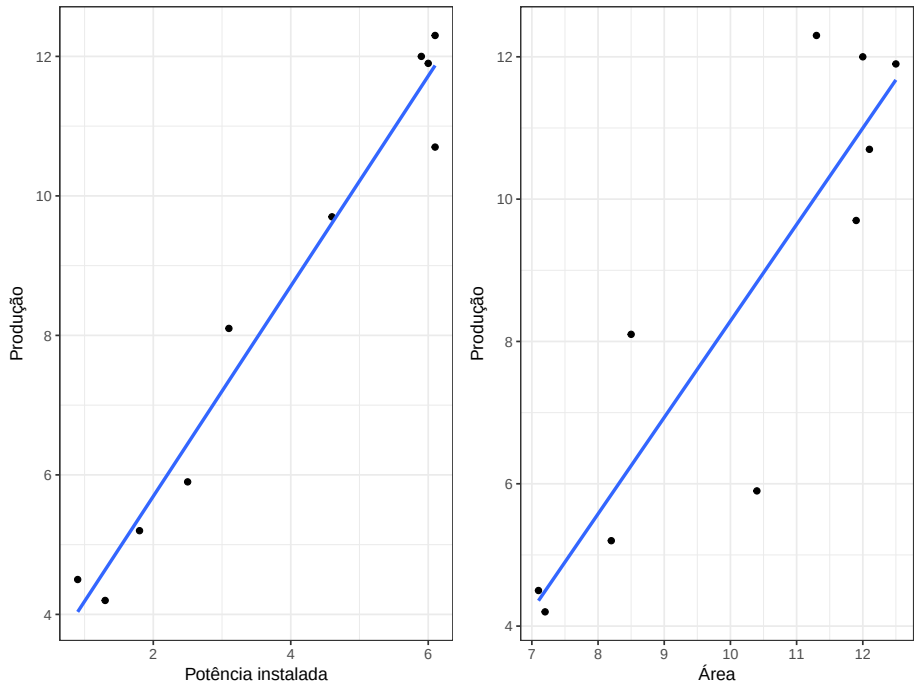


Figura 6.27: Gráficos de dispersão correspondentes aos dados da Tabela 6.8.

Estimativas dos coeficientes (com erros padrões entre parênteses) correspondentes ao intercepto, potência instalada e área construída, de um modelo de regressão linear múltipla ajustado aos dados são, respectivamente, $\widehat{\beta_0} = 4,41\ (1,74)$, $\widehat{\beta_1} = 1,75\ (0,26)$ e $\widehat{\beta_2} = -0,26\ (0,26)$. O coeficiente de determinação associado é $R^2 = 0,972$. Chama a atenção, o valor negativo do coeficiente relativo à área construída, pois o gráfico de dispersão da Figura 6.27 sugere uma associação positiva. A justificativa está no fato de as duas variáveis explicativas serem altamente correlacionadas (coeficiente de correlação de Pearson $= 0,93$) de forma que a contribuição de uma delas não acrescenta poder de explicação da produção média na presença da outra. Um teste da hipótese de que $\beta_2 = 0$ produz $p = 0,34$ sugerindo que esse coeficiente pode ser considerado nulo. Em particular, a potência instalada é suficiente para explicar a variação da produção média.

O ajuste de um modelo de regressão linear simples tendo unicamente a potência instalada como variável explicativa indica que o intercepto e o coeficiente associado à essa variável são estimados, respectivamente, por $\widehat{\beta_0} = 2,68\ (0,42)$ e $\widehat{\beta_1} = 1,50\ (0,10)$ com um coeficiente de determinação $R^2 = 0,9681$. Gráficos de resíduos padronizados correspondentes aos dois modelos estão apresentados na Figura 6.28 e corroboram a conclusão de que apenas a variável potência instalada é suficiente para a explicação da variação da produção média.

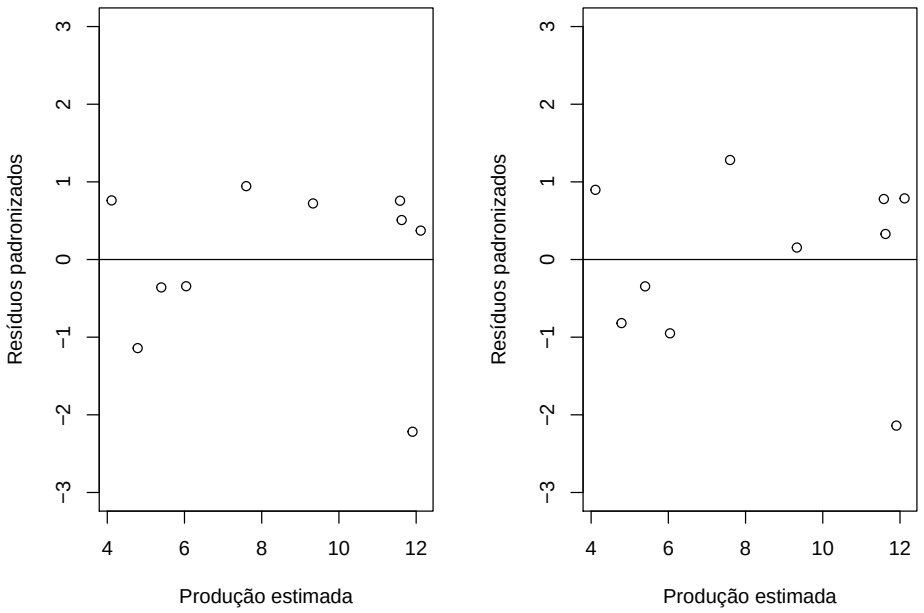


Figura 6.28: Gráficos de resíduos padronizados correspondentes aos modelos ajustados aos dados da Tabela 6.8 com duas (painel esquerdo) ou uma (painel direito) variável explicativa.

Note que o valor do coeficiente de determinação do modelo com duas variáveis explicativas $R^2 = 0,9723$ é maior do que aquele correspondente ao modelo que inclui apenas uma delas $R^2 = 0,9681$. Pela definição desse coeficiente, quanto mais variáveis forem acrescentadas ao modelo, maior será ele. Por esse motivo, convém utilizar o **coeficiente de determinação ajustado** que inclui uma penalidade pelo acréscimo de variáveis explicativas. Para o exemplo, temos $R^2_{aj} = 0,9644$ quando duas variáveis explicativas são consideradas e $R^2_{aj} = 0,9641$ quando apenas uma delas é incluída no modelo (ver Nota de Capítulo 3 para detalhes).

Uma outra ferramenta útil para avaliar a importância marginal de uma variável na presença de outras é o **gráfico da variável adicionada**. Consideremos o modelo de regressão linear múltipla

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i,\ i = 1, \ldots, n$$

com as suposições usuais. Para avaliar a importância marginal da variável X_2 na presença da variável X_1, o gráfico da variável adicionada é obtido por meio dos seguintes passos

i) Obtenha os resíduos $\widehat{e}_{1i}$ do modelo $y_i = \beta_0 + \beta_1 x_{1i} + e_{1i}$;

ii) Obtenha os resíduos $\widehat{d}_{1i}$ do modelo $x_{2i} = \gamma_0 + \gamma_1 x_{1i} + d_{1i}$;

iii) Construa o gráfico de dispersão de $\widehat{e}_{1i}$ em função de $\widehat{d}_{1i}$.

Uma tendência "relevante" nesse gráfico indica que a variável X_2 contribui para explicar a variação na média da variável resposta. Na realidade, a inclinação de uma reta ajustada aos valores de $\widehat{e}_{1i}$ em função de $\widehat{d}_{1i}$ é exatamente o coeficiente de X_2 no modelo original.

Para o exemplo da Tabela 6.8 o gráfico da variável adicionada está apresentado na Figura 6.29. O coeficiente da reta ajustada $(-0,26)$ é não significativo, sugerindo que a variável X_2 não precisa ser utilizada para explicar a variação na média da variável resposta. Compare a inclinação (negativa) da reta representada nessa figura com aquela (positiva) da reta representada no painel direito da Figura 6.27.

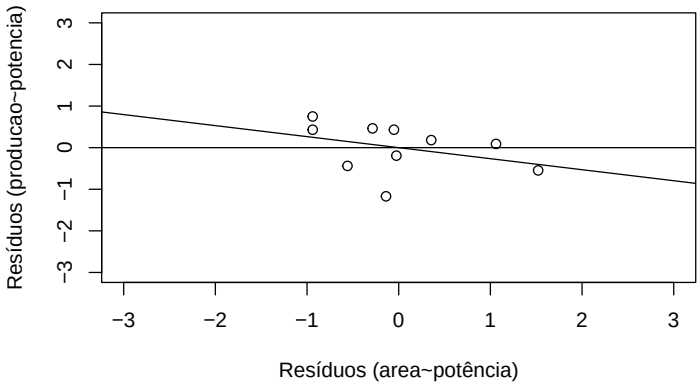


Figura 6.29: Gráfico da variável adicionada correspondente ao modelo ajustado aos dados da Tabela 6.8.

6.4 Regressão para dados longitudinais

Consideremos os dados do Exemplo 2.3, dispostos na Tabela 2.2 (disponíveis no arquivo `bezerros`) correspondentes a um estudo cujo objetivo é avaliar a variação de peso de bezerros entre a $12a$ e $26a$ semanas após o nascimento. Como cada animal é avaliado em 8 instantes (semanas 12, 14, 16, 18, 20, 22, 24 e 26), convém dispor os dados no formato da planilha 2.3 em que ficam caracterizadas tanto a variável resposta (peso) quanto a variável explicativa (tempo). O gráfico de dispersão correspondente está apresentado na Figura 6.30.

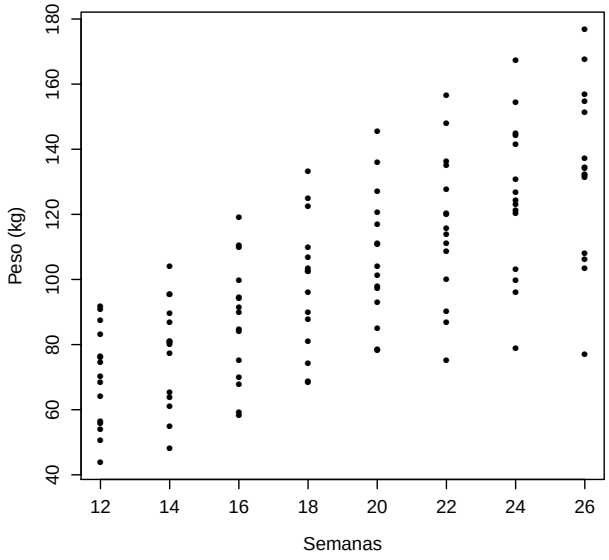


Figura 6.30: Gráfico de dispersão para os dados da Tabela 2.2.

Esse gráfico sugere que o crescimento dos animais pode ser representado pelo seguinte modelo de regressão:

$$y_{ij} = \alpha + \beta(x_j - 12) + e_{ij} \tag{6.27}$$

em que y_{ij} corresponde ao peso do i-ésimo animal no j-ésimo instante de observação, x_j corresponde ao número de semanas pós nascimento no j-ésimo instante de observação e e_{ij} têm média zero, variância constante σ^2 e são não correlacionados. Aqui o parâmetro α denota o peso esperado para animais na 12a semana pós nascimento e β corresponde ao ganho esperado de peso por semana.

Como cada animal é pesado várias vezes, a suposição de que os erros e_{ij} não são correlacionados pode não ser adequada, pois animais com peso acima ou abaixo da média na 12a semana tendem a manter esse padrão ao longo das observações. Para avaliar esse aspecto, convém construir um **gráfico de perfis** em que as observações realizadas num mesmo animal são ligadas por segmentos de reta, como indicado na Figura 6.31. A correlação entre as observações realizadas no mesmo animal fica evidenciada no gráfico do desenhista disposto na Figura 6.32.

Como no gráfico de perfis a variabilidade da resposta é similar em todos os instantes de observação e os perfis individuais têm aproximadamente as mesmas inclinações e além disso, no gráfico do desenhista podem-se notar correlações lineares com magnitudes semelhantes entre as medidas realizadas em cada par de instantes de observação, um modelo alternativo que

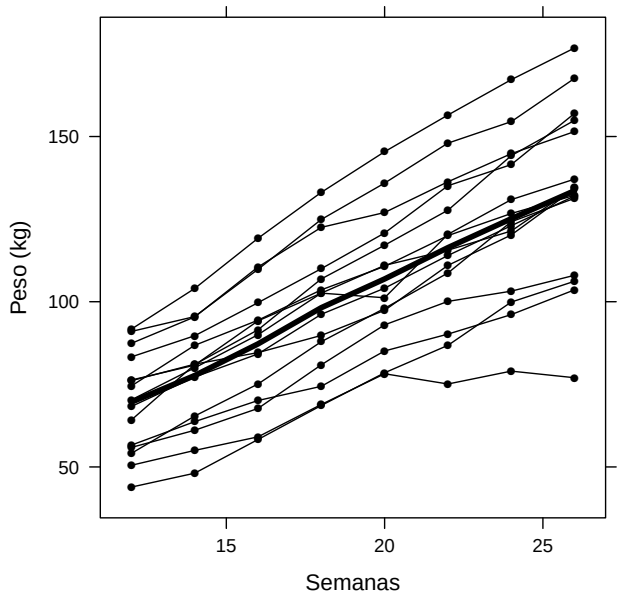


Figura 6.31: Gráfico de perfis para os dados da Tabela 2.2.

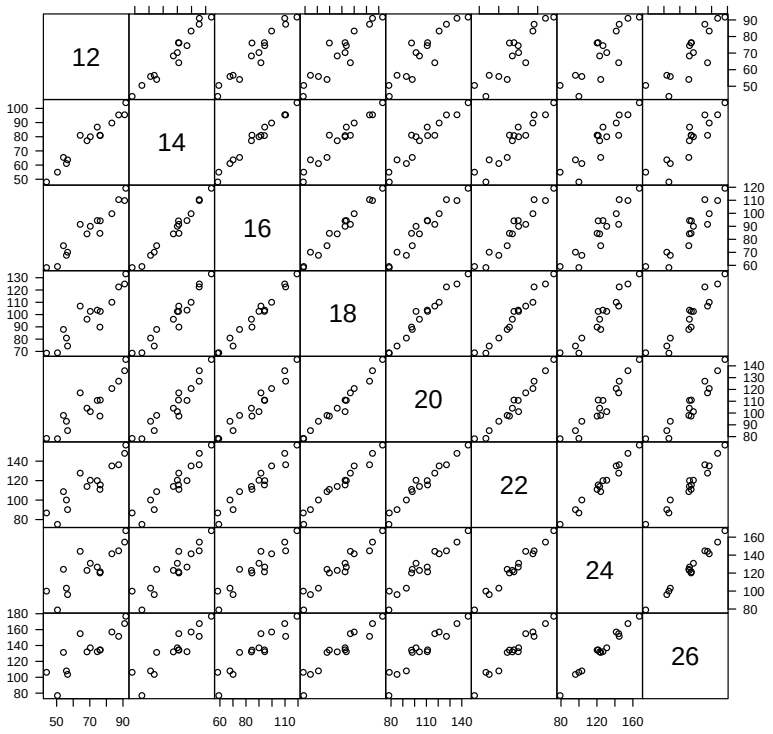


Figura 6.32: Gráfico do desenhista para os dados da Tabela 2.2.

incorpora essas características é um **modelo linear misto** dado por

$$y_{ij} = \alpha + \beta(x_j - 12) + a_i + e_{ij}, \tag{6.28}$$

em que os termos y_{ij}, x_j, α, β e e_{ij} são definidos como no modelo (6.27) e a_i é um **efeito aleatório** com média zero e variância σ_a^2, independente de e_{ij}. Esse modelo é homoscedástico, com variância de y_{ij} igual a $\sigma_a^2 + \sigma^2$ e covariância entre y_{ij} e y_{ik}, $j \neq k$ igual a σ_a^2. Essencialmente, esse modelo considera que o crescimento de cada bezerro pode ser modelado por uma reta com a mesma inclinação β, porém com intercepto $\alpha + a_i$ que varia de bezerro para bezerro. O intercepto tem um componente aleatório porque os animais constituem uma amostra de uma população para a qual se quer fazer inferência. Os parâmetros α e β constituem as características populacionais de interesse.

As estimativas dos parâmetros α e β obtidas do ajuste dos modelos (6.27) e (6.28) são iguais $\widehat{\alpha} = 69,9$ e $\widehat{\beta} = 4,7$, porém os erros padrões correspondentes são menores sob o modelo (6.28), nomeadamente, $5,6$ *versus* $7,8$ para $\widehat{\alpha}$ e $0,1$ *versus* $0,4$ para $\widehat{\beta}$.

Como existem três tipos de resíduos para essa classe de modelos, ferramentas diagnósticas são bem mais complexas do que aquelas apropriadas para os modelos lineares usais. Detalhes sobre a análise de modelos mistos podem ser obtidos em Singer et al. (2018).

6.5 Regressão Logística

Exemplo 6.9. O conjunto de dados apresentado na Tabela 6.9 (disponível no arquivo `inibina`) foi obtido de um estudo cuja finalidade era avaliar a utilização da inibina B como marcador da reserva ovariana de pacientes submetidas à fertilização *in vitro*. A variável explicativa é a diferença entre a concentração sérica de inibina B após estímulo com o hormônio FSH e sua concentração sérica pré estímulo e a variável resposta é a classificação das pacientes como boas ou más respondedoras com base na quantidade de oócitos recuperados. Detalhes podem ser obtidos em Dzik et al. (2000).

A diferença entre esse problema e aqueles estudados nas seções anteriores está no fato de a variável resposta ser dicotômica e não contínua. Se definirmos a variável Y com valor igual a 1 no caso de resposta positiva e igual a zero no caso de resposta negativa, a resposta média será igual à proporção $p = \mathrm{E}(Y)$ de pacientes com resposta positiva. Essencialmente, o objetivo da análise é modelar essa proporção como função da variável explicativa. Em vez de modelar essa resposta média, convém modelar uma função dela, a saber o logaritmo da chance de resposta positiva (ver Seção 4.2) para evitar estimativas de proporções com valores fora do intervalo (0, 1). O modelo correspondente pode ser escrito como

$$\log \frac{P(Y_i = 1|X = x)}{P(Y_i = 0|X = x)} = \alpha + \beta x_i,\ i = 1, \ldots, n. \tag{6.29}$$

Tabela 6.9: Concentração de inibina B antes e após estímulo hormonal em pacientes submetidas a fertilização *in vitro*

ident	resposta	inibpre	inibpos	ident	resposta	inibpre	inibpos
1	pos	54,03	65,93	17	pos	128,16	228,48
2	pos	159,13	281,09	18	pos	152,92	312,34
3	pos	98,34	305,37	19	pos	148,75	406,11
4	pos	85,30	434,41	20	neg	81,00	201,40
5	pos	127,93	229,30	21	neg	24,74	45,17
6	pos	143,60	353,82	22	neg	3,02	6,03
7	pos	110,58	254,07	23	neg	4,27	17,80
8	pos	47,52	199,29	24	neg	99,30	127,93
9	pos	122,62	327,87	25	neg	108,29	129,39
10	pos	165,95	339,46	26	neg	7,36	21,27
11	pos	145,28	377,26	27	neg	161,28	319,65
12	pos	186,38	1055,19	28	neg	184,46	311,44
13	pos	149,45	353,89	29	neg	23,13	45,64
14	pos	33,29	100,09	30	neg	111,18	192,22
15	pos	181,57	358,45	31	neg	105,82	130,61
16	pos	58,43	168,14	32	neg	3,98	6,46

pos: resposta positiva
neg: resposta negativa

ou equivalentemente (ver Exercício 20), como

$$P(Y_i = 1|X = x) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)},\ i = 1, \ldots, n. \tag{6.30}$$

Neste contexto, o parâmetro α é interpretado como o logaritmo da chance de resposta positiva para pacientes com $x_i = 0$ (concentrações de inibina pré e pós estímulo iguais) e o parâmetro β corresponde ao logaritmo da razão entre a chance de resposta positiva para pacientes com diferença de uma unidade na variável explicativa (ver Exercício 21).

O ajuste desse modelo é realizado pelo método de máxima verossimilhança. A função de verossimilhança a ser maximizada é

$$\ell(\alpha, \beta|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}$$

$$p(x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

Sua maximização pode ser concretizada por meio da maximização de seu

logaritmo

$$L(\alpha, \beta|\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \Big\{ y_i \log[p(x_i)] + (1 - y_i) \log[1 - p(x_i)] \Big\}.$$

Os estimadores de máxima verossimilhança de α e β correspondem à solução das **equações de estimação**

$$\sum_{i=1}^{n} \Big\{ y_i - \frac{\exp(\widehat{\alpha} + \widehat{\beta} x_i)}{1 + \exp(\widehat{\alpha} + \widehat{\beta} x_i)} \Big\} = 0 \quad \text{e} \quad \sum_{i=1}^{n} x_i \Big\{ y_i - \frac{\exp(\widehat{\alpha} + \widehat{\beta} x_i)}{1 + \exp(\widehat{\alpha} + \widehat{\beta} x_i)} \Big\} = 0.$$

Como esse sistema de equações não tem solução explícita, deve-se recorrer a métodos iterativos como o **método de Newton-Raphson**. Para detalhes o leitor poderá consultar Paulino e Singer (2006), por exemplo.

O uso da função glm() produz os resultados a seguir:

```
Call:
glm(formula = resposta ~ difinib, family = binomial, data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9770  -0.5594   0.1890   0.5589   2.0631

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.310455   0.947438  -2.439  0.01474 *
inib         0.025965   0.008561   3.033  0.00242 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 24.758  on 30  degrees of freedom
AIC: 28.758

Number of Fisher Scoring iterations: 6
```

As estimativas dos parâmetros (com erro padrão entre parênteses) α e β correspondentes ao modelo ajustado aos dados da Tabela 6.9 são, respectivamente, $\widehat{\alpha} = -2,31\ (0,95)$ e $\widehat{\beta} = 0,03\ (0,01)$. Consequentemente, a chance de resposta positiva para pacientes com mesmo nível de inibina B pré e pós estímulo hormonal é $\exp(\widehat{\alpha}) = 0,10\ (0,09)$. Essa chance fica multiplicada por $\exp(\widehat{\beta}) = 1,03\ (0,09)$ para cada aumento de uma unidade na diferença entre os níveis de inibina B pré e pós estímulo hormonal.[2]

[2]Os erros padrões de $\exp(\widehat{\alpha})$ e $\exp(\widehat{\beta})$ são calculados por meio do **método Delta**. Ver Nota de Capítulo 6.

A função predict() pode ser usada para estimar a probabilidade de que a resposta seja positiva, dados os valores da variável explicativa. Algumas dessas probabilidades estão indicadas abaixo:

```
      1         2         3         4         5         6
0.1190483 0.7018691 0.9554275 0.9988353 0.5797138 0.9588247
      7         8         9        10
0.8045906 0.8362005 0.9534173 0.8997726
```

Por exemplo, o valor 0,1190483 foi obtido calculando-se

$$P(X = 1|X = 11,90) = \frac{\exp\{-2,310455 + (0,025965)(11,90)\}}{1 + \exp\{-2,310455 + (0,025965)(11,90)\}}. \quad (6.31)$$

Para prever se a resposta vai ser positiva ou negativa, temos que converter essas probabilidades previstas em rótulos de classes,"positiva"/ ou "negativa". Considerando respostas positivas como aquelas cuja probabilidade seja maior do que 0,7, digamos, podemos utilizar a função `table()` para obter a seguinte tabela:

```
table(glm.pred,resposta)
```

```
          resposta
glm.pred negativa positiva
negativa       11        5
positiva        2       14
```

Os elementos da diagonal dessa tabela indicam os números de observações corretamente classificadas. Ou seja, a proporção de respostas corretas será (11+14)/32= 78%. Esse valor depende do limiar fixado, $0,7$, no caso. Um *default* usualmente fixado é $0,5$, e nesse caso, a proporção de respostas corretas vai aumentar.

A utilização de Regressão Logística nesse contexto de classificação será detalhada no Capítulo 10.

Uma das vantagens do modelo de regressão logística é que, com exceção do intercepto, os coeficientes podem ser interpretados como razões de chances e suas estimativas são as mesmas independentemente de os dados terem sido obtidos prospectiva ou retrospectivamente (ver Seção 4.2).

Quando todas as variáveis envolvidas são categorizadas, é comum apresentar os dados na forma de uma tabela de contingência e nesse caso, as estimativas também podem ser obtidas pelo método de mínimos quadrados generalizados.

Exemplo 6.10: Num estudo epidemiológico, 1448 pacientes com problemas cardíacos foram classificados segundo o sexo (feminino ou masculino), idade (< 55 anos ou ≥ 55 anos) e status relativo à hipertensão arterial (sem ou com). Por meio de um procedimento de cineangiocoronariografia, o grau de lesão das artérias coronarianas foi classificado como $< 50\%$ ou $\geq 50\%$. Os dados estão resumidos na Tabela 6.10.

Tabela 6.10: Frequência de pacientes avaliados em um estudo epidemiológico

		Hipertensão	Grau de lesão	
Sexo	Idade	arterial	$< 50\%$	$\geq 50\%$
Feminino	< 55	sem	31	17
Feminino	< 55	com	42	27
Feminino	≥ 55	sem	55	42
Feminino	≥ 55	com	94	104
Masculino	< 55	sem	80	112
Masculino	< 55	com	70	130
Masculino	≥ 55	sem	74	188
Masculino	≥ 55	com	68	314

Fonte: Singer, J.M. e Ikeda, K. (1996).

Nesse caso, um modelo de regressão logística apropriado (escrito de forma geral) para a análise é

$$\log\{P(Y_{ijk} = 1)/[1 - P(Y_{ijk} = 1)]\} = \alpha + \beta x_i + \gamma v_j + \delta w_k, \qquad (6.32)$$

$i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K$, em que $Y_{ijk} = 1$ se um paciente do sexo i ($i = 1$: feminino, $i = 2$: masculino), idade j ($j = 1$: < 55 , $j = 2$: ≥ 55) e status relativo à hipertensão k ($k = 1$: sem, $k = 2$: com) tiver lesão coronariana $\geq 50\%$ e $Y_{ijk} = 0$ em caso contrário. Aqui, I, J e K são iguais a 2 e $x_1 = 0$ e $x_2 = 1$ para pacientes femininos ou masculinos, respectivamente, $v_1 = 0$ e $v_2 = 1$ para pacientes com idades < 55 ou ≥ 55, respectivamente e $w_1 = 0$ e $w_2 = 1$ para pacientes sem ou com hipertensão, respectivamente. O parâmetro α corresponde ao logaritmo da chance de lesão coronariana $\geq 50\%$ para mulheres não hipertensas com menos de 55 anos (consideradas como referência); o parâmetro β corresponde ao logaritmo da razão entre a chance de lesão coronariana $\geq 50\%$ para homens não hipertensas com menos de 55 anos e a chance correspondente para mulheres com as mesmas características (de idade e de hipertensão); o parâmetro γ corresponde ao logaritmo da razão entre a chance de lesão coronariana $\geq 50\%$ para pacientes com 55 anos ou mais e a chance correspondente para pacientes com as mesmas características (sexo e hipertensão) e menos de 55 anos; o parâmetro δ corresponde ao logaritmo da razão entre a chance de lesão coronariana $\geq 50\%$ para pacientes hipertensos e a chance correspondente para pacientes com as mesmas características (de sexo e de idade) não hipertensos.

Usando-se o pacote ACD e a função funlinWLS(), as estimativas dos parâmetros obtidas por máxima verossimilhança (com erros padrões entre parênteses) são: $\widehat{\alpha} = -0,91(0,15)$, $\widehat{\beta} = 1,23(0,13)$, $\widehat{\gamma} = 0,67(0,12)$, $\widehat{\delta} = 0,41(0,12)$. Um intervalo de confiança aproximado com coeficiente de confiança de 95% correspondente à chance de lesão coronariana $\geq 50\%$ para mulheres não hipertensas com menos de 55 anos pode ser obtido por meio

Tabela 6.11: Estimativas (e intervalos de confiança de 95%) para a chance e razões de chances associadas aos dados da Tabela 6.10

	Estimativa	Limite inferior	Limite superior
Chance de lesão $\geq 50\%$ mulheres < 55 não hipertensas	0,40	0,30	0,54
Razão de chances para sexo masculino	3,43	2,69	4,38
Razão de chances para idade ≥ 55	1,95	1,55	2,48
Razão de chances para hipertensão	1,51	1,20	1,89

Tabela 6.12: Estimativas das chances de lesão coronariana para $\geq 50\%$ para pacientes com diferentes níveis dos fatores de risco obtidas com os dados da Tabela 6.10

Sexo	Idade	Hipertensão	Chance (lesão $\geq 50\%$)/(lesão $< 50\%$)
Fem	< 55	sem	R
Fem	< 55	com	R×1,51
Fem	≥ 55	sem	R×1,95
Fem	≥ 55	com	R×1,51×1,95
Masc	< 55	sem	R×3,43
Masc	< 55	com	R×3,43×1,51
Masc	≥ 55	sem	R×3,43×1,95
Masc	≥ 55	com	R×3,43×1,95×1,51

da exponenciação dos limites de um intervalo de confiança para o parâmetro α; o mesmo procedimento pode ser empregado para a obtenção de intervalos de confiança para as razões de chances associadas ao sexo, idade e status de hipertensão. Esses intervalos estão dispostos na Tabela 6.11.

Se os 1448 pacientes avaliados no estudo puderem ser considerados como uma amostra aleatória de uma população de interesse, a chance de lesão coronariana $\geq 50\%$ para uma mulher não hipertensa com idade < 55 é de 0,40 [IC(95%) = 0,30 a 0,54]. Independentemente dessa suposição, *i.e.*, mesmo que essa chance tenha um valor R desconhecido, ela fica multiplicada por 3,43 [IC(95%) = 2,69 a 4,38] para homens não hipertensos e de mesma idade, por 1,95 [IC(95%) = 1,55 a 2,48] para mulheres não hipertensas com idade ≥ 55 ou por 1,51 [IC(95%) = 1,20 a 1,89] para mulheres hipertensas com idade < 55. O modelo ainda permite estimar as chances para pacientes com diferentes níveis dos três fatores, conforme indicado na Tabela 6.12. Quando

o estudo não permite estimar a chance de lesão coronariana $\geq 50\%$ para o grupo de referência (neste caso, mulheres não hipertensas com idade < 55) como em estudos retrospectivos, as razões de chances estimadas continuam válidas. Nesse contexto, por exemplo, a chance de lesão coronariana $\geq 50\%$ para homens hipertensos com idade ≥ 55 é 1,95×1,51 a chance correspondente para homens não hipertensos com idade < 55. O cálculo do erro padrão dessa razão de chances depende de uma estimativa da matriz de covariâncias dos estimadores dos parâmetros do modelo e está fora do escopo deste texto. O leitor pode consultar Paulino e Singer (2006) para detalhes.

A avaliação da qualidade do ajuste de modelos de regressão é baseada em resíduos da forma $y_i - \widehat{y}_i$ em que y_i é a resposta observada para a i-ésima unidade amostral e $\widehat{y}_i$ é o correspondente valor ajustado, *i.e.* predito pelo modelo. Para regressão logística a avaliação do ajuste é mais complexa, pois os resíduos podem ser definidos de diferentes maneiras. Apresentamos alguns detalhes na Nota de Capítulo 7.

6.6 Regularização

Consideremos um exemplo [proposto em Bishop (2006)] cujo objetivo é ajustar um modelo de regressão polinomial a um conjunto de 10 pontos gerados por meio da expressão $y_i = sen(2\pi x_i) + e_i$ em que e_i segue um distribuição Normal com média nula e variância σ^2. Os dados estão representados na Figura 6.33 por pontos em azul. A curva verde corresponde a $y_i = sen(2\pi x_i)$; em vermelho estão representados os ajustes baseados em regressões polinomiais de graus, 0, 1, 3 e 9. Claramente, a curva baseada no polinômio do terceiro grau consegue reproduzir o padrão da curva geradora dos dados sem, no entanto, predizer os dados com total precisão. A curva baseada no polinômio de grau 9, por outro lado, tem um ajuste perfeito, mas não reproduz o padrão da curva utilizada para gerar os dados, Esse fenômeno é conhecido como sobreajuste.

O termo regularização refere-se a um conjunto de técnicas utilizadas para especificar modelos que se ajustem a um conjunto de dados evitando o **sobreajuste** (*overfitting*). Essencialmente, essas técnicas servem para ajustar modelos de regressão em que a função de perda contém um termo de penalização cuja finalidade é reduzir a influência de coeficientes responsáveis por flutuações excessivas.

Embora haja várias técnicas de regularização, consideraremos apenas a regularização L_2, ou **Ridge**, a regularização L_1 ou **Lasso** (*least absolute shrinkage and selection operator*) e uma mistura dessas duas, chamada de **Elastic net**.

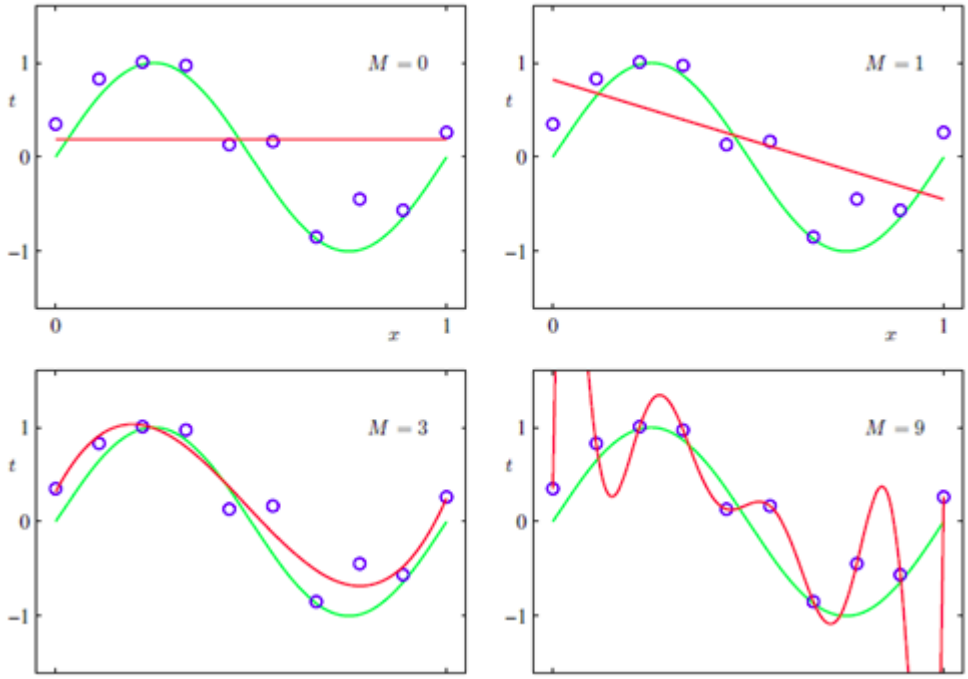


Figura 6.33: Ajuste de modelos polinomiais a um conjunto de dados hipotéticos.

O componente de regularização da técnica Lasso usa uma soma de valores absolutos dos parâmetros e um **coeficiente de penalização** que os encolhe para zero. Essa técnica serve para seleção de modelos, porque associa pesos nulos a parâmetros não significativos. Isso implica uma **solução esparsa**[3] Na regularização L_2, por outro lado, o termo de regularização usa uma soma de quadrados dos parâmetros e um coeficiente de penalização que força alguns pesos a serem pequenos, mas não os anula e consequentemente não conduz a soluções esparsas. Essa técnica de regularização não é robusta com relação a valores atípicos, pois pode conduzir a valores muito grandes do termo de penalização.

Neste capítulo vamos nos basear em Hastie et al. (2017), James et al. (2017) e Medeiros (2019).

6.6.1 Regularização L_2 (Ridge)

Consideremos o modelo de regressão

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_p x_{pt} + \varepsilon_t, \quad t = 1, 2, \ldots, n, \tag{6.33}$$

[3]Dizemos que um modelo é esparso se a maioria dos elementos do correspondente vetor de parâmetros é nula ou desprezável.

com as p variáveis preditoras reunidas no vetor $\mathbf{x}_t = (x_{1t}, \ldots, x_{pt})^\top$, y_t representando a variável resposta, ε_t indicando as inovações de média zero e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denotando o vetor de parâmetros a serem estimados. Supomos adicionalmente que $\beta_0 = 0$ e consideremos estimadores de mínimos quadrados (EMQ) penalizados da forma

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \tag{6.34}$$

em que λ é o coeficiente de regularização, que controla o número de parâmetros do modelo. Se $\lambda = \infty$, não há variáveis a serem incluídas no modelo e se $\lambda = 0$, obtemos os EMQ usuais. Dizemos que $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)$ é o **estimador *Ridge***. Pode-se mostrar que

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{6.35}$$

em que $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$ é a matriz de especificação do modelo e $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

Alguns resultados sobre as propriedades dessa classe de estimadores são:

1) Em geral, o estimador *Ridge* não é consistente. Sua consistência assintótica vale quando $\lambda = v\lambda_n \to \infty$, $\lambda_n/n \to 0$ e $p < n$.

2) O estimador *Ridge* é enviesado para os parâmetros não nulos.

3) A técnica de regularização *Ridge* não serve para a seleção de modelos.

4) A escolha do coeficiente de regularização λ pode ser feita via validação cruzada ou por meio de algum critério de informação. Detalhes são apresentados na Nota de Capítulo 10.

5) A técnica de regressão *Ridge* foi introduzida por Hoerl e Kennard (1970) para tratar do problema da multicolinearidade.

Obter o mínimo em (6.34) é equivalente a minimizar a soma de quadrados não regularizada sujeita à restrição

$$\sum_{j=1}^{p} \beta_j^2 \leq m, \tag{6.36}$$

para algum valor apropriado m, ou seja, é um problema de otimização com multiplicadores de Lagrange.

Na Figura (6.34) apresentamos um esquema com o valor ótimo do vetor $\boldsymbol{\beta}$, a região circular correspondente à restrição (6.36) e os círculos representando as curvas de nível da função erro não regularizada.

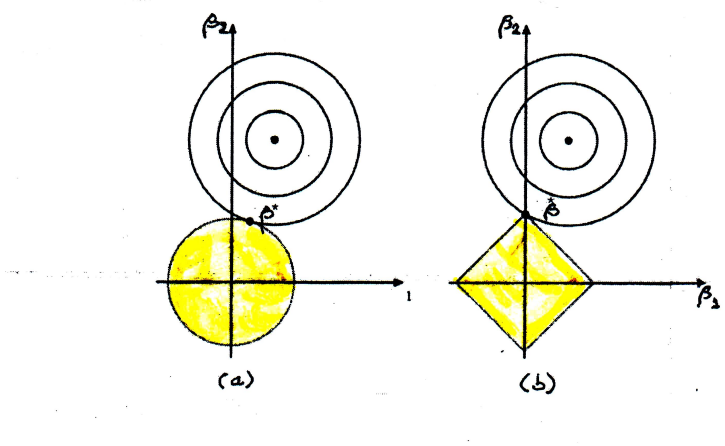


Figura 6.34: Esparsidade do modelo: (a) ridge; (b) lasso.

6.6.2 Regularização L_1 (Lasso)

Consideremos, agora, o **estimador Lasso**, obtido de

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{6.37}$$

Neste caso, a restrição (6.36) é substituída por

$$\sum_{j=1}^{p} |\beta_j| \leq m, \tag{6.38}$$

No painel (b) da Figura 6.34 (b), podemos observar que a regularização Lasso pode gerar uma solução esparsa, ou seja com $\beta_1^* = 0$.

Algumas propriedades estatísticas do estimador Lasso:

1) O estimador Lasso encolhe para zero os parâmetros que correspondem a preditores redundantes.

2) O estimador é enviesado para parâmetros não nulos.

3) Sob certas condições, o estimador Lasso seleciona as variáveis relevantes do modelo atribuindo pesos nulos aos respectivos coeficientes.

4) Quando $p = n$, ou seja, quando o número de variáveis preditoras é igual ao número de observações, a técnica Lasso corresponde à aplicação de um **limiar brando** (*soft threshold*) a $Z_j = \mathbf{x}_j^\top \mathbf{y}/n$, ou seja,

$$\widehat{\beta}_j(\lambda) = \text{sinal}(Z_j) \left(|Z_j| - \lambda/2 \right)_+, \tag{6.39}$$

em que $(x)_+ = \max\{x, 0\}$.

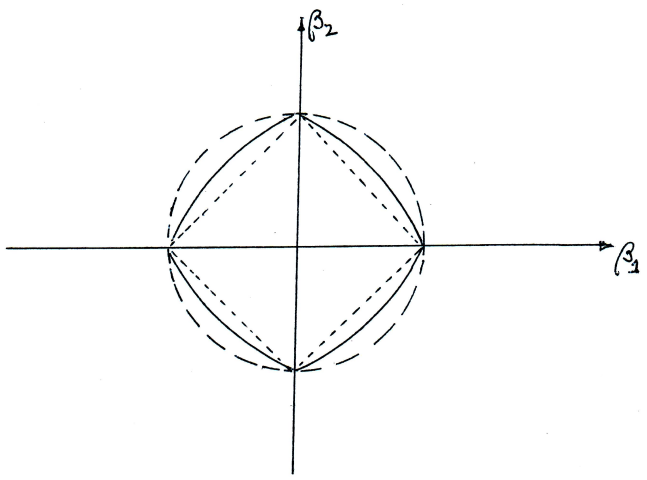


Figura 6.35: Geometria das restrições *Elastic Net* (curva contínua), *Ridge* (curva tracejada) e Lasso (curva pontilhada)

Para outras propriedades, veja Medeiros (2019) e Bühlmann and van de Geer (2011).

6.6.3 Outras propostas

O estimador *Elastic net* (EN) é

$$\widehat{\boldsymbol{\beta}}_{\text{EN}}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} \frac{1}{n}(y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda_2 \sum_{i=1}^{p} \beta_i^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i|. \quad (6.40)$$

Na Figura 6.35 apresentamos esquematicamente uma região delimitada pela restrição $J(\boldsymbol{\beta}) \leq m$, em que $J(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|$, para algum m, com $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, além daquelas delimitadas pelas restrições *Ridge* e Lasso.

Pode-se mostrar que sob determinadas condições, o estimador *Elastic Net* é consistente.

O estimador **Lasso adaptativo** (adaLASSO) é

$$\widehat{\boldsymbol{\beta}}_{AL}(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{i=1}^{p} w_i |\beta_i|, \quad (6.41)$$

em que $w_1, \ldots, w_p$ são pesos não negativos pré definidos. Usualmente, toma-se $w_j = |\tilde{\beta}_j|^{-\tau}$, para $0 < \tau \leq 1$ e $\tilde{\beta}_j$ é um estimador inicial (por exemplo o estimador Lasso).

O estimador **Lasso adaptativo** é consistente sob condições não muito fortes.

A função `adalasso` do pacote `parcor` do `R` pode ser usada para calcular esse estimador. O pacote `glmnet` do `R` pode ser usado para obter estimadores Lasso e *Elastic net* sob modelos de regressão linear, regressão logística e multinomial, regressão Poisson além de modelos de Cox. Para detalhes, veja Friedman et al. (2010).

Exemplo 6.11. Vamos considerar o conjunto de dados `esforco`, centrando o interesse na predição da variável resposta Y: VO2 (consumo de oxigênio em ml/(kg.min)) com base nas variáveis preditoras X_1: Idade (em anos), X_2: Peso (em kg), X_3: Superfície corpórea e X_4: IMC (índice de massa corpórea em kg/m2).

Ajustando o modelo via mínimos quadrados ordinários, obtemos os coeficientes:

```
   Intercept       Idade          Peso        Sup.Corp
  5.204200870  -0.002386615  -0.026465910   0.365381912
      IMC
 -0.011107867
```

Os coeficientes correspondentes obtidos por meio de regularização *Ridge* são

```
Intercept    5.185964640
Idade       -0.000133776
Peso        -0.006946405
Sup.Corp    -0.295094364
IMC         -0.022923850
```

O valor do coeficiente de regularização $\lambda = 0,82065$, mostra que as estimativas para os coeficientes de Idade e Peso foram encolhidas para zero, enquanto aquelas correspondentes à Sup.Corp tem peso maior do que as demais. Neste caso, a raiz quadrada do **erro quadrático médio** (*root mean squared error*) e o **coeficiente de determinação** são, respectivamente $RMSE = 0,928$ e $R^2 = 0,235$.

Os coeficientes correspondentes obtidos por meio de regularização Lasso são

```
Intercept    4.95828012
Idade        .
Peso        -0.01230145
Sup.Corp     .
IMC         -0.02011871
```

O valor do coeficiente de regularização $\lambda = 0,0257$ mostra que as estimativas os coeficientes Idade e Sup. Corp foram encolhidas para zero. Neste caso $RMSE$ e R^2 são, respectivamente, $0,927$ e $0,228$.

Os coeficientes correspondentes obtidos por meio de regularização *Elastic net* são

```
Intercept    4.985532570
Idade         .
Peso        -0.009099925
Sup.Corp    -0.097034844
IMC         -0.023302254
```

Os parâmetros de suavização estimados foram $\alpha = 0,1$ e $\lambda = 0,227$ e também indicam que o coeficiente associado à Idade foi encolhido para zero. Também obtemos $RMSE = 0,927$ e $R^2 = 0,228$ neste caso, mostrando que os três métodos de regularização têm desempenhos similares quando vistos pelas óticas do $RMSE$ e do R^2.

6.7 Notas de capítulo

1) **Inferência baseada em modelos de regressão linear simples**.

Para o modelo (6.1) fizemos a suposição de que os erros são não correlacionados, têm média 0 e variância constante σ^2. Geralmente, também se supõe que a variável explicativa X seja fixa. Se quisermos testar hipóteses sobre os parâmetros α e β ou construir intervalos de confiança para eles por meio de estatísticas com distribuições exatas, devemos fazer alguma suposição adicional sobre a distribuição dos erros. Usualmente, supõe-se que os e_i têm uma distribuição Normal. Se a distribuição dos erros tiver caudas mais longas (pesadas) do que as da distribuição Normal, os estimadores de mínimos quadrados podem se comportar de forma inadequada e estimadores robustos devem ser usados.

Como

$$\widehat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \sum_{i=1}^n w_i y_i,$$

com $w_i = (x_i - \overline{x}) / \sum_{i=1}^n (x_i - \overline{x})^2$, o estimador $\widehat{\beta}$ é uma função linear das observações y_i. O mesmo vale para $\widehat{\alpha}$. Utilizando esse resultado, pode-se demonstrar (veja a seção de exercícios) que

a) $\mathrm{E}(\widehat{\alpha}) = \alpha$ e $\mathrm{E}(\widehat{\beta}) = \beta$, ou seja, os EMQ são não enviesados.
b) $\mathrm{Var}(\widehat{\alpha}) = \sigma^2 \sum_{i=1}^n x_i^2 / [n \sum_{i=1}^n (x_i - \overline{x})]^2$.
c) $\mathrm{Var}(\widehat{\beta}) = \sigma^2 / \sum_{i=1}^n (x_i - \overline{x})^2$.
d) $\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = -\sigma^2 \overline{x} / \sum_{i=1}^n (x_i - \overline{x})^2$.

Com a suposição adicional de normalidade, pode-se mostrar que

e) $y_i \sim N(\alpha + \beta x_i, \sigma^2)$
f) as estatísticas

$$t_{\widehat{\alpha}} = \frac{\widehat{\alpha} - \alpha}{S} \sqrt{\frac{n \sum (x_i - \overline{x})^2}{\sum x_i^2}}$$

e

$$t_{\widehat{\beta}} = \frac{\widehat{\beta} - \beta}{S}\sqrt{\sum (x_i - \overline{x})^2}$$

têm distribuição t de Student com $(n-2)$ graus de liberdade. Nesse contexto, os resíduos padronizados, definidos em (6.9) também seguem uma distribuição t de Student com $(n-2)$ graus de liberdade. Daí a denominação alternativa de resíduos studentizados.

Um teorema importante conhecido como **Teorema de Gauss-Markov** (e que não depende da suposição de normalidade dos erros) afirma que os EMQ têm variância mínima na classe dos estimadores não enviesados que sejam funções lineares das observações y_i. Com esses resultados é possível testar as hipóteses $H_0 : \alpha = 0$ e $H_0 : \beta = 0$, bem como construir intervalos de confiança para esses parâmetros.

Quando os erros não seguem uma distribuição Normal, mas o tamanho da amostra é suficientemente grande, pode-se mostrar com o auxílio do **Teorema Limite Central** que sob certas condições de regularidade (usualmente satisfeitas na prática), os estimadores $\widehat{\alpha}$ e $\widehat{\beta}$ têm distribuições aproximadamente normais com variâncias que podem ser estimadas pelas expressões indicadas nos itens b) e c). Detalhes podem ser obtidos em Sen et al. (2009).

2) **Estimação e previsão sob modelos de regressão linear simples**.

Um dos objetivos da análise de regressão é fazer previsões sobre a variável resposta com base em valores das variáveis explicativas. Por simplicidade trataremos do caso de regressão linear simples. Uma estimativa para o valor esperado $\mathrm{E}(Y|X = x_0)$ da variável resposta Y dado um valor x_0 da variável explicativa é $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x_0$ e com base nos resultados apresentados na Nota de Capítulo 1 pode-se mostrar que a variância de $\widehat{y}$ é

$$\mathrm{Var}(\widehat{y}) = \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\Big].$$

Então os limites superior e inferior para um **intervalo de confiança** aproximado com coeficiente de confiança de 95% para o valor esperado de Y dado $X = x_0$ são

$$\widehat{y} \pm 1,96S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

com S^2 denotando uma estimativa de σ^2. *Grosso modo*, podemos dizer que esse intervalo deve conter o verdadeiro valor esperado de $\mathrm{E}(Y|X = x)$, *i.e.*, a média de Y para todas as observações em que $X = x_0$. Isso não significa que esperamos que o intervalo contenha o verdadeiro valor de Y, digamos Y_0 para uma unidade de investigação para a qual

$X = x_0$. Nesse caso precisamos levar em conta a variabilidade de $Y|X = x_0$ em torno de seu valor esperado $\mathrm{E}(Y|X = x_0)$.

Como $Y_0 = \widehat{y} + e_0$ sua variância é

$$\mathrm{Var}(Y_0) = \mathrm{Var}(\widehat{y}) + \mathrm{Var}(e_0) = \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x - \overline{x})^2}\Big] + \sigma^2$$

Então os limites superior e inferior de um **intervalo de previsão** (aproximado) para Y_0 são

$$\widehat{y} \pm 1,96S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

Note que se aumentarmos indefinidamente o tamanho da amostra, a amplitude do intervalo de confiança para o valor esperado tenderá para zero, porém a amplitude do intervalo de previsão correspondente a uma unidade específica tenderá para $2 \times 1,96 \times \sigma$.

3) **Coeficiente de determinação**.

Consideremos um conjunto de dados pareados $(x_1, y_1), \ldots, (x_n, y_n)$ de duas variáveis contínuas X e Y. Se não levarmos em conta a variável X para explicar a variabilidade da variável Y como no modelo de regressão linear simples, a melhor previsão para Y é $\overline{y}$ e uma estimativa da variância de Y é dada por $S^2 = \sum_{i=1}^n (y_i - \overline{y})^2/(n-1)$. Para relacionar esse resultado com aquele obtido por meio de um modelo de regressão linear para os mesmos dados, podemos escrever

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}).$$

Uma representação gráfica dessa relação está apresentada na Figura **??**.

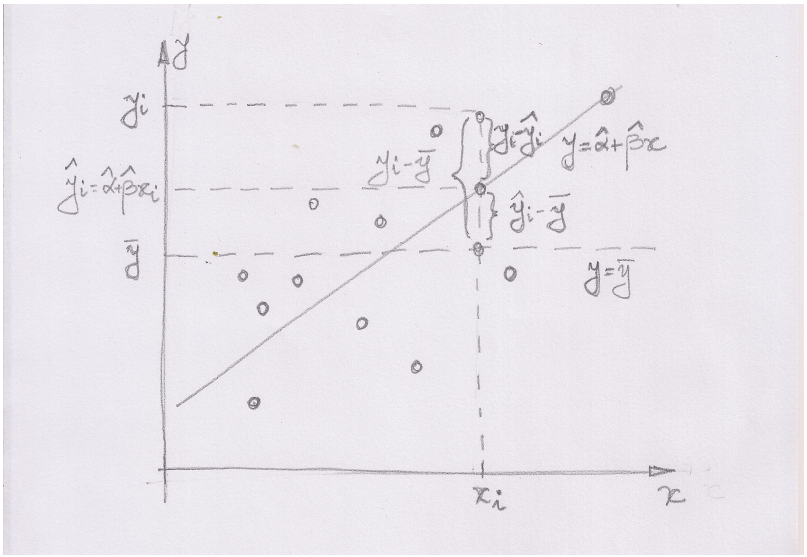


Figura 6.36: Representação gráfica da decomposição da soma de quadrados.

Pode-se mostrar (ver Exercício 15) que

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}\widehat{e}_i^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$$

ou, de forma abreviada,

$$SQTot=SQRes+SQReg.$$

Esse resultado indica que a soma de quadrados total ($SQTot$) pode ser decomposta num termo correspondente à variabilidade dos resíduos ($SQRes$) e em outro correspondente à regressão ($SQReg$). Quanto maior for esse último termo, maior é a evidência de que a variável X é útil para explicar a variabilidade da variável Y. Tendo em vista a expressão (6.6), é fácil ver que

$$SQReg=\widehat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Nesse contexto, a estatística $R^2=SQReg/SQTot$ corresponde à porcentagem da variabilidade de Y explicada pelo modelo, ou seja, pela introdução da variável X no modelo mais simples, $y_i=\mu+e_i$.

Como a soma de quadrados $SQReg$ (e consequentemente, o coeficiente R^2) sempre aumenta quando mais variáveis explicativas são introduzidas no modelo, convém considerar uma penalidade correspondente ao número de variáveis explicativas. Nesse sentido, para comparação de modelos com números diferentes de variáveis explicativas, costuma-se utilizar o coeficiente de determinação ajustado

$$R^2_{aj}=1-(1-R^2)\frac{n-1}{n-p-1}=1-\frac{SQRes/(n-p-1)}{SQTot/(n-1)}$$

em que p é o número de variáveis explicativas do modelo. Lembrando que

$$R^2=1-\frac{SQRes}{SQTot}=1-\frac{SQRes/n}{SQTot/n},$$

o coeficiente R^2_{aj} é obtido por meio de um aumento maior no numerador do que no denominador de R^2, com mais intensidade quanto maior for o número de variáveis explicativas.

4) **Distância de Cook**.

A distância de Cook é uma estatística que mede a mudança nos valores preditos pelo modelo de regressão quando eliminamos uma das observações. Denotando por $\widehat{\mathbf{y}}$ o vetor (de dimensão n) com os valores preditos obtidos do ajuste do modelo baseado nas n observações e por $\widehat{\mathbf{y}}^{(-i)}$ o correspondente vetor com valores preditos (de dimensão n)

obtido do ajuste do modelo baseado nas $n-1$ observações restantes após a eliminação da i-ésima, a distância de Cook é definida como

$$D_i = \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(-i)})^\top (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(-i)})}{(p+1)S}$$

em que p é o número de coeficientes de regressão e S é uma estimativa do desvio padrão. Pode-se mostrar que a distância de Cook (D_i) pode ser calculada sem a necessidade de ajustar o modelo com a omissão da i-ésima observação por meio da expressão

$$D_i = \frac{1}{p+1}\widehat{e}_i^2 \frac{h_{ii}}{(1-h_{ii})^2}.$$

lembrando que

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$$

Para o modelo de regressão linear simples, $p = 2$. Detalhes podem ser obtidos em Kutner et al. (2004).

5) **Influência local**.

Influência local é o efeito de uma pequena variação no valor da variável resposta nas estimativas dos parâmetros do modelo. Consideremos uma observação (x_j, y_j) e quantifiquemos o efeito de uma mudança de y_j para $y_j + \Delta y_j$ nos valores de $\widehat{\alpha}$ e $\widehat{\beta}$. Com esse propósito, observando que

$$\widehat{\beta} + \Delta\widehat{\beta}(y_j) = \frac{\sum_{i\neq j}(x_i - \overline{x})y_i + (x_j - \overline{x})(y_j + \Delta y_j)}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

podemos concluir que

$$\Delta\widehat{\beta}(y_j) = \frac{(x_j - \overline{x})\Delta y_j}{\sum_{i=1}^n (x_i - \overline{x})^2}. \tag{6.42}$$

Este resultado indica que, fixado Δy_j, a variação em $\widehat{\beta}$ é diretamente proporcional a $x_j - \overline{x}$ e inversamente proporcional a $(n-1)S^2$. Portanto, o efeito da variação no valor de y_j será grande se x_j estiver bastante afastado da média dos x_i e se a variabilidade dos x_i for pequena.

Para o intercepto, temos

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} = \frac{\sum_i y_i}{n} - \widehat{\beta}\overline{x},$$

logo, quando y_j é substituído por $y_j + \Delta y_j$, teremos

$$\widehat{\alpha} + \Delta\widehat{\alpha}(y_j) = \frac{\sum_{i\neq j} y_i + (y_j + \Delta y_j)}{n} - (\widehat{\beta} + \Delta\widehat{\beta})\overline{x},$$

e portanto

$$\Delta\widehat{\alpha}(y_j) = \frac{\Delta y_j}{n} - (\Delta\widehat{\beta})\overline{x},$$

ou ainda

$$\Delta\widehat{\alpha}(y_j) = \Big[\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2}\Big]\Delta y_j. \tag{6.43}$$

Se $x_j = \overline{x}$, então $\Delta\widehat{\beta} = 0$, mas $\Delta\widehat{\alpha} = \Delta y_j/n$, ou seja, Δy_j não afeta a inclinação mas afeta o intercepto. Gráficos de (6.42) e (6.43) em função dos índices de cada observação indicam para que pontos a variação nos valores da variável resposta tem maior influência nas estimativas dos parâmetros.

6) **Método Delta**.

Considere um parâmetro β para o qual se dispõe de um estimador $\widehat{\beta}$ cuja variância é $\sigma^2_{\widehat{\beta}}$ e suponha que haja interesse em obter a variância de uma transformação $g(\widehat{\beta})$. Por meio de uma expansão de Taylor, pode-se mostrar que

$$\mathrm{Var}[g(\widehat{\beta})] = [g'(\widehat{\beta})]^2\sigma^2_{\widehat{\beta}}$$

em que $g'(\widehat{\beta})$ denota a primeira derivada de g calculada no ponto $\widehat{\beta}$. Detalhes podem ser obtidos em Sen et al. (2009).

7) **Análise do ajuste de modelos de regressão logística**.

Nos casos em que todas as variáveis explicativas utilizadas num modelo de regressão logística são categorizadas, podemos agrupar as respostas y_i segundo os diferentes padrões definidos pelas combinações dos níveis dessas variáveis. Quando o modelo envolve apenas uma variável explicativa dicotômica (Sexo, por exemplo), há apenas dois padrões, nomeadamente, M e F). Se o modelo envolver duas variáveis explicativas dicotômicas (Sexo e Faixa etária com dois níveis, ≤ 40 anos e > 40 anos, por exemplo), estaremos diante de uma situação com quatro padrões, a saber, (F e ≤ 40), (F e > 40), (M e ≤ 40) e (M e > 40). A introdução de uma ou mais variáveis explicativas contínuas no modelo, pode gerar um número de padrões igual ao número de observações.

Consideremos um caso com p variáveis explicativas $\mathbf{x} = (x_1, \ldots, x_p)^\top$ e sejam M o número de padrões (correspondente ao número de valores distintos de $\mathbf{x}$) e m_j, $j = 1, \ldots, M$, o número de observações com o mesmo valor $\mathbf{x}_j$ de $\mathbf{x}$. Note que no caso mais comum, em que existe pelo menos uma variável contínua, $m_j \approx 1$ e $M \approx n$. Além disso, seja $\tilde{y}_j$ o número de respostas $Y = 1$ entre as m_j unidades amostrais com o mesmo valor $\mathbf{x}_j$. O valor ajustado $\widehat{y}_j$ correspondente a $\tilde{y}_j$ é

$$\widehat{y}_j = m_j\widehat{p}_j = m_j\frac{\exp(\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}})}$$

em que $\widehat{\boldsymbol{\beta}}$ é o estimador do vetor de parâmetros de modelo.

O **resíduo de Pearson** é definido como

$$\widehat{e}_j = \frac{\tilde{y}_j - m_j\widehat{p}_j}{\sqrt{m_j\widehat{p}_j(1-\widehat{p}_j)}}$$

e uma medida resumo para a avaliação do ajuste do modelo é a **estatística de Pearson**

$$Q_p = \sum_{j=1}^{M} \widehat{e}_j^2.$$

Para M suficientemente grande, a estatística Q_P tem distribuição assintótica χ^2 com $M-(p+1)$ graus de liberdade quando o modelo está bem ajustado.

Para evitar problemas com a distribuição assintótica de Q_P quando $M \approx n$, convém agrupar os dados de alguma forma. Hosmer and Lemeshow (1980) e Lemeshow e Hosmer (1982) sugerem que os dados sejam agrupados segundo percentis das probabilidades $\widehat{p}_i$, $i = 1, \ldots, n$ estimadas sob o modelo. Por exemplo, podem-se considerar $g = 10$ grupos, sendo o primeiro formado pelas unidades amostrais com os 10% menores valores das probabilidades estimados (ou seja, aquelas para as quais $\widehat{p}_i$ sejam menores os iguais ao primeiro decil; o segundo grupo deve conter as unidades amostrais para as quais $\widehat{p}_i$ estejam entre o primeiro e o segundo decil etc. O último grupo conterá as unidades amostrais cujas probabilidades estimadas sejam maiores que o nono decil. Com esse procedimento, cada grupo deverá conter $n_j^* = n/10$ unidades amostrais. A estatística proposta por esses autores é

$$\widehat{C} = \sum_{j=1}^{g} \frac{(o_j - n_j^*\overline{p}_j)^2}{n_j^*\overline{p}_j(1-\overline{p}_j)}$$

com $o_j = \sum_{i=1}^{c_j} y_i$ denotando o número de respostas $Y = 1$ dentre as unidades amostrais incluídas no j-ésimo grupo (c_j representa o número de padrões de covariáveis encontrados no j-ésimo grupo) e $\overline{p}_j = \sum_{i=1}^{c_j} m_i\widehat{p}_i/n_j^*$ denota a média das probabilidades estimadas no j-ésimo grupo. A estatística $\widehat{C}$ tem distribuição aproximada χ^2 com $g-2$ graus de liberdade quando o modelo estiver correto.

Os chamados **resíduos da desviância** (*deviance residuals*) são definidos a partir da logaritmo da função de verossimilhança e também podem ser utilizados com o propósito de avaliar a qualidade do ajuste de modelos de regressão logística. O leitor poderá consultar Hosmer and Lemeshow (2000) para detalhes.

8) **Regressão resistente**.

Os estimadores $\widehat{\alpha}$ e $\widehat{\beta}$ em (6.7) e (6.6) considerados para o ajuste do modelo (6.1) a um conjunto de dados $(x_i, y_i), i = 1, \ldots, n$ são baseados em $\overline{x}, \overline{y}$ e nos desvios em relação a essas médias. Esses estimadores

podem ser severamente afetados pela presença de observações discrepantes (*outliers*). Como alternativa, podemos considerar modelos de **regressão resistente**, em que os estimadores são baseados em medianas.

Para o ajuste desses modelos, inicialmente, dividimos o conjunto de n pontos em três grupos de tamanhos aproximadamente iguais. Chamemos esses grupos de E, C e D. Se $n = 3k$, cada grupo terá k pontos. Se $n = 3k + 1$, colocamos k pontos nos grupos E e D e $k + 1$ pontos no grupo C. Finalmente, se $n = 3k + 2$, colocamos $k + 1$ pontos nos grupos E e D e k pontos no grupo C.

Para cada grupo, obtemos um **ponto resumo**, cujas coordenadas são a mediana dos x_i e a mediana dos y_i naquele grupo. Denotemos esses pontos por $(x_E, y_E), (x_C, y_C), (x_D, y_D)$.

Os estimadores resistentes de β e α são dados por

$$b_0 = \frac{y_D - y_E}{x_D - x_E}, \tag{6.44}$$

e

$$a_0 = \frac{1}{3}\left[(y_E - b_0 x_E) + (y_C - b_0 x_C) + (y_D - b_0 x_D)\right]. \tag{6.45}$$

Convém notar as diferenças entre b_0 e (6.6) e entre a_0 e (6.7). A correspondente reta resistente ajustada é

$$\widetilde{y}_i = a_0 + b_0 x_i, \quad i = 1, \ldots, n. \tag{6.46}$$

Exemplo 6.11 Consideremos novamente os dados da Tabela 6.2 aos quais um modelo de regressão linear simples foi ajustado; tanto o gráfico de dispersão apresentado na Figura 6.7 quanto o gráfico de resíduos (Figura 6.8) revelam um ponto discrepante, $(2; 61, 1)$ que afeta as estimativas dos parâmetros do modelo. O gráfico de dispersão com a reta de mínimos quadrados e com a reta resistente está disposto na Figura 6.37.

Como nesse caso, como $n = 3 \times 5$, consideramos os grupos E, C e D com 5 pontos cada. Os pontos resumo são $(x_E, y_E) = (3, 0; 22, 2)$, $(x_C, y_C) = (8, 0; 7, 4)$ e $(x_D, y_D) = (13, 0; 13, 0)$ e as correspondentes estimativas resistentes são $b_0 = -0, 92$ e $a_0^* = 21, 56$. Portanto, a reta resistente estimada ou ajustada é

$$\widetilde{v}_t = 21, 56 - 0, 92t. \tag{6.47}$$

Esta reta não é tão afetada pelo ponto discrepante (que não foi eliminado da análise).

9) **Viés da regularização Ridge**.

Fazendo $\mathbf{R} = \mathbf{X}^\top \mathbf{X}$, a expressão do estimador *ridge* (6.34), obtemos

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = (\mathbf{I} + \lambda \mathbf{R}^{-1})^{-1} \widehat{\boldsymbol{\beta}}_{\text{MQ}}, \tag{6.48}$$

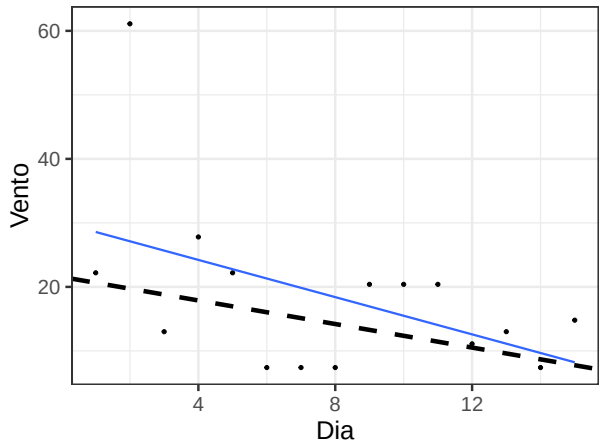


Figura 6.37: Gráfico de dispersão com retas de mínimos quadrados (linha cheia) e resistente (linha tracejada) correspondentes ao ajuste do modelo de regressão linear simples aos dados da Tabela 6.3.

em que $\widehat{\boldsymbol{\beta}}_{\text{MQ}}$ denota o estimador de mínimos quadrados ordinários e tomando a esperança condicional da expressão anterior, dada $\mathbf{X}$, obtemos

$$E(\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)) = (\mathbf{I} + \lambda \mathbf{R}^{-1})^{-1}\boldsymbol{\beta} \neq \boldsymbol{\beta}. \tag{6.49}$$

10) **Escolha do parâmetro λ**

A escolha do parâmetro de regularização λ pode ser baseada em **validação cruzada**, descrita na Seção 8.4 ou em algum critério de informação.

Pode-se provar que

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \mathbf{V}\text{diag}\left(\frac{d_1}{d_1^2+\lambda}, \frac{d_2}{d_2^2+\lambda}, \ldots, \frac{d_p}{d_p^2+\lambda}\right)\mathbf{U}^\top \mathbf{y},$$

em que $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ é a decomposição em valores singulares de $\mathbf{X}$ com $\mathbf{U}$ denotando uma matriz ortogonal de dimensão $n \times p$, $\mathbf{V}$ uma matriz ortogonal de dimensão $p \times p$ e $\mathbf{D}$ uma matriz diagonal com dimensão $p \times p$, contendo os correspondentes valores singulares $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ (raízes quadradas dos valores próprios de $\mathbf{X}^\top\mathbf{X}$.

Seja $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ uma grade de valores para λ. Podemos usar um critério de informação do tipo

$$\widehat{\lambda} = \arg\min_{\lambda \in \Lambda}[-\log \text{verossimilhança} + \text{penalização}],$$

como

$$AIC \quad = \quad \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{2}{n},$$

$$\begin{aligned} BIC &= \log[\widehat{\sigma}^2(\lambda)] + \mathrm{gl}(\lambda)\frac{\log n}{n}, \\ HQ &= \log[\widehat{\sigma}^2(\lambda)] + \mathrm{gl}(\lambda)\frac{\log\log n}{n}, \end{aligned}$$

em que $\mathrm{gl}(\lambda)$ é o número de graus de liberdade associado a λ, nomeadamente

$$\mathrm{gl}(\lambda) = \mathrm{tr}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\right] = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda},$$

e

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n - \mathrm{gl}(\lambda)}\sum_{t=1}^{n}[y_t - \widehat{\boldsymbol{\beta}}_{\mathrm{Ridge}}(\lambda)^\top\mathbf{x}_t]^2.$$

11) **Formulação geral do modelo de regressão**

O modelo de regressão múltipla (6.21) pode ser escrito na forma

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + e_i, i = 1, \ldots, n, \tag{6.50}$$

em que

$$f(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \tag{6.51}$$

com $\mathbf{x}_i = (x_{i1}, \ldots, x_{i,p})^\top$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

Esse modelo pode ser generalizado tomando

$$f(\mathbf{x}, \boldsymbol{\beta}) = \sum_{j=0}^{M-1}\beta_j\phi_j(\mathbf{x}), \tag{6.52}$$

em que $\phi_j(\cdot)$, $j = 0, \ldots, M-1$ são funções pertencentes a uma base de funções, com $\phi_0(x) = 1$. Essa formulação é útil no contexto de **redes neurais** (ver Capítulo ??).

Em notação matricial, temos

$$f(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^\top\boldsymbol{\phi}(\mathbf{x}), \tag{6.53}$$

com $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^\top$.

O caso $\phi(\mathbf{x}) = \mathbf{x}$ corresponde ao modelo de regressão linear múltipla. Outras bases comumente usadas são:

a) polinômios ($\phi_j(x) = x^j$);
b) splines;
c) gaussiana ($\phi_j(x) = \exp\{-\frac{(x-\mu_j)^2}{2s^2}\}$, com μ_j denotando parâmetros de posição e s denotando o parâmetro de dispersão;
d) sigmoide ($\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$ em que, $\sigma(a)$ pode ser qualquer uma das funções (a)–(d) da Seção 13.3.
e) Fourier [$\phi_j(x)$ é uma cossenoide (ver Morettin (2014)];
f) ondaletas [$\phi_j(x)$ é uma ondaleta (ver Morettin (2014)].

Toda a teoria de MQ assim como as técnicas de regularização podem ser aplicados a essa formulação mais geral.

6.8 Exercícios

1) Considere o modelo

$$y_i = \beta x_i + e_i, \quad i = 1, \ldots, n$$

em que $\mathrm{E}(e_i) = 0$ e $\mathrm{Var}(e_i) = \sigma^2$ são erros aleatórios não correlacionados.

   a) Obtenha o estimador de mínimos quadrados de β e proponha um estimador não enviesado para σ^2.

   b) Especifique a distribuição aproximada do estimador de β.

   c) Especifique um intervalo de confiança aproximado para o parâmetro β com coeficiente de confiança γ, $0 < \gamma < 1$.

2) Considere o modelo especificado no Exercício 1 e mostre que o parâmetro β corresponde à variação esperada para a variável Y por unidade de variação da variável X.

   Sugestão: Subtraia $\mathrm{E}(y_i|x_i)$ de $\mathrm{E}(y_i|x_i + 1)$.

3) Para investigar a associação entre tipo de escola (particular ou pública), cursada por calouros de uma universidade e a média no curso de Cálculo I, obtiveram-se os seguintes dados:

Escola	Média no curso de Cálculo I								
Particular	8,6	8,6	7,8	6,5	7,2	6,6	5,6	5,5	8,2
Pública	5,8	7,6	8,0	6,2	7,6	6,5	5,6	5,7	5,8

   Seja y_i a nota obtida pelo i-ésimo aluno, $x_i = 1$ se o aluno cursou escola particular e $x_i = -1$ se o aluno cursou escola pública, $i = 1, \ldots, 18$. Considere o modelo $y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, 18$ em que os e_i são erros aleatórios não correlacionados com $\mathrm{E}(e_i) = 0$ e $\mathrm{Var}(e_i) = \sigma^2$.

   i) Interprete os parâmetros α e β.
   ii) Estime α e β pelo método de mínimos quadrados. Obtenha também uma estimativa de σ^2.
   iii) Avalie a qualidade do ajuste do modelo por meio de técnicas de diagnóstico.
   iv) Construa intervalos de confiança para α e β.
   v) Com base nas estimativas obtidas no item ii), construa intervalos de confiança para os valores esperados das notas dos alunos das escolas particulares e públicas.
   vi) Ainda utilizando o modelo proposto, especifique e teste a hipótese de que ambos os valores esperados são iguais.
   vii) Repita os itens i)-vi) definindo $x_i = 1$ se o aluno cursou escola particular e $x_i = 0$ se o aluno cursou escola pública, $i = 1, \ldots, 18$.

4) Num estudo realizado na Faculdade de Medicina da Universidade de São Paulo foram colhidos dados de 16 pacientes submetidos a transplante intervivos e em cada um deles obtiveram-se medidas tanto do peso (g) real do lobo direito do fígado quanto de seu volume (cm^3) previsto pré operatoriamente por métodos ultrassonográficos. O objetivo é estimar o peso real por meio do volume previsto. Os dados estão dispostos na Tabela 6.13.

i) Proponha um modelo de regressão linear simples para analisar os dados e interprete seus parâmetros.

ii) Construa um gráfico de dispersão apropriado.

iii) Ajuste o modelo e utilize ferramentas de diagnóstico para avaliar a qualidade do ajuste.

iv) Construa intervalos de confiança para seus parâmetros.

v) Construa uma tabela com intervalos de confiança para o peso esperado do lobo direito do fígado correspondentes a volumes (estimados ultrassonograficamente) de 600, 700, 800, 900 e 1000 cm^3.

vi) Repita os itens anteriores considerando um modelo linear simples sem intercepto. Qual dos dois modelos você acha mais conveniente? Justifique a sua resposta.

Tabela 6.13: Peso (g) real e volume (cm^3) obtido ultrassonograficamente do lobo direito do fígado de pacientes submetidos a transplante

Volume USG (cm^3)	Peso real (g)	Volume USG (cm^3)	Peso real (g)
656	630	737	705
692	745	921	955
588	690	923	990
799	890	945	725
766	825	816	840
800	960	584	640
693	835	642	740
602	570	970	945

5) Os dados da Tabela 6.14 são provenientes de uma pesquisa cujo objetivo é propor um modelo para a relação entre a área construída de um determinado tipo de imóvel e o seu preço.

Tabela 6.14: Área (m^2) e Preço (R$) de imóveis

Imóvel	Área (m^2)	Preço (R$)
1	128	10.000
2	125	9.000
3	200	17.000
4	4.000	200.000
5	258	25.000
6	360	40.000
7	896	70.000
8	400	25.000
9	352	35.000
10	250	27.000
11	135	11.000
12	6.492	120.000
13	1.040	35.000
14	3.000	300.000

i) Construa um gráfico de dispersão apropriado para o problema.
ii) Ajuste um modelo de regressão linear simples e avalie a qualidade do ajuste (obtenha estimativas dos parâmetros e de seus erros padrões, calcule o coeficiente de determinação e construa gráficos de resíduos e um gráfico do tipo QQ).
iii) Ajuste o modelo linearizável (por meio de uma transformação logarítmica)

$$y = \beta x^{\gamma} e$$

em que y representa o preço e x representa a área e avalie a qualidade do ajuste comparativamente ao modelo linear ajustado no item ii); construa um gráfico de dispersão com os dados transformados.
iv) Utilizando o modelo com o melhor ajuste, construa intervalos de confiança com coeficiente de confiança (aproximado) de 95% para os preços esperados de imóveis com $200m^2$, $500m^2$ e $1000m^2$.

6) Os dados abaixo correspondem ao faturamento de empresas similares de um mesmo setor industrial nos últimos 15 meses.

mês	jan	fev	mar	abr	mai	jun	jul	ago
vendas	1,0	1,6	1,8	2,0	1,8	2,2	3,6	3,4

mês	set	out	nov	dez	jan	fev	mar
vendas	3,3	3,7	4,0	6,4	5,7	6,0	6,8

Utilize técnicas de análise de regressão para quantificar o crescimento do faturamento de empresas desse setor ao longo do período observado, Com essa finalidade:

a) Proponha um modelo adequado, interpretando todos os parâmetros e especificando as suposições.
b) Estime os parâmetros do modelo e apresente os resultados numa linguagem não técnica.

c) Utilize técnicas de diagnóstico para avaliar o ajuste do modelo.

7) A Tabela 6.15 contém dados obtidos de diferentes institutos de pesquisa coletados entre fevereiro de 2008 e março de 2010 e correspondem às porcentagens de eleitores favoráveis a cada um dos dois principais candidatos à presidência do Brasil.

a) Construa um diagrama de dispersão apropriado, evidenciando os pontos correspondentes a cada um dos candidatos.
b) Especifique um modelo polinomial de segundo grau, homocedástico, que represente a variação da preferência eleitoral de cada candidato ao longo do tempo.
c) Ajuste o modelo especificado no item anterior.
d) Avalie o ajuste do modelo e verifique, por meio de testes de hipóteses adequadas, se ele pode ser simplificado; em caso afirmativo, ajuste o modelo mais simples.
e) Com base no modelo escolhido, estime a porcentagem esperada de eleitores favoráveis a cada um dos candidatos em 3 de outubro de 2010 e construa um intervalo de confiança para a diferença entre essas porcentagens esperadas.
f) Faça uma crítica da análise e indique o que poderia ser feito para melhorá-la (mesmo não que não saiba implementar suas sugestões).

Tabela 6.15: Porcentagem de eleitores favoráveis

Fonte	Data	Dilma	Serra	Fonte	Data	Dilma	Serra
sensus	16/02/2008	4,5	38,2	sensus	13/08/2009	19	39,5
dataf	27/03/2008	3	38	ibope	04/09/2009	14	34
sensus	25/04/2008	6,2	36,4	sensus	14/09/2009	21,7	31,8
sensus	19/09/2008	8,4	38,1	ibope	20/11/2009	17	38
dataf	28/11/2008	8	41	vox	30/11/2009	17	39
sensus	30/11/2008	10,4	46,5	vox	07/12/2009	18	39
ibope	12/12/2008	5	42	dataf	14/12/2009	23	37
sensus	14/12/2008	13,3	42,8	vox	18/12/2009	27	34
dataf	30/01/2009	11	41	sensus	17/01/2010	27	33,2
sensus	19/03/2009	16,3	45,7	ibope	29/01/2010	25	36
dataf	27/03/2009	16	38	dataf	06/02/2010	28	32
sensus	28/05/2009	23,5	40,4	ibope	25/02/2010	30	35
ibobe	29/05/2009	18	38	dataf	27/03/2010	27	36
dataf	01/06/2009	17	36	vox	31/03/2010	31	34

8) Uma fábrica de cadeiras dispõe dos seguintes dados sobre sua produção mensal:

Número de cadeiras produzidas	105	130	141	159	160	172
Custos fixos e variáveis (R$)	1700	1850	1872	1922	1951	1970

a) Proponha um modelo de regressão linear simples para a relação entre o custo e o número de cadeiras produzidas e interprete os parâmetros.
b) Utilize um intervalo de confiança com coeficiente de confiança de 95% para estimar o custo esperado de produção para 200 cadeiras.
c) Admitindo que o preço de venda é de R$ 20,00 por unidade, qual a menor quantidade de cadeiras que deve ser produzida para que o lucro seja positivo?

9) Considere a seguinte reta de regressão ajustada a um conjunto de dados em que se pretendia estimar o volume de certos recipientes a partir de seus diâmetros: $Volumeesperado = 7,68 + 0,185Diametro$

   Podemos dizer que:

   a) O volume esperado não pode ser estimado a partir do diâmetro.
   b) O coeficiente de correlação linear entre as duas variáveis é nulo.
   c) Há um aumento médio de 0,185 unidades no volume com o aumento de uma unidade de diâmetro.
   d) O valor estimado do volume é 7,68 unidades para diâmetros iguais a 1 unidade.

10) No artigo intitulado "Estimativas do Valor Energético a partir de Características Químicas e Bromatológicas dos Alimentos" (Rev. Bras. Zootec., 30: 1837-1856, 2001) estudou-se a disponibilidade de energia de alimentos considerando os nutrientes digestíveis totais (NDT) e também as análises químicas e metabólicas das dietas. Nele se apresentam os gráficos apresentados nas Figuras 6.38 e 6.39:

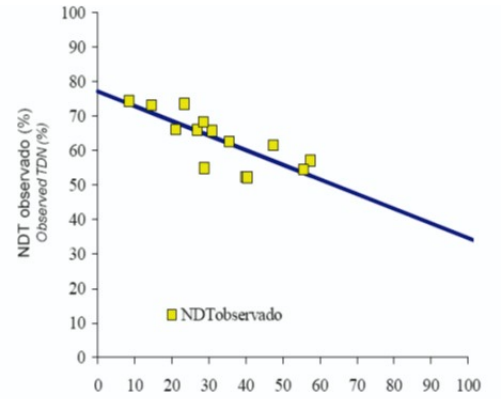


Figura 6.38: Relação entre o NDT e a fibra em detergente ácido (FDA) nas dietas totais. A linha contínua representa a reta obtida pela equação de regressão NDT=77,13-0,4250FDA (r^2=0,59 ; P<0,01).

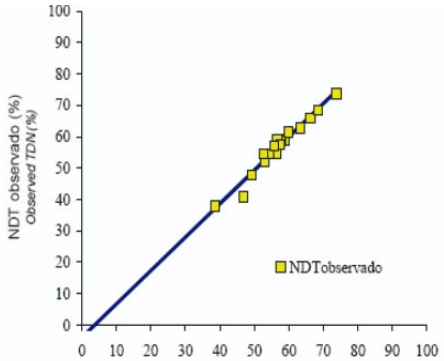


Figura 6.39: Relação entre o NDT e a disgestibilidade da matéria seca (DMS) nas dietas totais. A linha contínua representa a reta obtida pela equação de regressão NDT=3,84+1,064DMS (r^2=0,96 ; P<0,01).

a) Qual variável (FDA ou DMS) tem maior correlação linear com o NDT? Justifique.

b) Calcule o valor esperado de NDT para porcentagem da digestibilidade da matéria seca igual a 47.

11) Para avaliar o efeito da dose de uma certa droga na redução da pressão arterial (PA) o seguinte modelo de regressão foi ajustado a um conjunto de dados:

$$\text{Redução esperada da PA} = 2 + 0.3\ \text{Sexo} + 1.2\ (\text{dose} - 10)$$

em que Sexo=0 (Masculino) e Sexo = 1 (Feminino). Indique a resposta correta:

a) A redução esperada da PA (mmHg) para uma dose de 20 mg é igual para homens e mulheres.
b) Com dose de 10 mg, a redução de PA esperada para mulheres é menor do que para homens.
c) O coeficiente da variável Sexo não poderia ser igual a 0.3.
d) Uma dose de 20 mg reduz a PA esperada para homens de 12 mmHg
e) Nenhuma das anteriores.

12) O gráfico QQ da Figura 6.40 corresponde ao ajuste de um modelo de regressão linear múltipla.

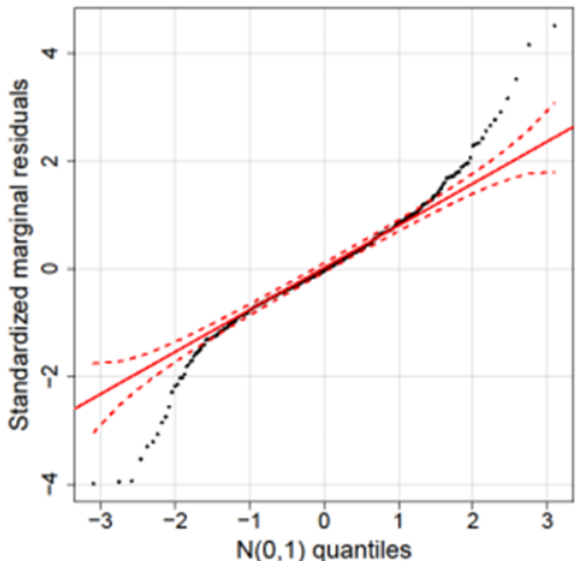


Figura 6.40: Gráfico QQ correspondente ajuste de um modelo de regressão linear múltipla.

Pode-se afirmar que:

a) Há indicações de que a distribuição dos erros é Normal.
b) Há evidências de que a distribuição dos erros é assimétrica.
c) Há evidências de que a distribuição dos erros tem caudas mais leves do que aquelas da distribuição Normal.

d) Há evidências de que a distribuição dos erros tem caudas mais pesadas que aquelas da distribuição Normal.
e) Nenhuma das anteriores.

13) Para o Exemplo 6.2, use a função `lm()` do `R` para ajustar o modelo contendo termos quadrático e cúbico.

14) Obtenha o modelo ajustado para o Exemplo 6.3, usando a função `lm()` e avalie a qualidade do ajuste utilizando todas as técnicas de diagnóstico discutidas neste capítulo.

15) Mostre que $SQTot = SQRes + SQReg$.

16) Os dados disponíveis no arquivo `profilaxia` são provenientes de um estudo realizado na Faculdade de Odontologia da Universidade de São Paulo para avaliar o efeito do uso contínuo de uma solução para bochecho no pH da placa bacteriana dentária. O pH da placa dentária retirada de 21 voluntários antes e depois de um período de uso de uma solução para bochecho foi avaliado ao longo de 60 minutos após a adição de sacarose ao meio em que as unidades experimentais foram colocadas.

a) Construa um gráfico de perfis para os dados obtidos antes do período de uso da solução para bochecho. Obtenha a matriz de covariâncias bem como o gráfico do desenhista correspondente.
b) Concretize as solicitações do item a) para os dados obtidos após a utilização da solução para bochecho.
c) Construa gráficos de perfis médios para os dados obtidos antes e depois da utilização da solução para bochecho colocando-os no mesmo painel.
d) Com base nos resultados dos itens a)-c), proponha um modelo de regressão polinomial que permita a comparação dos parâmetros correspondentes.

17) Os dados disponíveis no arquivo `esforco` são oriundos de um estudo realizado na Faculdade de Medicina da Universidade de São Paulo para avaliar pacientes com insuficiência cardíaca. Foram estudados 87 pacientes com algum nível de insuficiência cardíaca avaliada pelo critério NYHA, além de 40 pacientes controle (coluna K). Para cada paciente foram registradas algumas características físicas (altura, peso, superfície corporal, idade, sexo). Eles foram submetidos a um teste de esforço cardiopulmonar em cicloergômetro em que foram medidos a frequência cardíaca, o consumo de oxigênio, o equivalente ventilatório de oxigênio, o equivalente ventilatório de dióxido de carbono, o pulso de oxigênio e a pressão parcial de dióxido de carbono ao final da expiração, em três momentos diferentes: no limiar anaeróbio, no ponto de compensação respiratória e no pico do exercício.

Ajuste um modelo linear tendo como variável resposta o consumo de oxigênio no pico do exercício (coluna AW) e como variáveis explicativas a carga na esteira ergométrica (coluna AU), a classificação NYHA (coluna K) além de frequência cardíaca (coluna AV), razão de troca respiratória (coluna AX), peso (coluna H), sexo (coluna D) e idade (coluna F). Com essa finalidade, você deve:

a) Construir gráficos de dispersão convenientes.
b) Interpretar os diferentes parâmetros do modelo.
c) Estimar os parâmetros do modelo e apresentar os respectivos erros padrões.

d) Avaliar a qualidade de ajuste do modelo por meio de gráficos de diagnóstico (resíduos, QQ, distância de Cook, etc).

e) Identificar as variáveis significativas.

f) Reajustar o modelo com base nas conclusões do item (e) e avaliar o seu ajuste.

g) Apresentar conclusões evitando jargão técnico.

18) Para estudar a associação entre gênero (1=Masc, 0=Fem) e idade (anos) e a preferência (1=sim, 0=não) pelo refrigerante Kcola, o seguinte modelo de regressão logística foi ajustado aos dados de 50 crianças escolhidas ao acaso:

$$\log\left\{\frac{\pi_i(x_i, w_i)}{1-\pi_i(x_i, w_i)}\right\} = \alpha + \beta x_i + \gamma(w_i - 5),$$

em que x_i (w_i) representa o gênero (idade) da i-ésima criança e $\pi_i(x_i, w_i)$ a probabilidade de uma criança do gênero x_i e idade w_i preferir Kcola. As seguintes estimativas para os parâmetros foram obtidas:

Parâmetro	Estimativa	Erro padrão	Valor p
α	0,69	0,12	$< 0,01$
β	0,33	0,10	$< 0,01$
γ	-0,03	0,005	$< 0,01$

a) Interprete os parâmetros do modelo por intermédio de chances e razões de chances,

b) Com as informações acima, estime a razão de chances de preferência por Kcola correspondente à comparação de crianças do mesmo gênero com 10 e 15 anos,

c) Construa intervalos de confiança (com coeficiente de confiança aproximado de 95%) para $\exp(\beta)$ e $\exp(\gamma)$ e traduza o resultado em linguagem não técnica,

d) Estime a probabilidade de meninos com 15 anos preferirem Kcola.

19) Mostre que as expressões (6.29) e (6.30) são equivalentes e que garantem que a probabilidade de que $Y = 1$ estará no intervalo (0, 1) independentemente dos valores de α, β e x_i.

20) Mostre que o parâmetro β no modelo (6.29) corresponde ao logaritmo da razão de chances de resposta positiva para pacientes com diferença de uma unidade na variável explicativa.

21) Os dados da Tabela 6.16 contem dados de uma investigação cujo objetivo era estudar a relação entre a duração de diabete e a ocorrência de retinopatia (uma moléstia dos olhos). Ajuste um modelo log-linear para avaliar a intensidade dessa relação.

Sugestão: Considere o ponto médio de cada intervalo como valor da variável explicativa.

Tabela 6.16: Frequências de retinopatia

Duração da	Retinopatia	
Diabete (anos)	Sim	Não
0 - 2	17	215
3 - 5	26	218
6 - 8	39	137
9 - 11	27	62
12 - 14	35	36
15 - 17	37	16
18 - 20	26	13
21 - 23	23	15

22) Considere os dados dispostos na Tabela 6.17 correspondentes a preços de ações da Telebrás e do índice da Bolsa de Valores de São Paulo (IBV), de 2 de janeiro a 24 de fevereiro de 1995, extraídos do CD-mercado.

Tabela 6.17: Preços de ações da Telebrás e Índice da Bolsa de Valores de São Paulo (2/1/1995 a 24/2/1995)

Obs	Telebrás	IBV	Obs	Telebrás	IBV
1	34,99	43,19	20	30,41	38,85
2	32,09	39,68	21	31,34	39,90
3	32,56	40,37	22	30,78	38,98
4	30,31	38,27	23	31,44	39,44
5	28,91	36,28	24	30,59	38,30
6	26,10	32,70	25	28,63	36,37
7	28,25	34,99	26	27,60	35,56
8	30,41	38,41	27	26,38	34,01
9	32,00	41,04	28	25,26	33,08
10	31,25	40,56	29	24,98	32,95
11	32,37	42,10	30	24,56	31,92
12	30,87	40,79	31	23,02	30,69
13	28,63	38,09	32	20,96	28,64
14	29,56	38,62	33	22,45	30,23
15	28,44	37,58	34	21,61	29,62
16	29,28	38,40	35	19,74	27,93
17	29,84	39,27	36	20,49	28,72
18	28,35	37,84	37	23,02	32,17
19	27,32	35,81	38	23,48	32,71

a) Construa um gráfico de dispersão para os dados da Tabela 6.17 e identifique um modelo de regressão a ser ajustado aos dados interpretando os seus parâmetros.

b) Ajuste o modelo identificado no item a) por meiodo método de mínimos quadrados e avalie a qualidade de seu ajuste por meio de ferramentas

de diagnóstico.

23) Obtenha os estimadores *Ridge*, Lasso e *Elastic net* para os dados do Exemplo 6.7.

24) Idem, para os dados do Exemplo 6.11, agora com a variável resposta sendo FC (frequência cardíaca).

Capítulo 7

Análise de Sobrevivência

> All models are wrong, but some are useful.
>
> George Box

7.1 Introdução

Análise de Sobrevivência lida com situações em que o objetivo é avaliar o tempo decorrido até a ocorrência de um evento, como a morte ou cura de pacientes submetidos a um certo tratamento, a quebra de um equipamento mecânico ou o fechamento de uma conta bancária. Em Engenharia, esse tipo de problema é conhecido sob a denominação de Análise de Confiabilidade.

Nesse contexto, duas características são importantes: a definição do tempo de sobrevivência e do evento, também chamado de **falha**.[1] Nosso objetivo aqui é apresentar os principais conceitos envolvidos nesse tipo de análise. O leitor pode consultar Colosimo e Giolo (2006) ou Lee and Wang (2003) entre outros para uma exposição mais detalhada.

Exemplo 7.1: Num estudo realizado no Instituto de Ciências Biológicas (ICB) da Universidade de São Paulo, o objetivo era verificar se lesões em áreas do sistema nervoso de ratos influenciam o padrão de memória. Com essa finalidade, três grupos de ratos foram submetidos a diferentes tipos de cirurgias, a saber,

> GRUPO 1: em que lesões pequenas foram induzidas no giro denteado dorsal (região supostamente envolvida com memória espacial);
>
> GRUPO 2: em que lesões pequenas foram induzidas no giro denteado ventral;
>
> GRUPO 3: (controle) em que apenas o trauma cirúrgico (sem lesões induzidas) foi aplicado.

Após a recuperação da cirurgia, os ratos foram submetidos a um treinamento em que eram deixados em uma piscina de água turva contendo uma plataforma fixa. Se não encontrasse a plataforma em 2 minutos, o rato era conduzido até ela. Após uma semana, mediu-se o tempo até o rato encontrar a plataforma. Nesse estudo, a variável resposta é o tempo até o encontro da plataforma (evento ou falha). A origem do tempo é o instante em que o animal é colocado na piscina.

[1] Apesar dessa terminologia, falha pode ter tanto uma conotação negativa, como a morte de um paciente, quanto positiva, como a sua cura.

Um dos problemas encontrados em estudos de sobrevivência é que nem sempre o instante de ocorrência do evento e consequentemente, o tempo exato de sobrevivência são conhecidos. Essa característica é conhecida como **censura**. No entanto, sabe-se que o tempo é maior que um determinado valor chamado de **tempo de censura**. Ver Nota de Capítulo 1. No caso de estudos na área de saúde, possíveis razões para censura são

i) o evento não ocorre antes do fim do estudo;

ii) há perda de contacto com o paciente durante o estudo;

iii) o paciente sai do estudo por outros motivos (morte por outra razão/fim do tratamento devido a efeitos colaterais etc.).

No Exemplo 7.1, a censura ocorreu para os animais que não encontraram a plataforma em 2 minutos.

Por esse motivo, a variável resposta de estudos de sobrevivência é definida pelo par (T, δ) em que T é o tempo associado a cada unidade amostral e δ é um indicador de censura, com valor 1 quando ocorre a falha e 0 em caso contrário (censura). Um exemplo de organização de dados dessa natureza está apresentado na Tabela 7.1.

Tabela 7.1: Modelo para organização de dados censurados

Unidade amostral	Tempo	Censura
A	5.0	1
B	12.0	0
C	3.5	0
D	8.0	0
E	6.0	0
F	3.5	1

Um esquema indicando a estrutura de dados de sobrevivência está disposto na Figura 7.1 em que t_0 e t_c indicam, respectivamente, os instantes de início e término do estudo. Os casos com $\delta = 1$ indicam falhas e aqueles com $\delta = 0$ indicam censuras.

Para caracterizar a variável resposta (que é positiva) usualmente emprega-se a **função de sobrevivência** definida como

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)$$

em que $F(t)$ é a função distribuição acumulada da variável T. Essencialmente, a função de sobrevivência calculada no instante t é a probabilidade de sobrevivência por mais do que t. Uma representação gráfica da função de sobrevivência está apresentada na Figura 7.2.

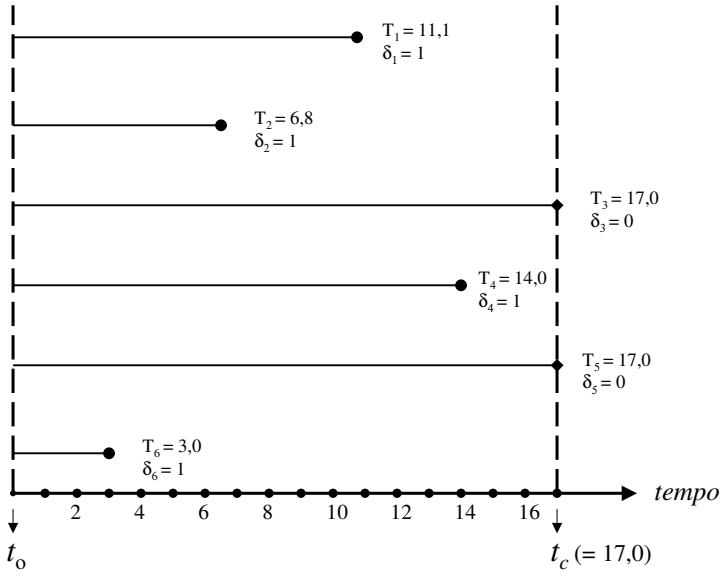


Figura 7.1: Representação esquemática de dados de sobrevivência.

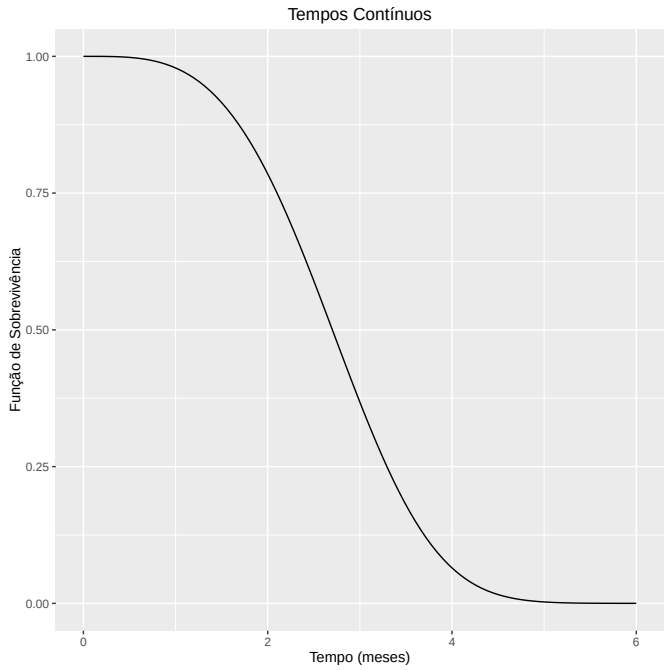


Figura 7.2: Função de sobrevivência teórica.

Na prática, como os tempos em que ocorrem falhas são medidos como variáveis discretas, a função de sobrevivência tem o aspecto indicado na Figura 7.3.

Outra função de interesse na análise de dados de sobrevivência é a **função de**

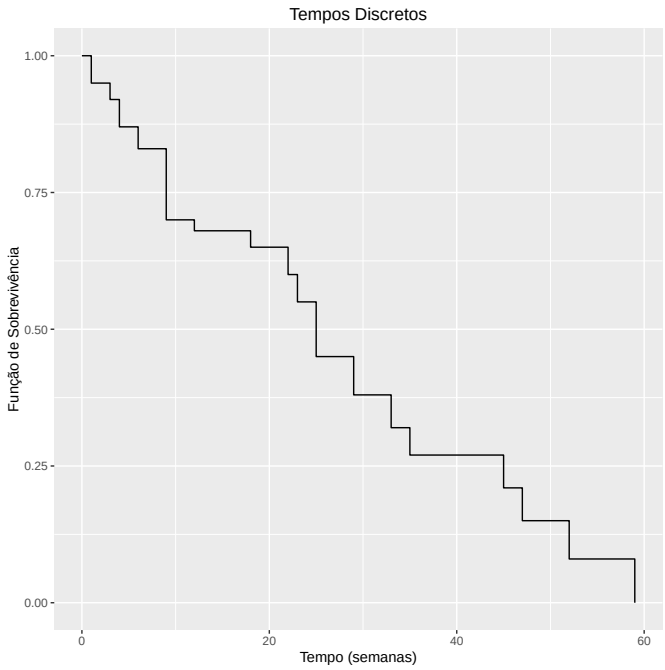


Figura 7.3: Função de sobrevivência observada.

risco (*hazard function*) definida como

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt) | T \geq t}{dt} \approx \frac{P(T = t)}{P(T \geq t)}.$$

Essencialmente, essa função corresponde ao "potencial instantâneo de ocorrência do evento de interesse por unidade de tempo, dado que a falha não ocorreu até o instante t", ou seja, "ao risco de ocorrência do evento de interesse no instante t para uma unidade amostral ainda não sujeita ao evento". Note que $h(t) \geq 0$ e não tem um valor máximo (pode ser infinito). Na prática, essa função dá uma ideia do comportamento da taxa condicional de falha e fornece informação para a escolha de um modelo probabilístico adequado ao fenômeno estudado.

Exemplos de funções de risco com diferentes padrões estão apresentados na Figura 7.4. No painel a), o risco de falha é constante e corresponde ao risco para pessoas sadias, por exemplo; nesse caso, um modelo probabilístico adequado é o **modelo exponencial**. No painel b), o risco de falha cresce com o tempo e usualmente é empregado para representar o risco para pacientes com alguma doença grave; um modelo probabilístico adequado é o **modelo Weibull**. No painel c), o risco de falha decresce com o tempo e usualmente é empregado para representar riscos pós cirúrgicos; um modelo probabilístico adequado nesse caso também é um modelo Weibull. No painel d), inicialmente o risco de falha cresce e posteriormente decresce, sendo adequado para situações em que um tratamento tem um certo tempo para fazer efeito, por exemplo; um modelo probabilístico adequado nesse caso, é o **modelo log normal**.

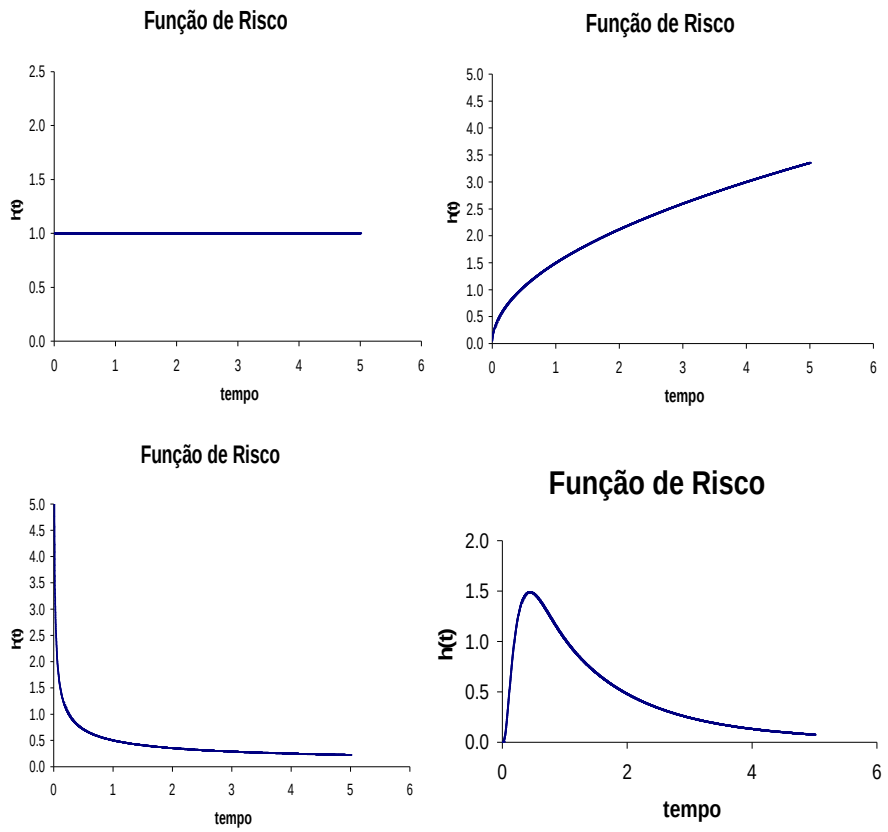


Figura 7.4: Exemplos de funções de risco

As funções de sobrevivência e de risco contêm a mesma informação e cada uma delas pode ser obtida a partir da outra por meio das relações

$$h(t) = -\frac{S'(t)}{S(t)} \text{ e } S(t) = \exp[-\int_0^t h(s)ds]$$

em que $S'(t)$ indica a derivada de S calculada no instante t.

Os objetivos operacionais da Análise de Sobrevivência são:

a) estimar e interpretar a função de sobrevivência;

b) interpretar funções de risco;

c) comparar funções de sobrevivência (ou funções de risco);

d) averiguar a contribuição de fatores de interesse (variáveis explicativas) para a ocorrência de falhas.

7.2 Estimação da função de sobrevivência

Para dados não censurados, a função distribuição empírica da variável T é

$$\overline{F}(t) = \frac{\text{número de observações} \leq t}{\text{número de observações}}$$

e consequentemente, um estimador da função de sobrevivência é $\widehat{S}(t) = 1 - \overline{F}(t)$. Para dados censurados, o **estimador de Kaplan-Meier** também conhecido como **estimador do lmite de produtos** é o mais utilizado na prática e é baseado na representação da sobrevivência num instante t como um produto de probabilidades de sobrevivência a intervalos de tempo disjuntos anteriores a t. Consideremos um exemplo em que o tempo até a cura de uma moléstia é medido em dias e que ocorreram falhas nos instantes $t = 2$, $t = 5$ e $t = 8$; a função de sobrevivência calculada no dia 10 (aqui interpretada como a probabilidade de cura após o décimo dia) pode ser calculada a partir de

$$\begin{aligned} S(10) &= P(T > 10) = P(T > 10 \cap T > 8) = P(T > 10|T > 8)P(T > 8) \\ &= P(T > 10|T > 8)P(T > 8|T > 5)P(T > 5) \\ &= P(T > 10|T > 8)P(T > 8|T > 5)P(T > 5|T > 2)P(T > 2). \end{aligned}$$

Lembrando que $t_{(0)} = t_0 = 0$ corresponde ao início do estudo, e que $S(0) = P(T > 0) = 1$, podemos generalizar esse resultado, obtendo

$$S[t_{(j)}] = \prod_{i=1}^{j} P[T > t_{(i)}|P(T > t_{(i-1)}].$$

Na prática, para a estimação da função de sobrevivência, os instantes $t_{(j)}$ de interesse são aqueles em que ocorreram falhas ou censuras. Definindo $R[t_{(i)}]$ como o número de unidades em risco no instante $t_{(i)}$ e M_i como o número de falhas ocorridas exatamente nesse instante, uma estimativa da probabilidade de que uma unidade sobreviva ao instante $t_{(i)}$ é

$$P(T > t_{(i)}) = \{R[t_{(i)}] - M_i\}/R[t_{(i)}] = 1 - M_i/R[t_{(i)}].$$

Nesse contexto, o estimador de Kaplan-Meier é definido como

$$\widehat{S}(t) = 1 \text{ se } t < t_{(1)}$$

ou

$$\widehat{S}(t) = \prod_{t_{(i)} < t} \{1 - M_i/R[t_{(i)}]\} \text{ se } t_{(i)} < t.$$

A variância desse estimador pode ser estimada como

$$\text{Var}[\widehat{S}(t)] = [\widehat{S}(t)]^2 \sum_{t_{(i)} < t} \frac{M_i}{R[t_{(i)}]\{R[t_{(i)}] - M_i\}}.$$

Exemplo 7.2: Consideremos um conjunto de $n = 21$ unidades para as quais os tempos de falhas ou censuras (representadas por +) são 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+. Para efeito do cálculo do estimador de Kaplan-Meier, convém dispor os dados no formato da Tabela 7.2.

Tabela 7.2: Formato apropriado para cálculo do estimador de Kaplan-Meier (Exemplo 7.2)

j	Tempo $t_{(j)}$	Falhas em $t_{(j)}$	Censuras em $t_{(j)}$	Unidades em risco ($R[t_{(j)}]$)	$\widehat{S}[t_{(j)}]$
0	0	0	0	21	1
1	6	3	1	21	$1 \times 18/21 = 0,86$
2	7	1	1	17 = 21-(3+1)	$0,86 \times 16/17 = 0,81$
3	10	1	2	15 = 17-(1+1)	$0,81 \times 14/15 = 0,75$
4	13	1	0	12 = 15-(1+2)	$0,75 \times 11/12 = 0,69$
5	16	1	3	11 = 12-(1+0)	$0,69 \times 10/11 = 0,63$
6	22	1	0	7 = 11-(1+3)	$0,63 \times 6/7 = 0,54$
7	23	1	5	6 = 7-(1+0)	$0,54 \times 5/6 = 0,45$

Um gráfico da função de sobrevivência estimada pelo método de Kaplan-Meier está apresentado na Figura 7.5. Os "saltos" representam as falhas e as cruzes representam as censuras. Esse gráfico pode ser obtido por meio das funções `Surv`, `survfit` e `ggsurvplot` dos pacotes `survival` e `survminer` do repositório `R`.

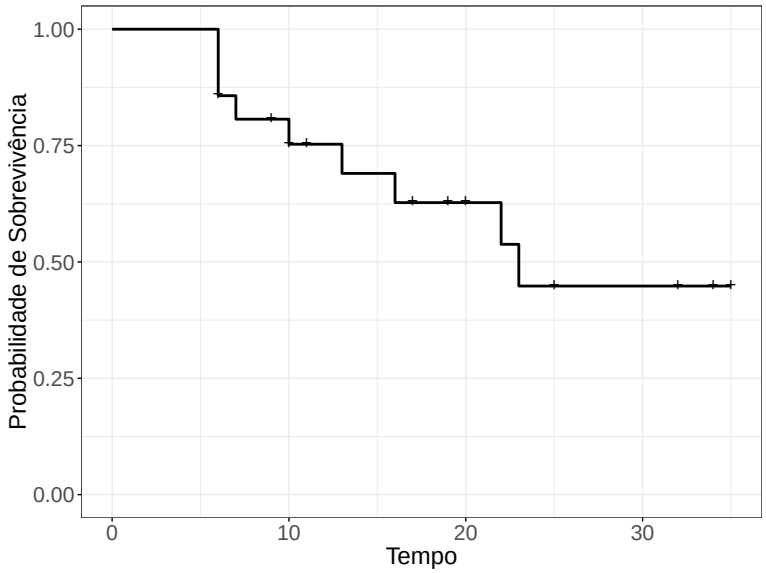


Figura 7.5: Curva de sobrevivência estimada para o Exemplo 7.2.

A área sob a curva baseada no estimador de Kaplan-Meier é um estimador do **tempo médio de acompanhamento** limitado à duração do estudo e definido como

$$\mu = \int_0^{t_c} S(t)dt.$$

Além disso, um estimador do **tempo mediano de sobrevivência** é

$$T_{med} = \{\inf t : S(t) \leq 0.5\}$$

ou seja, é o menor valor de t para o qual o valor da função de sobrevivência é menor ou igual a 0,5. De uma forma mais geral, um estimador do p-ésimo quantil $(0 < p < 1)$ do tempo de sobrevivência é $T_p = \{\inf\, t : S(t) \leq 1 - p\}$. Para o Exemplo 7.2, o tempo médio de acompanhamento é 23,29 (com erro padrão 2,83) e o tempo mediano de sobrevivência é 23,00 (com erro padrão 5,26).

Estimativas para as curvas de sobrevivência referentes ao Exemplo 7.1 estão dispostas na Figura 7.6 e estatísticas daí decorrentes, na Tabela 7.3

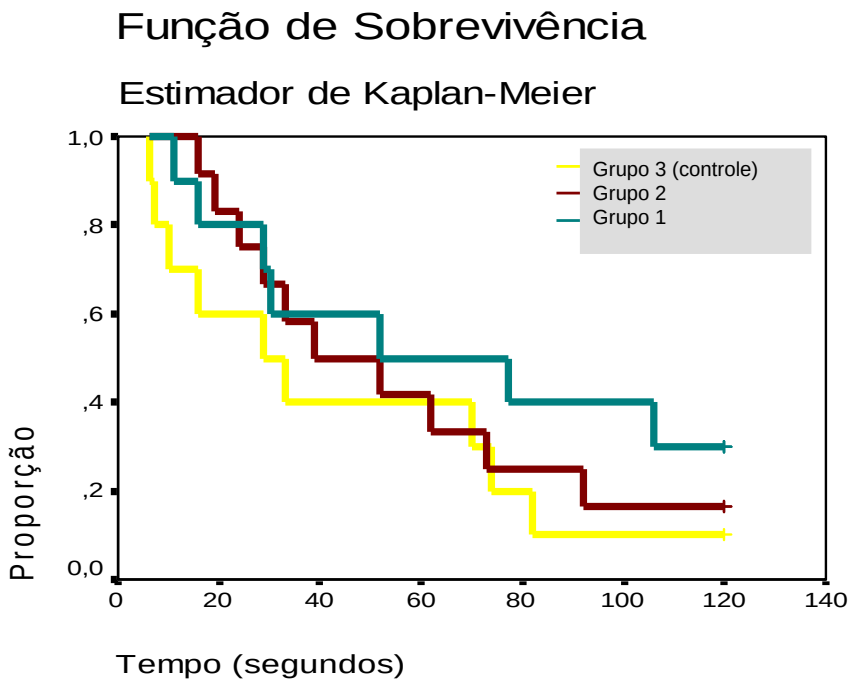


Figura 7.6: Curva de sobrevivência estimada para o Exemplo 7.1.

Tabela 7.3: Estatísticas descritivas com erros padrões entre parênteses (Exemplo 7.1)

Tratamento	Censuras	Tempo Médio	Primeiro Quartil	Tempo Mediano	Terceiro Quartil
Grupo 1	30%	68,1 (13,7)	-	52,0 (37,2)	29,0 (18,8)
Grupo 2	17%	56,6 (10,3)	73,0 (22,5)	39,0 (16,5)	24,0 (7,5)
Grupo 3	10%	44,7 (11,8)	74,0 (5,1)	29,0 (13,4)	10,0 (4,4)

Em muitos casos, os arquivos com dados de sobrevivência contêm as datas de início do estudo e ocorrência do evento de interesse ou de censura e essas datas precisam ser transformadas em intervalos de tempo.

Exemplo 7.3: Os dados disponíveis no arquivo `hiv` foram obtidos de um estudo cujo objetivo era avaliar o efeito do uso de drogas intravenosas no tempo de sobrevivência de pacientes HIV positivos e têm o formato indicado na Tabela 7.4.

Tabela 7.4: Formato dos dados correspondentes ao Exemplo 7.3

ident	datainicio	datafim	idade	droga	delta
1	15mai90	14out90	46	0	1
2	19set89	20mar90	35	1	0
3	21abr91	20dez91	30	1	1
4	03jan91	04abr91	30	1	1
⋮	⋮	⋮	⋮	⋮	⋮
98	02abr90	01abr95	29	0	0
99	01mai91	30jun91	35	1	0
100	11mai89	10jun89	34	1	1

Nesse exemplo, a variável delta =1 indica a ocorrência do evento e delta = 0, uma censura.

No exemplo, a primeira dificuldade é ler as datas no formato indicado utilizando alguma função do repositório `R`. Uma sugestão é utilizar o comando *find/replace* ou equivalente na própria planilha em que os dados estão disponíveis e substituir **jan** por **/01/**, por exemplo. Em seguida pode-se utilizar a função `as.Date` para transformar as datas no formato **dd/mm/aa** no número de dias desde 01 de janeiro de 1970, com datas anteriores assumindo valores negativos. Consequentemente, o intervalo de tempo entre as datas de início do estudo e aquela de ocorrência do evento ou de censura pode ser calculada por diferença, deixando os dados no formato indicado na Tabela 7.1. A partir daí podem-se utilizar as mesmas funções empregadas para análise do dados do Exemplo 7.2 para gerar as curvas de Kaplan-Meier dispostas na Figura 7.7.

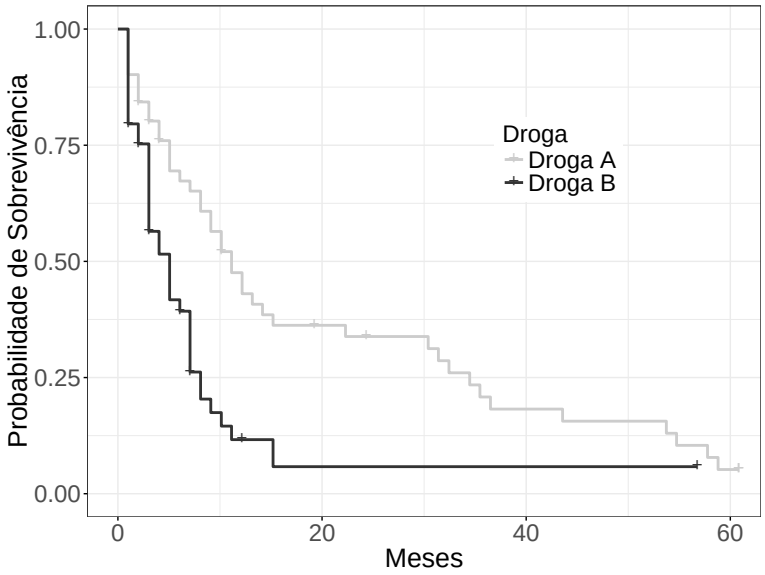


Figura 7.7: Curvas de sobrevivência estimadas para o Exemplo 7.3.

7.3 Comparação de curvas de sobrevivência

Um dos problemas oriundos de estudos como aquele descrito no Exemplo 7.1 é a comparação das curvas de sobrevivência associadas aos tratamentos. Para efeito didático, simplifiquemos o problema, restringindo-nos à comparação das curvas de sobrevivência de dois grupos. Essencialmente, queremos saber se, com base nas curvas de Kaplan-Meier, $\widehat{S}_1(t)$ e $\widehat{S}_2(t)$ obtidas de de duas amostras podemos concluir que as curvas de sobrevivência $S_1(t)$ e $S_2(t)$, associadas às populações de onde as amostras foram selecionadas, são iguais. Uma alternativa disponível para esse propósito é o teste *log rank*, baseado na comparação de valores esperados e observados.

Sejam t_j, $j = 1, \ldots, J$ os tempos em que ocorreram falhas em qualquer dos dois grupos. Para cada um desses tempos, sejam R_{1j} e R_{2j} os números de unidades em risco nos grupos 1 e 2, respectivamente e seja $R_j = R_{1j} + R_{2j}$. Similarmente, sejam O_{1j} e O_{2j}, respectivamente, os números de falhas nos grupos 1 e 2 no tempo t_j e seja $O_j = O_{1j} + O_{2j}$. Dado o número de falhas (em ambos os grupos) ocorridas no tempo t_j é O_j, a estatística O_{1j} tem uma distribuição hipergeométrica quando a hipótese de igualdade das funções de sobrevivência é verdadeira. Sob essas condições, o valor esperado e a variância de O_{1j} são, respectivamente,

$$\mathrm{E}(O_{1j}) = E_{1j} = O_{1j}\frac{O_j}{R_j} \quad \text{e} \quad \mathrm{Var}(O_{1j}) = V_j = \frac{O_j(R_{1J}/R_j)(R_j - O_j)}{R_j - 1}.$$

A estatística *log rank* de teste,

$$LR = \frac{\sum_{j=1}^{J}[O_{1j} - E_{1j}]^2}{\sum_{j=1}^{J} V_j}$$

tem uma distribuição aproximada χ^2_1 (qui quadrado com um grau de liberdade) sob a hipótese nula.

Extensões desse teste para a comparação de três ou mais curvas de sobrevivência assim como outros testes construídos para os mesmos propósitos podem ser encontrados nas referências citadas no início deste capítulo.

7.4 Regressão para dados de sobrevivência

Problemas em que o objetivo é avaliar o efeito de variáveis explicativas na distribuição do tempo de falhas (sobrevivência) são similares àqueles tratados no Capítulo 6 com a diferença de que a variável resposta (tempo) só pode assumir valores positivos. A distribuição adotada deve ser escolhida entre aquelas que têm essa característica como as distribuições exponencial, Weibull, log normal ou Birnbaum-Saunders entre outras. Modelos nessa classe são chamados **modelos paramétricos** e geralmente são expressos na forma do **modelo de tempo de falha acelerado** (*accelerated failure time models)*,

$$\log(T) = \alpha + \mathbf{x}^\top \boldsymbol{\beta} + \sigma e$$

em que α e $\boldsymbol{\beta}$ são parâmetros, $\mathbf{x}$ é um vetor com valores de variáveis explicativas, $\sigma > 0$ é uma constante conhecida e e é um erro aleatório com distribuição de forma conhecida. Com uma única variável explicativa dicotômica com valores 0 ou 1, o modelo é

$$\log(T) = \alpha + \beta x + \sigma e.$$

O tempo de falha para uma unidade com $x = 0$ é $T_0 = \exp(\alpha + \sigma e)$; para uma unidade com $x = 1$, o tempo de falha é $T_1 = \exp(\alpha + \beta + \sigma e)$. Então, se $\beta > 0$, teremos $T_1 > T_0$; por outro lado, se $\beta < 0$, teremos $T_1 < T_0$ o que implica que a covariável x **acelera** ou **desacelera** o tempo de falha. A relação entre algumas distribuições para T e $\log(T)$ está indicada na Tabela 7.5.

Tabela 7.5: Relação entre algumas distribuições para T e $\log(T)$

Distribuição de	
T	$\log(T)$
exponencial	Valores extremos
Weibul	Valores extremos
log logística	logística
log normal	normal

Esses modelos podem ser ajustados por meio do método da máxima verossimilhança. Mais detalhes podem ser obtidos nas referências citadas no início do capítulo.

Uma alternativa são os **modelos semipraramétricos** em que se destaca o **modelo de riscos proporcionais** (*proportional hazards model*) também conhecidos como **modelos de regressão de Cox** e expressos como

$$h(t|\mathbf{X} = \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x})$$

em que $h_0(t)$ representa a função de risco basal, *i.e.*, para $\mathbf{X} = \mathbf{0}$ e $\exp(\boldsymbol{\beta}^\top \mathbf{x})$ é a função de risco relativo com parâmetros $\boldsymbol{\beta}$ e cujo valor no ponto $\mathbf{x}$ corresponde ao quociente entre o risco de falha para uma unidade com variáveis explicativas iguais a $\mathbf{x}$ e o o risco de falha para uma unidade com variáveis explicativas iguais a $\mathbf{0}$.

Essa classe de modelos é uma das mais utilizadas na análise de dados de sobrevivência e tem as seguintes vantagens

i) não requer a especificação da forma da função de risco;

ii) os resultados obtidos por meio da formulação mais simples (com apenas dois grupos)são equivalentes àqueles obtidos com o teste *log rank*;

iii) permite a avaliação de várias variáveis explicativas simultaneamente.

Consideremos um estudo em que pacientes com as mesmas características são submetidos de forma aleatória a dois tratamentos: placebo ($x = 0$) e ativo ($x = 1$). Então, sob o modelo de Cox temos:

$$\frac{h(t|x = 1)}{h(t|x = 0)} = \frac{h_0(t) \exp(\alpha + \beta)}{h_0(t) \exp(\alpha)} = \exp(\beta)$$

indicando que para qualquer valor de t o risco relativo de falha é constante. Daí a denominação de riscos proporcionais. Por essa razão, o modelo de Cox só deve ser considerado nessa situação. Ver Nota de Capítulo 2. Uma ferramenta útil para avaliação dessa suposição é o gráfico das curvas de sobrevivência obtido por intermédio do estimador de Kaplan-Meier. Análise de resíduos também pode ser utilizada com esse propósito.

7.5 Notas de Capítulo

1) **Tipos de censura**

Três tipos de censura podem ser consideradas em estudos de sobrevivência:

a) **censura à direita**, para a qual se conhece o instante em que uma característica de interesse (por exemplo, contaminação pelos vírus HIV) ocorreu porém a falha (por exemplo, morte do paciente) não foi obervada após a inclusão da unidade no estudo.

b) **censura à esquerda**, para a qual não se conhece o instante de ocorrência da característica de interesse porém a falha foi observada após a inclusão da unidade no estudo.

c) **censura intervalar**, para a qual não se conhece o instante em que a falha ocorreu, mas sabe-se que ocorreu num intervalo de tempo conhecido.

2) Para situações em que os riscos não são proporcionais, algumas alternativas podem ser consideradas para o modelo de Cox, lembrando que não são isentas de dificuldades de interpretação. Entre elas, destacamos

a) Determinar instantes de tempo em que ocorre mudança no padrão da sobrevivência.

b) Ajustar modelos diferentes para intervalos de tempo distintos.

c) Refinar o modelo com a inclusão de variáveis explicativas dependentes do tempo.

7.6 Exercícios

1) Num estudo realizado no Instituto do Coração da FMUSP, candidatos a transplante foram acompanhados durante o período de espera por um coração. O tempo até o evento de interesse (aqui chamado de tempo de sobrevivência) foi definido como o número de dias decorridos entre a primeira consulta de avaliação e o procedimento cirúrgico. Para detalhes, consulte Pedroso de Lima et al. (2000). Entre possíveis fatores que poderiam influenciar o tempo até o transplante está a presença de insuficiência tricúspide. Para avaliar a importância desse fator, foram construídas curvas de sobrevivência pelo método de Kaplan-Meier e realizada uma análise baseada no modelo de riscos proporcionais de Cox, com ajuste por sexo, idade e etiologia. Os resultados estão indicados na Figura 7.8 e na Tabela 7.6.

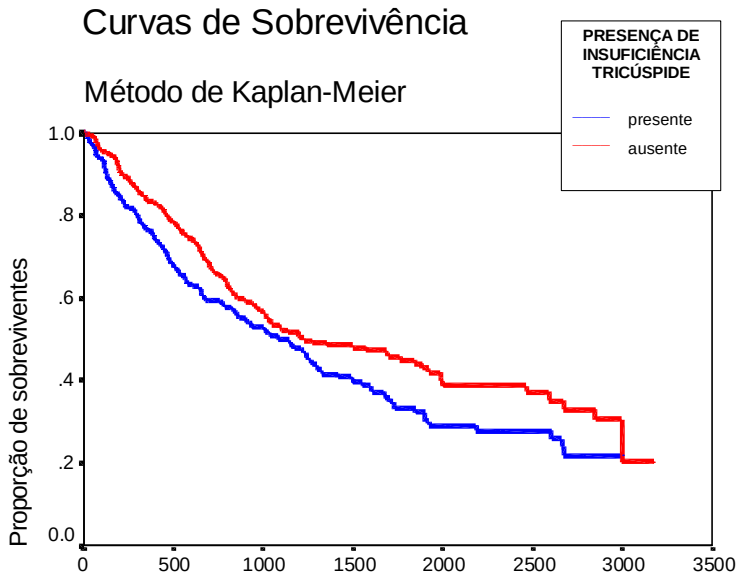


Figura 7.8: Curva de sobrevivência estimada para o estudo de transplante cardíaco

Tabela 7.6: Resultados para a variável explicativa "Insuficiência tricúspide"obtidos por meio do modelo de Cox para o estudo de transplante cardíaco.

Número		Risco	Intervalo de confiança (95%)	
de casos	Valor p	relativo	lim inferior	lim superior
868	0,039	1,25	1,01	1,54

a) Estime descritivamente a proporção de pacientes com e sem insuficiência tricúspide cujo tempo até a ocorrência do transplante é de 1500 dias.

b) Existem evidências de que a presença de insuficiência tricúspide contribui para um pior prognóstico? Justifique sua resposta.

c) Interprete o risco relativo apresentado na Tabela 7.6.

d) Qual a razão para se incluir um intervalo de confiança na análise.

2) Os dados da Tabela 7.7 foram extraídos de um estudo cuja finalidade era avaliar o efeito da contaminação de um estuário por derramamento de petróleo na fauna local. Cada um de oito grupos de 32 siris (*Calinectes danae*) foi submetido a um tratamento obtido da classificação cruzada dos níveis de dois factores, a saber, Contaminação por petróleo (sim ou não) e Salinidade de aclimatação (0.8%, 1.4%, 2.4%, 3.4%). Os animais foram observados por setenta e duas horas e o número de sobreviventes foi registado a cada 12 horas. Detalhes podem ser encontrados em Paulino e Singer (2006).

Tabela 7.7: Dados de sobrevivência de siris

		Tempo (horas)					
Grupo	Salinidade	12	24	36	48	60	72
Petróleo	0.8%	30	26	20	17	16	15
	1.4%	32	31	31	29	27	22
	2.4%	32	30	29	26	26	21
	3.4%	32	30	29	27	27	21
Controle	0.8%	31	27	25	19	18	18
	1.4%	32	31	31	31	31	30
	2.4%	32	31	31	28	27	26*
	3.4%	32	32	30	30	29*	28

$*$ = um animal foi retirado do estudo

a) Para cada um dos oito tratamentos, construa tabelas com o formato da Tabela 7.8.

Tabela 7.8: Dados de sobrevivência de siris do grupo Controle submetido à salinidade 3,4% no formato de tabela atuarial

Intervalo	Em risco	Sobre-viventes	Mortos	Retirados do estudo
0 - 12	32	32	0	0
12 - 24	32	32	0	0
24 - 36	32	30	2	0
36 - 48	30	30	0	0
48 - 60	30	29	0	1
60 -72	29	28	1	0

b) Construa curvas de sobrevivência obtidas por meio do estimador de Kaplan-Meier.

c) Utilize testes *log-rank* para avaliar o efeito da contaminação por petróleo e da salinidade na sobrevivência dos siris.

3) O arquivo `sondas` contém dados de pacientes com câncer que recebem um de dois tipos de sondas (protpla e WST) para facilitar o fluxo de fluidos do órgão. Uma possível complicação do uso dessas sondas é que após algum tempo pode ocorrer obstrução. O número de dias até a obstrução (ou censura devido ao término do estudo/óbito - informação dada na coluna evento) é apresentado na coluna rotulada "evento".

a) Construa curvas de sobrevivência para pacientes submetidos a cada um dos tipos de sonda. Coloque as curvas em um mesmo gráfico e, a partir delas, obtenha o tempo médio e o tempo mediano para obstrução em cada tipo de sonda. Comente os resultados.

b) Utilize o teste *log-rank* para comparar as duas curvas.

c) Defina dois grupos de pacientes com base na idade mediana denotando-os "jovens" e "idosos". Construa 4 estratos, formados pela combinação

dos níveis de idade e tipo de sonda e obtenha as curvas de sobrevivência correspondentes.

PARTE II: APRENDIZADO SUPERVISIONADO

A ideia fundamental do aprendizado supervisionado é utilizar preditores (dados de entrada ou *inputs*) para prever uma ou mais respostas (dados de saída ou *outputs*), que podem ser quantitativas ou qualitativas (categorias, atributos ou fatores). O caso de respostas qualitativas corresponde a problemas de **classificação** e aquele de respostas quantitativas, a problemas de **previsão**. Nos Capítulos 8, 9 e 10, consideramos métodos utilizados para a classificação de unidades de investigação em dois ou mais grupos (cujos elementos são de alguma forma parecidos entre si) com base em preditores. Por exemplo, pode-se querer classificar clientes de um banco como bons ou maus pagadores de um empréstimo com base nos salários, idades, classe social etc. Esses métodos envolvem tanto técnicas clássicas de regressão logística, função discriminante linear, método do vizinho mais próximo, quanto aqueles baseados em árvores e em algoritmos de suporte vetorial (*support vector machines*). O Capítulo 11 é dedicado a problemas de previsão, ou seja, em que se pretende prever o **valor esperado** de uma variável resposta ou o **valor específico** para uma unidade de investigação. Por exemplo, pode haver interesse em prever o saldo médio de clientes de um banco com base em salários, idades, classe social etc. A previsão pode ser concretizada seja por meio de técnicas de regressão seja por meio de métodos baseados em árvores e em algoritmos de suporte vetorial.

Uma fronteira de decisão linear (FDL) pode ser uma reta, no caso de duas variáveis, ou, em geral, um hiperplano.

Há diversas maneiras pelas quais FDL podem ser obtidas, dentre as quais destacamos:

i) Ajuste de modelos de regressão linear para as variáveis indicadoras de grupos. Esta abordagem faz parte de uma classe de métodos que modelam **funções discriminantes** $\delta_k(x)$ para cada classe e classifica x na classe com o maior valor de sua função discriminante. Uma dessas funções é a *função discriminante linear de Fisher*. Veja a Seção 8.3.

ii) Métodos que modelam a probabilidade a posteriori $P(G = k|X = x)$, em que $G(x)$ é um preditor com valores num conjunto discreto $\mathcal{G}$. Se essa probabilidade for uma função linear, obteremos uma FDL.

iii) Outro método popular usa regressão logística, estudado na no Capítulo 6.

Classificação por meio de técnicas clássicas

8.1 Introdução

De modo genérico, vamos designar a variável preditora por X (que pode ser escalar ou vetorial) e a resposta (indicadora de uma classe) por Y. Os dados serão indicados por (x_i, y_i), $i = 1, \ldots, n$. A ideia é usar os dados para obter agrupamentos cujos elementos sejam de alguma forma parecidos entre si (com base em alguma medida obtida a partir da variável preditora) e depois utilizar essa medida para classificar um ou mais novos elementos (para os quais dispomos apenas dos valores da variável preditora) em uma das classes. Esse é o conjunto de **dados para previsão**. Se tivermos d variáveis preditoras e uma resposta dicotômica (*i.e.*, duas classes), um **classificador** é uma função que mapeia um espaço d-dimensional sobre $\{-1, 1\}$.

Formalmente, seja (X, Y) um vetor aleatório, de modo que $X \in \mathbb{R}^d$ e $Y \in \{-1, 1\}$. Então, um classificador é uma função $g : \mathbb{R}^d \to \{-1, 1\}$ e a **função erro** ou **risco** é a probabilidade de erro, $L(g) = P\{g(X) \neq Y\}$.

A acurácia de um estimador de g, digamos $\widehat{g}$, pode ser medida pelo estimador de $L(g)$, chamado de **taxa de erros**, que é a proporção de erros gerados pela aplicação de $\widehat{g}$ às observações do conjunto de dados, ou seja,

$$\widehat{L}(\widehat{g}) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \widehat{y}_i), \tag{8.1}$$

com $\widehat{y}_i = \widehat{g}(x_i)$ indicando o rótulo (-1 ou 1) da classe prevista por meio de $\widehat{g}$. Se $I(y_i \neq \hat{y}_i) = 0$, a i-ésima observação estará classificada corretamente.

Sob o enfoque de aprendizado automático (AA), o objetivo é comparar diferentes modelos para identificar aquele com menor taxa de erros. Nesse contexto, dispomos de um conjunto de **dados de treinamento** (x_i, y_i), $i = 1, \ldots, n$ e de um conjunto de dados de teste, cujo elemento típico é (x_0, y_0). O interesse é minimizar a **taxa de erro de teste** associada ao conjunto de observações teste que pode ser estimada por

$$\text{Média}[I(y_0 \neq \widehat{y_0})],$$

em que a média é calculada relativamente aos elementos do conjunto de dados de teste. O classificador (ou modelo) ótimo é aquele que minimiza (8.1). Com o objetivo de classificar os elementos do conjunto de dados classificação deve-se ajustar o classificador ótimo ao conjunto de dados disponíveis (treinamento e teste)

e utilizar a estimativa $\widehat{g}$ daí obtida para classificar os elementos do conjunto de dados para classificação.

Quando dispomos de apenas um conjunto de dados, podemos recorrer ao processo de validação cruzada (ver Nota de Capítulo 2) para dividi-lo em conjuntos de dados de treinamento e de dados de teste.

Neste capítulo, concretizaremos o processo de classificação por meio de técnicas clássicas como regressão logística, o método da função discriminante linear de Fisher e o método do vizinho mais próximo com o objetivo de classificação. Outras técnicas serão consideradas nos capítulos subsequentes.

8.2 Classificação por regressão logística

Juntamente com os modelos de regressão múltipla, os modelos de **regressão logística** estudados no Capítulo 6 estão entre os mais utilizados com o objetivo de classificação. Para ilustrá-los, consideremos o seguinte exemplo.

Exemplo 8.1: Os dados da Tabela 8.1 são extraídos de um estudo realizado no Hospital Universitário da Universidade de São Paulo com o objetivo de avaliar se algumas medidas obtidas ultrassonograficamente poderiam ser utilizadas como substitutas de medidas obtidas por métodos de ressonância magnética, considerada como padrão áureo, para avaliação do deslocamento do disco da articulação temporomandibular (doravante referido simplesmente como disco). Distâncias cápsula-côndilo (em mm) com boca aberta ou fechada (referidas, respectivamente, como distância aberta ou fechada no restante do texto) foram obtidas ultrassonograficamente de 104 articulações e o disco correspondente foi classificado como deslocado (1) ou não (0) segundo a avaliação por ressonância magnética. A variável resposta é o *status* do disco (1 = deslocado ou 0 = não). Mais detalhes podem ser obtidos em Elias et al. (2006).

Com intuito didático, voltemos aos dados da Tabela 8.1 e consideremos um modelo logístico para a chance de deslocamento do disco, tendo apenas a distância aberta como variável explicativa. Nesse contexto, o modelo (6.29) corresponde a

$$\log[\theta(x_i;\alpha,\beta)]/[1-\theta(x_i;\alpha,\beta)] = \alpha + x_i\beta \tag{8.2}$$

$i = 1,\ldots,104$ em que $\theta(x_i;\alpha,\beta)$ representa a probabilidade de deslocamento do disco quando o valor da distância aberta é x_i, α denota o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0$ e β é interpretado como a variação no logaritmo da chance de deslocamento do disco por unidade de variação da distância aberta. Consequentemente, a razão de chances do deslocamento do disco correspondente a uma diferença de d unidades da distância aberta será $\exp(d\times\beta)$. Como não temos dados correspondentes a distâncias abertas menores que 0,5, convém substituir os valores x_i por valores "centrados", ou seja por $x_i^* = x_i - x_0$. Uma possível escolha para x_0 é o mínimo de x_i, que é 0,5. Essa transformação na variável explicativa altera somente a interpretação do parâmetro α que passa a ser o logaritmo o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0,5$.

Usando a função `glm()` obtemos os seguintes resultados:

```
Call:
glm(formula = deslocamento ~ (distanciaAmin), family = binomial,
```

Tabela 8.1: Dados de um estudo odontológico

Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco
2.2	1.4	0	0.9	0.8	0	1.0	0.6	0
2.4	1.2	0	1.1	0.9	0	1.6	1.3	0
2.6	2.0	0	1.4	1.1	0	4.3	2.3	1
3.5	1.8	1	1.6	0.8	0	2.1	1.0	0
1.3	1.0	0	2.1	1.3	0	1.6	0.9	0
2.8	1.1	1	1.8	0.9	0	2.3	1.2	0
1.5	1.2	0	2.4	0.9	0	2.4	1.3	0
2.6	1.1	0	2.0	2.3	0	2.0	1.1	0
1.2	0.6	0	2.0	2.3	0	1.8	1.2	0
1.7	1.5	0	2.4	2.9	0	1.4	1.9	0
1.3	1.2	0	2.7	2.4	1	1.5	1.3	0
1.2	1.0	0	1.9	2.7	1	2.2	1.2	0
4.0	2.5	1	2.4	1.3	1	1.6	2.0	0
1.2	1.0	0	2.1	0.8	1	1.5	1.1	0
3.1	1.7	1	0.8	1.3	0	1.2	0.7	0
2.6	0.6	1	0.8	2.0	1	1.5	0.8	0
1.8	0.8	0	0.5	0.6	0	1.8	1.1	0
1.2	1.0	0	1.5	0.7	0	2.3	1.6	1
1.9	1.0	0	2.9	1.6	1	1.2	0.4	0
1.2	0.9	0	1.4	1.2	0	1.0	1.1	0
1.7	0.9	1	3.2	0.5	1	2.9	2.4	1
1.2	0.8	0	1.2	1.2	0	2.5	3.3	1
3.9	3.2	1	2.1	1.6	1	1.4	1.1	0
1.7	1.1	0	1.4	1.5	1	1.5	1.3	0
1.4	1.0	0	1.5	1.4	0	0.8	2.0	0
1.6	1.3	0	1.6	1.5	0	2.0	2.1	0
1.3	0.5	0	4.9	1.2	1	3.1	2.2	1
1.7	0.7	0	1.1	1.1	0	3.1	2.1	1
2.6	1.8	1	2.0	1.3	1	1.7	1.2	0
1.5	1.5	0	1.5	2.2	0	1.6	0.5	0
1.8	1.4	0	1.7	1.0	0	1.4	1.1	0
1.2	0.9	0	1.9	1.4	0	1.6	1.0	0
1.9	1.0	0	2.5	3.1	1	2.3	1.6	1
2.3	1.0	0	1.4	1.5	0	2.2	1.8	1
1.6	1.0	0	2.5	1.8	1			

Dist aberta: distância cápsula-côndilo com boca aberta (mm)
Dist fechada: distância cápsula-côndilo com boca fechada (mm)
Desloc disco: deslocamento do disco da articulação temporomandibular (1=sim, 0=não)

```
    data = disco)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5240  -0.4893  -0.3100   0.1085   3.1360

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.8593     1.1003  -5.325 1.01e-07 ***
distanciaAmin   3.1643     0.6556   4.827 1.39e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.11  on 103  degrees of freedom
Residual deviance:  71.60  on 102  degrees of freedom
AIC: 75.6

Number of Fisher Scoring iterations: 6
```

Estimativas (com erros padrões entre parênteses) dos parâmetros desse modelo ajustado por máxima verossimilhança aos dados da Tabela 8.1, são, $\widehat{\alpha} = -5,86$ $(1,10)$ e $\widehat{\beta} = 3,16$ $(0,66)$ e então, segundo o modelo, uma estimativa da chance de deslocamento do disco para articulações com distância aberta $x = 0,5$ (que corresponde à distância aberta transformada $x^* = 0,0$ é $\exp(-5,86) = 0,003$; um intervalo de confiança (95%) para essa chance pode ser obtido exponenciando os limites (LI e LS) do intervalo para o parâmetro α, nomeadamente,

$$\begin{aligned} LI &= \exp[\widehat{\alpha} - 1,96EP(\widehat{\alpha})] = \exp(-5,86 - 1,96 \times 1,10) = 0,000 \\ LS &= \exp[\widehat{\alpha} + 1,96EP(\widehat{\alpha})] = \exp(-5,86 + 1,96 \times 1,10) = 0,025. \end{aligned}$$

Os limites de um intervalo de confiança para a razão de chances correspondentes a um variação de uma unidade no valor da distância aberta podem ser obtidos de maneira similar e são 6,55 e 85,56.

Substituindo os parâmetros α e β por suas estimativas $\widehat{\alpha}$ e $\widehat{\beta}$ em (8.2) podemos estimar a probabilidade de sucesso (deslocamento do disco, no exemplo sob investigação); por exemplo, para uma articulação cuja distância aberta seja 2,1 (correspondente à distância aberta transformada igual a 1,6), a estimativa dessa probabilidade é

$$\widehat{\theta} = \exp(-5,86 + 3,16 \times 1,6)/[1 + \exp(-5,86 + 3,16 \times 1,6)] = 0,31.$$

Lembrando que o objetivo do estudo é substituir o processo de identificação de deslocamento do disco realizado via ressonância magnética por aquele baseado na medida da distância aberta por meio de ultrassonografia, podemos estimar as probabilidades de sucesso para todas as articulações e identificar um **ponto de corte** d_0 segundo o qual, distâncias abertas com valores acima dele sugerem decidirmos pelo deslocamento do disco e distâncias abertas com valores abaixo dele sugerem a decisão oposta. Obviamente, não esperamos que todas as decisões tomadas dessa forma sejam corretas e consequentemente, a escolha do ponto de corte deve ser feita com o objetivo de minimizar os erros (decidir pelo deslocamento quando ele não existe ou *vice versa*).

Nesse contexto, um contraste entre as decisões tomadas com base em um determinado ponto de corte d_0 e o padrão áureo definido pela ressonância magnética para todas as 104 articulações pode ser resumido por meio da Tabela 8.2 em que as frequências da diagonal principal correspondem a decisões corretas e aquelas da diagonal secundária às decisões erradas.

Tabela 8.2: Frequência de decisões para um ponto de corte d_0

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	n_{11}	n_{12}
distância aberta d_0	não	n_{21}	n_{22}

O quociente $n_{11}/(n_{11}+n_{21})$ é conhecido como **sensibilidade** do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco está realmente deslocado (ver Seção 4.2 para mais detalhes). O quociente $n_{22}/(n_{12}+n_{22})$ é conhecido como **especificidade** do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco realmente não está deslocado. A situação ideal é aquela em que tanto a sensibilidade quanto a especificidade do processo de decisão são iguais a 100%.

O problema a resolver é determinar o ponto de corte d_{max} que gere o melhor equilíbrio entre sensibilidade e especificidade. Com essa finalidade, podemos construir tabelas com o mesmo formato da Tabela 8.2 para diferentes pontos de corte e um gráfico cartesiano entre a sensibilidade e especificidade obtida de cada uma delas. Esse gráfico, conhecido como **curva ROC** (do termo inglês *Receiver Operating Characteristic*) gerado para os dados da Tabela 8.1 está apresentado na Figura 8.1.

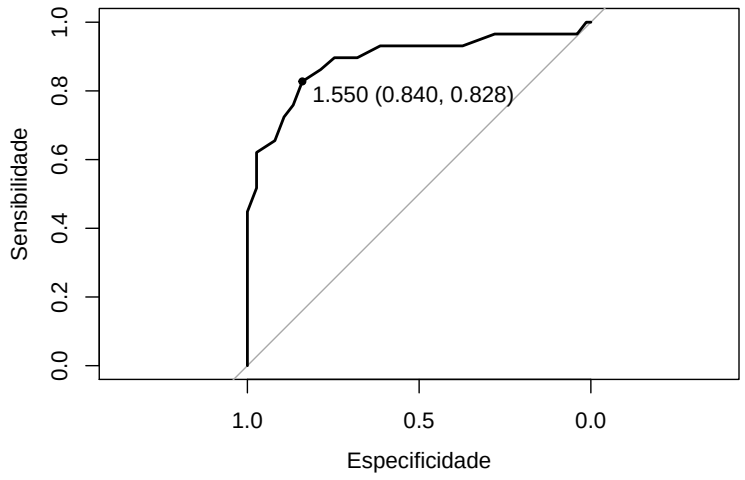


Figura 8.1: Curva ROC para os dados da Tabela 8.1 baseada no modelo (8.2) com distância aberta como variável explicativa.

O ponto de corte ótimo é aquele mais próximo do vértice superior esquerdo (em que tanto a sensibilidade quanto a especificidade seriam iguais a 100%.

Para o exemplo, esse ponto está salientado na Figura 8.1 e corresponde à distância aberta com valor $d_{max} = 2,05$ $(= 1,55 + 0,5)$. A sensibilidade e a especificidade associadas à decisão baseada nesse ponto de corte, são, respectivamente, 83% e 84% e as frequências de decisões corretas estão indicadas na Tabela 8.3. Com esse procedimento de decisão a porcentagem de acertos (**acurácia**) é 84%

Tabela 8.3: Frequência de decisões para um ponto de corte para distância aberta $d_{max} = 2,05$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	24	12
distância aberta $d_{max} = 2,05$	não	5	63

$[= (24 + 63)/104]$. A porcentagem de **falsos positivos** é 17% $[= 5/(5 + 29)]$ e a porcentagem de **falsos negativos** é 16% $[= 12/(12 + 63)]$.

Um gráfico de dispersão com o correspondente ponto de corte baseado apenas na distância aberta está apresentado na Figura 8.2 com símbolos vermelhos indicando casos com deslocamento do disco e em preto indicando casos sem deslocamento. Os valores de ambas as distâncias foram ligeiramente alterados para diminuir a superposição nos pontos.

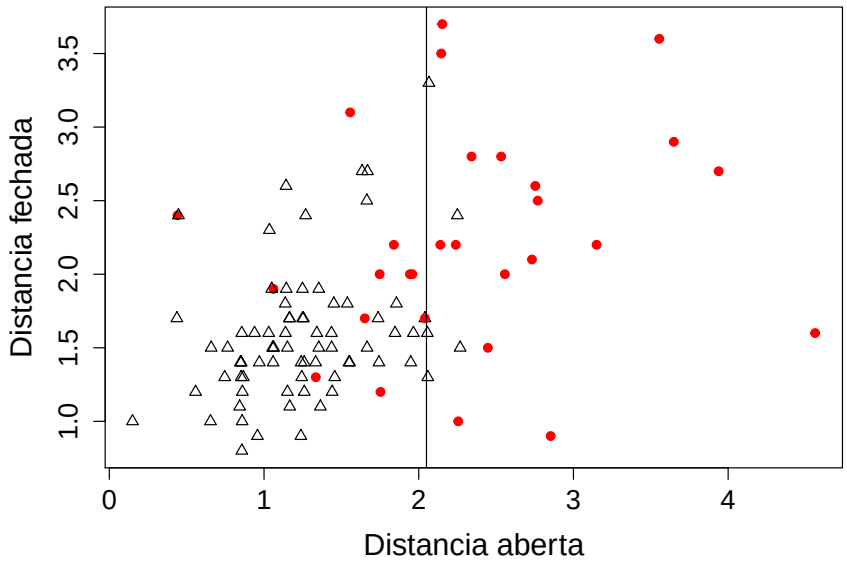


Figura 8.2: Gráfico de dispersão para os dados da Tabela 8.1 com ponto de corte baseado apenas na distância aberta.

Uma análise similar, baseada na distância fechada (transformada por meio da subtração de deu valor mínimo (0,4) gera a curva ROC apresentada na Figura 8.3

e frequências de decisões apresentada na Tabela 8.4.

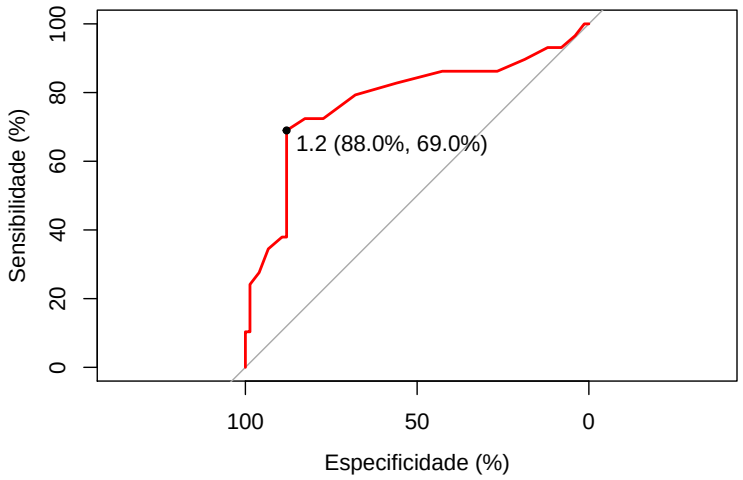


Figura 8.3: Curva ROC para os dados da Tabela 8.1 baseada no modelo (8.2) com distância fechada como variável explicativa.

Tabela 8.4: Frequência de decisões para um ponto de corte para distância fechada $d_{max} = 1,6$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	20	9
distância fechada $d_{max} = 1,6$	não	9	66

A acurácia associada a processo de decisão baseado apenas na distância fechada, 83% [= (20 + 66)/104] é praticamente igual àquela obtida com base apenas na distância aberta; no entanto aquele processo apresenta um melhor equilíbrio entre sensibilidade e especificidade (83% e 84%, respectivamente, *versus* 69% e 88%).

Se quisermos avaliar o processo de decisão com base nas observações das distâncias aberta e fechada simultaneamente, podemos considerar o modelo

$$\log[\theta(x_i;\alpha,\beta,\gamma)]/[1-\theta(x_i;\alpha,\beta,\gamma)] = \alpha + x_i\beta + w_i\gamma \qquad (8.3)$$

$i = 1,\ldots,104$ em que w_i corresponde à distância fechada observada na i-ésima articulação. Neste caso, γ corresponde à razão entre a chance de deslocamento do disco para articulações com distância fechada $w+1$ e a chance de deslocamento do disco para articulações com distância fechada w para aquelas com mesmo valor da distância aberta; uma interpretação similar vale para o parâmetro β. Estimativas dos parâmetros (com erros padrões entre parênteses) do modelo (8.3) obtidas após a transformação das variáveis explicativas segundo o mesmo figurino adotado nas análises univariadas são $\widehat{\alpha} = -6,38\ (1,19)$, $\widehat{\beta} = 2,83\ (0,67)$ e $\widehat{\gamma} = 0,98\ (0,54)$. A estimativa do parâmetro γ é apenas marginalmente significativa, ou seja a inclusão da

variável explicativa distância fechada não acrescenta muito poder de discriminação além daquele correspondente à distância aberta. Uma das razões para isso é que as duas variáveis são correlacionadas (com coeficiente de correlação de Pearson igual a 0,46. A determinação de pontos de corte para modelos com duas ou mais variáveis explicativas é bem mais complexa do que no caso univariado e não será abordada neste texto. Para efeito de comparação com as análises anteriores, as frequências de decisões obtidas com os pontos de corte utilizados naquelas estão dispostas na Tabela 8.5, e correspondem a uma sensibilidade de 62%, especificidade de 97% e acurácia de 88%.

Tabela 8.5: Frequência de decisões correspondentes a pontos de corte $d_{max} = 2,05$ para distância aberta e $d_{max} = 1,6$ para distância fechada

		Deslocamento real do disco	
		sim	não
Decisão baseada em	sim	18	2
ambas as distâncias	não	11	73

Numa segunda análise, agora sob o paradigma de aprendizado automático (AA), a escolha do modelo ótimo é baseada apenas nas porcentagens de classificação correta (acurácia) obtidas por cada modelo num conjunto de dados de teste a partir de seu ajuste a um conjunto de dados de treinamento. Como neste caso não dispomos desses conjuntos *a priori*, podemos recorrer à técnica de **validação cruzada** mencionada na Seção 1.3 e detalhada na nota de Capítulo 2. Neste exemplo, utilizamos validação cruzada de ordem 5 com 5 repetiçoes (VC5/5), em que o conjunto de dados é dividido em dois, cinco vezes, gerando cinco conjuntos de dados de treinamento e de teste. A análise é repetida cinco vezes em cada conjunto e a a acurácia média obtida das 25 análises serve de base para a escolha do melhor modelo. Comparamos quatro modelos de regressão logística, os dois primeiros com apenas uma das variáveis preditoras (distância aberta ou distância fechada), o terceiro com ambas incluídas aditivamente e o último com ambas as distâncias e sua interação. A análise pode ser concretizada por meio do pacote `caret`. Os resultados estão dispostos na Tabela 8.6 tanto para validação cruzada VC5/5 quanto para validação cruzada LOOCV.

Tabela 8.6: Acurácia obtida por validação cruzada para as regressões logísticas ajustados as dados do Exemplo 8.1

Modelo	Variáveis	Acurácia VC5/5	Acurácia LOOCV
1	Distância aberta	84,8 %	84,6 %
2	Distância fechada	75,2 %	74,0 %
3	Ambas (aditivamente)	85,7 %	85,6 %
4	Ambas + Interação	83,6 %	83,6 %

Com ambos os critérios, o melhor modelo é aquele que inclui as duas variáveis preditoras de forma aditiva. Para efeito de classificar uma nova observação (para a qual só dispomos dos valores das variáveis preditoras, o modelo selecionado deve

ser ajustado ao conjunto de dados original (treinamento + teste) para obtenção dos coeficientes do classificador. Os comandos e a saída associada ao ajuste desse modelo aos 5 conjuntos de dados gerados para validação cruzada e no conjunto completo seguem. A seleção obtida por meio do AA corresponde ao modelo (8.3). Embora a variável Distância fechada seja apenas marginalmente significativa, sua inclusão aumenta a proporção de acertos (acurácia) de 84% no modelo que inclui apenas Distância aberta para 86%. A estatísca Kappa apresentada juntamente com a acurácia serve para avaliar a concordância entre o processo de classificação e a classificação observada (veja a Seção 4.2).

```
> set.seed(369321)
> train_control =
               trainControl(method="repeatedcv", number=5, repeats=5)
> model3 = train(deslocamento ~ distanciaAmin + distanciaFmin,
           data=disco, method="glm", family=binomial,
           trControl=train_control)
> model3
Generalized Linear Model

104 samples
  2 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 5 times)
Summary of sample sizes: 83, 83, 84, 83, 83, 83, ...
Resampling results:

  Accuracy   Kappa
  0.8573333  0.6124102

> disco$predito3 = predict(model3, newdata=disco, type="raw")
> summary(model3$finalModel)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.82771  -0.45995  -0.28189   0.07403   2.82043

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.3844     1.1932  -5.351 8.76e-08 ***
distanciaAmin   2.8337     0.6676   4.245 2.19e-05 ***
distanciaFmin   0.9849     0.5383   1.830   0.0673 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.107  on 103  degrees of freedom
Residual deviance:  67.991  on 101  degrees of freedom
AIC: 73.991
Number of Fisher Scoring iterations: 6
```

```
> table(disco$deslocamento, disco$predito3)
    0  1
  0 72  3
  1 12 17
```

Como a divisão do conjunto original nos subconjuntos de treinamento e de teste envolve uma escolha aleatória, os resultados podem diferir (em geral de forma desprezável) para diferentes aplicações dos mesmos comandos, a não ser que se especifique a semente do processo aleatório de divisão por meio do comando `set.seed()`.

8.3 Função discriminante linear de Fisher

Consideremos novamente o caso de duas classes (ou populações), $\mathcal{G}_1$ e $\mathcal{G}_2$ para as quais pretendemos obter um classificador com base em um vetor de variáveis preditoras, $\mathbf{X} = (X_1, \ldots, X_p)^\top$.

A ideia de Fisher é considerar uma combinação linear $Y = \boldsymbol{\ell}^\top \mathbf{X}$, com $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_p)^\top$ de modo que o conjunto de variáveis preditoras seja transformado numa variável escalar Y. Sejam μ_{1Y} e μ_{2Y}, respectivamente, as médias de Y obtidas dos valores de $\mathbf{X}$ associadas aos dados $\mathcal{G}_1$ e $\mathcal{G}_2$. A regra para classificação consiste em selecionar a combinação linear que maximiza a distância quadrática entre essas duas médias, relativamente à variabilidade dos valores de Y.

Uma suposição adicional e, às vezes, irrealista, é que as matrizes de covariâncias

$$\boldsymbol{\Sigma}_i = \mathrm{E}(\mathbf{X} - \boldsymbol{\mu}_i)(\mathbf{X} - \boldsymbol{\mu}_i)^\top, \tag{8.4}$$

$i = 1, 2$, em que $\boldsymbol{\mu}_1 = \mathrm{E}(\mathbf{X}|\mathcal{G}_1)$ e $\boldsymbol{\mu}_2 = \mathrm{E}(\mathbf{X}|\mathcal{G}_2)$, sejam iguais para as duas classes, isto é, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Consequentemente,

$$\sigma^2_Y = \mathrm{Var}(\boldsymbol{\ell}^\top \mathbf{X}) = \boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}$$

é igual para ambas as classes. Então

$$\mu_{1Y} = \mathrm{E}(Y|\mathcal{G}_1) = \boldsymbol{\ell}^\top \boldsymbol{\mu}_1 \;\text{ e }\; \mu_{2Y} = \mathrm{E}(Y|\mathcal{G}_2) = \boldsymbol{\ell}^\top \boldsymbol{\mu}_2$$

e a razão

$$\begin{aligned} \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma^2_Y} &= \frac{(\boldsymbol{\ell}^\top \boldsymbol{\mu}_1 - \boldsymbol{\ell}^\top \boldsymbol{\mu}_2)^2}{\boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}} = \frac{\boldsymbol{\ell}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\ell}}{\boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}} \\ &= \frac{(\boldsymbol{\ell}^\top \boldsymbol{\delta})^2}{\boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}}, \end{aligned} \tag{8.5}$$

com $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ é maximizada se

$$\boldsymbol{\ell} = c\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{8.6}$$

para todo $c \neq 0$. No caso $c = 1$, obtemos a **função discriminante linear de Fisher**

$$Y = \boldsymbol{\ell}^\top \mathbf{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}. \tag{8.7}$$

e o valor máximo da razão (8.5) é $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$.

Para uma nova observação $\mathbf{x}_0$, sejam $y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_0$ e

$$\mu = \frac{\mu_{1Y} + \mu_{2Y}}{2} = \frac{1}{2}(\boldsymbol{\ell}^\top \boldsymbol{\mu}_1 + \boldsymbol{\ell}^\top \boldsymbol{\mu}_2) \tag{8.8}$$

(o ponto médio entre as médias univariadas associadas às duas classes). Em virtude de (8.7), esse ponto médio pode ser expresso como

$$\mu = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{2}, \tag{8.9}$$

Consequentemente (veja o Exercício 8.1)

$$\mathrm{E}(Y_0|\mathcal{G}_1) - \mu \geq 0 \;\text{ e }\; \mathrm{E}(Y_0|\mathcal{G}_2) - \mu < 0.$$

e uma **regra de classificação** é

Classifique $\mathbf{x}_0$ em $\mathcal{G}_1$ se $y_0 \geq \mu$,
Classifique $\mathbf{x}_0$ em $\mathcal{G}_2$ se $y_0 < \mu$.

Normalmente, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ são desconhecidas e têm que ser estimadas a partir de amostras de $\mathcal{G}_1$ e $\mathcal{G}_2$, denotadas por $\mathbf{X}_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1,n_1}]$, uma matriz com dimensão $p \times n_1$ e $\mathbf{X}_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{n_2}]$, uma matriz com dimensão $p \times n_2$. Com os dados dessas amostras, podemos obter estimativas $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}_1$ e $\mathbf{S}_2$, das médias e da matriz de covariâncias comum $\boldsymbol{\Sigma}$, para a qual um estimador não enviesado é

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \tag{8.10}$$

A função discriminante estimada é $\widehat{\boldsymbol{\ell}}^\top \mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}_p \mathbf{x}$ e a regra de classificação é:

Classifique a observação $\mathbf{x}_0$ em $\mathcal{G}_1$ se $y_0 - \widehat{\mu} \geq 0$,
Classifique a observação $\mathbf{x}_0$ em $\mathcal{G}_2$ se $y_0 - \widehat{\mu} < 0$,

em que $\widehat{\mu} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)$. Nas demais expressões, os parâmetros são substituídas pelas respectivas estimativas.

Outra suposição comumente adotada é que as variáveis preditoras têm distribuição Normal multivariada. Nesse caso, a solução encontrada por meio da função discriminante linear de Fisher é ótima. A técnica aqui abordada pode ser generalizada para três ou mais classes.

Exemplo 8.2: Consideremos os dados do arquivo `inibina`, analisados por meio de regressão logística no Exemplo 6.9. Um dos objetivos é classificar as pacientes como tendo resposta positiva ou negativa ao tratamento com inibina. A análise por meio da função discriminante linear de Fisher pode ser concretizada por meio da função `lda()` do pacote `MASS`. Os comandos e resultados da aplicação dessa função estão indicados abaixo

```
lda(inibina$resposta ~ inibina$difinib, data = inibina)
Prior probabilities of groups:
negativa positiva
 0.40625  0.59375
Group means:
         inibina$difinib
negativa        49.01385
positiva       202.70158

Coefficients of linear discriminants:
                       LD1
inibina$difinib 0.007050054
```

O resultado indica que 41% das observações correspondem a respostas negativas e 59% a respostas positivas. As médias dos grupos são as médias do preditor (diferença entre as concentrações de inibpre e inibpos) em cada classe, 49.0 e 202.7 respectivamente usadas como estimativas de μ_1 e μ_2.

O coeficiente da função discriminante corresponde à combinação linear de `difinib` = `inibpos-inibpre` usada na função de decisão de forma que se $0.00705 \times$ `difinib` for grande, o classificador linear preverá uma resposta positiva e se for pequeno, preverá uma resposta negativa.

Uma tabela relacionando a classificação predita com os valores reais da resposta e o gráfico da Figura 8.4 com os valores da função discriminante calculada para cada uma das observações podem ser obtidos por meio dos comandos

```
> predito <- predict(fisher)
> table(predito$class, inibina$resposta)

          negativa positiva
  negativa        9        2
  positiva        4       17
> ldahist(predito$x[,1], g=inibina$resposta)
```

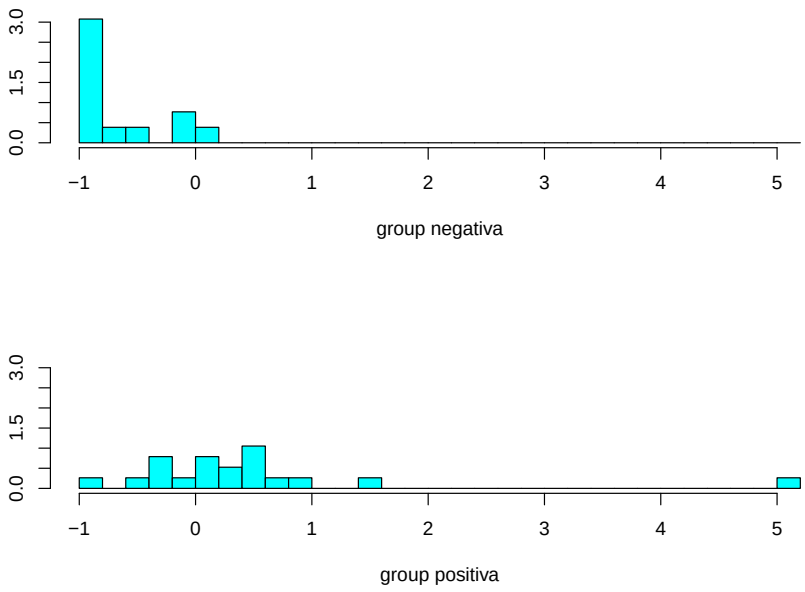


Figura 8.4: Gráfico de classificação para cada grupo

.

indicando que a probabilidade de classificação correta é 81%, ligeiramente superior ao que foi conseguido com o emprego de regressão logística (ver Exemplo 6.9).

8.4 Classificador bayesiano e do vizinho mais próximo

Pode-se mostrar que (8.5) é minimizada por um classificador que associa cada observação à classe mais provável, dados os valores dos preditores. O **classificador de Bayes** associa cada observação de teste com o valor do preditor x_0 à classe j de forma que

$$P(Y = j|X = x_0) \tag{8.11}$$

seja a maior possível. No caso de duas classes, a observação será associada à Classe 1 se $P(Y = 1|X = x_0) > 0,5$ e à Classe 2, se $P(Y = 0|X = x_0) < 0,5$. A **fronteira de Bayes** é $P(Y = 1|X = x_0) = 0,5$.

O classificador de Bayes produz a menor taxa de erro de teste possível, dada por $1-\max_j P(Y = j|X = x_0)$. A **taxa de erro de Bayes global** é $1-\mathrm{E}(\max_j P(Y = j|X))$, obtida com base na média de todas as taxas de erro sobre todos os valores possíveis de j.

Na prática como não conhecemos a distribuição condicional de Y, dado X, precisamos estimar essa probabilidade condicional, o que pode ser efetivado por meio de um método conhecido por **K-ésimo vizinho mais próximo** (*K-nearest neighbor*, KNN). O algoritmo associado a esse método é:

i) Fixe K e uma observação teste x_0;

ii) Identifique K pontos do conjunto de dados de treinamento que sejam os mais próximos de x_0 segundo alguma medida de distância; denote esse conjunto por $\mathcal{V}_0$;

iii) Estime a probabilidade condicional de que a observação teste pertença à Classe j como a fração dos pontos de $\mathcal{V}_0$ cujos valores de Y sejam iguais a j, ou seja, como

$$P(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{V}_0} I(y_i = j). \tag{8.12}$$

iv) classifique x_0 na classe asociada à maior probabilidade.

A função `knn()` do pacote `caret` pode ser utilizada com essa finalidade.

Exemplo 8.3. Consideremos, novamente, os dados do arquivo `inibina` utilizando a variável `difinib` como preditora e adotemos a estratégia de validação cruzada por meio do método LOOCV. Além disso, avaliemos o efeito de considerar entre 1 e 5 vizinhos mais próximos no processo de classificação. Os comandos necessários para a concretização da análise são

```
set.seed(2327854)
trControl <- trainControl(method  = "LOOCV")

fit <- train(resposta ~ difinib,
             method     = "knn",
             tuneGrid   = expand.grid(k = 1:5),
             trControl  = trControl,
             metric     = "Accuracy",
             data       = inibina)
fit
```

Os resultados correspondentes são

```
k-Nearest Neighbors

32 samples
 1 predictor
 2 classes: 'negativa', 'positiva'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 31, 31, 31, 31, 31, 31, ...
Resampling results across tuning parameters:

  k  Accuracy  Kappa
  1  0.71875   0.4240000
  2  0.78125   0.5409836
  3  0.81250   0.6016598
  4  0.78125   0.5409836
  5  0.81250   0.6016598

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Segundo o esse processo, o melhor resultado (com K=5 vizinhos) gera uma acurácia (média) de 81.3%. A tabela de classificação obtida por meio do ajuste do modelo final ao conjunto de dados original, juntamente com estatísticas descritivas pode ser obtida por meio dos comandos

```
predito <- predict(fit)
confusionMatrix(predito, inibina$resposta)
```

que geram os seguintes resultados

```
Confusion Matrix and Statistics
          Reference
Prediction negativa positiva
  negativa        9        1
  positiva        4       18

               Accuracy : 0.8438
                 95% CI : (0.6721, 0.9472)
    No Information Rate : 0.5938
    P-Value [Acc > NIR] : 0.002273
                  Kappa : 0.6639
 Mcnemar's Test P-Value : 0.371093

            Sensitivity : 0.6923
            Specificity : 0.9474
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.8182
             Prevalence : 0.4062
         Detection Rate : 0.2812
   Detection Prevalence : 0.3125
      Balanced Accuracy : 0.8198

       'Positive' Class : negativa
```

8.5 Notas de capítulo

1) **Validação cruzada**

Validação cruzada é a denominação atribuída a um conjunto de técnicas utilizadas para avaliar o erro de previsão de modelos estatísticos. O erro de previsão é uma medida da precisão com que um modelo pode ser usado para prever o valor de uma nova observação *i.e.*, uma observação diferente daquelas utilizados para o ajuste do modelo.

Em modelos de regressão o erro de previsão é definido como $EP = \mathrm{E}(y - \widehat{y})^2$ em que y representa uma nova observação e $\widehat{y}$ é a previsão obtida pelo modelo. O **erro quadrático médio** dos resíduos pode ser usado como uma estimativa do erro de previsão (EP), mas tende, em geral, a ser muito otimista, ou seja, a subestimar o seu verdadeiro valor. Uma razão é que os mesmos dados são utilizados para ajustar e avaliar o modelo.

No processo de validação cruzada, o modelo é ajustado a um subconjunto dos dados (**dados de treinamento**) e o resultado é empregado num outro subconjunto (**dados de teste**) ou de validação para avaliar se ele tem um bom desempenho ou não.

O seguinte algoritmo (Efron e Tibshirani, 1993), conhecido por **LOOCV** (de *Leave-One-Out Cross Validation*) bastante utilizado nesse processo é o seguinte:

1) Dadas n observações, $y_1, \ldots, y_n$, o modelo é ajustado n vezes, em cada uma delas eliminando uma observação e o valor previsto para essa observação, denotado por $\widehat{y}_{-i}$, é calculado com base no resultado obtido com as demais $n - 1$.

2) O erro de previsão é estimado por

$$VC_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{-i})^2. \tag{8.13}$$

Como alternativa para (8.13) pode-se considerar

$$VC_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_{-i}}{1 - h_i}\right)^2, \tag{8.14}$$

em que h_i é a **alavanca** (*leverage*), definida por

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}. \tag{8.15}$$

Na chamada **validação cruzada de ordem** k (*k-fold cross validation*) o conjunto de dados original é subdividido em dois, sendo um deles utilizado como conjunto de treinamento e o segundo como conjunto de teste (ou validação). Esse processo é repetido k vezes (usualmente, considera-se $k = 5$ ou $k = 10$) com conjuntos de treinamento e validação diferentes como mostra o esquema indicado na Figura 8.5 para o caso $k = 5$.

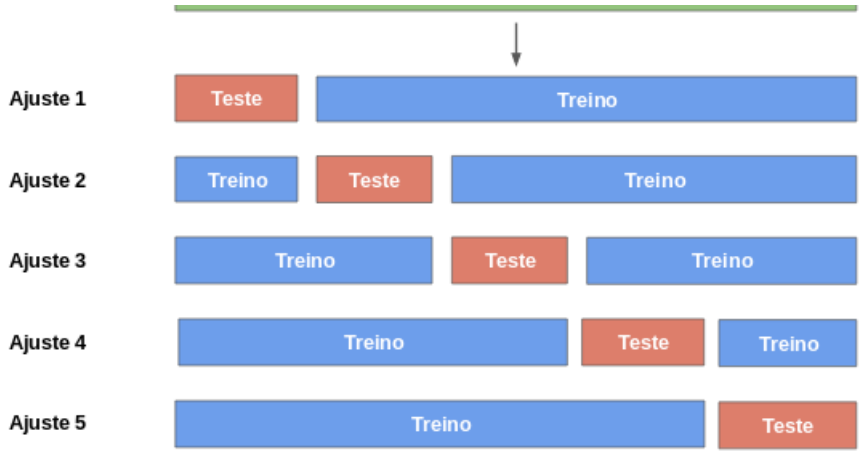


Figura 8.5: Representação esquemática da divisão dos dados para validação cruzada de ordem $k = 5$.

O correspondente erro de previsão é estimado por

$$VC_{(k)} = \frac{1}{k}\sum_{i=1}^{k} EQM_i. \tag{8.16}$$

em que o erro quadrático médio (EQM) obtido no i-ésimo ajuste, $i = 1, \ldots, k$, $EQM_i = \sum(y_j - \widehat{y}_j)^2/n_i$ com y_j, $\widehat{y}_j$ e n_i denotando, respectivamente os valores observado e predito para a j-ésima observação e o número de observações no i-ésimo conjunto de teste.

Nos casos em que o interesse é classificação, o EQM é substituído pela taxa de erros obtida quando o classificador $\widehat{g}$ obtido do ajuste do modelo aos dados de treinamento é utilizado nas observações do conjunto de dados de teste, calculada como

$$\text{Média}[I(y_0 \neq \widehat{y_0})], \tag{8.17}$$

em que y_0 denota uma observação do conjunto de dados de teste e $\widehat{y_0}$ o valor predito por meio do classificador $\widehat{g}$ e a média é calculada em relação a todos os dados desse conjunto.

8.6 Exercícios

1) Prove a relação (8.9).

2) Reanalise os dados do Exemplo 8.2 por meio da função discriminante linear de Fisher incluindo ambas as variáveis preditoras.

3) Reanalise os dados do Exemplo 8.3 considerando duas varáveis preditoras `inibpre` e `inibpos` e validação cruzada de ordem k com diferentes valores de k. Avalie o efeito da semente nos resultados.

4) Analise os dados do arquivo `endometriose2` por meio das três técnicas de classificação descritas neste capítulo, utilizando validação cruzada com diferentes parâmetros. Compare-as e discuta o resultado.

5) Use os métodos discutidos nesse capítulo (regressão logística, função discriminante linear de Fisher e KNN) para os dados do arquivo `iris`.

6) Use os métodos discutidos nesse capítulo (regressão logística, função discriminante linear de Fisher e KNN) para os dados do arquivo `cifose`.

Capítulo 9

Classificação por algoritmos de suporte vetorial

> Life is complicated, but not uninteresting.
>
> Jerzy Neyman

9.1 Introdução

Algoritmos de Suporte Vetorial (ASV), conhecidos na literatura anglo-saxônica como *Support Vector Machines* (SVM) foram introduzidas por V. Vapnik e co-autores trabalhando no AT & T Bell Laboratories e englobam técnicas úteis para classificação, com inúmeras aplicações, dentre as quais, destacamos reconhecimento de padrões, classificação de imagens, reconhecimentos de textos escritos à mão, expressão de genes em DNAs etc.[1] Em particular, Cortes and Vapnik (1995) desenvolveram essa classe de algoritmos para classificação binária.

Vapnik and Chervonesky (1964, 1974) foram, talvez, os primeiros a usar o termo **Aprendizado com Estatística** (*Statistical Learning*) em conexão com problemas de reconhecimento de padrões e inteligência artificial. Algoritmos de suporte vetorial são generalizações não lineares do algoritmo *Generalized Portrait*, desenvolvido por Vapnik e Chervonesky (1964). Um excelente tutorial sobre o tema pode ser encontrado em Smola and Schölkopf (2004). Outras referências importantes são Vapnik (1995, 1998), Hastie at al. (2017) e James et al. (2017).

Os algoritmos de suporte vetorial competem com outras técnicas bastante utilizadas, como Modelos Lineares Generalizados (MLG), Modelos Aditivos Generalizados (MAG), Redes Neurais, modelos baseados em árvores etc. A comparação com esses métodos é baseada em três fatores: interpretabilidade do modelo usado, desempenho na presença de valores atípicos e poder preditivo. Por exemplo, os MLG têm baixo desempenho na presença de valores atípicos, valor preditivo moderado e boa interpretabilidade. Por outro lado, os ASV têm desempenho moderado na presença de valores atípicos, alto poder preditivo e baixa interpretabilidade.

A abordagem de Cortes and Vapnik (1995) para o problema de classificação baseia-se nas seguintes premissas (Meyer, 2018):

[1] Embora a tradução literal do termo proposto por Vapnik seja **Máquinas** de Suporte Vetorial, optamos por utilizar **Algoritmos** de Suporte Vetorial para que o leitor não pense que algum tipo de máquina esteja ligado a essas técnicas. Aparentemente, Vapnik utilizou esse termo para enfatizar o aspecto computacional intrínseco à aplicação dos algoritmos.

a) **Separação de classes**: procura-se o melhor hiperplano separador (ver Nota de Capítulo 1) entre as classes, maximizando-se a **margem** entre os pontos mais próximos das duas classes. Os pontos sobre as fronteiras dessas classes são chamados **vetores suporte** (*support vectors*).

b) **Superposição de classes**: pontos de uma classe que estão no outro lado do hiperplano separador são ponderados com baixo peso para reduzir sua influência.

c) **Não linearidade**: quando não pudermos encontrar um separador linear, utilizamos um **kernel**[2] para mapear os dados de entrada em um espaço de dimensão mais alta [chamado de **espaço característico**, (*feature space*)] de tal forma que nesse espaço, são construídos os hiperplanos separadores. O sucesso das aplicações dos ASV depende da escolha desse *kernel*. Os *kernels* mais populares são: Gaussiano, polinomial, de base exponencial (*exponential radial basis*), *splines* e, mais recentemente, aqueles baseados em ondaletas. Veja a Nota de Capítulo 5.

d) **Solução do problema**: o problema envolve otimização quadrática e pode ser resolvido com técnicas conhecidas.

Essencialmente, um ASV é implementado por um código computacional que realiza essas tarefas. No repositório `R` há pacotes como o `e1071` e a função `svm` desenvolvidos com essa finalidade. Outras alternativas são o pacote `kernlab` e a função `ksvm`.

9.2 Fundamentação dos algoritmos de suporte vetorial

Nesta seção, apresentaremos as ideias básicas sobre algoritmos de suporte vetorial (ASV), concentrando-nos no problema de classificação dicotômica, *i.e.*, em que as unidades amostrais devem ser classificadas em uma de duas classes possíveis. Para ideias sobre o caso de mais de duas classes, veja a Nota de Capítulo 6. Adotaremos uma abordagem heurística, mais próxima daquela usualmente empregada no Aprendizado com Estatística ou Aprendizado Automático, deixando para as Notas de Capítulo 3, 4 e 5 a abordagem original (e mais formal) dos ASV.

Seja $\mathcal{X}$ o **espaço dos dados (ou dos padrões)**; em geral, $\mathcal{X} = \mathbb{R}^d$ e seja a resposta $y \in \{-1, 1\}$. Por exemplo, podemos ter dados de várias variáveis explicativas (idade, peso, taxa de colesterol etc.) e uma variável resposta (doença cardíaca, com $y = 1$ em caso afirmativo e $y = -1$ em caso negativo) observadas em vários indivíduos (o **conjunto de treinamento**). Nesse caso, o problema de classificação consiste na determinação de dois subconjuntos (classes) de $\mathcal{X}$, um dos quais estará associado a indivíduos com doença cardíaca. O classificador indicará em qual das classes deveremos incluir novos indivíduos (**o conjunto de previsão**) para os quais conhecemos os valores das variáveis explicativas. A escolha do classificador é feita com base em seu desempenho num **conjunto de dados de teste**. Quando esse conjunto não está disponível, costuma-se usar a técnica de **validação cruzada**. Ver Nota de Capítulo 2 do Capítulo 8).

Vamos considerar três situações:

[2]Optamos por manter a palavra em inglês em vez de utilizar núcleo, que é a tradução em português.

1) As classes são perfeitamente separáveis por uma fronteira linear; nesse caso, o separador (hiperplano) é conhecido como **classificador de margem máxima** (CMM). Para duas variáveis, o separador é uma reta; para três variáveis, o separador é um plano. No caso de p variáveis, o separador é um **hiperplano** de dimensão $p-1$. A Figura 9.1 é um exemplo. Note que podemos ter mais de uma reta separando as duas classes.

2) Não há um hiperplano que separe as duas classes, como no exemplo apresentado na Figura 9.2, que corresponde à Figura 9.1 com pontos trocados de lugar. O separador, neste caso é o **classificador de margem flexível** (CMF).

3) Um separador linear não conduz a resultados satisfatórios exigindo a definição de fronteiras de separação não lineares. Para isso, recorremos ou a funções não lineares das observações ou a *kernels*, para mapear o espaço dos dados em um espaço de dimensão maior. O separador, neste caso é o **classificador de margem não linear** (CMNL).

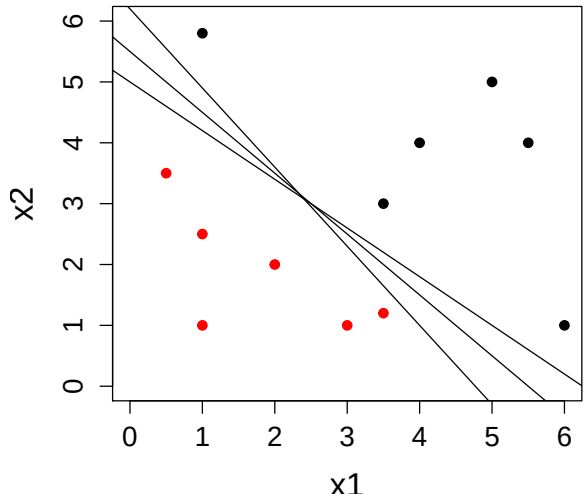


Figura 9.1: Dois conjuntos de pontos perfeitamente separáveis por um hiperplano (reta)

.

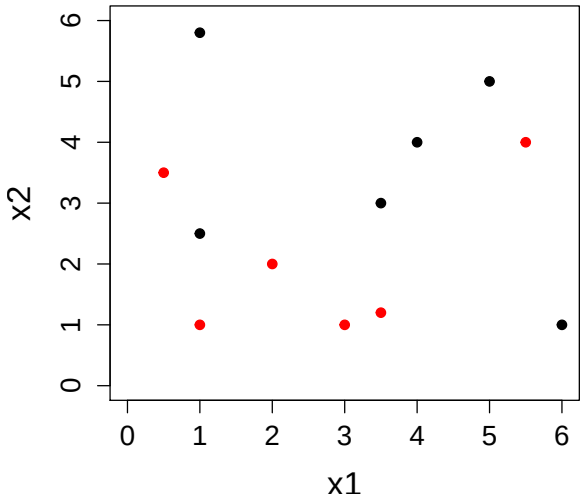


Figura 9.2: Dois conjuntos de pontos não separáveis por um hiperplano (reta)

.

9.3 Classificador de margem máxima

No caso de duas variáveis, o hiperplano é uma reta com equação $\alpha + \beta_1 X_1 + \beta_2 X_2 = 0$. Essa reta separa o plano em duas regiões, uma em que $\alpha+\beta_1 X_1+\beta_2 X_2 > 0$ e outra em que $\alpha + \beta_1 X_1 + \beta_2 X_2 < 0$.

Consideremos agora o caso com n observações das variáveis $X_1, \ldots, X_p$, dispostas na forma de uma matriz $\mathbf{X}$, de ordem $n \times p$. Seja $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, o vetor correspondente à i-ésima coluna de $\mathbf{X}$. Além disso, sejam $y_1, \ldots, y_n \in \{-1, 1\}$, variáveis respostas indicadoras de duas classes. O conjunto de dados de treinamento é $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Dado um vetor $\mathbf{x}_0 = (x_{10}, \ldots, x_{p0})^\top$ de variáveis preditoras associada a uma unidade amostral, o objetivo é classificá-la em uma das duas classes.

Queremos desenvolver um classificador usando um hiperplano separador no espaço $\mathbb{R}^p$ com base no conjunto de treinamento $\mathcal{T}$. Definindo $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, teremos

$$\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i > 0, \quad \text{se} \quad y_i = 1, \tag{9.1}$$

$$\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i < 0, \quad \text{se} \quad y_i = -1. \tag{9.2}$$

Seja

$$f(\mathbf{x}) = \alpha + \boldsymbol{\beta}^\top \mathbf{x}. \tag{9.3}$$

Então, classificaremos $\mathbf{x}_0$ a partir do sinal de $f(\mathbf{x}_0) = \alpha + \boldsymbol{\beta}^\top \mathbf{x}_0$; se o sinal for positivo, $\mathbf{x}_0$ será classificado na Classe 1 (para a qual $y = 1$, digamos), e se o sinal for negativo, na Classe 2 (para a qual $y = -1$). Em qualquer situação, $y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq 0$.

Como vimos, podem existir infinitos hiperplanos separadores, se os dados de treinamento estiverem perfeitamente separados. A sugestão de Vapnik e colaboradores é escolher um hiperplano que esteja o mais afastado das observações de treinamento, chamado de **hiperplano de margem máxima**. A **margem** é a menor distância entre o hiperplano e os pontos de treinamento.

O classificador de margem máxima (CMM) é a solução (se existir) do seguinte problema de otimização:

$$\begin{aligned} &\text{maximizar}_{(\alpha,\boldsymbol{\beta})} \;\; m(\alpha,\boldsymbol{\beta}) \\ &\quad \text{sujeito a} \\ &\sum_{i=1}^{p} \beta_i^2 = 1 \\ y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) &\geq m(\alpha,\boldsymbol{\beta}), \;\; i = 1,\ldots,n. \end{aligned} \tag{9.4}$$

Dizemos que $m = m(\alpha, \boldsymbol{\beta})$ é a **margem** do hiperplano e cada observação estará do lado correto do hiperplano se $m > 0$.

Os chamados **vetores suporte** são definidos pelos pontos cujas distâncias ao hiperplano separador sejam iguais à margem e se situam sobre as **fronteiras de separação** que são retas paralelas cujas distâncias ao hiperplano separador é igual à margem. O classificador depende desses vetores, mas não das demais observações.

A distância m do hiperplano separador a um ponto do conjunto de treinamento é

$$m = |f(\mathbf{x})|/||\boldsymbol{\beta}||,$$

em que o denominador indica a norma do vetor $\boldsymbol{\beta}$.

Como o interesse está nos pontos que são corretamente classificados, devemos ter $y_i f(\mathbf{x}_i) > 0$, $i = 1,\ldots,n$. Então

$$\frac{y_i f(\mathbf{x}_i)}{||\boldsymbol{\beta}||} = \frac{y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i)}{||\boldsymbol{\beta}||}, \tag{9.5}$$

e queremos escolher α e $\boldsymbol{\beta}$ de modo a maximizar essa distância. A margem máxima é encontrada resolvendo

$$\text{argmax}_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{||\boldsymbol{\beta}||} \text{min}_i \left[y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \right] \right\}. \tag{9.6}$$

A solução de (9.6) é complicada e sua **formulação canônica** pode ser convertida num problema mais fácil por meio do uso de multiplicadores de Lagrange. Veja a Nota de Capítulo 3 para mais detalhes sobre esse problema.

Exemplo 9.1. Consideremos os 12 pontos dispostos na Figura 9.1, sendo 6 em cada classe. Usando a função `svm` do pacote `e1071` e o comando `summary(svm.model)` obtemos o seguinte resultado

```
Call:
svm(formula = type ~ ., data = my.data, type = "C-classification",
kernel = "linear", scale = FALSE)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.5
Number of Support Vectors:  3
 ( 1 2 )
Number of Classes:  2
Levels:
 -1 1
```

Observe que a função usa o kernel linear, que corresponde ao CMM. As opções `cost` e `gamma` serão explicadas adiante. Os coeficientes do hiperplano separador, que nesse caso é uma reta, podem ser obtidos por meio dos comandos

```
alpha = svm.model$rho
beta = t(svm.model$coefs) %*% svm.model$SV
```

e são

```
> alpha
[1] 5.365853
> beta
          x1        x2
[1,] -0.8780489 -1.097561
```

A equação do hiperplano separador, disposto na Figura 9.3 é $5,366 - 0,878X_1 - 1,098X_2 = 0$. Na mesma figura, indicamos as fronteiras de separação e os vetores suporte, dados pela solução de (9.4).[3] Neste caso há três vetores suporte (indicados por círculos azuis), um na Classe 1 (ponto em vermelho) e dois na Classe 2 (pontos em preto). Os demais pontos estão em lados separados, delimitados pelas fronteiras de separação (não há pontos entre as fronteiras). A margem é $m = 0,71$.

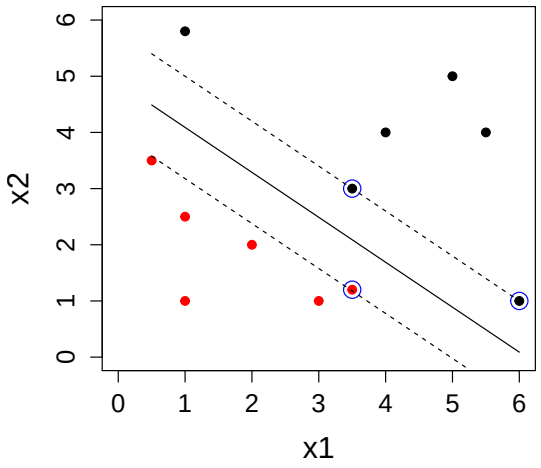


Figura 9.3: Hiperplano (reta) separador, margem, fronteiras e vetores suporte.

Consideremos agora dois novos pontos, $(\mathbf{x}_0^*, y = 1)$ e $(\mathbf{x}_1^*, y = 2)$, o primeiro na Classe 1 o segundo na Classe 2 e vamos classificá-los, usando o algoritmo. No contexto de Aprendizado com Estatística, o conjunto dos dois pontos é o chamado conjunto de dados de teste e serve para avaliar a capacidade de predição do modelo. Por meio da função `predict`, obtemos a Figura 9.4, que mostra a classificação correta de ambos os pontos (representados nas cores verde e azul).

[3]Note que os coeficientes $\beta_1 = 0,8780489$ e $\beta_2 = -1,097561$ não satisfazem a restrição indicada em (9.4), pois foram obtidos por meio da formulação canônica do problema em que a restrição é imposta ao numerador de (9.5). Para detalhes, consulte a Nota de Capítulo 3.

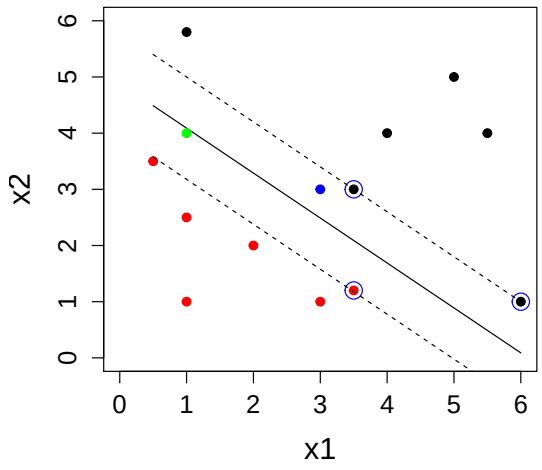


Figura 9.4: Classificação de novos pontos indicados pelas cores verde e azul.

Se o problema acima não tiver solução não existirá hiperplano separador, como é o caso apresentado na Figura 9.2. Nesse caso precisamos recorrer a um classificador que **quase** separa as duas classes. É o que veremos na próxima seção.

9.4 Classificador de margem flexível

Se não existir um hiperplano separador, como aquele do Exemplo 8.1, observações podem estar do lado errado da margem ou mesmo do hiperplano, correspondendo nesse caso a classificações erradas.

O **classificador de margem flexível** (CMF), também conhecido como **classificador baseado em suporte vetorial**[4] é escolhido de modo a classificar corretamente a maioria das observações, o que se consegue com a introdução de **variáveis de folga**, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$, no seguinte problema de otimização:

$$\begin{aligned} &\text{maximizar}_{(\alpha,\boldsymbol{\beta},\boldsymbol{\xi})} \;\; m(\alpha,\boldsymbol{\beta},\boldsymbol{\xi}) \\ &\text{sujeito a} \\ &\sum_{i=1}^{p} \beta_i^2 = 1 \\ &y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq m(\alpha,\boldsymbol{\beta},\boldsymbol{\xi})(1-\xi_i), \\ &\xi_i \geq 0, \quad \sum_{i=1}^{n} \xi_i \leq C. \end{aligned} \tag{9.7}$$

em que C é uma constante positiva. Mais detalhes sobre a constante C serão apresentados posteriormente.

As variáveis de folga permitem que observações estejam do lado errado da margem ou do hiperplano. Pontos tais que $\xi_i = 0$ são corretamente classificados e estão sobre a fronteira de separação ou do lado correto da fronteira. Pontos para

[4]Embora esse tipo de classificador seja conhecido como *support vector classifier* ou *soft margin classifier*, optamos por denominá-lo "classificador de margem flexível" para diferenciá-lo do "classificador de margem máxima", que também é baseado em vetores suporte.

os quais $0 < \xi_i \leq 1$ estão dentro da fronteira da margem, mas do lado correto do hiperplano, e pontos para os quais $\xi_i > 1$ estão do lado errado do hiperplano e serão classificados erroneamente. Veja a Figura 9.5, extraída de Bishop (2006) em que m está normalizada apropriadamente. Mais detalhes podem ser obtidos nas Notas de Capítulo 3 e 4.

Como o objetivo é maximizar a margem, podemos minimizar

$$C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||\boldsymbol{\beta}||^2, \tag{9.8}$$

em que $C > 0$ controla o equilíbrio entre a penalidade das variáveis de folga e a margem.

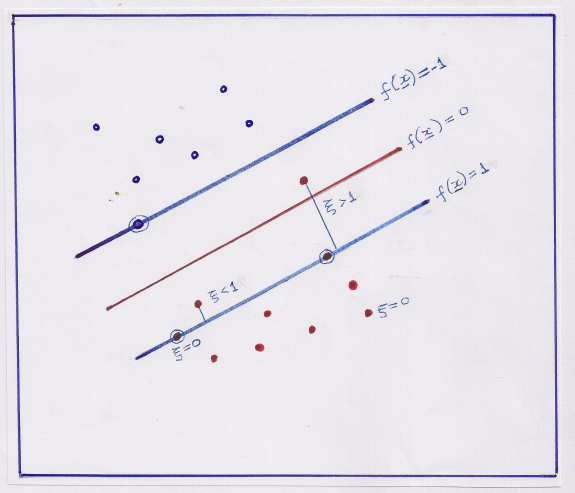


Figura 9.5: Detalhes sobre o classificador de margem flexível.

Como qualquer ponto classificado erroneamente satisfaz $\xi_i > 1$, um limite superior para o número de classificações errôneas é $\sum_{i=1}^{n} \xi_i$. No limite, quando $C \to \infty$, obtemos o CMM. O objetivo, então é minimizar(9.8) sujeito a (9.7). Veja a Nota de Capítulo 4.

A constante $C \geq 0$ deve ser escolhida apropriadamente e determina o número de violações (classificações erradas) permitidas pelo algoritmo. Se $C = 0$, então não há violações e $\xi_1 = \ldots = \xi_n = 0$. Se C aumenta, a margem fica mais larga e o contrário ocorre se C decresce. O valor de C tem a ver com a relação viés–variância: quando a constante C é pequena, o viés é pequeno e a variância é grande; se C é grande, o viés é grande e a variância é pequena. Veja o Exercício 1. Pode-se dizer que C representa o **custo** do classificador.

A constante C normalmente é escolhida por **validação cruzada** (veja a Nota de Capítulo 2 do Capítulo 8). O pacote `e1071` tem uma função, `tune()`, que realiza esse procedimento para diferentes valores de C, com o intuito de escolher o melhor modelo. Veja o Exercício 5.

Exemplo 9.1. continuação Consideremos agora os dados dispostos na Figura 9.3 em que as duas classes não são perfeitamente separáveis. Nesse caso, a utilização da função `tune()` do pacote `e1071` gera o seguinte resultado, indicando que a melhor opção é considerar $C = 0.5$ e `gamma` $= 4$.

```
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
 gamma cost
   0.5    4
- best performance: 0.5
- Detailed performance results:
  gamma cost error dispersion
1   0.5    4  0.50  0.4714045
2   1.0    4  0.60  0.4594683
3   2.0    4  0.70  0.4216370
4   0.5    8  0.65  0.4743416
5   1.0    8  0.65  0.4743416
6   2.0    8  0.70  0.4216370
7   0.5   16  0.65  0.4743416
8   1.0   16  0.65  0.4743416
9   2.0   16  0.70  0.4216370
```

Com esses parâmetros, as funções `svm` e `summary` geram o seguinte resultado, indicando que há 8 vetores suporte, 4 em cada classe.

```
svm(formula = type ~ ., data = my.data, type = "C-classification",
kernel = "linear", gamma = 0.5, cost = 4, scale = FALSE)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  4
      gamma:  0.5
Number of Support Vectors:  8
 ( 4 4 )
Number of Classes:  2
Levels:
 -1 1
```

Um gráfico indicando os vetores suporte e as regiões de classificação correspondentes está apresentado na Figura 9.6.

A equação do hiperplano classificador é $3,760 - 0,676x_1 - 0,704x_2 = 0$ ou equivalentemente, $x_2 = 3,760/0,704 - 0,676/0,704x_1 = 5,339 - 0,960x_1$. A margem correspondente é $m = (0,676^2 + 0,704^2)^{1/2} = 0,976$. Para detalhes, consulte as Notas de Capítulo 3 e 4.

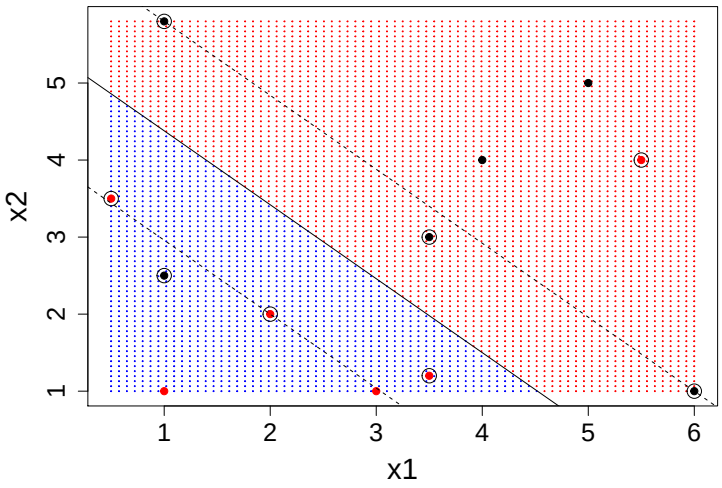


Figura 9.6: Vetores suporte para os dados da Figura 9.3

.

Com os comandos `svm.pred <- predict(svm.model, my.data)` e `table(svm.pred, ys)` podem-se obter uma tabela com as classificações certas e erradas assim como as classificações determinadas pelo algoritmo. No exemplo, há 2 classificações erradas conforme indicado na Tabela 9.1.

Tabela 9.1: Coordenadas e classificação dos pontos do Exemplo 8.1 com observações trocadas e classificação predita pelo algoritmo

observação	x1	x2	y	y predito
1	0.5	3.5	1	1
2	1.0	1.0	1	1
3	1.0	2.5	-1	**1**
4	2.0	2.0	1	1
5	3.0	1.0	1	1
6	3.5	1.2	1	1
7	1.0	5.8	-1	-1
8	3.5	3.0	-1	-1
9	4.0	4.0	-1	-1
10	5.0	5.0	-1	-1
11	5.5	4.0	1	**-1**
12	6.0	1.0	-1	-1

Exemplo 9.2. Os dados do arquivo `tipofacial` forma extraídos de um estudo odontológico realizado pelo Dr. Flávio Cotrim Vellini. Um dos objetivos era utilizar medidas entre diferentes pontos do crânio para caracterizar indivíduos com diferentes tipos faciais, a saber, braquicéfalos, mesocéfalos e dolicocéfalos. O conjunto de dados contém observações de 11 variáveis em 101 pacientes. Para efeitos

didáticos, utilizaremos apenas a altura facial e a profundidade facial como variáveis preditoras. A Figura 9.7 mostra os três grupos (correspondentes à classificação do tipo facial).

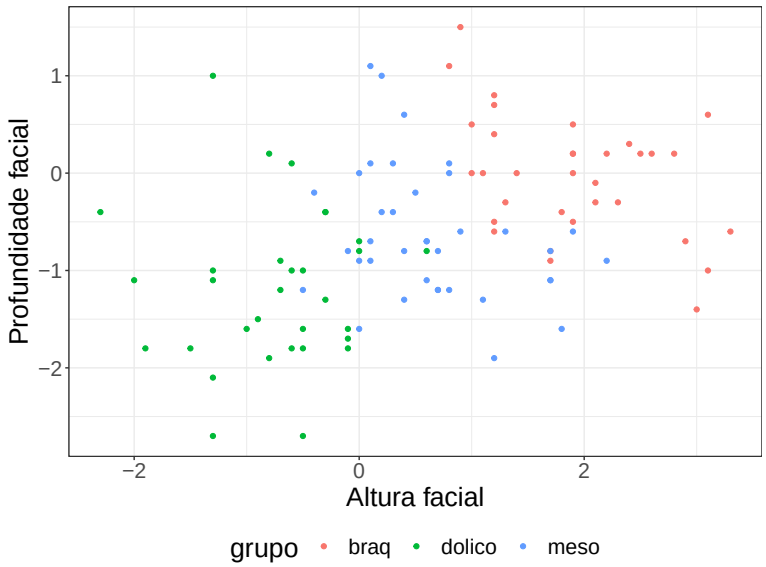


Figura 9.7: Gráfico de dispersão com identificação dos três tipos faciais.

Utilizando a função `tune.svm()` do pacote `e1071` por meio dos seguintes comandos

```
> escolhaparam <- tune.svm(grupo ~ altfac + proffac, data = face,
                   gamma = 2^(-2:2), cost = 2^2:5,
                   na.action(na.omit(c(1, NA))))
> summary(escolhaparam)
```

obtemos os resultados, apresentados abaixo, que indicam que as melhores opções para os parâmetros C e `gamma` (obtidas por meio de validação cruzada de ordem 10) para o classificador de margem flexível são $C = 4$ e `gamma`=2.

```
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
 gamma cost
     2    4
- best performance: 0.1281818
- Detailed performance results:
   gamma cost     error dispersion
1   0.25    4 0.1481818  0.1774759
2   0.50    4 0.1681818  0.1700348
3   1.00    4 0.1681818  0.1764485
4   2.00    4 0.1281818  0.1241648
5   4.00    4 0.1581818  0.1345127
6   0.25    5 0.1481818  0.1774759
7   0.50    5 0.1681818  0.1700348
8   1.00    5 0.1481818  0.1503623
```

```
9   2.00    5 0.1281818  0.1148681
10  4.00    5 0.1772727  0.1453440
```

Por intermédio da função `svm` com os parâmetros $C = 4$ e `gamma`=2 obtemos o seguinte resultado com o classificador de margem flexível:

```
svm.model <- svm(grupo ~ altfac + proffac, data = face,
                 kernel = "linear", gamma=2, cost=4)
summary(svm.model)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  4
Number of Support Vectors:  43
 ( 12 10 21 )
Number of Classes:  3
Levels:
 braq dolico meso
```

A tabela de classificação obtida com os comandos apresentados abaixo, indica o número de classificações certas e erradas.

```
svm.pred  <- predict(svm.model, face)
table(pred = svm.pred, true = face\$grupo)

        true
pred     braq dolico meso
  braq     26      0    2
  dolico    0     28    4
  meso      7      3   31
```

Na Figura 9.8 apresentamos o gráfico de classificação correspondente, obtido por meio do comando

```
plot(svm.model, face, proffac ~ altfac, svSymbol = 4, dataSymbol = 4,
     cex.lab=1.8, main="", color.palette = terrain.colors)
```

Uma das características importantes dos classificadores baseados em vetores suporte é que apenas as observações que se situam sobre a margem ou do lado errado da mesma afetam o hiperplano. Observações que se situam no lado correto da margem podem ser alteradas (mantendo suas classificações) sem que o hiperplano separador seja afetado.

Figura 9.8: Classificação do tipo facial obtida pelo classificador de margem flexível (símbolos vermelhos = dolicocéfalos, verdes = mesocéfalos, pretos = braquicéfalos).

9.5 Classificador de margem não linear

Na seção anterior apresentamos um algoritmo de classificação (CMF), usado quando as fronteiras são lineares. Para fronteiras não lineares, precisamos aumentar a dimensão do espaço de dados (ou espaço característico) por meio de outras funções, polinomiais ou não, para determinar as fronteiras de separação. No caso de duas variáveis, X_1 e X_2, por exemplo, poderíamos considerar o espaço determinado por X_1, X_2, X_1^2, X_2^3. Uma alternativa mais conveniente e mais atrativa para aumentar a dimensão do espaço característico consiste na utilização de *kernels*.

Pode-se demonstrar que um classificador linear como aquele definido em (9.7) depende somente dos vetores suporte e pode ser escrito na forma

$$f(\mathbf{x}) = \sum_{i \in S} \gamma_i < \mathbf{x}, \mathbf{x}_i > + \delta, \tag{9.9}$$

em que S indica o conjunto dos vetores suporte, os γ_i são funções de α e $\boldsymbol{\beta}$ e $< \mathbf{x}, \mathbf{y} >$ indica o produto interno dos vetores $\mathbf{x}$ e $\mathbf{y}$. Uma das vantagens de se utilizar *kernels* na construção de classificadores é que eles dependem somente dos vetores suporte e não de todas as observações o que implica uma redução considerável no custo computacional.

O classificador CMF usa um *kernel* linear, da forma

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} x_{ik} x_{jk} = \mathbf{x}_i^\top \mathbf{x}_j.$$

Se quisermos usar um CMF em um espaço característico de dimensão maior, podemos incluir polinômios de grau maior ou mesmo outras funções na definição do classificador. Os *kernels* mais utilizados na prática são:

a) lineares: $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{x}_2$;
b) polinomiais: $K(\mathbf{x}_1, \mathbf{x}_2) = (a + \mathbf{x}_1^\top \mathbf{x}_2)^d$;
c) radiais: $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma ||\mathbf{x}_1 - \mathbf{x}_2||^2\right)$, com $\gamma > 0$ constante.
d) tangentes hiperbólicas: $K(\mathbf{x}_1, \mathbf{x}_2) = \tanh(\theta + k\mathbf{x}_1^\top \mathbf{x}_2)$.

Os **classificadores CMNL** são obtidos combinando-se CMF com *kernels* não lineares, de modo a obter

$$f(\mathbf{x}) = \alpha + \sum_{i \in S} \gamma_i K(\mathbf{x}, \mathbf{x}_i) + \delta. \tag{9.10}$$

Exemplo 9.3. Consideremos uma análise alternativa para dados do Exemplo 9.2, utilizando um *kernel* polinomial, de grau 3. Os comandos e resultados reanálise dos dados por meio do classificador de margem não linear são:

```
escolhaparam <- tune.svm(grupo ~ altfac + proffac, data = face,
                         kernel = "polynomial", degree=3,
                         gamma = 2^(-1:2), cost = 2^2:6)
> summary(escolhaparam)

Parameter tuning of ‘svm’:
- sampling method: 10-fold cross validation

- best parameters:
 degree gamma cost
      3   0.5    4
- best performance: 0.1681818

- Detailed performance results:
   degree gamma cost     error dispersion
1       3   0.5    4 0.1681818 0.09440257
2       3   1.0    4 0.1772727 0.12024233
3       3   2.0    4 0.1872727 0.11722221
4       3   4.0    4 0.1872727 0.11722221
5       3   0.5    5 0.1972727 0.11314439
6       3   1.0    5 0.1772727 0.12024233
7       3   2.0    5 0.1872727 0.11722221
8       3   4.0    5 0.1872727 0.11722221
9       3   0.5    6 0.1872727 0.12634583
10      3   1.0    6 0.1772727 0.12024233
11      3   2.0    6 0.1872727 0.11722221
12      3   4.0    6 0.1872727 0.11722221

 svm.model <- svm(grupo ~ altfac + proffac, data=face,
                  type='C-classification', kernel='polynomial',
                  degree=3, gamma=1, cost=4, coef0=1, scale=FALSE)
summary(svm.model)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  4
```

```
     degree:  3
     coef.0:  1
Number of Support Vectors:  40
 ( 11 10 19 )Number of Classes:  3
Levels:
 braq dolico meso
```

A tabela de classificação é

```
        true
pred     braq dolico meso
  braq     29      0    4
  dolico    0     26    3
  meso      4      5   30
```

O gráfico correspondente está apresentado na Figura 9.9.

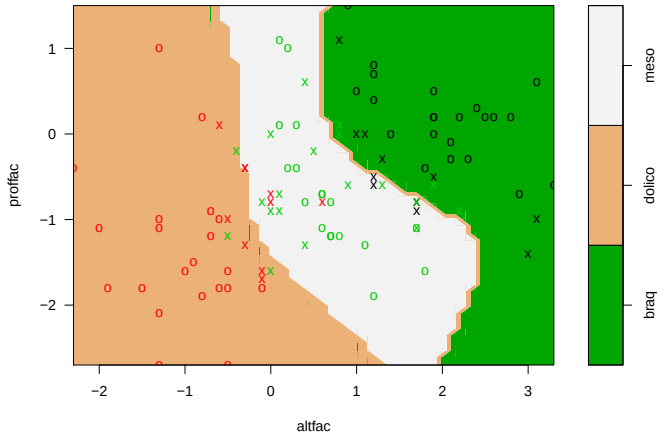


Figura 9.9: Classificação do tipo facial obtida pelo classificador de margem não linear (símbolos vermelhos = dolicocéfalos, verdes = mesocéfalos, pretos = braquicéfalos).

Neste caso, o número de classificações erradas (9) é menor do que no caso do classificador de margem flexível (16).

Com base nesses resultados, podemos classificar indivíduos para os quais dispomos apenas dos valores das variáveis preditoras. Com essa finalidade, consideremos o seguinte conjunto de previsão com 4 indivíduos:

```
  paciente altfac proffac
1      102    1.4     1.0
2      103    3.2     0.1
3      104   -2.9    -1.0
4      105    0.5     0.9
```

Por meio dos seguintes comandos

```
svm.model <- svm(grupo ~ altfac + proffac, data=face, type='C-classification',
                 kernel='polynomial', degree=3, gamma=1, cost=4, coef0=1,
                 scale=FALSE, probability=TRUE)
prednovos  <- predict(svm.model, teste, probability=TRUE)
```

obtemos a tabela com as probabilidades de classificação de cada um dos 4 indivíduos

```
     1      2      3      4
  braq   braq dolico   meso
attr(,"probabilities")
         braq       dolico         meso
1 0.954231749 0.0193863931 0.0263818582
2 0.961362058 0.0006154201 0.0380225221
3 0.008257919 0.9910764215 0.0006656599
4 0.254247666 0.1197179567 0.6260343773
Levels: braq dolico meso
```

O processo classifica os indivíduos 102 e 103 como braquicéfalos, o indivíduo 103 como dolicocéfalo e o 104, como mesocéfalo.

9.6 Notas de Capítulo

1) **Hiperplano separador**

Um hiperplano definido num espaço de dimensão p é um **subespaço** de dimensão $p-1$ definido por

$$\alpha + \beta_1 X_1 + \ldots + \beta_p X_p = 0. \tag{9.11}$$

Um ponto com coordenadas $(x_1, \ldots, x_p)$ satisfazendo (9.11) situa-se no hiperplano. Se $\alpha + \beta_1 x_1 + \ldots + \beta_p x_p > 0$, esse ponto situa-se num lado do hiperplano e se $\alpha + \beta_1 x_1 + \ldots + \beta_p x_p < 0$, o ponto situa-se no outro lado desse hiperplano. Dessa forma, o hiperplano separa o espaço p dimensional em duas metades.

2) **CMM – Classificador de margem máxima**

Nesta seção, baseada em Bishop (2006), procuramos detalhar o algoritmo utilizado para obtenção do hiperplano separador

$$f(\mathbf{x}) = \alpha + \boldsymbol{\beta}^\top \mathbf{x}. \tag{9.12}$$

Consideremos o espaço característico $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ e as respostas $y_1, \ldots, y_n$ com $y_i \in \{-1, 1\}$, definindo o conjunto de treinamento. Novos dados $\mathbf{x}_0$ são classificados de acordo com o sinal de $f(\mathbf{x}_0)$.

Suponha que exista um hiperplano separador, de modo que α e $\boldsymbol{\beta}$ são tais que $f(\mathbf{x}) > 0$, para pontos com $y = +1$ e $f(\mathbf{x}) < 0$, para pontos com $y = -1$, de modo que $yf(\mathbf{x}) > 0$, para qualquer dado de treinamento.

Podem existir muitos hiperplanos que separam as classes exatamente, como na Figura 9.1. O CMM tem como objetivo maximizar a margem que é a menor distância entre o hiperplano e qualquer ponto do conjunto de treinamento.

Para entender o procedimento de otimização, considere a distância de um ponto $\mathbf{x}$ ao hiperplano cuja equação é $f(\mathbf{x}) = 0$, nomeadamente

$$d = |f(\mathbf{x})|/||\boldsymbol{\beta}||,$$

em que denominador indica a norma do vetor $\boldsymbol{\beta}$. Como o interesse está nos pontos que são corretamente classificados, devemos ter $y_i f(\mathbf{x}_i) > 0$, $i = 1, \ldots, n$. Logo, a distância entre qualquer ponto $\mathbf{x}_i$ e o hiperplano é

$$\frac{y_i f(\mathbf{x}_i)}{||\boldsymbol{\beta}||} = \frac{y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i)}{||\boldsymbol{\beta}||}. \tag{9.13}$$

A margem é a distância do hiperplano ao ponto $\mathbf{x}$ mais próximo e queremos escolher α e $\boldsymbol{\beta}$ de modo a maximizar essa distância. A margem máxima é obtida por meio da resolução de

$$\text{argmax}_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{||\boldsymbol{\beta}||} \min \left[y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \right] \right\}. \tag{9.14}$$

A solução de (9.14) é complicada mas é possível obtê-la por meio da utilização de **multiplicadores de Lagrange**. Note que se multiplicarmos α e $\boldsymbol{\beta}$ por uma constante, a distância de um ponto $\mathbf{x}$ ao hiperplano separador não se altera (veja o Exercício 1). Logo podemos considerar a transformação $\alpha^* = \alpha/f(\mathbf{x})$ e $\boldsymbol{\beta}^* = \boldsymbol{\beta}/f(\mathbf{x})$ e para o ponto mais próximo do hiperplano, digamos $\mathbf{x}^*$, obtendo

$$y^*(\alpha + \boldsymbol{\beta}^\top \mathbf{x}^*) = 1, \tag{9.15}$$

e consequentemente, $d = ||\boldsymbol{\beta}||^{-1}$. Desse modo, todos os pontos do conjunto de treinamento satisfarão

$$y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq 1, \quad i = 1, \ldots, n. \tag{9.16}$$

Esta relação é chamada **representação canônica do hiperplano separador**. Dizemos que há uma **restrição ativa** para os pontos em que há igualdade; para os pontos em que vale a desigualdade, dizemos que há uma **restrição inativa**. Como sempre haverá um ponto que está mais próximo do hiperplano, sempre haverá uma restrição ativa.

Então, o problema de otimização implica maximizar $||\boldsymbol{\beta}||^{-1}$, que é equivalente a minimizar $||\boldsymbol{\beta}||^2$. Na linguagem de Vapnik (1995), isso equivale a escolher $f(\mathbf{x})$ de maneira que seja a mais achatada (*flat*) possível, que por sua vez implica que $\boldsymbol{\beta}$ deve ser pequeno. Isso corresponde à resolução do problema de **programação quadrática**

$$\text{argmin}_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\boldsymbol{\beta}||^2 \right\}, \tag{9.17}$$

sujeito a (9.16). O fator $1/2$ é introduzido por conveniência.

Com esse objetivo, para cada restrição em (9.16), introduzimos os multiplicadores de Lagrange $\lambda_i \geq 0$, obtendo a função lagrangeana

$$L(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} - \sum_{i=1}^{n} \lambda_i [y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) - 1], \tag{9.18}$$

em que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$. O sinal negativo no segundo termo de (9.18) justifica-se por que queremos minimizar em relação a α e $\boldsymbol{\beta}$ e maximizar em relação a $\boldsymbol{\lambda}$.

Derivando L em relação a $\boldsymbol{\beta}$ e a $\boldsymbol{\lambda}$, obtemos

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i \quad \text{e} \quad \sum_{i=1}^{n} \lambda_i y_i = 0. \tag{9.19}$$

Eliminando α e $\boldsymbol{\beta}$ em (9.18) e usando (9.19), obtemos a chamada **representação dual** do problema da margem máxima, no qual maximizamos

$$\tilde{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{9.20}$$

com respeito a $\boldsymbol{\lambda}$, sujeito às restrições

$$\begin{aligned} \lambda_i &\geq 0, \quad i = 1, \ldots, n, \qquad (9.21) \\ \sum_{i=1}^{b} \lambda_i y_i &= 0. \qquad (9.22) \end{aligned}$$

Em (9.20), $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ é um *kernel* linear, que será estendido para algum *kernel* mais geral com a finalidade de ser aplicado a espaços característicos cuja dimensionalidade excede o número de dados como indicado na Seção 9.2.3,. Esse *kernel* deve ser positivo definido.

Para classificar um novo dado $\mathbf{x}_0$ usando o modelo treinado, avaliamos o sinal de $f(\mathbf{x}_0)$, que por meio de (9.19), pode ser escrito como

$$f(\mathbf{x}_0) = \alpha + \sum_{i=1}^{n} \lambda_i y_i K(\mathbf{x}_0, \mathbf{x}_i). \qquad (9.23)$$

Pode-se demonstrar (veja Bishop, 2006), que esse tipo de otimização restrita satisfaz certas condições, chamadas de **condições de Karush-Kuhn-Tucker** (KKT), que implicam

$$\begin{aligned} \lambda_i &\geq 0, \\ y_i f(\mathbf{x}_i) - 1 &\geq 0, \\ \lambda_i (y_i f(\mathbf{x}_i) - 1) &= 0. \end{aligned} \qquad (9.24)$$

Para cada ponto, ou $\lambda_i = 0$ ou $y_i f(\mathbf{x}_i) = 1$. Um ponto para o qual $\lambda_i = 0$ não aparece em (9.23) não tem influência na classificação de novos pontos.

Os pontos restantes são chamados **vetores suporte** e satisfazem $y_i f(\mathbf{x}_i) = 1$; logo esses pontos estão sobre as fronteiras do espaço separador, como na Figura 9.3. O valor de α pode ser encontrado a partir de

$$y_i \left(\sum_{j \in S} \lambda_j y_j K(\mathbf{x}_j \mathbf{x}_j) + \alpha \right) = 1, \qquad (9.25)$$

em que S é o conjunto dos vetores suporte. Multiplicando essa expressão por y_1, observando que $y_1^2 = 1$ e tomando a média de todas as equações sobre S, obtemos

$$\alpha = \frac{1}{n_S} \sum_{i \in S} \left(y_i - \sum_{j \in S} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}_j) \right), \qquad (9.26)$$

em que n_S é o número de vetores suporte.

3) **CMF – Classificador de margem flexível**

Vamos considerar agora, o caso em que as duas classes podem se sobrepor. Precisamos modificar o CMM para permitir que alguns pontos do conjunto de treinamento sejam classificados erroneamente. Para isso introduzimos uma penalidade, que cresce com a distância ao hiperplano separador. Isso é conseguido pela introdução de **variáveis de folga** (*slack*) $\xi_i \geq 0, i = 1, \ldots, n$, uma para cada dado.

Então, $\xi_i = 0$ para pontos sobre ou dentro da fronteira correta [delimitada por $f(\mathbf{x}) = -1$ e $f(\mathbf{x}) = 1$] e ξ_i dado pela distância do ponto à fronteira, para os outros pontos. Assim, um ponto que estiver sobre o hiperplano $f(\mathbf{x}) = 0$ terá $\xi_i = 1$ e pontos com $\xi_i > 1$ são classificados erroneamente. Nesse caso, a restrição (9.16) será substituída por

$$y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \tag{9.27}$$

com $\xi_i \geq 0$. Pontos para os quais $\xi_i = 0$ são corretamente classificados e estão sobre a fronteira da margem ou do lado correto da fronteira da margem. Pontos para os quais $0 < \xi_i \leq 1$ estão dentro da fronteira da margem, mas do lado correto do hiperplano, e pontos para os quais $\xi_i > 1$ estão do lado errado do hiperplano e são classificados erroneamente. Veja a Figura 9.5.

Nesse contexto, estamos diante de uma **margem flexível** ou **suave**. O objetivo é maximizar a margem e, para isso, minimizamos

$$C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||\boldsymbol{\beta}||^2, \tag{9.28}$$

em que $C > 0$ controla o balanço entre a penalidade das variáveis de folga e a margem.

Como qualquer ponto classificado erroneamente satisfaz $\xi_i > 1$, segue-se que $\sum_{i=1}^{n} \xi_i$ é um limite superior do número de classificações errôneas. No limite, quando $C \to \infty$, obtemos o CMM.

Para minimizar (9.28) sujeito a (9.27) e $\xi_i > 0$ consideramos o lagrangeano

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \frac{1}{2}||\boldsymbol{\beta}||^2 + C \sum_{i=1}^{n} \xi_i \\ &\quad - \sum_{i=1}^{n} \lambda_i [y_i f(\mathbf{x}_i) + \xi_i - 1] - \sum_{i=1}^{n} \mu_i \xi_i, \end{aligned} \tag{9.29}$$

em quel $\lambda_i \geq 0, \mu_i \geq 0$ são multiplicadores de Lagrange. Derivando (9.30 com relação a $\boldsymbol{\beta}, \alpha, \xi_i$, obtemos

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{n} \lambda_i y_i = 0 \tag{9.30}$$

e

$$\lambda_i = C - \mu_i. \tag{9.31}$$

Substituindo (9.30) - (9.31 em (9.30), temos

$$\widetilde{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{9.32}$$

que é uma expressão idêntica ao caso separável, com exceção das restrições, que são diferentes. Como $\lambda_i \geq 0$ são multiplicadores de Lagrange e como $\mu_i \geq 0$, de (9.31) segue que $\lambda_i \leq C$. Logo, precisamos maximizar (9.32) com respeito às variáveis duais λ_i, sujeito a

$$0 \leq \lambda_i \leq C, \tag{9.33}$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0, \quad i = 1, \ldots, n. \tag{9.34}$$

Novamente, estamos diante de um problema de programação quadrática. A previsão para um novo ponto $\mathbf{x}$ é obtida avaliando o sinal de $f(\mathbf{x})$ em (9.12). Substituindo (9.30) em (9.12) obtemos

$$f(\mathbf{x}) = \alpha + \sum_{i=1}^{n} \lambda_i y_i K(\mathbf{x}, \mathbf{x}_i). \tag{9.35}$$

Dados para os quais $\lambda_i = 0$ não contribuem para (9.35). Os dados restantes formam os vetores de suporte. Para esses, $\lambda_i > 0$ e, por (9.37) abaixo, devem satisfazer

$$y_i f(\mathbf{x}_i) = 1 - \xi_i. \tag{9.36}$$

No caso de CMF, as condições de KKT são dadas por

$$\begin{aligned} \lambda_i \geq 0, \quad y_i f(\mathbf{x}_i) - 1 + \xi_i \geq 0, \\ \lambda_i (y_i f(\mathbf{x}_i) - 1 + \xi_i) = 0, \qquad (9.37) \\ \mu_i \geq 0, \quad \xi_i \geq 0, \\ \mu_i \xi_i = 0, \quad i = 1, \ldots, n. \qquad (9.38) \end{aligned}$$

Procedendo como no caso de CMM, obtemos

$$\alpha = \frac{1}{N_{\mathcal{M}}} \sum_{i \in \mathcal{M}} \left(y_i - \sum_{j \in S} \lambda_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{9.39}$$

em que $\mathcal{M}$ é o conjunto do pontos tais que $0 < \lambda_i < C$.

Se $\lambda_i < C$, então, por (9.31), $\mu_i > 0$ e por (9.38), temos $\xi = 0$ e tais pontos estão na fronteira de separação. Pontos com $\lambda_i = C$ estão dentro da fronteira de separação e podem ser classificados corretamente se $\xi_i \leq 1$ e erroneamente se $\xi_i > 1$.

4) **Classificador de margem não linear**

Seja $\mathcal{X}$ o conjunto de dados (ou de padrões). A função $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ é um *kernel* se existir um espaço vetorial com produto interno, $\mathcal{H}$ (usualmente um espaço de Hilbert) e uma aplicação $\Phi : \mathcal{X} \to \mathcal{H}$, tal que, para todos $x, y \in \mathcal{X}$, tivermos

$$K(x, y) =< \Phi(x), \Phi(y) > . \tag{9.40}$$

Φ é a aplicação característica e $\mathcal{H}$, o espaço característico.

Por exemplo, tomemos $\mathcal{X} = \mathbb{R}^2$ e $\mathcal{H} = \mathbb{R}^3$ e definamos

$$\begin{aligned} \Phi : \mathbb{R}^2 &\to \mathbb{R}^3, \\ (x_1, x_2) &\to (x_1^2, x_2^2, \sqrt{2} x_1 x_2). \end{aligned}$$

Então, se $x = (x_1, x_2)$ e $y = (y_1, y_2)$, é fácil verificar que $< \Phi(x), \Phi(y) >=< x, y >$; logo $K(x, y) =< \Phi(x), \Phi(y) >=< x, y >$ é um *kernel*.

Para tornar o algoritmo de suporte vetorial não linear, notamos que ele depende somente de produtos internos entre os vetores de $\mathcal{X}$; logo, é suficiente conhecer $K(\mathbf{x}, \mathbf{x}^\top) =< \Phi(\mathbf{x}), \Phi(\mathbf{x}^\top) >$, e não de Φ explicitamente. Isso permite formular o problema de otimização, substituindo (9.30) por

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i) \tag{9.41}$$

e $f(\mathbf{x})$ dado por (9.10). Agora, $\boldsymbol{\beta}$ não é mais dado explicitamente como antes. Também, o problema de otimização é agora realizado no espaço característico e não em $\mathcal{X}$.

Os *kernels* a serem usados têm que satisfazer certas condições de admissibilidade. Veja Smola e Schölkopf (2004) para detalhes. Os *kernels* mencionados na Seção 9.2.3 são admissíveis.

5) **Classificação com mais de duas classes**

Para casos em que há mais de duas classes, duas abordagens são possíveis:

a) Classificação **uma contra uma** (*one-versus-one*)
Se tivermos K classes, são construídos $\binom{K}{2}$ classificadores, cada um com duas classes. Para uma observação teste $\mathbf{x}$, contamos quantas vezes essa observação é associada a cada uma das K classes. O classificador final é obtido associando a observação teste à classe que recebeu mais associações dentre as $\binom{K}{2}$ classificações duas a duas.

(b) Classificação **uma contra todas** (*one-versus-all*)
Consideramos K classificadores, cada vez comparando uma classe com as restantes $K-1$ classes. Sejam $\alpha_k, \beta_{1k}, \ldots, \beta_{pk}$ os parâmetros associados ao k-ésimo classificador. Para uma observação teste $\mathbf{x}^*$, vamos associá-la à classe para a qual $\alpha_k + \beta_{1k}{x_1}^* + \ldots + \beta_{pk}{x_p}^*$ seja a maior possível.

Os resultados obtidos podem ser inconsistentes. Veja Bishop (2006), para outras sugestões.

9.7 Exercícios

1) No contexto do classificador de margem máxima, mostre que se multiplicarmos os coeficientes α e β por uma constante, a distância de qualquer ponto ao hiperplano separador não se alterá.

2) Explique a relação entre valores de C em (9.9) e viés–variância.

3) Reanalise os dados do Exemplo 9.3 usando um *kernel* radial.

4) No Exemplo 9.3 foram usados 2 atributos. Reanalise-os o exemplo, usando 4 atributos e note que a acurácia da classificação melhora sensivelmente.

5) Considere os pontos da Figura 9.2. Use o CMF e o CBK.

6) Use função `tune()` do `e1071` para escolher o melhor modelo para os Exemplos 9.2 e 9.3.

Capítulo 10

Classificação por árvores e florestas

10.1 Introdução

Modelos baseados em árvores foram desenvolvidos na década de 1980 por Leo Breiman e associados e são bastante utilizados tanto para classificação quanto para previsão. Esses modelos são baseados numa segmentação do espaço gerado pelas variáveis explicativas (preditoras) em algumas regiões em que ou a moda (no caso de variáveis respostas categorizadas) ou a média (no caso de variáveis contínuas) é utilizada como predição. A definição dessas regiões é baseada em alguma medida de erro de previsão (ou de classificação). Em geral, as árvores são construídas a partir de um conjunto de **observações de treinamento** e testadas em um conjunto de **observações de teste** . Os modelos de árvores de decisão são conceitualmente e computacionalmente simples e bastante populares em função de sua interpretabilidade, apesar de serem menos precisos que modelos de regressão, por exemplo. Generalizações dos modelos originais, conhecidos como **florestas aleatórias** (*random forests*) costumam apresentar grande precisão, mesmo quando comparados com modelos lineares, porém pecam pela dificuldade de interpretação. A referência básica para esse tópico é Breiman et al. (1984). O texto de Hastie and Tibshirani (1990) também contém resultados sobre o tema.

Para predizer o valor de uma variável resposta Y (no caso de variáveis contínuas) ou classificar as observações em uma de suas categorias (no caso de variáveis categorizadas) a partir de um conjunto de variáveis preditoras $X_1, \ldots, X_p$, o algoritmo usado na construção de árvores de decisão consiste essencialmente na determinação das regiões (retângulos mutuamente exclusivos) em que o espaço das variáveis preditoras é particionado. A metodologia desenvolvida em Breiman et al. (1994),conhecida como **CART** (de *Classification And Regression Trees*) é baseada na seguinte estratégia:

a) Considere uma partição do espaço das variáveis preditoras (conjuntos dos possíveis valores de $X_1, \ldots, X_p$) em M regiões, $R_1, \ldots, R_M$.

b) Para cada observação pertencente a R_j, o previsor (ou categoria) de Y (que designaremos $\widehat{Y}_{R_j}$) será a moda (no caso discreto), a média (no caso contínuo) ou a porcentagem (no caso categorizado) dos valores de Y correspondentes àqueles com valores de $X_1, \ldots, X_p$ em R_j.

Embora a partição do espaço das variáveis preditoras seja arbitrária, usualmente ela é composta por retângulos p-dimensionais que devem ser construídos de modo a

minimizar alguma medida de erro de previsão ou de classificação (que explicitaremos posteriormente). Como esse procedimento geralmente não é computacionalmente factível dado o número de partições possíveis, mesmo com p moderado, usa-se uma **divisão binária recursiva** (*recursive binary splitting, RBS*) que é uma abordagem "de cima para baixo e gananciosa" (*top-down and greedy*) segundo James et al. (2013). A primeira locução justifica-se pelo fato de o procedimento ter início no topo da árvore (em que as observações estão todas na mesma região do espaço das variáveis preditoras) e a segunda, porque a melhor decisão é tomada em cada passo, sem avaliar se uma decisão melhor não poderia ser tomada num passo futuro.

Dado o vetor de variáveis preditoras $\mathbf{X} = (X_1, \ldots, X_p)^\top$, o algoritmo consiste dos passos:

i) Selecione uma variável preditora X_j e um limiar (ou ponto de corte) t, de modo que a divisão do espaço das variáveis preditoras nas regiões $\{\mathbf{X} : X_j < t\}$ e $\{\mathbf{X} : X_j \geq t\}$ corresponda ao menor erro de predição (ou de classificação).

(ii) Para todos os pares (j, t), considere as regiões

$$R_1(j,t) = \{\mathbf{X} : X_j < t\}, \quad R_2(j,t) = \{\mathbf{X} : X_j \geq t\}$$

e encontre o par (j, s) que minimiza o erro de predição (ou de classificação) adotado.

iii) Repita o procedimento, agora dividindo uma das duas regiões encontradas, obtendo três regiões; depois divida cada uma dessas três regiões minimizando o erro de predição (ou de classificação).

iv) Continue o processo até que algum critério de parada (obtenção de um número mínimo fixado de observações em cada região, por exemplo) seja satisfeito.

10.2 Árvores para classificação

Quando a variável resposta Y é categorizada, o objetivo é identificar a classe mais provável (**classe modal**) associada aos valores $\mathbf{X} = (X_1, \ldots, X_p)^\top$ das variáveis preditoras. Neste caso, uma medida de erro de classificação, comumente denominada **taxa de erros de classificação** (TEC) é a proporção de observações do conjunto de treinamento que não pertencem à classe modal. Outras medidas de erro de classificação, como **índice de Gini** ou **entropia cruzada** também podem ser usadas. Veja a Nota de Capítulo 1.

Admitamos que a variável resposta tenha K classes e que o espaço de variáveis preditoras seja particionado em M regiões. Designando por $\widehat{p}_{mk}$, a proporção de observações de treinamento da m-ésima região, $m = 1, \ldots, M$, pertencentes à k-ésima classe $k = 1, \ldots, K$, a taxa de erros de classificação dos elementos pertencentes à m-ésima região é

$$TEC_m = 1 - \max_k(\widehat{p}_{mk}).$$

Utilizaremos um exemplo para descrever o processo de construção de uma árvore de decisão. Vários pacotes (`tree`, `partykit`, `rpart`) podem ser utilizados com esse propósito. Cada um desses pacotes é regido por parâmetros que controlam diferentes aspectos da construção das árvores e não pretendemos discuti-los. O leitor interessado deverá consultar os manuais correspondentes com o objetivo de nortear uma seleção adequada para problemas específicos.

Exemplo 10.1 Consideremos novamente os dados analisados no Exemplo 9.2, disponíveis no arquivo `tipofacial`, extraídos de um estudo cujo objetivo era avaliar se duas ou mais medidas ortodônticas poderiam ser utilizadas para classificar indivíduos segundo o tipo facial (braquicéfalo, mesocéfalo ou dolicocéfalo). Como no Exemplo 9.2, para efeitos didáticos consideramos apenas duas variáveis preditoras, correspondentes a duas distâncias de importância ortodôntica, nomeadamente, a altura facial (`altfac`) e a profundidade facial (`proffac`). Os comandos do pacote `partykit` (com os parâmetros *default*) para a construção da árvore de classificação e do gráfico correspondente seguem juntamente com os resultados

```
> facetree <- ctree(grupo ~ altfac + proffac, data=face)

Model formula:
grupo ~ altfac + proffac

Fitted party:
[1] root
|   [2] altfac <= -0.1: dolico (n = 31, err = 9.7%)
|   [3] altfac > -0.1
|   |   [4] altfac <= 0.8: meso (n = 28, err = 14.3%)
|   |   [5] altfac > 0.8
|   |   |   [6] proffac <= -0.6: meso (n = 17, err = 41.2%)
|   |   |   [7] proffac > -0.6: braq (n = 25, err = 0.0%)

Number of inner nodes:    3
Number of terminal nodes: 4
> plot(facetree)
```

A variável preditora principal e o correspondente ponto de corte que minimiza a taxa de erros de classificação são `altfac` e $t = -0,1$, com $TEC_1 = 9,7\%$. Para observações com valores `altfac` $\leq -0,1$ (região R_1), classificamos o indivíduo como dolicocéfalo. Para valores de `altfac` $> -0,1$, a classificação depende do valor de `proffac`. Nesse caso, se `altfac` estiver entre $-0,1 \leq 0,8$, (região R_2), classificamos o indivíduo como mesocéfalo com $TEC_2 = 14,3\%$; se, por outro lado, `altfac` $> 0,8$ e `proffac` $\leq -0,6$, também classificamos o indivíduo como mesocéfalo (região R_3), com $TEC_3 = 41,2\%$; agora, se `altfac` $> 0,8$ e `proffac` $> -0,6$, o indivíduo deve ser classificado como braquicéfalo (região R_4), com $TEC_4 = 0,0\%$.

Na Figura 10.1, os símbolos ovais, que indicam divisões no espaço das variáveis preditoras são chamados de **nós internos** e os retângulos, que indicam as divisões finais são conhecidos por **nós terminais** ou **folhas** da árvore. Neste exemplo, temos 3 nós internos e 4 nós terminais. Os segmentos que unem os nós são os **galhos** da árvore. Os gráficos de barras apresentados em cada nó terminal indicam a frequência relativa com que as observações que satisfazem as restrições definidoras de cada galho são classificadas nas diferentes categorias da variável resposta.

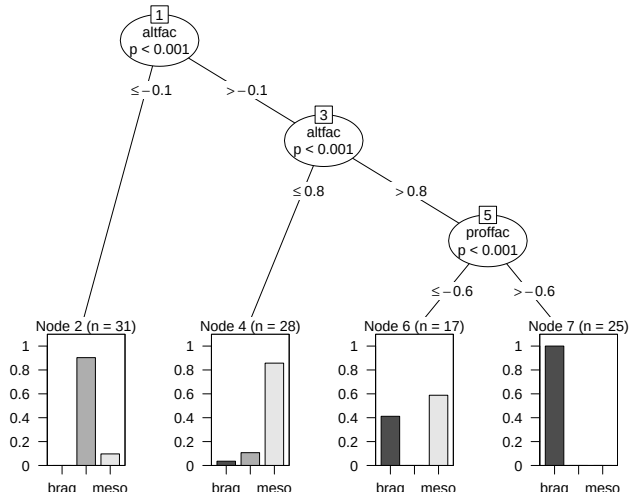


Figura 10.1: Árvore de decisão para os dados do Exemplo 10.1.

As regiões em que o espaço das variáveis preditoras foi particionado estão indicadas na Figura 10.2.

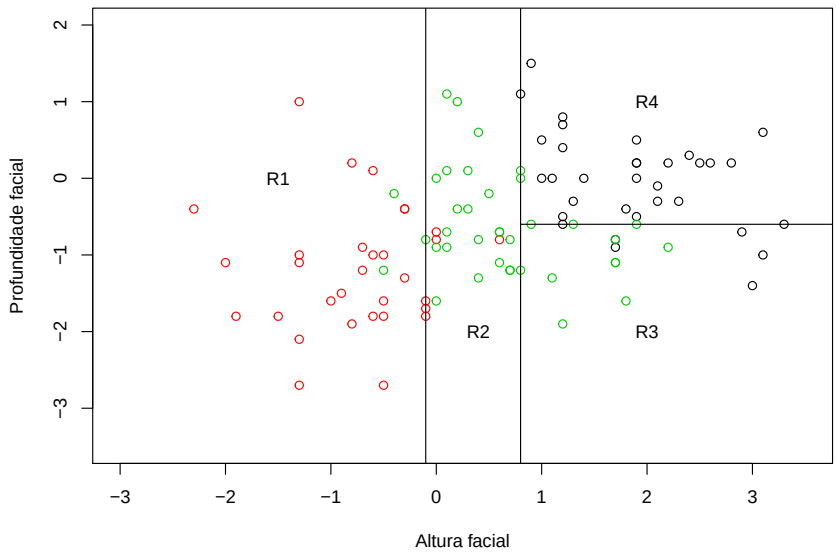


Figura 10.2: Partição do espaço das variáveis preditoras do Exemplo 10.1 (círculos vermelhos = dolicocéfalos, verdes = mesocéfalos, pretos = braquicéfalos).

Uma tabela com as classificações originais e preditas por meio da árvore de classificação é obtida por meio do comando `predict`

```
table
          braq dolico meso
  braq      25      0    0
  dolico     0     28    3
  meso       8      3   34
```

e indica uma taxa de erros de classificação de 13,9% ($16/101 \times 100\%$).

Um dos problemas associados à construção de árvores de decisão está relacionado com o **sobreajuste** (*overfitting*). Se não impusermos uma regra de parada para a construção dos nós, o processo é de tal forma flexível que o resultado final pode ter tantos nós terminais quantas forem as observações, gerando uma árvore em que cada observação é classificada perfeitamente. Para contornar esse problema, pode-se considerar o procedimento conhecido como **poda**, que engloba técnicas para limitar o número de nós terminais das árvores. A ideia que fundamenta essas técnicas está na construção de árvores com menos nós e, consequentemente com menor variância e interpretabilidade. O preço a pagar é um pequeno aumento no viés. Para detalhes, consulte a Nota de Capítulo 2.

Exemplo 10.2 Consideremos agora os dados do arquivo `coronarias` provenientes de um estudo cujo objetivo era avaliar fatores prognósticos para lesão obstrutiva coronariana (`LO3`) com categorias 1 :≥ 50% ou 0 :< 50%. Embora tenham sido observadas cerca de 70 variáveis preditoras, aqui trabalharemos com `SEXO` (0=fem, 1=masc), `IDADE1` (idade), `IMC` (índice de massa corpórea), `DIAB` (diabetes: 0=não, 1=sim), `TRIG` (concentração de triglicérides) e `GLIC` (concentração de glicose). Com propósito didático eliminamos casos em que havia dados omissos em alguma dessas variáveis, de forma que 1034 pacientes foram considerados na análise.

Os comandos do pacote `rpart` para a construção da árvore de classificação por meio de validação cruzada com os resultados correspondentes seguem

```
> rpart(formula = LO3 ~ GLIC + SEXO + IDADE1 + DIAB + TRIG + IMC,
    data = coronarias3, method = "class", xval = 20, minsplit = 10,
    cp = 0.005)
> printcp(lesaoobs)

Variables actually used in tree construction:
[1] GLIC   IDADE1 IMC    SEXO   TRIG

Root node error: 331/1034 = 0.32012
n= 1034
         CP nsplit rel error  xerror     xstd
1 0.0453172      0   1.00000 1.00000 0.045321
2 0.0392749      3   0.85801 0.97281 0.044986
3 0.0135952      4   0.81873 0.88218 0.043733
4 0.0090634      6   0.79154 0.87915 0.043687
5 0.0075529      7   0.78248 0.88822 0.043823
6 0.0060423     11   0.75227 0.92749 0.044386
7 0.0050000     13   0.74018 0.97885 0.045062
```

O erro relativo mede o erro de classificação dos dados de treinamento obtido por intermédio do modelo. O termo rotulado `xerror` é o erro de predição obtido por validação cruzada e o correspondente erro padrão é rotulado `xstd`. Cada linha da tabela CP representa um nível diferente da árvore. O erro de classificação obtido por validação cruzada tende a aumentar, pelo menos após o nível ótimo. Uma regra de parada prática consiste em selecionar o nível para o qual `rel error + xstd < xerror`. O gráfico correspondente, obtido por intermédio do comando

```
> rpart.plot(lesaoobs, clip.right.labs = TRUE, under = FALSE,
             extra = 101, type=4)
```

está disposto na Figura 10.3

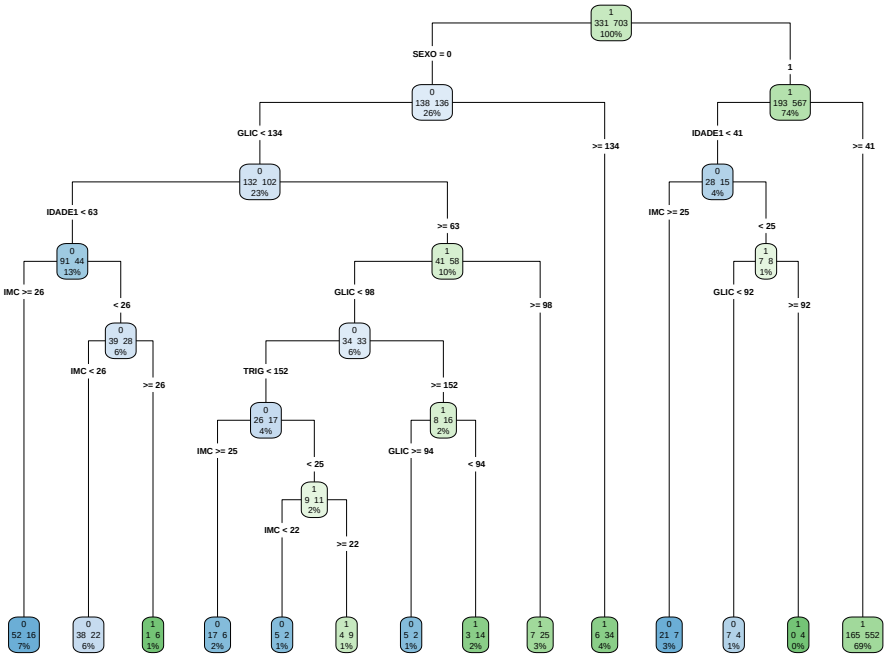


Figura 10.3: Árvore de decisão para os dados do Exemplo 10.2.

A tabela com as classificações originais e preditas é obtida por meio do comando

```
> table(coronarias3$LO3, predict(lesaoobs, type="class"))
     0   1
  0 145 186
  1  59 644
```

e indica um erro de classificação de $23,4\% = (186 + 59)/1034$.

O parâmetro CP (*complexity parameter*) serve para controlar o tamanho da árvore e corresponde ao menor incremento no custo do modelo necessário para a adição de uma nova variável. Um gráfico com a variação desse parâmetro com o número de nós, obtido com o comando `plotcp()` pode ser visto na Figura 10.4.

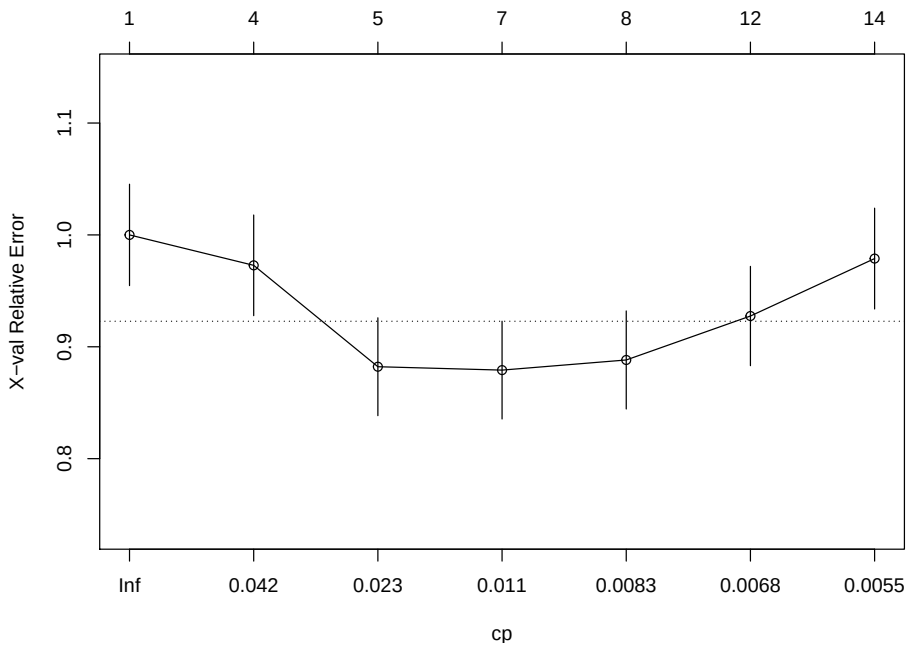


Figura 10.4: Gráfico CP para o ajuste da árvore aos dados do Exemplo 10.2.

Em geral, quanto mais níveis a árvore apresenta, menor será o erro de classificação, mas com o risco de sobreajuste (*overfitting*). O parametro CP pode ser usado para controlar esse sobreajuste sugerindo que a árvore deva ser podada (ver Nota de Capítulo 2.

Na Figura 10.4, procura-se o nível para o qual o erro relativo obtido por validação cruzada é mínimo. Para o exemplo, esse nível é 4, sugerindo que a árvore obtida no exemplo deve ser podada. A poda juntamente com o gráfico da árvore podada e a tabela com os correpsondentes valores preditos podem ser obtidos com os comandos

```
> lesaoobspoda <- prune(lesaoobs, cp = 0.015 ,"CP", minsplit=20, xval=25)
> rpart.plot(lesaoobspoda, clip.right.labs = TRUE, under = FALSE,
             extra = 101, type=4)
> rpart.rules(lesaoobspoda, cover = TRUE)
  LO3                                                cover
 0.33 when SEXO is 0 & IDADE1 <  63 & GLIC <  134     13%
 0.35 when SEXO is 1 & IDADE1 <  41                    4%
 0.59 when SEXO is 0 & IDADE1 >= 63 & GLIC <  134     10%
 0.77 when SEXO is 1 & IDADE1 >= 41                   69%
 0.85 when SEXO is 0                & GLIC >= 134      4%
```

A árvore podada está representada graficamente na Figura 10.5.

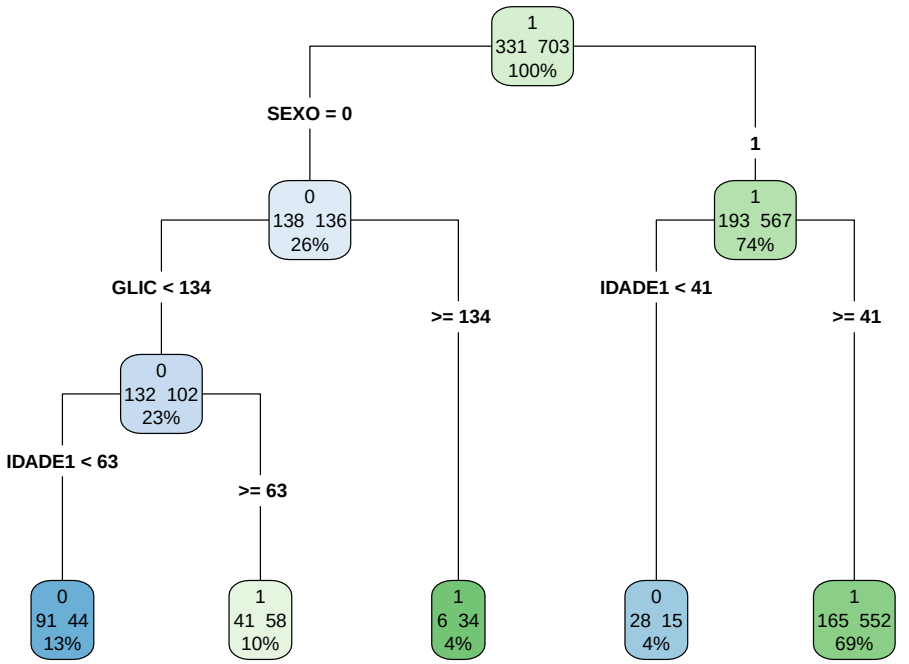


Figura 10.5: Gráfico CP para o ajuste da árvore (podada) aos dados do Exemplo 10.2.

O número indicado na parte superior de cada nó da Figura 10.5 indica a classe modal na qual são classificadas as observações; o valor à esquerda no centro do nó representa a frequência de classificações na classe $LO3 = 0$ e o valor à direita corresponde à frequência das classificações na classe $LO3 = 1$. Na última linha aparece a porcentagem de observações correspondentes ao nó. A tabela de classificação obtida a partir da árvore podada é

```
    0   1
0 119 212
1  59 644
```

e indica um erro de classificação de $26,2\%$, ligeiramente maior que o erro obtido com a árvore original, bem mais complexa.

10.3 Bagging, boosting e florestas

De um modo geral, árvores de decisão produzem resultados com grande variância, ou seja, dependendo de como o conjunto de dados é subdividido em conjuntos de treinamento e de teste, as árvores produzidas podem ser diferentes. As técnicas que descreveremos nesta seção tem a finalidade de reduzir essa variância.

10.3.1 Bagging

De modo geral, a técnica de **agregação bootstrap** (*bootstrap aggregating*) ou, simplesmente **bagging**, é um método para gerar múltiplas versões de um previsor

(ou classificador) a partir de vários conjuntos de treinamento e, com base nessas versões, construir um previsor (ou classificador) agregado.

A ideia básica é considerar um conjunto de **previsores (classificadores) fracos** de modo a obter um **previsor (classificador) forte**. Classificadores fracos têm taxas de erro de classificação altas. No caso binário, por exemplo, isso corresponde a uma taxa próxima de 0,50, que seria obtida com uma decisão baseada num lançamento de moeda. Um classificador forte, por outro lado, tem uma taxa de erro de classificação baixa.

Como em geral não dispomos de vários conjuntos de treinamento, a alternativa é utilizar réplicas bootstrap do conjunto de treinamento acessível para a obtenção das versões do preditor (ou classificador) que serão agregadas. Detalhes sobre a técnica bootstrap podem ser obtidos na Nota de Capítulo 3.

Para facilitar a exposição, consideremos um problema de regressão cujo objetivo é prever uma variável resposta quantitativa y a partir de um conjunto de treinamento $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. Nesse caso, a técnica consiste em obter B réplicas bootstrap desse conjunto, para cada conjunto de treinamento bootstrap, b, determinar o previsor de y, digamos $\widehat{f}^b(\mathbf{x})$, e agregá-los obtendo o previsor

$$\widehat{f}_{\text{bag}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}^b(\mathbf{x}).$$

No caso de classificação, para uma determinada observação de treinamento, $\mathbf{x}$, calcula-se $\widehat{c}^b(\mathbf{x})$ em cada uma das B árvores geradas e adota-se a classe k^* correspondente àquela com maior ocorrência como o valor do classificador agregado $\widehat{c}_{\text{bag}}(\mathbf{x})$. Esse procedimento é conhecido como escolha pelo **voto majoritário**). Especificamente,

$$\widehat{c}_{\text{bag}}(\mathbf{x}) = \text{argmax}_k[\#\{(b|\widehat{c}^b(\mathbf{x}) = k\}]$$

em que $\#\{A\}$ denota a cardinalidade do conjunto A.

O número de réplicas bootstrap sugerido por Breiman (1996) é cerca de 25. Veja a Nota de Capítulo 2.

Exemplo 10.3 A técnica bagging pode ser aplicada aos dados do Exemplo 10.2 por meio dos comandos

```
> set.seed(054)
>
> # train bagged model
> lesaoobsbag <- bagging(
+   formula = LO3 ~ SEXO + IDADE1+ GLIC,
+   data = coronarias3,
+   nbagg = 200,
+   coob = TRUE,
+   control = rpart.control(minsplit = 20, cp = 0.015)
+ )
>
> lesaoobspred <- predict(lesaoobsbag, coronarias3)
> table(coronarias3$LO3, predict(lesaoobsbag, type="class"))

     0   1
  0 117 214
  1  78 625
```

Variando a semente do processo aleatório, as taxas de erros de classificação giram em torno de 27% a 28%.

Quatro das 100 árvores obtidas em cada réplica bootstrap podem ser obtidas por meio dos comandos e estão representadas na Figura 10.6

```
as.data.frame(coronarias4)
clr12 = c("#8dd3c7","#ffffb3","#bebada","#fb8072")
n = nrow(coronarias4)
par(mfrow=c(2,2))
sed=c(1,10,22,345)
for(i in 1:4){
  set.seed(sed[i])
  idx = sample(1:100, size=n, replace=TRUE)
  cart =  rpart(LO3 ~ DIAB + IDADE1 + SEXO, data=coronarias4[idx,], model=TRUE)
  prp(cart, type=1, extra=1, box.col=clr12[i])
}
```

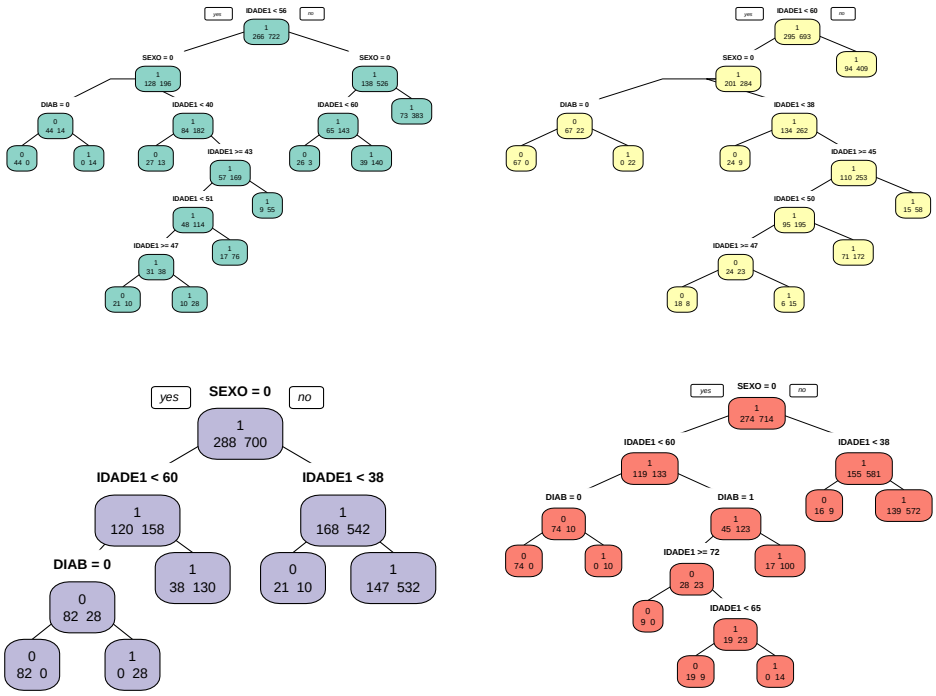


Figura 10.6: Exemplos de árvores obtidas por bagging para os dados do Exemplo 10.2.

10.3.2 Boosting

O objetivo do procedimento **boosting** é reduzir o viés e a variância em modelos utilizados para aprendizado supervisionado.

Diferentemente da técnica bagging, em que B árvores são geradas independentemente por meio de bootstrap, com cada observação tendo a mesma probabilidade

de ser selecionada em cada um dos conjuntos de treinamento, no procedimento boosting, as B árvores são geradas **sequencialmente** a partir de um único conjunto de treinamento, com probabilidades de seleção (pesos) diferentes atribuídos às observações. Observações mal classificadas em uma árvore recebem pesos maiores para seleção na árvore subsequente (obtida do mesmo conjunto de treinamento), com a finalidade de dirigir a atenção aos casos em que a classificação é mais difícil.

Em ambos os casos, o classificador final é obtido por meio da aplicação dos B classificadores fracos gerados com as diferentes árvores, por meio do voto majoritário. Além dos pesos atribuídos às observações no processo de geração dos classificadores fracos, o procedimento boosting atribui pesos a cada um deles, em função das taxas de erros de classificação de cada um. Essencialmente, o classificador forte pode ser expresso como

$$\widehat{c}_{\text{boost}}(\mathbf{x}) = \sum_{b=1}^{B} \widehat{c}^{b}(\mathbf{x}) w(b) \tag{10.1}$$

em que $w(b)$ é o peso atribuído ao classificador $\widehat{c}^{b}(\mathbf{x})$.

Se por um lado, o procedimento bagging raramente reduz o viés quando comparado com aquele obtido com uma única árvore de decisão, por outro, ele tem a característica de evitar o sobreajuste. Essas características são invertidas com o procedimento boosting.

Existem vários algoritmos para a implementação de boosting. O mais usado é o algoritmo conhecido como `AdaBoost` (de *adaptive boosting*), desenvolvido por Freund e Schapire (1997). Dada a dificuldade do processo de otimização de (10.1), esse algoritmo considera um processo iterativo de otimização que produz bons resultados embora não sejam ótimos. A ideia é adicionar o melhor classificador fraco numa determinada iteração ao classificador fraco obtido na iteração anterior, ou seja, considerar o processo

$$\widehat{c}^{b}_{boost}(\mathbf{x}) = \widehat{c}^{b-1}_{boost}(\mathbf{x}) + \widehat{c}^{b}(\mathbf{x}) w(b)$$

em que $\widehat{c}^{b}_{boost}(\mathbf{x})$ é o classificador com o melhor incremento no desempenho relativamente ao classificador $\widehat{c}^{b-1}_{boost}(\mathbf{x})$. Nesse contexto, a escolha de $[\widehat{c}^{b}(\mathbf{x}), w(b)]$ deve satisfazer

$$[\widehat{c}^{b}(\mathbf{x}), w(b)] = \text{argmin}_{[c^{b}(\mathbf{x}), w(b)]}\{E[\widehat{c}^{b-1}_{boost}(\mathbf{x}) + \widehat{c}^{b}(\mathbf{x}) w(b)]\}$$

em que $E[c^{b}]$ denota o erro de classificação do classificador c^{b}.

Consideremos, por exemplo, um problema de classificação binária baseado num conjunto de treinamento com N observações. No algoritmo **AdaBoost**, o classificador sempre parte de um único nó (conhecido como **toco** (*stump*) em que cada observação tem peso $1/N$. O ajuste por meio desse algoritmo é realizado por meio dos seguintes passos

1) Ajuste o melhor classificador (fraco) com os pesos atuais e repita os passos seguintes para os $B - 1$ classificadores subsequentes.

2) Calcule o valor do peso a ser atribuído ao classificador fraco corrente a partir de alguma métrica que indique quanto esse classificador contribui para o classificador forte corrente.

3) Atualize o classificador forte adicionando o classificador fraco multiplicado pelo peso calculado no passo anterior.

4) Com esse classificador forte, calcule os pesos atribuídos às observações de forma a indicar quais devem ser o foco da próxima iteração (pesos atribuídos às observações mal classificadas devem ser maiores que pesos atribuídos a observações bem classificadas.

Os algoritmos para implementação de boosting têm 3 parâmetros: i) o número de árvores, B, que pode ser determinado por validação cruzada (ver Nota de Capítulo 1 do Capítulo 8); ii) um **parâmetro de encolhimento** (*shrinkage*), $\lambda > 0$, pequeno, da ordem de 0,01 ou 0,001, que controla a velocidade do aprendizado; iii) o número de divisões em cada árvore, d; como vimos, o AdaBoost, usa $d = 1$ e, em geral, esse valor funciona bem.

O pacote `gbm` implementa o boosting e o pacote `adabag` implementa o algoritmo AdaBoost.

Exemplo 9.5. Vamos utilizar os dados do Exemplo 12.4 e usar o pacote `gbm`. Com o comando `summary()` obtemos a influência relativa dos preditores e a figure respectiva, Figura 9.8.

```
summary(boost.esteira)
var    rel.inf
CARGA  43.46858
IMC    24.22825
Idade  16.84917
Peso   15.45401
```

Vemos que os preditores CARGA e IMC são os mais importantes. Podemos produzir gráficos parciais de dependência para essas duas variáveis (Figura 9.9). Vemos que o consumo de oxigênio cresce com a CARGA e diminui com o IMC.

Figura 9.8: Influência elativa para os preditores do Exemplo 9.5.

Figura 9.9: Consumo de oxigênio versus CARGA e IMC.

Agora usamos o modelo via *boosting* para prever VO2 para o conjunto teste.

```
yhat.boost=predict(boost.esteira,newdata=esteira2[-train,],n.trees=5000)
mean((yhat.boost-boston.test)^2)
[1] 2.063152e-05
```

Vemos que o EQM obtido via *boosting* é consideravelmente menor do que aquele obtido via *bagging*.

10.3.3 Florestas aleatórias

Tanto bagging quanto boosting quanto florestas aleatórias têm o mesmo objetivo: diminuir a variância e o viés de procedimentos baseados em árvores de decisão. Enquanto os dois primeiros enfoques são baseados em um conjunto de B árvores utilizando o mesmo conjunto de p variáveis preditoras em cada um deles, o enfoque conhecido por **florestas aleatórias** utiliza diferentes conjuntos das variáveis preditoras na construção de cada árvore. Pode-se dizer que esse procedimento acrescenta bagging ao conjunto das p variáveis preditoras e nesse sentido, introduz

mais aleatoriedade e diversidade no processo de construção do modelo agregado. Na construção de um novo nó, em vez de escolher a melhor variável dentre as p disponíveis no conjunto de treinamento, o algoritmo de florestas aleatórias seleciona a melhor delas dentre um conjunto de $m < P$ selecionadas ao acaso. Usualmente, escolhe-se $m \approx \sqrt{p}$.

Intuitivamente, a utilização de florestas aleatórias para tomada de decisão corresponde à síntese da opinião de indivíduos com diferentes fontes de informação sobre o problema em questão.

Formalmente, para cada árvore j, um vetor aleatório θ_j é gerado, independentemente de vetores prévios $\theta_1, \ldots, \theta_{j-1}$, mas com a mesma distribuição. A árvore usando o conjunto de treinamento e θ_j resulta num classificador $\hat{f}_j(\mathbf{x}, \theta_j)$. Cada árvore vota na classe mais popular para o vetor $\mathbf{x}$.

A acurácia das árvores aleatórias é tão boa quanto a do *AdaBoost* e, às vezes, melhor. O resultado obtido por intermédio do algoritmo de árvores aleatórias em geral é mais robusto com relação a valores atípicos e ruído além de ser mais rápido do que bagging e boosting.

Para mais informação e resultados teóricos sobre AA, veja Breiman (2001).

O pacote randomForest do R implementa florestas. Nesse caso, o número de preditores tem que ser menor do que p.

Exemplo 9.6. Vamos usar o conjunto de dados esteira2 e as variáveis CARGA e IMC para crescer uma floresta. Obtemos o sumário abaixo:

```
Call:
 randomForest(formula = VO2 ~ Idade + CARGA + IMC + Peso,
              data = esteira2, mtry = 2, importance = TRUE,
subset = train)
              Type of random forest: regression
              Number of trees: 500
              No. of variables tried at each split: 2

          Mean of squared residuals: 4.164505
           Percentage Var explained: 52.23
```

Obtendo-se previsões para o conjunto teste, temos o sumário abaixo:

```
yhat.rf=predict(rf.esteira,newdata=esteira2[-train,])
mean((yhat.rf-esteira.test)^2)
[1] 3.713565
```

O EQM de previsão via floresta é menor do que aquele via *bagging*, mas maior do que o via *boosting*.

Um sumário da importância de cada preditor é dado abaixo:

```
summary(boost.esteira)
var  rel.inf
CARGA  43.46858
IMC    24.22825
Idade  16.84917
Peso   15.45401
```

e um gráfico está na Figura 9.10. Novamente, as variáveis CARGA e IMC são as mais importantes.

Concluindo, sobre os três algoritmos, *bagging*, floresta e *boosting*, o último é o que apresentou o maior poder preditivo.

Figura 9.10: Importância relativa os preditores do Exemplo 9.6.

10.4 Notas de Capítulo

1) **Critérios para avaliação de árvores de classificação**

Como alternativa para as taxas de erros de classificação, pode-se usar o **índice de Gini** definido como

$$G_m = \sum_{k=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk}),$$

que, essencialmente, corresponde à soma das variâncias das proporções de classificação em cada classe. Quando o valor de $\widehat{p}_{mk}$ para um dado nó m e uma dada categoria k estiver próximo de 1 e para as demais categorias estiver próximo de zero, o índice de Gini correspondente estará próximo de zero, indicando que para esse nó, uma grande proporção das observações será classificada em uma das K categorias. Quanto mais concentradas em uma categoria forem as classificações em um dado nó tanto maior será o o seu grau de **pureza**.

Outra medida utilizada com o mesmo propósito e que tem características similares àquelas do coeficiente de Gini é a **entropia cruzada**, definida como

$$ET_m = \sum_{k=1}^{K} \widehat{p}_{mk} \log(\widehat{p}_{mk}).$$

Para detalhes, consulte James et al. (2017).

2) **Poda de árvores**

Normalmente, árvores com muitos nós terminais apresentam bom desempenho no conjunto de treinamento mas podem estar sujeitas a sobreajuste, e não produzir boas classificações no conjunto de teste. Nesse contexto, árvores com um número menor de regiões (ou subdivisões) constituem uma boa alternativa produzindo resultados com menor variância e melhor interpretação. O procedimento chamado **poda** (*pruning*) pode ser usado com esse fim. A poda pode ser realizada na própria construção da árvore (**pré poda**) ou após sua finalização (**pós poda**). No primeiro caso, a poda é obtida por meio da especificação de um critério de parada, como a determinação do número mínimo de observações em cada nó terminal.

A poda, propriamente dita, consiste na construção de uma árvore com muitos nós terminais e segundo algum critério e na eliminação de alguns deles, obtendo uma árvore menor. Essencialmente, o procedimento consiste em construir a árvore até que o decréscimo no critério de avaliação (taxa de classificações erradas, por exemplo) gerada em cada divisão exceda algum limiar (em geral alto), obtendo-se uma **sub-árvore**.

Usar esse procedimento até se obter a menor SQR pode ser não factível e o que se faz é considerar uma sequência de árvores indexada por um parâmetro de poda $\alpha \geq 0$: para cada α temos uma subárvore A tal que

$$\sum_{j=1}^{|A|} \sum_{i:\mathbf{x}_i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |A|, \tag{10.2}$$

seja o menor possível, onde $|A|$ é o número de nós terminais de A, R_j é a região (retângulo) correspondente ao j-ésimo nó terminal e $\hat{y}_{R_j}$ é a resposta prevista associada à região R_j. O valor de α é escolhido por validação cruzada.

Voltemos ao Exemplo 12.3, dados de Iris, e vamos aplicar uma poda. Primeiramente, o comando `cv.tree` realiza CV. Os comandos apropriados do pacote `tree` são:

```
> tree.iris=tree(Species~., iris)
> cv.iris=cv.tree(tree.iris, FUN=prune.misclass)
> names(cv.iris)
[1] "size"   "dev"    "k"      "method"
> cv.iris
$size
[1] 6 4 3 2  1

$dev
[1] 5 5 9 99 114

$k
[1] -Inf    0    2   44   50

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Acima, `dev` corresponde ao erro de VC e as árvores com 6 e 4 nós terminais têm o menor erro (5). A Figura 9.11 mostra a taxa de erro como função do tamanho (`size`) e de `k` (que corresponde a α em (12.5)). Vemos que a taxa de erro da VC diminui com o tamanho da árvore e aumenta com o valor de `k` (ou α).

```
> par(mfrow=c(1,2))
> plot(cv.iris$size,cv.iris$dev,type="b")
> plot(cv.iris$k,cv.iris$dev,type="b")
```

Agora, vamos usar a função `prune.misclass()` para efetuar a poda da árvore, tomando aquela com 4 nós terminais.

```
> prune.iris=prune.misclass(tree.iris,best=4)
> plot(prune.iris)
> text(prune.iris,pretty=0)
```

Obtemos a Figura 9.12, onde vemos que a árvore simplificou-se.

Figura 9.11: Taxa de erro em função de `dev` e `k`.

Figura 9.12: Árvore resultante da poda da Figura 9.6.

Para ver qual tamanho de árvore resultante tem a melhor acurácia, podemos usar a função `predict()`. Veja o Exercício 7.

3) **Bootstrap**

Com o progresso de métodos computacionais e com capacidade cada vez maior de lidar com grandes conjuntos de dados, o cálculo de erros padrões, vieses etc., pode ser concretizado sem recorrer a uma teoria, que muitas vezes é muito complicada ou simplesmente não existe. Um desses métodos é o chamado **bootstrap**, introduzido por B. Efron, em 1979. A ideia que fundamenta o método bootstrap é reamostrar o conjunto de dados disponível para estimar um parâmetro θ, com o fim de criar dados replicados. A partir dessas replicações, podemos avaliar a variabilidade de um estimador proposto para θ, sem recorrer a cálculos analíticos.

Considere um conjunto de dados $\mathbf{x} = (x_1, \ldots, x_n)$ a ser utilizado para estimar a mediana populacional, Md, por meio da mediana amostral $\text{md}(\mathbf{x}) = \text{med}(x_1, \ldots, x_n)$. Seja $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ uma amostra aleatória simples, com reposição, de tamanho n dos dados $\mathbf{x}$, chamada de **amostra bootstrap**. Por exemplo, suponha que $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$. Um amostra bootstrap é, por exemplo, $\mathbf{x}^* = (x_4, x_3, x_3, x_1, x_2)$.

Repita o processo de amostragem, gerando B amostras bootstrap independentes, denotadas $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$). Para cada amostra bootstrap, calcule uma réplica bootstrap do estimador proposto, ou seja, de $\text{md}(\mathbf{x}^*)$, obtendo

$$\text{md}(\mathbf{x}_1^*), \ldots, \text{md}(\mathbf{x}_B^*). \tag{10.3}$$

O estimador bootstrap do erro padrão de $\text{md}(\mathbf{x})$ é definido como

$$\widehat{\text{e.p.}}_B(\text{md}) = \left[\frac{\sum_{b=1}^{B} (\text{md}(\mathbf{x}_b^*) - \text{md}(\cdot))^2}{B-1} \right]^{1/2},$$

com

$$\text{md}(\cdot) = \frac{\sum_{b=1}^{B} \text{md}(\mathbf{x}_b^*)}{B},$$

ou seja, o estimador bootstrap do erro padrão da mediana amostral é o desvio padrão amostral do conjunto (10.3).

A questão que se apresenta é: qual deve ser o valor de B, ou seja, quantas amostras bootstrap devemos gerar para estimar erros padrões de estimadores? A experiência indica que um valor razoável é $B = 200$.

No caso geral de um estimador $\widehat{\theta} = t(\mathbf{x})$, o algoritmo bootstrap para estimar o erro padrão de $\widehat{\theta}$ é o seguinte:

1) Selecione B amostras bootstrap independentes $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$, cada uma consistindo de n valores selecionados com reposição de $\mathbf{x}$. Tome $B \approx 200$.
2) Para cada amostra bootstrap $\mathbf{x}_b^*$ calcule a réplica bootstrap

$$\widehat{\theta}^*(b) = t(\mathbf{x}_b^*), \quad b = 1, \ldots, B.$$

3) O erro padrão de $\widehat{\theta}$ é estimado pelo desvio padrão das B réplicas

$$\widehat{\text{e.p.}}_B = \left[\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2\right]^{1/2},$$

com

$$\widehat{\theta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*(b).$$

Para mais detalhes, consulte Efron and Tibshirani (1993).

10.5 Uma aplicação

Nesta seção vamos aplicar os métodos estudados neste capítulo e nos capítulos 8 (MSV) e 11(boosting). Vamos usar o conjunto de dados endometriose2, que possui 1872 indivíduos e 24 variáveis. Dessas, consideraremos as variáveis idade, dismenorreia, esterelidade, raça, instrução e proctoragia como preditores e endometriose como resposta (sim ou não).

10.5.1 Regressão logística

Usando a função glm() vemos que, dessas variáveis, idade, esterilidade (sim), raça (negra) e proctoragia (sim) são significativas. Usando-as, obtemos:

```
Call:
glm(formula = endometriose ~ idade + esterelidade +
    raca + proctoragia, family = binomial, data = endo)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6609  -0.7932  -0.6366   1.1637   2.1372

Coefficients:
                 Estimate Std. Error  zvalue  Pr(>|z|)
(Intercept)      0.948140   0.416559   2.276   0.0228 *
idade           -0.035571   0.006685  -5.321 1.03e-07 ***
esterelidadesim  0.716594   0.118297   6.058 1.38e-09 ***
racabranca      -0.805057   0.351769  -2.289   0.0221 *
racanegra       -1.559232   0.362713  -4.299 1.72e-05 ***
proctoragiasim   0.976514   0.244560   3.993 6.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2155.9  on 1870  degrees of freedom
Residual deviance: 2025.0  on 1865  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 2037
```

```
Number of Fisher Scoring iterations: 4
```

A matriz fornecendo os números de classificados corretamente e não corretamente resulta

```
          endometriose
glm.pred  nao  sim
     nao  1336  463
     sim   44   29
```

da qual obtemos que a proporção de classificação correta é 0,73. Incluindo todas as variáveis, esta proporção sobe para $0,86$, deconsiderando variáveis com coeficientes não significativos.

10.5.2 Função discriminate linear

Usando a função lda() obtemos o resultados:

```
         Length Class  Mode
prior      2     -none- numeric
counts     2     -none- numeric
means     10     -none- numeric
scaling    5     -none- numeric
lev        2     -none- character
svd        1     -none- numeric
N          1     -none- numeric
call       4     -none- call
terms      3     terms  call
xlevels    3     -none- list
na.action  1     omit   numeric
> lda.fit
Call:
lda(endometriose ~ idade + esterelidade + raca + proctoragia,
    data = endo, )

Prior probabilities of groups:
     nao      sim
0.737039 0.262961

Group means:
       idade   esterelidadesim  racabranca  racanegra
nao 35.60986       0.2081218    0.6185642   0.3691080
sim 33.15041       0.3597561    0.7479675   0.2154472
     proctoragiasim
nao     0.03045685
sim     0.07113821
```

```
Coefficients of linear discriminants:
                        LD1
idade           -0.05637592
esterelidadesim  1.26760067
racabranca      -1.66975999
racanegra       -2.80536761
proctoragiasim   1.81867093
```

Na Figura 10.4, vemos o gráfico dos dois grupos.

Figura 10.4: Grupos para a função discriminante linear.

A matriz de classificação fica

```
          endometriose
lda.class  nao  sim
      nao 1334  456
      sim   45   36
```

da qual obtemos 0,73 como proporção de classificação correta, igual ao caso de regressão logística.

10.5.3 Método do vizinho mais próximo

10.5.4 MSV

10.5.5 Boosting

10.6 Exercícios

1) Para avaliar o desempenho de uma árvore, dividimos o conjunto de observações em dois, um de treinamento e outro de teste e samos a função predict() e depois a função table. Avalie o poder preditivo para a árvore do Exemplo 9.1.
2) Para o Exemplo 9.3, use somente os comprimentos de pétalas e sépalas para construir a árvore. Obtenha as regiões e faça o seu gráfico, usando o comando partition.tree() do pacote tree.
3) Escreva, formalmente, as regiões determinadas na Figura 9.5.
4) Descreva, formalmente, 5 regiões que determinam as árvores das Figuras 11.3 e 11.6.
5) Efetue a poda do exemplo dado na Seção 11.5, tomando o tamanho da árvore igual a 6. Veja o que acontece e compare com o caso dado.
6) Efetue a poda das árvores das Figuras 11.3 e 11.4.
7) Separe o conjunto Iris em um conjunto de treinamento (100 observações, por exemplo) e um conjunto teste (50 observações). Com as funções predict() e table(), veja o que acontece com as taxas de erro de classificação quando efetuamos uma poda.
8) Use *bagging*, *boosting* e floresta para os conjuntos de dados A e B. Qual revela-se o melhor em termos de previsão, para cada conjunto?

Referências

Afiune, J. Y. (2000). Avaliação ecocardiográfica evolutiva de recém-nascidos pré-termo, do nascimento até o termo. Tese de Doutorado, Faculdade de Medicina da USP.

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

Bartlett, M.S. (1937). The statistical conception of mental factors. *British Journal of Psychology,* **28**, 97-104.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* **327**, 307–310. doi:10.1016/S0140-6736(86)90837-8

Blei, D. M. Smyth, P. (2017). Science and data science. PNAS, **114**, 8689–8692.

Box, G.E.P. and Müller, M.E.(1958). A note on the generation of random normal deviates. *The Annals of Statistics,* **29**, 610–611.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* **26**: 211-252.

Box, G. E. P. and Müller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, **29**, 610–611.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society. Series B,* **45**, 47–50.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, **16**, 199–231.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Chapman & Hall/CRC.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Berlin: Springer.

Bussab, W.O. e Morettin, P.A. (2017). *Estatística Básica, 9a Edição*. São Paulo: Saraiva.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

Chambers, J.M., Cleveland, W.S., Kleiner, B and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. London: Chapman and Hall.

Chambers, J. M. and Hastie, T. J. (eds.) (1992). *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

Chambers, J. M. (1993). Greater or lesser Statistics: A choice for future research. *Statistics and Computing*, **3**, 182–184.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician*, bf 49, 327–335.

Chollet, F. (2018). *Deep Learning with R*. Manning.

Cleveland, W. M. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W. M. (1985). *The Elements of Graphing Data*. Monterey: Wadsworth.

Cleveland, W. M. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.

Cleveland, W. M. (2001). Data Science: An action plan for expanding the technical areas of the field of Statistics. *International Statistical Revue*, **69**, 21–26.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* **20**, 37–46. doi:10.1177/001316446002000104

Colosimo, E.A. e Giolo, S.R. (2006). *Análise de Sobrevivência Aplicada*. São Paulo: Blücher.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273–297.

Dempster, A.P., Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Donoho, D. (2017). 50 years of Data Science. *Journal of Computational and Graphical Statistics*, **26**, 745–766.

Durbin, J. and Watson, G.S. (1950). Testing for serial correlation in least squares regression, I. *Biometrika* , **37**, 409-428.

Durbin, J. and Watson, G.S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika* , **38**, 159-178.

Durbin, J. and Watson, G.S. (1971). Testing for serial correlation in least squares regression, III. *Biometrika* , **58**, 1-19.

Dzik A., Lambert-Messerlian, G.,Izzo, V.M., Soares, J.B., Pinotti, J.A. and Seifer, D.B. (2000). Inhibin B response to EFORT is associated with the outcome of oocyte retrieval in the subsequent in vitro fertilization cycle. *Fertility and Sterility,*, **74**, 1114-1117.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapmand and Hall

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference.* New York: Cambridge University Press.

Elias, F.M., Birman, E.G., Matsuda, C.K., Oliveira, I.R.S. and Jorge, W.A. (2006). Ultrasonographic findings in normal temporomandibular joints. *Brazilian Oral Research,* **20**, 25-32.

Embrecths, P., Lindskog, F. and McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. In Handbook of Heavy Tailed Distributions in Finance, ed. S. Rachev, Elsevier, Ch. 8, 329–384.

Ehrenberg, A. S. C. (1981). The problem of numeracy. *The American Statistician*, **35**, 67-71.

Ferreira, J.E., Takecian, P.L., Kamaura, L.T., Padilha, B. and Pu, C. (2017). Dependency Management with WED-flow Techniques and Tools: A Case Study. *Proceedings of the IEEE 3rd International Conference on Collaboration and Internet Computing*, 379-388. doi:10.1109/CIC.2017.00055

Fletcher, R. (1987). *Practical Methods of Optimization, Second Edition.* New York: Wiley.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **57**, 453–476.

Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, **55**, 119–139.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Fuller, W.A. (1996). *Introduction to Statistical Time Series, Second Edition*. New York: Wiley.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo.* Boca Raton: Chapman & Hall.

Gartner, Inc. (2005). *Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007.* http://www.gartner.com.2005.

Gegembauer, H. V. (2010). *Análise de Componentes Independentes com Aplicações em Séries Temporais Financeiras.* Dissertação de Mestrado, IME-USP.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoraion of images.*IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6** , 721–741.

Giampaoli, V., Magalhães, M.N., Fonseca, F.C. e Anorozo, N.F. (2008). Relatório de análise estatística sobre o projeto: "Avaliação e pesquisa: Investigando as dificuldades em Matemática no Ensino Fundamental da Rede Municipal da cidade de São Paulo – 2ª fase". São Paulo, IME-USP. (RAE – CEA – 08P27).

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practices*. London: Chapman and Hall.

Goldfeld, S.M., Quandt, R.E. and Trotter, H.F. (1966). Maximisation by quadratic hill-climbing. *Econometrica*, **34**, 541-551.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. The MIT Press.

Graedel, T, and Kleiner, B. (1985). Exploratory analysis of atmospheric data. In *Probability, Statistics and Decision Making in Atmospheric Sciences, A.H. Murphy and R.W. Katz, eds), pp 1-43*. Boulder: Westview Press.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods.* New York: Wiley.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* London: Chapman & Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd Edition, Springer.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Hinkley, B. (1977). On quick choice of probability transformations. *Applied Statsistics*, **26**, 67-69.

Hebb, D.O. (1949). *The organization of behavior*. New York: Wiley.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic rgression model. *Communications in Statistics*, **A10**, 1043–1069.

Hosmer, D.W. and Lemeshow, S. (2013). *Applied Logistic Regression, Third ed.* New York: John Wiley.

Hyvärinen, A. and Oja, E. (1997). A fast fixed point algorithm for independent component analysis. *Neural Computation*, **9**, 1483–1492.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithm for independent component analysis. *IEEE Transactions on Neural Network*, **10**, 626–634.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Alorithms and applications. *Neural Networks*, **13**, 411–430.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*. New York: Wiley.

Jaiswal, S. (2018). K-Means Clustering in R Tutorial. Disponível em `https://www.datacamp.com/community/tutorials/k-means-clustering-r`.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Johnson, N.L. and Leone, F.C. (1964). *Statistics and Experimental Design in Engineering and Physical Sciences, Vols 1, 2*. New York: Wiley.

Jordan, M. I. (2019). Artificial inteligence – The revolution hasn't heppened yet. *Harvard Data Science Review*, Issue 1.1.

Kleijnen, J. and Groenendall, W. (1994). *Simulation: A Statistical Perspective*. Chichester: Wiley.

Kutner, M.H., Neter, J., Nachtsheim, C.J. and Li, W. (2004). *Applied Linear Statistical Models. 5th ed.* New York: McGraw-Hill/Irwin. ISBN-10: 007310874X, ISBN-13: 978-0073108742.

Lee, E.T. and Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis, 3rd. Edition*. New York: Wiley

Lemeshow, S. and Hosmer, D.W. (1982). The use of goodness-of-fit statisticsin the development of logistic regression models. *American Journal of Epidemiology*, **115**, 92-106.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Louis, T.A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society, Series B,* **44**, 98-130.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute ,* **22**, 719-748.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C.E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Ausgust 31, 1955.

McCulloch, W. S. and Pitts, W. A. (1943). Logical calculus of the ideas immanent in nervous activity. *Butt. math. Biophysics*, **S**, 115-133.

McGill, R., Tukey, J. W. and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, **32**, 12-16.

Medeiros, M. C. (2019). *Machine Learning Theory and Econometrics*. Lecture Notes.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society. Series B,* **51**, 127–138.

Meng, X.L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267-278.

Metropolis, N. and Ulam, S.(1949). The Monte Carlo method. *Journal of the American Statistical Association,* **44**, 335–341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092 .

Meyer, D. (2018). Support vector machines. The interface to libsvm in package e1071. FH technikum Wien, Austria.

Miller, R.G. and Halpern, J.H. (1982). Regression via censored data. *Biometrika*, **69**, 521-531.

Morettin, P. A. (2014). *Ondas e Ondaletas: da Análise de Fourier à Análise de Ondaletas de Séries Temporais*. São Paulo: EDUSP.

Morettin, P.A. and Tolói, C.M.C. (2018). *Análise de Séries Temporais,* 3a Edição, Volume 1. São Paulo: Blücher.

Morrison, D.F. (1976). *Multivariate Statistical Methods, 2nd Ed.* New York: McGraw-Hill.

Müller, P. (1992). Alternatives to the Gibbs sampling scheme. *Technical report*. Institute of Statistics and decision Sciences, Duke University, Durham.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

Paulino, C.D. e Singer, J.M. (2006). *Análise de Dados Categorizados.* São Paulo: Blücher.

Pedroso de Lima, A.C., Singer, J.M. e Fusaro, E.R. (2000). *Relatório de análise estatística sobre o projeto "Prognóstico de pacientes com insuficiência cardíaca encaminhados para tratamento cirúrgico."* São Paulo: Centro de Estatística Aplicada do IME/USP.

Powell, M.J.D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, **7**, 155-162.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rainardi, V. (2008). *Building a Data Warehouse with Examples in SQL Server*. Apress (Springer). doi: 10.1007/978-1-4302-0528-9

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Second Edition. New York: Springer.

Roberts, G. O. and Smith, A. F. M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications*, **49**, 207–216.

Rosenblatt, F. (1958). The perceptron: A theory of statistical separability in cog-

nitive systems. Buffalo: Cornell Aeronautical Laboratory, Inc. Rep. No. VG-1196-G-1.

Ross, S.(1997). *Simulation, 2nd Ed.*, New York: Academic Press.

Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473-489.

Schmee, J. and Hahn, G.J. (1979). A simple method for regression analysis with censored data. *Technometrics,*, **21**, 417-432.

Sen, P.K., Singer, J.M. and Pedroso-de-Lima, A.C. (2009), *From finite sample to asymptotic methods in Statistics*. Cambridge: Cambridge University Press.

Singer, J.M. and Andrade, D.F. (1997). Regression models for the analysis of pretest/posttest data. *Biometrics,* **53**, 729-735.

Singer, J.M. e Ikeda, K. (1996). Relatório de Análise Estatística sobre o projeto "Fatores de risco na doença aterosclerótica coronariana". São Paulo, SP, IME-USP, 1996, 28p. (CEA-RAE-9608).

Singer, J.M., Rocha, F.M.M e Nobre, J.S. (2018). *Análise de Dados Longitudinais*. Versão parcial preliminar.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, **14**, 199–222.

Sobol, I.M.(1976). *Método de Monte Carlo.* Moscow: Editorial MIR.

Stone, J. V. (2004). *Independent Component Analysis: A Tutorial Introduction*. MIT Press.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, **21**, 65–66.

Tanner, M.A. (1996). *Tools for Statistical Inference, 3rd Ed.*. New York: Springer.

Thurstone, L.L. (1947). *Multiple Factor Analysis: A development and expansion of vectors of the mind.*. Chicago: University of Chicago Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B (methodological), **58**, 267–288.

Trevino, A. (2016). Introduction to K-means Clustering. Disponível em `https://www.datascience.com/blog/k-means-clustering`.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, **33**, 1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Turing, A. (1950). Computing machinery and intelligence". *Mind*, LIX (236).

Vapnik, V. and Chervonenkis, A. (1964). A note on a class of perceptrons. *Automation and Remote Contro*, **25**.

Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern recognition* [in Russian]. Moskow: Nauka.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

Vapnik, V. (1998). *SAtatistical Learning Theory.* New York: Wiley.

Viera, J. and Garrett, J.M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine,* **37**, 360-263

von Neumann, J.(1951). Various techniques used in connection with random digits, Monte Carlo Method. *U.S. National Bureau of Standards Applied Mathematica Series*, **12**, 36–38.

Wayne, D.W. (1990). *Applied Nonparametric Statistics, Second Edition* . Boston: PWS-Kent. ISBN 0-534-91976-6.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association,* **85**, 699-704.

Wilks, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, **2**, 163–195

Witzel, M.F., Grande, R.H.M. and Singer, J.M. (2000). Bonding systems used for sealing: evaluation of microleakage. *Journal of Clinical Dentistry* , **11**, 47-52.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics,* **11**, 95-103.

Zan, A.S.C.N. (2005). Ultra-sonografia tridimensional: determinação do volume do lobo hepático direito no doador para transplante intervivos. Tese de doutorado. São Paulo: Faculdade de Medicina, Universidade de São Paulo.

Zerbini, T., Gianvecchio, V.A.P., Regina, D., Tsujimoto, T., Ritter, V. and Singer, J.M. (2018). Suicides by hanging and its association with meteorological conditions in São Paulo, Brazil. *Journal of Forensic and Legal Medicine,* **53**, 22-24. doi: dx.doi.org/10.1016/j.jflm.2017.10.010

Índice