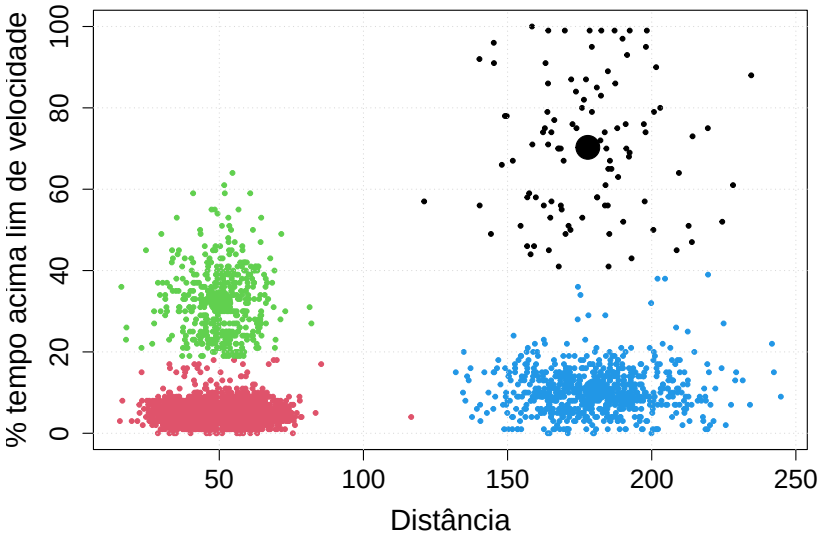

Estatística e Ciência de Dados

Versão preliminar

agosto de 2021

Pedro A. Morettin

Julio M. Singer

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

Conteúdo

Prefácio viii

1 Estatística, Ciência de Dados e Megadados **1**
1.1 Introdução 1
1.2 Aprendizado com Estatística 3
1.3 Aprendizado automático 4
1.4 Uma cronologia do desenvolvimento da Estatística 6
1.4.1 Probabilidades 6
1.4.2 Estatística 7
1.4.3 Estatística e computação 8
1.5 Notação e tipos de dados 8
1.6 Paradigmas para o aprendizado com Estatística 10
1.6.1 Aprendizado supervisionado 10
1.6.2 Aprendizado não supervisionado 13
1.7 Este livro 14
1.8 Conjuntos de dados 17
1.9 Notas de capítulo 20

PARTE I: ANÁLISE EXPLORATÓRIA DE DADOS **23**

2 Preparação dos dados **25**
2.1 Considerações preliminares 25
2.2 Planilhas de dados 28
2.3 Construção de tabelas 33
2.4 Construção de gráficos 35
2.5 Notas de capítulo 36
2.6 Exercícios 40

3 Análise de dados de uma variável **45**
3.1 Introdução 45
3.2 Distribuições de frequências 46
3.2.1 Variáveis qualitativas 48
3.2.2 Variáveis quantitativas 50
3.3 Medidas resumo 56

3.3.1 Medidas de posição 56
3.3.2 Medidas de dispersão 58
3.3.3 Medidas de forma 61
3.4 *Boxplots* 65
3.5 Modelos probabilísticos 67
3.6 Dados amostrais 70
3.7 Gráficos QQ 72
3.8 Desvio padrão e erro padrão 76
3.9 Intervalo de confiança e tamanho da amostra 79
3.10 Transformação de variáveis 81
3.11 Notas de capítulo 84
3.12 Exercícios 87

4 Análise de dados de duas variáveis 97
4.1 Introdução 97
4.2 Duas variáveis qualitativas 98
4.3 Duas variáveis quantitativas 111
4.4 Uma variável qualitativa e outra quantitativa 124
4.5 Notas de capítulo 131
4.6 Exercícios 140

5 Análise de dados de várias variáveis 149
5.1 Introdução 149
5.2 Gráficos para três variáveis 150
5.3 Gráficos para quatro ou mais variáveis 163
5.4 Medidas resumo multivariadas 164
5.5 Tabelas de contingência de múltiplas entradas 165
5.6 Notas de capítulo 168
5.7 Exercícios 179

6 Análise de Regressão 187
6.1 Introdução 187
6.2 Regressão linear simples 190
6.3 Regressão linear múltipla 209
6.4 Regressão para dados longitudinais 221
6.5 Regressão logística 224
6.6 Notas de capítulo 231
6.7 Exercícios 241

7 Análise de Sobrevivência 251
7.1 Introdução 251
7.2 Estimação da função de sobrevivência 255
7.3 Comparação de curvas de sobrevivência 262
7.4 Regressão para dados de sobrevivência 263
7.5 Notas de capítulo 265
7.6 Exercícios 266

PARTE II: APRENDIZADO SUPERVISIONADO **271**

8 Regularização e Modelos Aditivos Generalizados **273**
8.1 Introdução 273
8.2 Regularização 274
8.2.1 Regularização L_2 (*Ridge)* 275
8.2.2 Regularização L_1 (*Lasso*) 276
8.2.3 Outras propostas 278
8.3 Modelos aditivos generalizados (GAM) 285
8.4 Notas de capítulo 295
8.5 Exercícios 299

Referências **301**

Índice Remissivo **310**

Prefácio

Com a ampla divulgação de uma "nova" área de trabalho conhecida como Ciência de Dados (*Data Science*), muitas universidades estrangeiras criaram programas para o seu ensino, primeiramente no formato de MBA e em seguida como mestrados regulares. Esses programas, que, atualemente, incluem doutorado e até mesmo graduação foram surpreendentemente, introduzidos em escolas de Engenharia, Economia e, em vários casos, em escolas especialmente criadas para abrigar o "novo" domínio mas não em Departamentos de Estatística. De certa forma, muitos estatísticos sentiram-se perplexos e imaginaram-se diante de algo totalmente diferente do seu ofício. No entanto, uma visão mais crítica mostra que Ciência de Dados consiste principalmente na aplicação de algumas técnicas estatísticas a problemas que exigem grande capacidade computacional. Muitos modelos empregados nesse "novo" campo estavam disponíveis (e esquecidos) na literatura estatística há décadas e não vinham sendo aplicados em grande escala em virtude de limitações computacionais. Árvores de decisão, por exemplo, amplamente utilizadas em Ciência de Dados, foram introduzidas na década de 1980. Outro tópico, conhecido como Algoritmos de Suporte Vetorial (*Support Vector Machines*) que não fazia parte da metodologia estudada por estatísticos, por necessitar de grande capacidade computacional para sua aplicação, está disponível na literatura desde a década de 1990.

Hoje, programas de MBA e cursos de extensão em Ciência de Dados têm surgido no Brasil e atualmente, no Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP) estuda-se a possibilidade de criar um Mestrado em Estatística com ênfase em Ciência de Dados. A maior dificuldade é encontrar professores interessados em se readequar a esse novo paradigma.

Originalmente, pretendíamos escrever um texto sobre Análise Exploratória de Dados para utilização na disciplina de Estatística Descritiva, ministrada no bacharelado em Estatística do IME-USP. Tendo em vista as considerações acima, decidimos ampliar o escopo da obra para incluir tópicos que normalmente não são abordados nos cursos de graduação e pós-graduação em Estatística. Desta forma, além de apresentar fundamentos de Análise Exploratória de Dados (extremamente importantes para quem

pretende se aventurar em Ciência de Dados), o objetivo do texto inclui o preenchimento de uma lacuna na formação de alunos de graduação e pós-graduação em Estatística proveniente da falta de exposição a tópicos que envolvem a interação entre Estatística e Computação. Num contexto que resume as principais aplicações de Ciência de Dados, ou seja, de previsão, classificação, redução da dimensionalidade e agrupamento, apresentamos as ideias que fundamentam os algoritmos de suporte vetorial, árvores de decisão, florestas aleatórias e redes neurais. A exposição é focada em diversos exemplos e os detalhes mais técnicos são apresentados em notas de capítulo.

Embora sejam tópicos que tangenciam os métodos apresentados no texto, mas que são amplamente utilizados no ajuste de modelos estatísticos, incluímos apêndices com conceitos básicos de simulação e otimização.

Em resumo, o texto pode ser útil para diferentes disciplinas de graduação e pós-graduação em Estatística assim como para profissionais de diferentes áreas que tenham interesse em conhecer os fundamentos estatísticos da Ciência de Dados.

São Paulo, agosto de 2021

Os autores

Capítulo 1

Estatística, Ciência de Dados e Megadados

> Data is not an end in itself but a means to an end. More is not always better if it comes with increased costs.
>
> Faraway and Augustin, in "When small data beats big data."

1.1 Introdução

Atualmente, os termos *Data Science* (**Ciência de Dados**) e *Big Data* (**Megadados**)[1] são utilizados em profusão, como se envolvessem 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 **Análise Exploratória de Dados** (*Exploratory Data Analysis*), o primeiro a chamar a atenção para o que hoje é conhecido como 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 modelagem ou a inferência estatística.

Basta uma procura simples na *internet* para identificar novos centros de Ciência de Dados em várias universidades ao redor do mundo, com programas de mestrado, doutorado e mesmo de graduação. O interessante é que

[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/`. As exceções são as expressões para as quais não há uma tradução oficial (*boxplot*, por exemplo) ou acrônimos usualmente utilizadas por pacotes computacionais (MSE, de *Mean Squared Error*, por exemplo).

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 Ciência de Dados! Certamente isso é um equívoco. Como ressalta Donoho (2017), se uma das principais características dessa área é analisar grandes conjuntos de dados (megadados), há mais de 200 anos 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 tradicional, seja ela frequentista ou bayesiana, trata somente de pequenos volumes de dados, conhecidos como **microdados** (*small data*). Essa interpretação errônea vem do fato de que muitos livros didáticos incluem conjuntos de dados de pequeno ou médio porte para permitir que as técnicas de análise apresentadas possam ser aplicadas pelos leitores, mesmo utilizando calculadoras, planilhas de cálculo ou pacotes estatísticos. Nada impede que esses métodos sejam aplicados a grandes volumes de dados a não ser pelas inerentes dificuldades computacionais. Talvez seja este aspecto de complexidade computacional, aquele que mascara os demais componentes daquilo que se entende por Ciência de Dados, em que, na maioria dos casos, o interesse é dirigido para o desenvolvimento de algoritmos cuja finalidade é "aprender" a partir dos dados, muitas vezes subestimando as características estatísticas.

Em particular, Efron and 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 do estimador é

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

Esse algoritmo, no entanto, não contempla uma questão importante, que é saber quão acurado e preciso é este estimador. Admitindo-se que a amostra tenha sido colhida segundo um procedimento adequado, a metodologia estatística permite mostrar que $\mathrm{E}(\overline{x}) = \mu$, ou seja, que o estimador $\overline{x}$ é **não enviesado** e que o seu **erro padrão** é

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

o que implica a sua consistência e estabelece as bases para uma inferência estatística (frequentista) adequada. Em resumo, a utilização da média amostral como estimador não pode prescindir de uma avaliação estatística. Esses autores também 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 as relações entre Estatística e Ciência de Dados sob três perspectivas: estatística, computacional e humana. Segundo os autores, Ciência de Dados é uma filha da Estatística e da Ciência da Computação. A Estatística serviria à Ciência de Dados guiando a coleta e análise de dados complexos; a Ciência da Computação, desenvolvendo algoritmos que, por exemplo, distribuem conjuntos enormes de dados por múltiplos processadores (proporcionando velocidade de cálculo) ou armazenado-os adequadamente em equipamentos com grande capacidade de memória. Sob a perspectiva humana, a Ciência de Dados 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 e Ciência de Dados. Por exemplo, sobre a *Big Data Meme*, diz que se pode rejeitar o termo megadados como um critério para uma distinção séria entre Estatística e Ciência de Dados, o que está de acordo com 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 megadados por décadas.

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

Alguns termos muito utilizados hoje em dia são *Statistical Learning* (**Aprendizado com Estatística**) e *Machine Learning* (**Aprendizado com Máquina ou Automático**). Esses termos estão associados à 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.

1.2 Aprendizado com Estatística

O que hoje se entende como aprendizado com Estatística envolve duas classes de técnicas, denominadas **aprendizado supervisionado** e **aprendizado não supervisionado**.

O **aprendizado supervisionado** está relacionado com metodologia de-

senvolvida essencialmente para **previsão** e **classificação**. No âmbito de previsão, o objetivo é utilizar **variáveis preditoras** (sexo, classe social, renda, por exemplo) observadas em várias **unidades** (clientes de um banco, por exemplo) para "adivinhar" valores de uma **variável resposta numérica** (saldo médio, por exemplo) de novas unidades. O problema de classificação consiste em usar as variáveis preditoras para indicar em que categorias de uma **variável resposta qualitativa** (bons e maus pagadores, por exemplo) as novas unidades são classificadas.

Sob um ponto de vista mais amplo, além desses objetivos, a Estatística tradicional adiciona metodologia direcionada ao entendimento das relações entre as características dos dados disponíveis e aqueles de uma população da qual se supõe que os dados foram obtidos. Essa metodologia mais ampla é o que se entende por **Inferência Estatística**.

No **aprendizado não supervisionado**, dispomos apenas um conjunto de variáveis, sem distinção entre preditoras e respostas e o objetivo é descrever **associações** e **padrões** entre essas variáveis, **agrupá-las** com o objetivo identificar características comuns a conjuntos de unidades de investigação ou desenvolver métodos para combiná-las e assim **reduzir sua dimensionalidade**. Essas combinações de variáveis podem ser utilizadas como novas variáveis preditoras em problemas de previsão ou classificação.

1.3 Aprendizado automático

Inteligência Artificial também é um rótulo que aparece frequentemente na mídia escrita e falada e que tem gerado amplo interesse para analistas de dados. Esse termo suscita questões do tipo: no futuro computadores tornar-se-ão inteligentes e a raça humana será substituída por eles? 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 hoje rotulado como inteligência artificial, nada mais é do que aquilo que chamamos de aprendizado automático (*machine learning*). O trecho abaixo é extraído do artigo mencionado.

Artificial Intelligence is the mantra of the current era. The phrase is intoned by technologists, academicians, journalists, and venture capitalists alike. As with many phrases that cross over from technical academic fields into general circulation, there is significant misunderstanding accompanying use of the phrase. However, this is not the classical case of the public not understanding the scientists — here the scientists are often as befuddled as the public. The idea that our era is somehow seeing the emergence of an intelligence in silicon that rivals our own entertains all of us, enthralling us and frightening us in equal measure. And, unfortunately, it distracts us.

O autor distingue três tipos de inteligência artificial:

i) **Inteligência artificial imitativa da humana** (*Human-imitative ar-*

tificial inteligence), que se refere à noção que a entidade inteligente deva se parecer como nós, fisicamente ou mentalmente.

ii) **Aumento de inteligência** (*Intelligence augmentation*), segundo a qual a análise de dados e computação são usados para criar serviços que aumentam a inteligência e criatividade humanas.

iii) **Infraestura inteligente** (*Intelligent infrastructure*), referindo-se à vasta rede de dados, entidades físicas e aparato computacional que dão suporte para tornar o ambiente humano mais seguro e interessante (*smart TV, smart phone, smart house*).

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, que podem ser avaliadas pelo teste de Turing (*Turing test*). A primeira página do artigo está na Nota de Capítulo 1.

O tema 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 inteligência artificial está relacionada com um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos (Chollet, 2018) e consequentemente, intimamente ligada ao desenvolvimento da computação (ou programação de computadores). Até a década de 1980, a programação clássica era apenas baseada em um sistema computacional (um computador ou um conglomerado (*cluster*) de computadores) ao qual se alimentavam 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 (ou algoritmo) pode ser programada em alguma linguagem (Fortran, C, R, Python etc.). A maioria dos pacotes estatísticos existentes funciona dessa maneira.

A partir da década de 1990, a introdução do conceito de aprendizado automático criou um novo paradigma para analisar dados oriundos de reconhecimento de imagens, voz, escrita etc. Problemas dessa natureza são dificilmente solucionáveis sem o recente avanço na capacidade computacional. A ideia subjacente é **treinar** um sistema computacional 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 aos chamados conjuntos de **dados de treinamento**, aplicados a um conjunto de **dados de**

validação e comparados segundo algum critério de desempenho com o objetivo de escolher o melhor para prever ou classificar futuras unidades de investigação.

Convém ressaltar que o objetivo do aprendizado automático não é o mesmo daquele considerado na análise de regressão usual, em que se pretende entender como cada variável preditora está associada com a variável resposta. O objetivo do aprendizado automático é selecionar o modelo que produz melhores previsões, mesmo que as variáveis selecionadas com essa finalidade não sejam aquelas consideradas numa análise padrão.

Quando esses dois conjuntos de dados (treinamento e validação) 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 sistema computacional com o outro servindo para validação. Calcula-se então alguma medida do erro de previsão obtido ao se aplicar o resultado do ajuste do modelo obtido com os dados de treinamento aos dados de validação. Essa subdivisão (em conjuntos de treinamento e de validação) é repetida várias vezes, ajustando o modelo a cada conjunto de dados de treinamento, utilizando os resultados para previsão com os dados de validação 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 1 no Capítulo 8). O modelo selecionado deve ser ajustado ao conjunto de dados completo (treinamento + validação) 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 das variáveis preditoras).

1.4 Uma cronologia do desenvolvimento da Estatística

Embora a terminologia "aprendizado com Estatística" seja recente, a maioria dos conceitos subjacentes foi desenvolvida a partir do século 19. Essencialmente, aprendizado com Estatística e aprendizado automático tratam dos mesmos tópicos, mas utilizaremos o primeiro quando os métodos do segundo são tratados com técnicas estatísticas apropriadas.

1.4.1 Probabilidades

As origens da teoria de probabilidades remontam a 1654, com Fermat (1601-1665) e Pascal (1632-1662), que trataram de jogos de dados, baralho 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** na última década do Século 18 (1795) e usou-o regularmente em cálculos astronômicos depois de 1801. Foi Legendre (1752-1833), todavia, quem primeiro publicou, sem justificação, detalhes sobre o método no apêndice de seu livro "Nouvelles Méthodes pour la Détermination des Orbites des Comètes". Gauss (1809) apresentou 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 (1749-1827) desenvolveu o Teorema de Bayes independentemente, em 1774. Em 1812 e 1814 deu a interpretação bayesiana para probabilidade e fez aplicações científicas e 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 (*Markov Chain Monte Carlo*) e filtros de partículas). As contribuições de Jeffreys (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 (1894-1981). Fisher, em 1936, propôs 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 dessa teoria. Segundo Stigler (1990), o artigo de Fisher (1922), "On the mathematical foundation of theoretical statistics", publicado na *Phil. Trans. Royal Society, A* foi o artigo mais influente sobre Teoria Estatística no Século 20. Neyman e Pearson (1933), por sua vez, 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, Efron introduz a técnica ***Bootstrap*** e em 1970, Hoerl e Kennard introduzem a **Regressão em crista** (*Ridge rgression*). 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*) considerado em Breiman et al. (1984). Tibshirani (1996) 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 aprendizado com Estatística estão diretamente relacionados com avanços na área computacional. Até 1960, os métodos estatísticos precisavam ser implementados em máquinas de calcular manuais ou elétricas. Entre 1960 e 1980, apareceram as máquinas de calcular eletrônicas e os computadores de grande porte, como o IBM 1620, CDC 360, VAX etc., que trabalhavam com cartões perfurados e discos magnéticos. A linguagem FORTRAN predominava.

A partir de 1980 apareceram os computadores pessoais, supercomputadores, computação paralela, computação na nuvem (*cloud computation*), 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. Em 1984 surgiu 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 passou a ser a linguagem preferida dos estatísticos.

Métodos de aprendizado com Estatística não usualmente considerados em programas de graduação e pós-graduação em Estatística surgiram recentemente (na realidade, não tão recentemente; veja a citação de Vapnik no início desse capítulo), estão atraindo a atenção de um público mais amplo e são englobados no que hoje chamamos de Ciência de Dados. Tais métodos incluem **Algoritmos de suporte vetorial** (*Support Vector Machines*), **Árvores de decisão** (*Decision Trees*), **Florestas aleatórias** (*Random Forests*), ***Bagging***, ***Boosting*** etc. Outros métodos mais tradicionais 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

Introduzimos, agora, a notação usada no livro. Denotamos por $\mathbf{X}$, uma matriz com dimensão $n \times p$, contendo as **variáveis preditoras** ou **explicativas**; n indica o número de unidades de observação ou amostrais (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 colunas de $\mathbf{X}$, vetores com dimensão $n \times 1$ são denotadas por $\mathbf{x}_1, \ldots, \mathbf{x}_p$. As linhas de $\mathbf{X}$, vetores com dimensão $p \times 1$ são denotadas por $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$.

Então

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

em que $\mathbf{x}^\top$ denota o vetor $\mathbf{x}$ transposto.

Denotamos por $\mathbf{y} = (y_1, \ldots, y_n)^\top$ o vetor cujos elementos são os valores da **variável resposta**. No caso de análise estatística supervisionada, y_i pode representar um valor de uma variável numérica em problemas de predição ou o **rótulo** da i-ésima classe, num problema de classificação. Consequentemente os dados são os pares $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

Uma das características que têm sido objeto de discussão em Ciência de Dados está relacionada com o volume de dados, especialmente quando se trata dos chamados megadados (*big data*). Nesse contexto, as seguintes estruturas podem ser consideradas:

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.

Uma representação pictórica dessas estruturas de dados está apresentada na Figura 1.1.

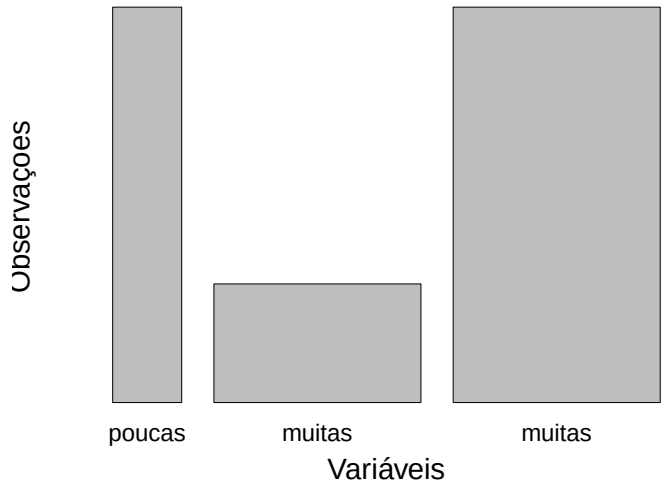

Figura 1.1: Estruturas de dados

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*, mídias sociais etc.

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 de modelos lineares a dados de alta dimensão pode ser tratado por meio técnicas de redução da dimensionalidade, regularização ou métodos bayesianos. Para modelos não lineares, árvores de decisão e redes neurais são técnicas mais adequadas.

1.6 Paradigmas para o aprendizado com Estatística

1.6.1 Aprendizado supervisionado

Previsão

Dados o vetor $\mathbf{y}$ com os valores da variável resposta e a matriz $\mathbf{X}$ com os correspondentes valores das variáveis preditoras, o modelo de regressão em **Aprendizado 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 aprendizado com Estatística é encontrar métodos para estimar f e usar o modelo (1.3) para fazer previsões ou em alguns casos, inferência sobre a população de onde os dados foram extraídos. A **previsão** para y_i é $\widehat{y}_i = \widehat{f}(\mathbf{x}_i)$ em que $\widehat{f}$ é a estimativa da função f, chamada de **previsor**. A acurácia de $\widehat{\mathbf{y}}$ como previsor de $\mathbf{y}$ depende dos seguintes dois tipos de 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 aprendizado com Estatística 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.

A acurácia do previsor $\widehat{f}$ é definida como

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

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.

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** mas há outros métodos disponíveis, como os **Algoritmos de Suporte Vetorial** (*Support Vector Machines - SVM*) ou **Árvores de Decisão**. O ajuste de um modelo de regressão por mínimos quadrados, por exemplo, pode ser pobre, como no Exemplo 6.7 do Capítulo 6 (veja a Figura 6.23). Nesse caso, pode-se tentar ajustar modelos mais flexíveis, escolhendo outras formas funcionais para f, incluindo aí modelos não lineares. Todavia, modelos mais flexíveis podem envolver a estimação de um grande número 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 de f com boa acurácia. 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 aprendizado com Estatística.

No caso de modelos de regressão, a medida mais usada para a avaliação da acurácia do modelo é o **Erro Quadrático Médio** (*Mean Squared Error*, MSE), definido por

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

em que $\widehat{f}(\mathbf{x}_i)$ é o valor predito da resposta para a i-ésima observação. Outra medida comumente utilizada para avaliar o ajuste de modelos de previsão é a **raiz quadrada** do erro quadrático médio (*Root Mean Squared Error*, $RMSE$). O erro quadrático médio calculado no conjunto de treinamento que produz o preditor $\widehat{f}$ é chamado **erro quadrático médio de treinamento**. Em geral, estamos mais interessados na acurácia do ajuste para os dados de validação e nesse caso podemos calcular o **erro quadrático médio de validação**,

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

que é o erro de previsão quadrático médio para as observações do conjunto de dados de validação, 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 preditores $\widehat{f}$ por meio da minimização de (1.5), calcular o correspondente erro quadrático médio no conjunto de validação via (1.6) e escolher o modelo para o qual esse valor é mínimo. Muitas vezes, usa-se **validação cruzada** em que o único conjunto de dados disponível é repetidamente dividido em dois subconjuntos, um deles servindo para treinamento e o outro para validação (ver Nota de Capitulo 1 do Capítulo 8).

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

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

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 validação 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 aprendizado com Estatística mais flexíveis têm viés baixo e variância grande. Na maioria dos casos, o erro quadrático médio de treinamento é menor que o erro quadrático médio de validação e o gráfico desses valores de validação em função do número de parâmetros de diferentes modelos, em geral, apresenta uma forma de U, resultante da competição entre viés e variância.

Classificação

Problemas de **classificação** são aqueles em que as respostas $y_1, \ldots, y_n$ são qualitativas. Formalmente, no caso de duas classes, 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 erros de treinamento**, que é a proporção de erros gerados pela aplicação do classificador $\widehat{g}$ às observações do conjunto 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 erros de validação**

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

para as observações do conjunto de validação, representadas por $(\mathbf{x}_0, y_0)$. Um bom classificador tem a taxa de erros de classificação (1.9) pequena. 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, uma alternativa é classificar a observação de validação 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 $1 - \max_j P(y = j|\mathbf{x} = \mathbf{x}_0),\ j = 1, -1$. 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 calculada 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 essa distribuição condicional. Detalhes serão apresentados no Capítulo 9.

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

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

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 9.

1.6.2 Aprendizado não supervisionado

Análise de agrupamentos

Nesta categoria de técnicas incluímos aquelas cujo objetivo é agrupar os elementos do conjunto de dados segundo com alguma medida de distância entre as variáveis preditoras, de modo que observações de um mesmo grupo tenham uma "pequena" distância entre elas.

Nos casos em que as as variáveis preditoras $\mathbf{x}_1, \ldots, \mathbf{x}_n$ pertencem a um espaço euclidiano p-dimensional (peso e altura, no caso bidimensional, por exemplo), a distância (euclidiana) definida por

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + \ldots + (x_{jp} - x_{jp})^2}$$

é utilizada. Casos em que as variáveis preditoras pertencem a espaços não euclidianos (palavras num texto, por exemplo), outras distâncias são consideradas.

Os algoritmos utilizados para a implementação das técnicas de agrupamento podem partir de um único grupo com todos os elementos do conjunto de dados e prosseguir subdividindo-o até que um número pré-fixado de grupos seja obtido ou considerar, inicialmente, cada elemento como um grupo e prosseguir combinando-os até a obtenção do número de grupos desejados.

Redução de dimensionalidade

O objetivo das técnicas consideradas nesta classe é reduzir a dimensionalidade de observações multivariadas com base em sua estrutura de dependência. Essas técnicas são usualmente aplicadas em conjuntos de dados com um grande número de variáveis e baseiam-se na obtenção de poucos **fatores**, obtidos como funções das variáveis observadas, que conservem, pelo menos aproximadamente, sua estrutura de covariância. Esses poucos fatores podem substituir as variáveis originais em análises subsequentes, servindo, por exemplo, como variáveis preditoras em modelos de regressão. Por esse motivo, a interpretação dessas novas variáveis é muito importante.

Dentre as técnicas mais utilizadas com essa finalidade, incluímos a **análise de componentes principais** e a **análise de componentes independentes** (ambas proporcionando a redução da dimensionalidade dos dados).

1.7 Este livro

Um dos maiores problemas oriundos da disseminação indiscriminada das técnicas utilizadas em Ciência de Dados 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 o principal componente da Ciência de Dados é 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.

A (bem-vinda) popularização das técnicas utilizadas em 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.

Embora muitos tópicos relacionados com Ciência dos Dados sejam abordados neste livro, o foco será na metodologia estatística que envolve a coleta, a organização, o resumo e a 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.

Há situações que um conjunto menor de dados pode ser mais útil do que um conjunto maior. Isto tem a ver com o equilíbrio entre viés e variância, que é um aspecto muito importante e realçado neste texto. O viés proveniente da análise de megadados pode ter consequências severas. Um exemplo apesentado em Meng (2014) mostra que em alguns casos, uma pequena amostra aleatória simples pode apresentar um erro quadrático médio menor do que

uma amostra administrativa (obtida observacionalmente, por exemplo) com 50% da população. Outro problema associado com megadados é o custo de sua obtenção. Para uma discussão desse problema, veja Faraway e Augustin (2018).

Além disso, segundo esses autores, a inferência estatística funciona melhor em conjuntos de dados menores. Métodos de aprendizado automático tendem a não fornecer medidas de incerteza, focalizando em melhores previsões e classificações. Finalmente, esses autores ressaltam um ponto já exposto acima: é melhor usar conjuntos de dados menores para o ensino, com a finalidade de garantir que os estudantes adquiram habilidades e conceitos e não somente aspectos computacionais, comuns quando se utilizam métodos dirigidos para a análise de megadados. Para uma discussão sobre essa dualidade *small data vs big data*, veja Lindstrom (2016).

Gostaríamos de fechar essa discussão com o texto abaixo, de Faraway e Augustin (2018), do qual a citação desse capítulo é parte:

Data is not an end in itself but a means to an end. The end is increased understanding, better calibrated prediction etc. More is not always better if this comes with increased costs. Data is sometimes viewed as something fixed that we have to deal with. It might be better to view it as a resource. We do not aim to use as many resources as possible. We try to use as few resources as possible to obtain the information we need. We have seen the benefit of big data but we are now also realizing the extent of associated damage. The modern environmental movement started in reaction to the excess of resource extraction. It advocates an approach that minimizes the use of resources and reduces the negative externalities. We believe the same approach should be taken with data: Small is beautiful.

O plano do livro é o seguinte. Nos Capítulos 2 a 7 apresentamos as bases para **análise exploratória de dados**. O Capítulo 2 é dedicado à preparação dos dados, geralmente apresentados de forma inadequada para análise. Nos Capítulos 3 a 5 dedicamo-nos 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. Técnicas de regressão, essenciais para o entendimento da associação entre uma ou mais variáveis explicativas e uma variável resposta são discutidas no Capítulo 6. O Capítulo 7 trata de técnicas de análise de sobrevivência, que essencialmente considera modelos de regressão em que a variável resposta é o tempo até a ocorrência de um evento de interesse.

Os Capítulos 8, 9, 10, 11 e 12 incluem os tópicos de **aprendizado supervisionado**. Em particular, no Capítulo 8, tratamos de extensões dos modelos de regressão bastante utilizadas para previsão e classificação, incluindo modelos de regularização e modelos aditivos generalizados. O Capítulo 9 é dirigido a técnicas clássicas como regressão logística, função discriminante

de Fisher etc., empregadas para classificação. Os Capítulos 10 e 11 abordam outras técnicas utilizadas para classificação ou previsão como os algoritmos de suporte vetorial, árvores e florestas além de redes neurais.

O aprendizado **não supervisionado** é abordado nos Capítulos 13 e 14, em que apresentamos métodos utilizados para agrupar dados e reduzir a dimensão do conjunto de variáveis disponíveis por meio de combinações delas. Neste contexto, não há distinção entre variáveis preditoras e respostas e o objetivo é o entendimento da estrutura de associação entre elas. Com essa finalidade, consideramos análise de agrupamentos, análise de componentes principais e análise de componentes independentes.

Conceitos básicos de otimização numérica, simulação e técnicas de dados aumentados são apresentados nos Apêndices A, B e C.

O texto poderá ser utilizado em programas de bacharelado em Estatística (especialmente na disciplina de Estatística Descritiva com os capítulos 1 a 8 ou 9 e na disciplina de Estatística Aplicada, com a inclusão dos capítulos restantes). Além disso, poderá ser utilizado como introdução à Ciência de Dados em programas de pós-graduação e também servirá para "cientistas de dados" que tenham interesse nos aspectos que fundamentam análise de dados.

Embora muitos cálculos necessários para uma análise estatística possam ser concretizados por meio de calculadoras ou planilhas eletrônicas, 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 *site*

`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, gam, MASS, mgcv, 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 planilhas Excel em arquivos disponíveis no formato

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

Por exemplo, no *site*

`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 *sites* 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	Resistência de adesivos dentários
antracose	Depósito de fuligem em pulmões
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
cidades	Dados demográficos de cidades brasileiras
coronarias	Fatores de risco na doença coronariana
covid	Internações por causas respiratórias em SP
disco	Deslocamento do disco temporomandibular
distancia	Distância para distinguir objeto em função da idade
empresa	Dados de funcionários de uma empresa
endometriose	Estudo sobre endometriose
endometriose2	Estudo sobre endometriose (1500 pacientes)
entrevista	Comparação intraobservadores em entrevista psicológica
esforco	Respostas de cardíacos em esteira ergométrica
esquistossomose	Testes para diagnóstico de esquistossomose
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
freios	Estudo de sobrevivência envolvendo pastilhas de freios
hiv	Sobrevivência de pacientes HIV
inibina	Utilização de inibina como marcador de reserva ovariana
lactato	Concentração de lactato de sódio em atletas
manchas	Número de manchas solares
morfina	Estudo sobre concentração de morfina em cabelos
municipios	Populações dos 30 maiores municípios do Brasil
neonatos	Pesos de recém nascidos
palato	Estudo sobre efeito de peróxido de hidrogênio na em palatos de sapos
piscina	Estudo de sobrevivência experimental com ratos
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	Rehabilitação de pacientes de infartos
rotarod	Tempo com que ratos permanecem em cilindro rotativo
salarios	Salários de profissionais em diferentes países
socioecon	Variáveis socioeconômicas para setores censitários de SP
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 de 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 statistician working alone is a statistician making mistakes.

David Byar

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. As unidades de investigação são os entes (indivíduos, animais, escolas, cidades etc.) em que as variáveis serão observadas. Num estudo em que se pretende avaliar a relação entre peso e altura de adultos, as unidades de investigação são os adultos e as variáveis a serem observadas são peso e altura. Em outro estudo, em que se pretenda comparar agências bancárias com relação ao desempenho medido em termos de diferentes variáveis, as unidades de investigação são as agências e as variáveis podem ser, por exemplo, saldo médio dos clientes, número de funcionários e total de depósitos em cadernetas de poupança.

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. Num modelo de regressão, por exemplo, isso corresponde a expressar a média da variável resposta como função dos valores de uma ou mais variáveis explicativas.

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 de estudos de 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

[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.

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.

Para se ter uma ideia de como é a função f para essa população, 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. O resultado sugere o consumo médio de indivíduos com renda nula é 1,44 e que esse consumo aumenta de 0,73 para cada incremento de uma unidade na renda. Detalhes sobre essa interpretação serão discutidos no Capítulo 6. Para diferentes populações poderemos ter curvas (modelos) diferentes para relacionar Y e C.

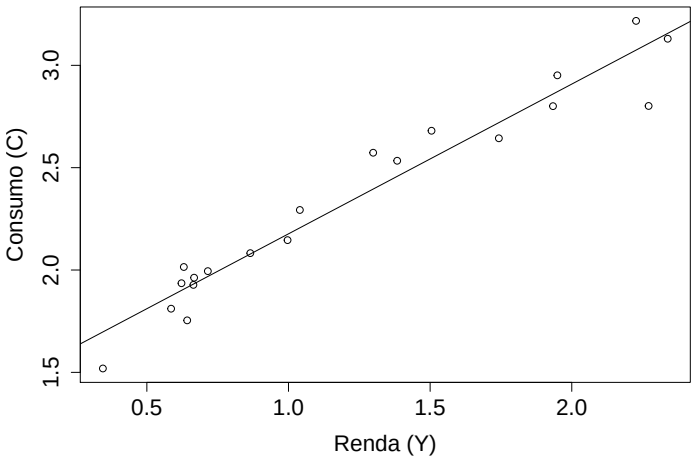

Figura 2.1: Relação entre renda e consumo de 20 indivíduos.

2.2 Planilhas de dados

Planilhas de dados (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 de dados 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 da construção de uma planilha de dados consiste na elaboração de um dicionário com a especificação das variáveis, que envolve

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

ii) a atribuição de rótulos às variáveis (esses rótulos devem ser preferencialmente mnemônicos e grafados 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) a indicação, quando pertinente, de limites (inferiores ou superiores) para facilitar a identificação de erros; por exemplo, o valor mínimo para aplicação num fundo de ações é de R$ 2000,00;

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

Algumas recomendações para a construção de planilhas eletrônicas 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[2];

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 2.2: Consideremos os dados 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	

Um exemplo de planilha de dados contendo observações das variáveis descritas na Tabela 2.1 está representado na Figura 2.2[3] Observe que neste conjunto de dados, as unidades de investigação são os pares (gestante+recém nascido). Há variáveis observadas tanto na gestante quanto no recém nascido e espera-se alguma dependência entre as observações realizadas no mesmo par mas não entre observações realizadas em pares diferentes.

[2]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.

[3]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**.

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

Figura 2.2: Planilha com dados referentes ao Exemplo 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.

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 de dados 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 o rótulo da mesma unidade amostral (animal, neste caso) é repetido 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.

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.

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 (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.

Exemplo 2.4: Os dados da Tabela 2.3 foram extraídos de um estudo sobre gestações gemelares com medidas de várias características anatômicas de fetos de gestantes com diferentes idades gestacionais (semanas).

Tabela 2.3: Diâmetro biparietal medido ultrassonograficamante (cm)

	Idade	Diâmetro biparietal	
Gestante	gestacional	Feto 1	Feto 2
1	28	7,8	7,5
2	32	8,0	8,0
3	25	5,8	5,9
⋮	⋮	⋮	⋮
34	32	8,5	7,2
35	19	3,9	4,1

Embora as medidas tenham sido observadas nos dois fetos, a unidade de investigação é a gestante. Espera-se que os diâmetros biparietais observados nos dois fetos da mesma gestante sejam dependentes. Esse tipo de dados tem uma **estrutura hierárquica** ou **por conglomerados**, em que os dois fetos estão aninhados na gestante. Tanto dados com essa natureza quanto dados longitudinais são casos particulares daquilo que se chama **dados com medidas repetidas** que essencialmente são dados em que a mesma variável

resposta é observada duas ou mais vezes em cada unidade de investigação.

Em geral, dados armazenados em planilhas eletrônicas devem ser apropriadamente transformados para análise por algum *software* estatístico (como o `R`). Nesse contexto, convém ressaltar que esses *software* tratam variáveis **numéricas** (peso, por exemplo) e **alfanuméricas** (bairro, por exemplo) de forma diferente. Além disso, observações omissas também requerem símbolos específicos. Cuidados nessa formatação dos dados apresentados em planilhas eletrônicas são importantes para evitar problemas na análise.

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 dificultar 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. Além disso, quanto mais casas decimais forem incluídas, mais difícil é a comparação. Por exemplo, compare a Tabela 2.4 com a Tabela 2.5.

   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 de se calcular as porcentagens pode ser útil para responder diferentes questões. Esse tópico será discutido no Capítulo 4.

Tabela 2.4: 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 2.5: 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 2.6 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 2.6 com a Tabela 2.7.

Tabela 2.6: Intenção de voto (%)

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

Tabela 2.7: 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 e use mais do que uma linha para acomodar rótulos longos. Avalie a Tabela 2.8.

Tabela 2.8: Sensibilidade dentinária pré- e pós-operatória

Material	Dentina	Sensibilidade pré-operatória	Sensibilidade pós-operatória		Total
			Ausente	Presente	
Single Bond	Seca	Ausente	22	1	23
		Presente	3	6	9
		Subtotal	25	7	32
Single Bond	Úmida	Ausente	12	10	22
		Presente	7	4	11
		Subtotal	19	14	33
Prime Bond	Seca	Ausente	10	6	16
		Presente	12	3	15
		Subtotal	22	9	31
Prime Bond	Úmida	Ausente	5	13	18
		Presente	11	3	14
		Subtotal	16	16	32

8) Não analise a tabela descrevendo-a, mas sim comentando as principais tendências sugeridas pelos dados. Por exemplo, os dados apresentados na Tabela 2.4 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. Exemplos serão apresentados ao longo do texto.

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 de análise de dados com diversas natureza (imagens, textos etc.). 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 necessário 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 por meio 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 fácil 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 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 in-

dividual 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 disponível no arquivo `morfina.doc`.

   A planilha morfina.xls, disponível no arquivo `morfina` foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo.

   a) Com base nessa planilha, apresente um dicionário com a especi-

ficação das variáveis segundo as indicações da Seção 2.2 e construa a planilha correspondente.

b) Com as informações disponíveis, construa tabelas para as variáveis sexo, raça, grau de instrução e tipo de lesão segundo as sugestões da Seção 2.3.

2) A Figura 2.5 foi extraída de um estudo sobre atitudes de profissionais de saúde com relação a cuidados com infecção hospitalar. Critique-a e reformule-a para facilitar sua leitura, lembrando que a comparação de maior interesse é entre as diferentes categorias profissionais.

	WHOQOL					
	Escores da avaliação da qualidade de vida					
Categoria profissional (n)	**Domínio I-Físico**	**Domínio II-Psicológico**	**Domínio III-Relações sociais**	**Domínio IV-Meio ambiente**	**Qualidade de vida global**	**Percepção geral da saúde**
Médico (42)						
Média (DP)	15,0 (2,8)	14,5 (2,4)	14,4 (3,2)	13,6 (1,6)	13,6 (3,2)	14,2 (2,0)
Mediana (min-max)	15,1 (6,3-20,0)	14,7 (9,3-18,7)	14,7 (5,3-20,0)	13,8 (9,5-17,5)	14,0 (6-20)	14,2 (8,5-18,3)
Enfermeiro (43)						
Média (DP)	14,7 (1,9)	14,6 (2,4)	14 (2,4)	12,7 (2,0)	14,2 (3,0)	13,9 (1,8)
Mediana (min-max)	14,9 (10,3-20,0)	15,0 (9,3-19,3)	14,7 (8,0-18,7)	12,8 (8,5-17,5)	14,0 (6,0-20)	14,2 (9,8-18,2)
Auxiliar de enfermagem (58)						
Média (DP)	14,6 (2,1)	15,5 (1,9)	14,7 (1,9)	11,8 (2,0)	14,3 (2,9)	13,9 (1,6)
Mediana (min-max)	14,9 (10,9-18,3)	15,3 (12,0-19,3)	14,7 (9.3-18,7)	12,0 (7,5-16,5)	15,0 (6,0-20,0)	13,7 (10,3-17,5)
Técnico de enfermagem (23)						
Média (DP)	15,2 (2,2)	15,7 (2,1)	15,6 (2,5)	12,5 (2,4)	15,7 (1,7)	14,5 (1,9)
Mediana (min-max)	14,9 (10,9-20,0)	16,0 (10,0-18,7)	16,0 (10,7-20,0)	12,5 (8,0-18,5)	16,0 (12,0-18,0)	15,1 (10,0-19,1)
TOTAL (166)						
Média (DP)	14,8 (2,3)	15,0 (2,3)	14,5 (2,5)	12,6 (2,1)	14,3 (2,9)	14,1 (1,8)
Mediana (min-max)	14,9 (6,3-20,0)	15,3 (9,3-19,3)	14,7 (5,3-20,0)	12,5 (7,5-18,5)	14,0 (6,0-20,0)	14,2 (8,5-19,1)

DP: Desvio Padrão
WHOQOL: World Health Organization Quality of Life Group

Figura 2.5: Tabela com escores de avaliação de qualidade de vida.

3) Utilize as sugestões para construção de planilhas apresentadas na Seção 2.2 com a finalidade de preparar os dados do arquivo `empresa` para análise estatística.

4) 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). Delineie uma planilha apropriada para a coleta e análise estatística dos dados, rotulando-a adequadamente.

5) Utilizando os dados do arquivo `esforco`, prepare uma planilha `Excel` num formato conveniente para análise pelo `R`. Inclua apenas as variáveis Idade, Altura, Peso, Frequência cardíaca e VO2 no repouso além do quociente VE/VCO2, as correspondentes porcentagens relativamente ao máximo, o quociente VO2/FC no pico do exercício e data do óbito. Importe a planilha `Excel` que você criou utilizando comandos `R` e obtenha as características do arquivo importado (número de casos, número de observações omissas etc.)

6) 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.

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.

7) 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 2.2, indicando claramente

   a) a definição das variáveis,

   b) os rótulos para as colunas da planilha.

.

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.

8) A planilha disponível no arquivo `cidades` contém informações demográficas de 3554 municípios brasileiros.

   a) Importe-a para permitir a análise por meio do *software* `R`, indicando os problemas encontrados nesse processo além de sua solução.

   b) Use o comando `summary` para obter um resumo das variáveis do arquivo.

   c) Classifique cada variável como numérica ou alfanumérica e indique o número de observações omissas de cada uma delas.

9) 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.

Capítulo 3

Análise de dados de uma variável

You see, but you do not observe.

Sherlock Holmes to Watson in A Scandal in Bohemia.

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) A distribuição de frequências dos dados pode ser considerada simétrica?

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 resumo.

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.

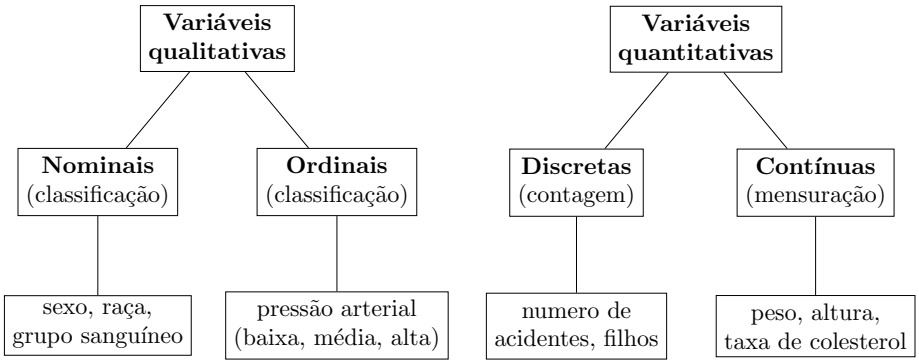

Figura 3.1: Classificação de variáveis.

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 o conjunto de dados o apresentado na Tabela 3.1 obtido de um questionário respondido por 50 alunos de uma disciplina ministrada na Fundação Getúlio Vargas em São Paulo. Os dados estão disponíveis no arquivo `ceagfgv`.

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

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.

3.2.1 Variáveis qualitativas

Uma tabela contendo as frequências (absolutas e/ou relativas) de unidades de investigação classificadas em cada categoria de uma variável qualitativa indica sua distribuição de frequências. A frequência absoluta corresponde ao número de unidades (amostrais ou populacionais) em cada classe e a frequência relativa indica a porcentagem correspondente. 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 qualitativas ordinais pode-se acrescentar uma coluna com as frequências relativas acumuladas que podem ser 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.

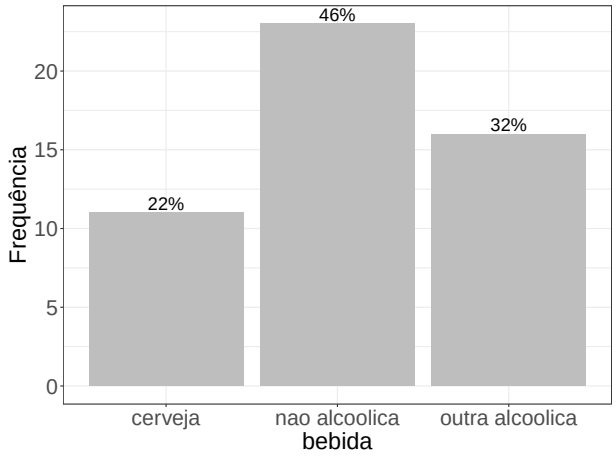

Figura 3.2: Gráfico de barras para Bebida preferida.

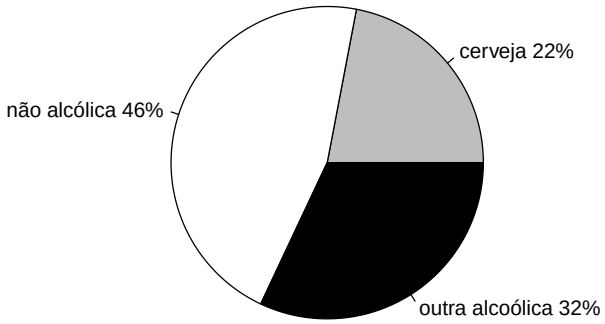

Figura 3.3: Gráfico tipo pizza para Bebida preferida.

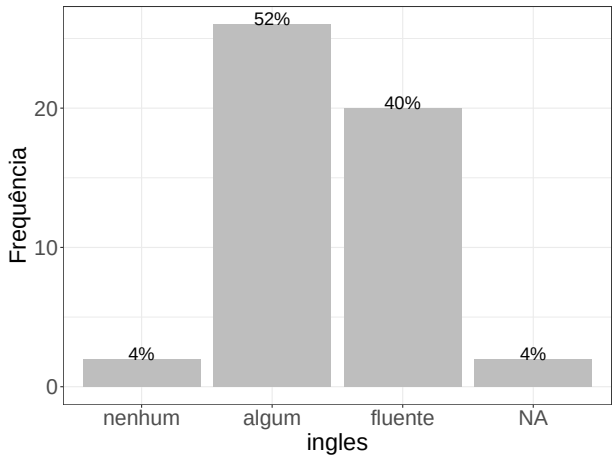

Figura 3.4: Gráfico de barras para Fluência em inglês.

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 frequências muito pequenas (em geral iguais a 1) nas 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.

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.

Alternativamente a tabelas com o formato da Tabela 3.4, vários gráficos podem ser utilizados para representar a distribuição de frequências das variáveis quantitativas 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.

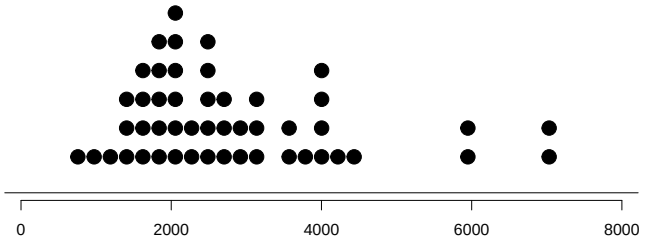

Figura 3.5: Gráfico de dispersão unidimensional para Salário (Exemplo 3.1).

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 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 primeiro dígito como o ramo e o segundo como folha, desprezando as dezenas. 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 e é composto de retângulos contíguos cuja área total é 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) a frequência absoluta n_k, que corresponde ao número de elementos cujos valores pertencem à classe k, $k = 1, \ldots, K$;

b) a frequência relativa $f_k = n_k/n$, que é a proporção de elementos cujos valores pertencem à classe k, $k = 1, \ldots, K$;

c) a densidade de frequência $d_k = f_k/h_k$ que representa a proporção de valores pertencentes à classe k por unidade de comprimento h_k de cada classe, $k = 1, \ldots, K$.

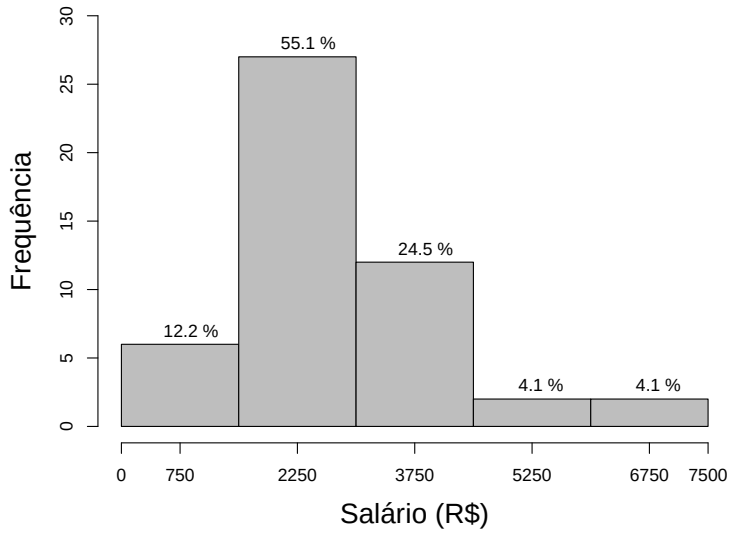

Figura 3.7: Histograma para a variável salário (R$).

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`.

[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

Ordenemos os valores $x_1, \ldots, x_{30}$ das populações dos 30 municípios, 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, devemos considerar intervalos de classe semi-abertos. A Tabela 3.6 contém a distribuição de frequências para a variável X que representa população. 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 também 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.

Tabela 3.6: Distribuição de frequências para a variável 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()`.

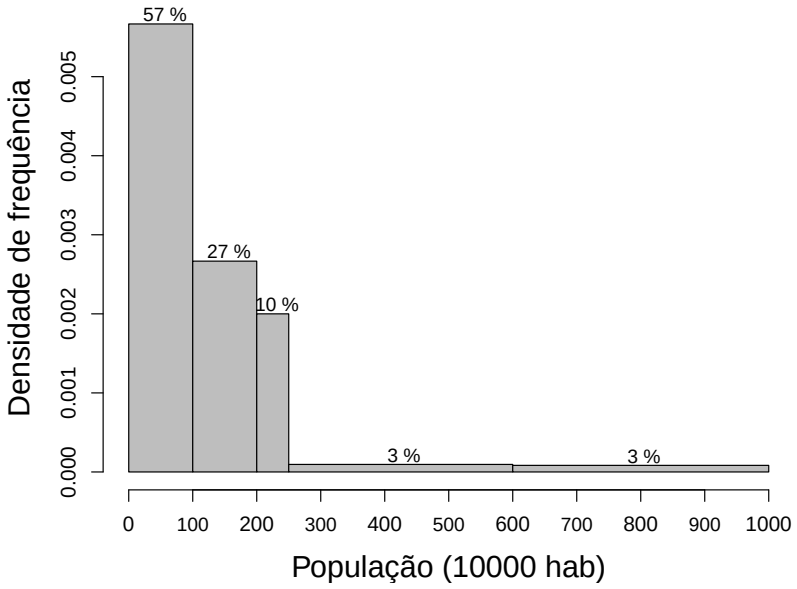

Figura 3.8: Histograma para a variável População (10000 habitantes).

O gráfico de ramo-e-folhas para os dados da Tabela 3.5 está apresentado na Figura 3.9. Por meio desse 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.

```
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 dos 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 | 11111222222223333344444
   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 forma, 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

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

em que $\widetilde{x}_k$ 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 $\widetilde{x}_k$ por x_k em (3.2).

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

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

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:

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

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, $\overline{x} = 34{,}9$, $\text{md}(x) = (14 + 15)/2 = 14{,}5$, 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 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-p) 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 $[p_{i+1}, Q(p_{i+1})]$ num gráfico cartesiano de p *versus* $Q(p)$.

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

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

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.

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

Com essa informação, podemos calcular outros quantis; por 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 $f_9 = 0{,}5$,

$$Q(0{,}62) = [0{,}30 \times x_{(6)} + 0{,}70 \times x_{(7)}] = (0{,}3 \times 15 + 0{,}7 \times 18) = 17{,}1,$$

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 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 para variáveis contínuas 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 de alguma função positiva 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

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

Neste caso, a função a que nos referimos é quadrática. Para uma tabela de frequências (com K classes), a expressão para cálculo da variância é

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

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

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

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 original. Uma medida com essa propriedade é a raiz quadrada positiva da variância, conhecida por **desvio padrão**.

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 amostras 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 (2017).

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. O desvio padrão amostral é usualmente denotado por S.

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

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

Neste caso, a função a que nos referimos acima é a função "valor absoluto".

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

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

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 $S^2 = 4313{,}9$; $S = 65{,}7$; $\text{dm}(x) = 37{,}0$; $d_Q = 23 - 7 = 16$; $S^2(0{,}20) = 34{,}3$; $S(0{,}20) = 5{,}9$ 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, dma/0,6745 é um estimador não enviesado para σ. A constante 0,6745 é obtida por meio de considerações assintóticas. No exemplo, 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 forma

Embora sejam menos utilizadas na prática que as medidas de posição e dispersão, as medidas de **assimetria** (*skewness*) e **curtose** 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, que, em geral, é baseada na relação entre o segundo e o terceiro **momentos centrados**, cujos correspondentes amostrais são respectivamente,

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

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:

$$\frac{\sqrt{n(n-1)}}{n-2}g_1. \tag{3.15}$$

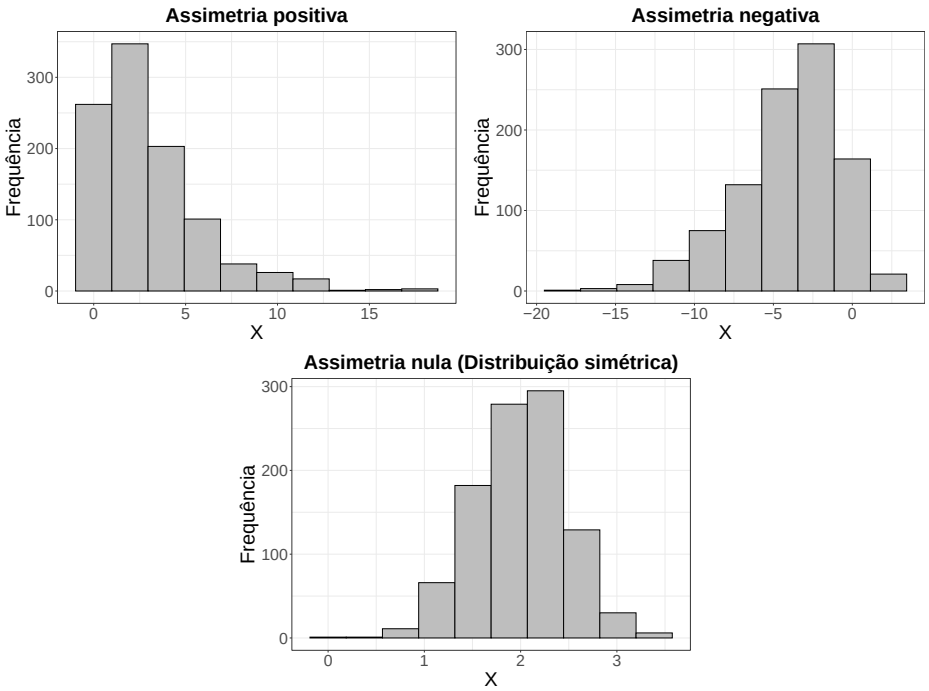

Figura 3.11: Histogramas com assimetria positiva e negativa e nula.

As principais propriedades desses coeficientes são

i) seu sinal reflete a direção da assimetria (sinal negativo corresponde a assimetria à direita e sinal positivo corresponde a assimetria à esquerda);

ii) comparam a assimetria dos dados com aquela da distribuição normal, que é simétrica;

iii) valores mais afastados do zero indicam maiores magnitudes de assimetria e consequentemente, maior afastamento da distribuição normal;

iv) a estatística indicada em (3.15) 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**, estimado por

$$Sk_2 = 3[\overline{x} - med(x_1, \ldots, x_n)]/S.$$

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. 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.

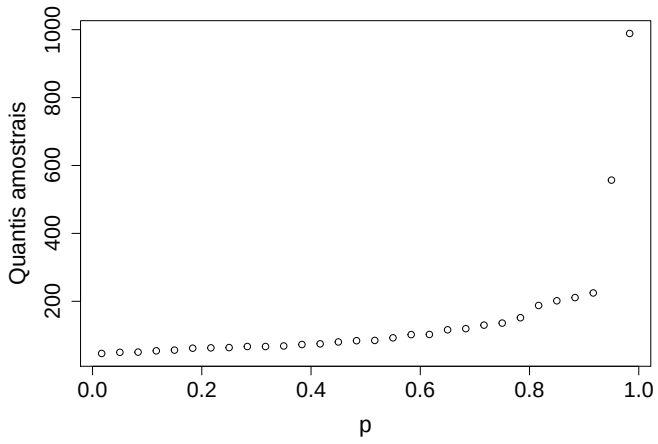

Figura 3.12: Gráfico de quantis para População (10000 habitantes).

A Figura 3.13 contém o gráfico de quantis para a variável `IMC` do arquivo `esteira` sugerindo que a distribuição correspondente é aproximadamente simétrica, com pequeno desvio na cauda superior.

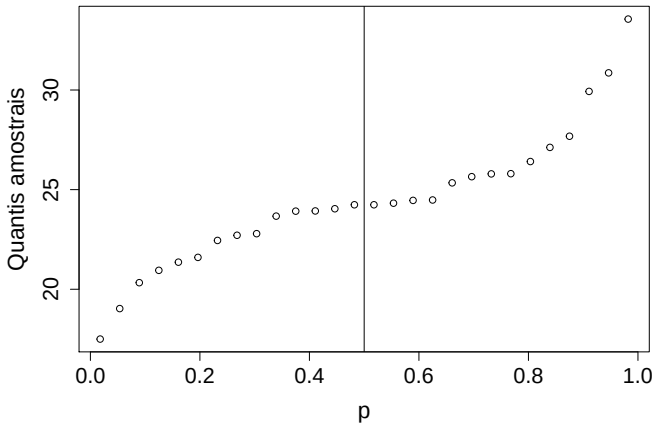

Figura 3.13: Gráfico de quantis para a variável `IMC` do arquivo `esteira`.

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. Para uma distribuição simétrica (ou aproximadamente simétrica), espera-se que

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

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

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

A condição a) nos diz que a **dispersão inferior** é igual (ou aproximadamente igual) à **dispersão superior**. 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 sua esquerda.

Além disso, se uma distribuição for (aproximadamente) simétrica, vale a relação

$$Q_2 - x_{(i)} \approx x_{(n+1-i)} - Q_2, \quad i = 1, \ldots, [(n+1)/2], \tag{3.16}$$

em que $[x]$ representa o maior inteiro contido em x. Definindo $u_i = Q_2 - x_{(i)}$ e $v_i = x_{(n+1-i)} - Q_2$, 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 3.14, na qual podemos observar que a maioria dos pontos está acima da reta $u = v$, representada pela linha pontilhada. Essa disposição mostra que a distribuição correspondente apresenta uma assimetria à direita.

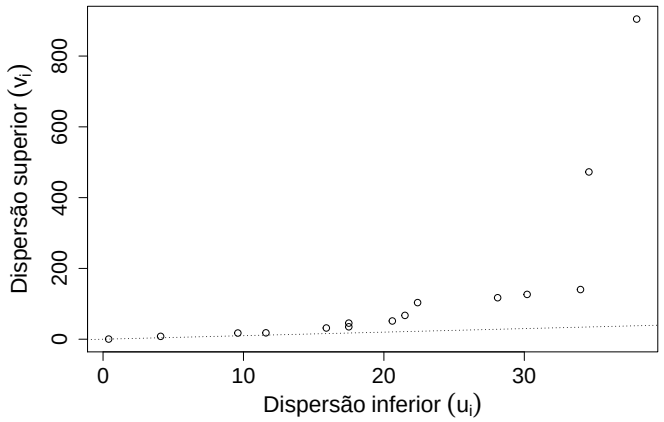

Figura 3.14: Gráfico de simetria para População (10000 habitantes).

Na Figura 3.15 apresentamos o gráfico de simetria para a variável `IMC` do arquivo `esteira`; esse gráfico corrobora as conclusões obtidas por meio da Figura 3.13.

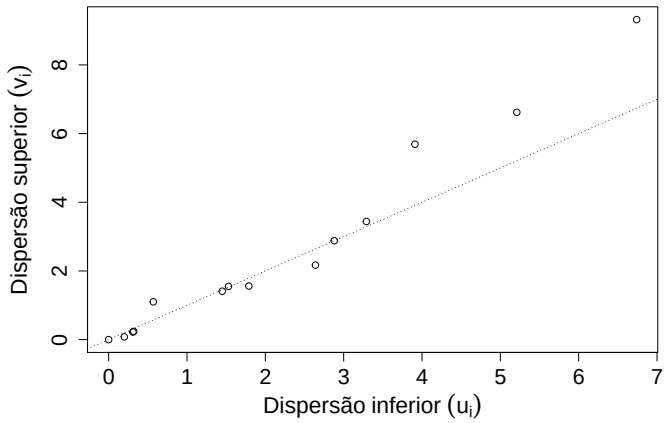

Figura 3.15: Gráfico de simetria para a variável `IMC` do arquivo `esteira`.

Outra medida de interesse para representar a distribuição de frequências de uma variável é a **curtose**, que está relacionada com as frequências relativas em suas caudas. Essa medida envolve momentos centrados de quarta ordem.

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

$$K(X) = \mathrm{E}\left[\frac{(X-\mu)^4}{\sigma^4}\right]. \tag{3.17}$$

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

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 a média μ é estimada por $\overline{X}$. Substituindo os momentos centrados de X pelos respectivos momentos centrados amostrais, obtemos um estimador da curtose, nomeadamente

$$\widehat{K}(X) = \frac{m_4}{m_2^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\widehat{\sigma}}\right)^4, \tag{3.18}$$

em que $\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$. Consequentemente, 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,

$$\widehat{K} \approx \mathcal{N}(3, 24/n), \tag{3.19}$$

ou seja, $\widehat{K}$ tem uma distribuição aproximadamente normal com média 3 e variância $24/n$.

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 base determinada por Q_1 e Q_3, como indicado na Figura 3.16. 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 $\max[x_{(1)}, Q_1 - 1{,}5 * d_Q]$ com d_Q representando a distância interquartis. Pontos colocados acima do limite superior ou abaixo do limite inferior, considerados **valores atípicos** ou **discrepantes** (*outliers*) são representados por algum símbolo ($*$, por exemplo).

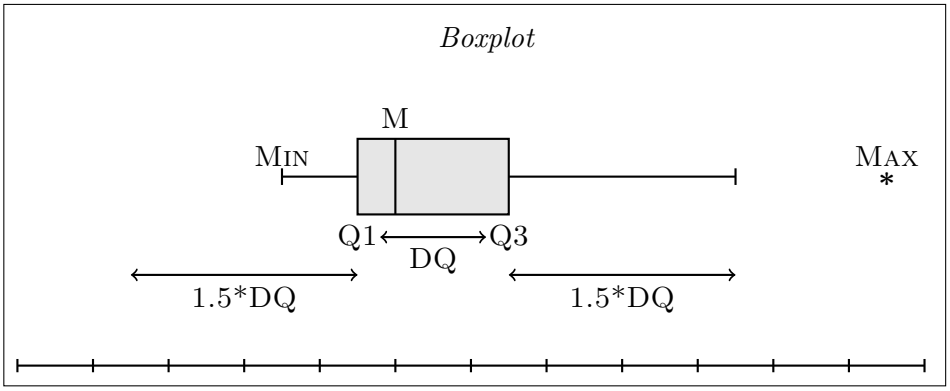

Figura 3.16: Detalhes para a construção de *boxplots*.

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 3.17 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 do *boxplot*, 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 3.18 apresentamos *boxplots* correspondentes àqueles da Figura 3.17 com os dentes (*notchs*) incorporados.

[3]Note que tanto a orientação horizontal (como na Figura 3.16) quanto a vertical (como na Figura 3.17) podem ser empregadas na construção dos *boxplots*.

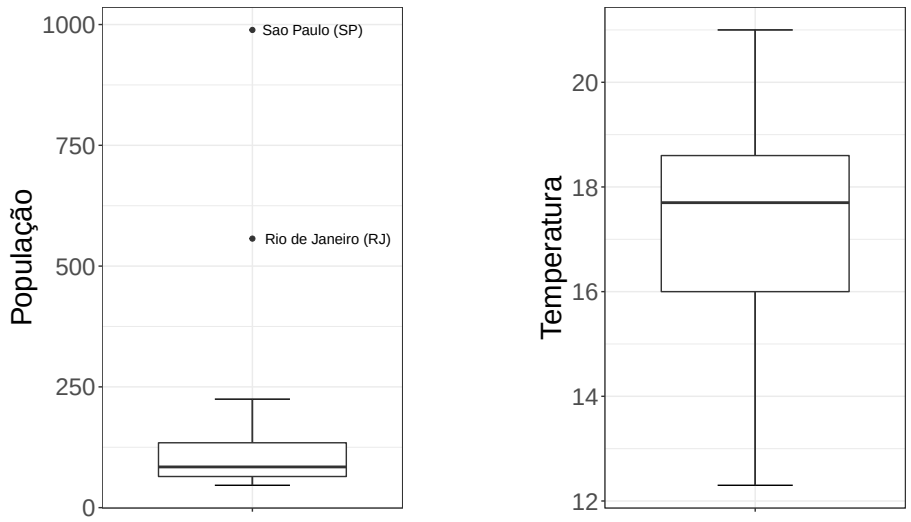

Figura 3.17: *Boxplots* para os dados dos Exemplos 3.2 (População) e 3.3 (Temperatura).

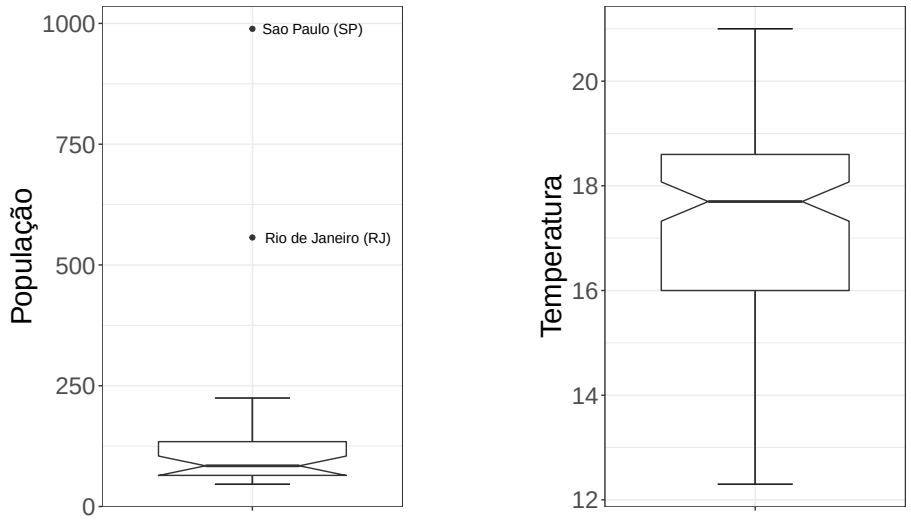

Figura 3.18: *Boxplots* dentados para os dados dos Exemplos 3.2 (População) e 3.3 (Temperatura).

3.5 Modelos probabilísticos

Um dos objetivos da Estatística é fazer inferência (ou tirar conclusões) sobre a distribuição de alguma variável em uma determinada população a partir de dados de 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 40000 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 com base num modelo probabilístico. Esse procedimento é denominado **inferência estatística**. No exemplo acima, poderíamos escolher uma amostra de 300 empregados da empresa e analisar a distribuição dos salários dessa amostra com o objetivo de tirar conclusões sobre a distribuição dos salários da população de 40000 empregados. Um dos objetivos da inferência estatística é quantificar a incerteza nessa generalização.

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 de uma população 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 descrita por um determinado modelo probabilístico representado por uma função, nosso problema reduz-se a estimar os **parâmetros** que a caracterizam. Para a distribuição normal, esses parâmetros são o valor esperado, usualmente representado por μ e a variância, usualmente representada por σ^2.

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

b) os parâmetros que caracterizam cada distribuição;

c) o valor esperado, representado por $\mathrm{E}(X)$ e a variância, representada por $\mathrm{Var}(X)$ de cada uma delas.

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	$\frac{1}{\sigma\sqrt{2\pi}}\exp\{(\frac{x-\mu}{\sigma})^2\},\ -\infty < x < \infty$	$\mu,\ \sigma$	μ, σ^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	$0,\ n/(n-2)$

Lembrando que num histograma, as frequências relativas correspondem a áreas de retângulos, num modelo probabilístico, as probabilidades correspondem a áreas sob regiões delimitadas pela função $f(x)$, como indicado na Figura 3.19.

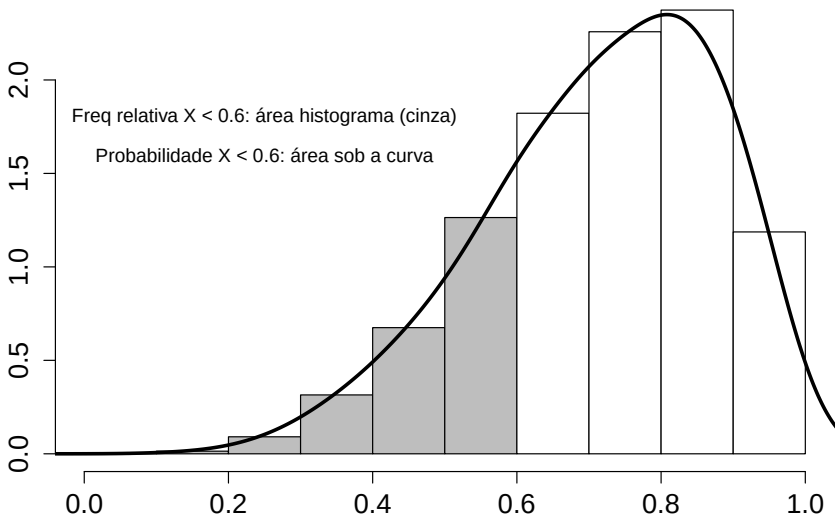

Figura 3.19: Aproximação de um histograma por uma função densidade de probabilidade.

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

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. 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 amostra aleatória simples 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 explicativas (preditoras) são controladas para produzir valores de uma variável resposta. As observações da resposta para as diferentes combinações dos níveis das variáveis explicativas constituem uma amostra estratificada. Exemplos são apresentados na Seção

4.4.

A não ser quando explicitamente indicado, para propósitos inferenciais, neste texto consideraremos os dados como provenientes de uma amostra aleatória simples.

Denotemos por $x_1, \ldots, x_n$ os valores efetivamente observados das variáveis $X_1, \ldots, X_n$[4]. Além disso, 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** $X_{(1)}, \ldots, X_{(n)}$.

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

$$F_n(x) = \frac{n(x)}{n}, \quad \forall x \in \mathbb{R}, \tag{3.20}$$

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

Com finalidade ilustrativa, considere novamente as observações do Exemplo 3.4 sem o valor 220. O gráfico de F_n que é essencialmente uma função em escada, com "saltos" de magnitude $1/n$ em cada $x_{(i)}$, nomeadamente

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

está disposto na Figura 3.20.

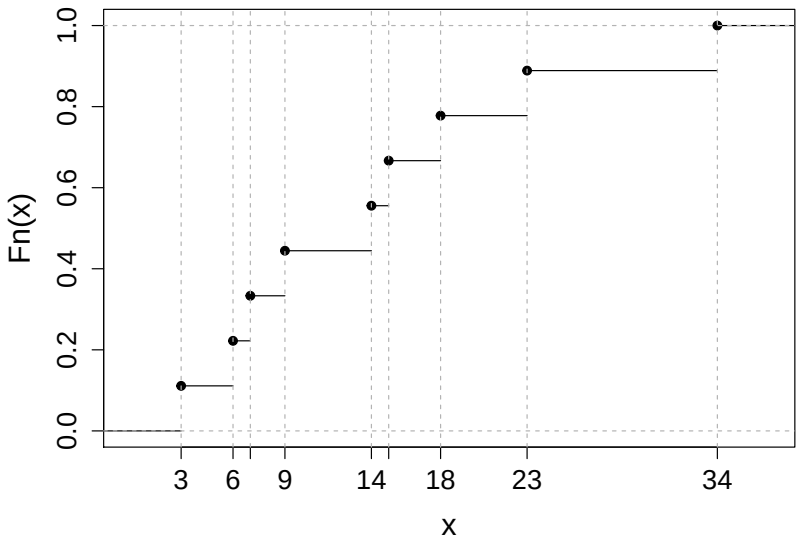

Figura 3.20: Função distribuição empírica para os dados do Exemplo 3.4 (sem o valor 220).

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

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 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 quando os quantis forem padronizados, *i.e.*,

$$Q^\star(p_i) = [Q(p_i) - \overline{x}]/dp(x).$$

Para quantis não padronizados, os pontos no gráfico estarão dispostos em torno de uma reta com inclinação diferente de 1.

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. Veja a Nota de Capítulo 5 deste capítulo e a Nota de Capítulo 5 do Capítulo 4.

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 3.21.

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

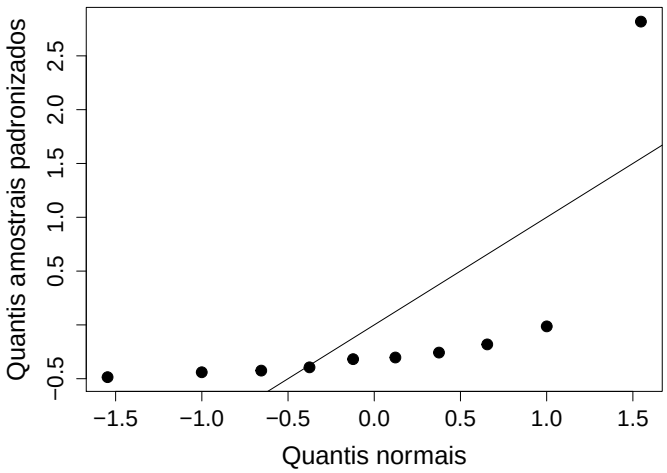

Figura 3.21: Gráfico QQ Normal para os dados do Exemplo 3.4 (com reta $y = x$).

Um exame da Figura 3.21 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 3.22 e indica que as evidências contrárias ao modelo normal são menos aparentes.

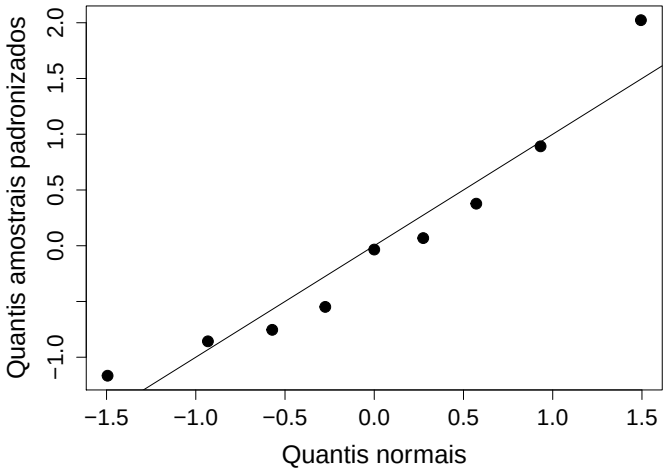

Figura 3.22: Gráfico QQ normal para os dados do Exemplo 3.4 com a eliminação do ponto 220 e reta $y = x$.

Um exemplo de gráfico QQ para uma distribuição amostral com 100 dados gerados a partir de uma distribuição normal padrão está apresentado na Figura 3.23.

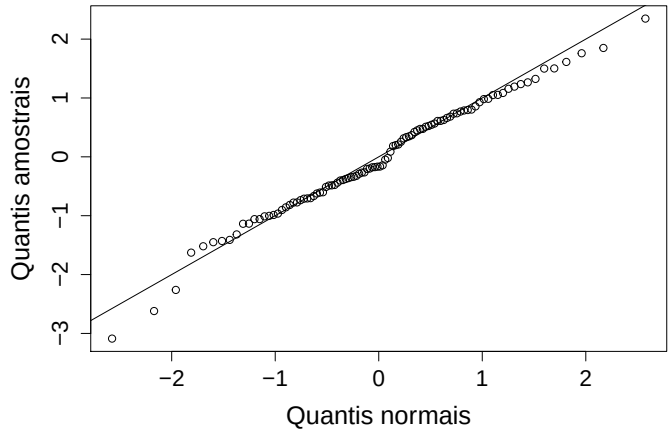

Figura 3.23: Gráfico QQ Normal para 100 dados gerados de uma distribuição normal padrão.

Embora os dados correspondentes aos quantis amostrais da Figura 3.23 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 3.24.

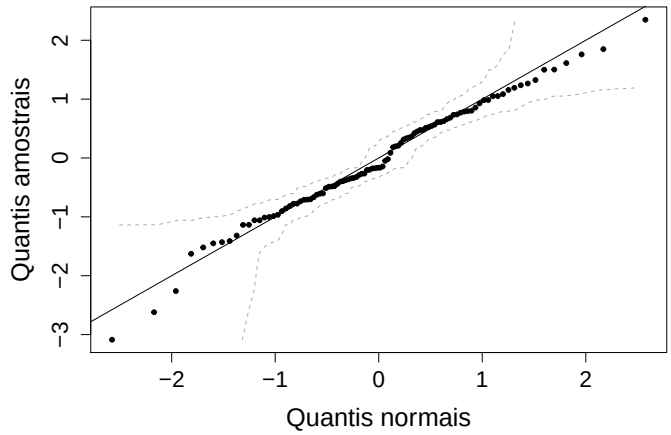

Figura 3.24: Gráfico QQ Normal para 100 dados gerados de uma distribuição normal padrão com bandas de confiança.

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 3.25.

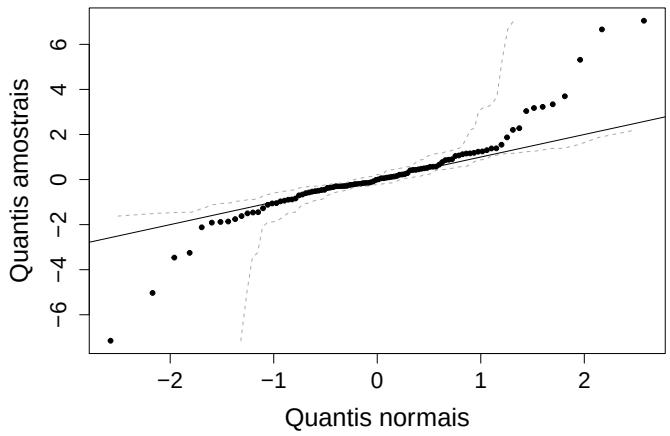

Figura 3.25: Gráfico QQ Normal para 100 dados gerados de uma distribuição t com 2 graus de liberdade.

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 3.26.

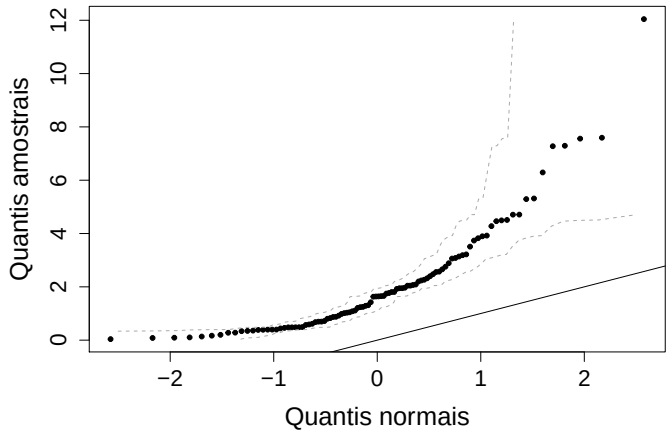

Figura 3.26: Gráfico QQ normal para 100 dados gerados de uma distribuição qui-quadrado com 2 graus de liberdade.

O gráfico QQ correspondente, agora obtido por meio dos quantis de uma distribuição qui-quadrado com 2 graus de liberdade é apresentado na Figura 3.27.

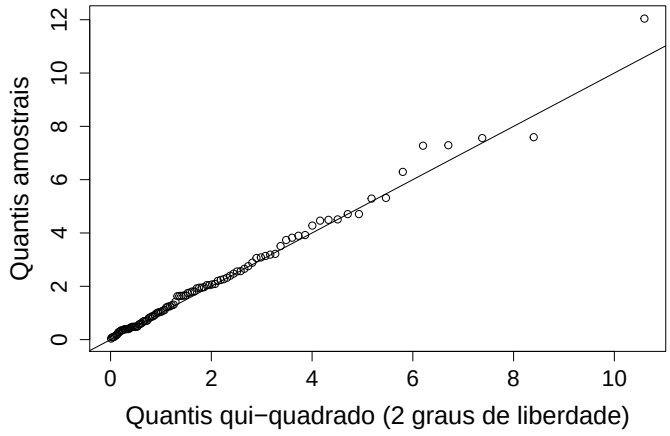

Figura 3.27: Gráfico QQ qui-quadrado para 100 dados gerados de uma distribuição qui-quadrado com 2 graus de liberdade.

3.8 Desvio padrão e erro padrão

Considere uma população para a qual a variável X tem valor esperado 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 o valor esperado 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.*, o valor esperado do conjunto das médias amostrais é igual ao valor esperado 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 seu valor esperado, que é o mesmo da variável original) será a sua distribuição. O desvio padrão da variável $\overline{X}$ é conhecido como **erro padrão** (da média).

Lembremos que, dada uma amostra $X_1, \ldots, X_n$, a variância σ^2 da variável X é estimada por

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Ao se aumentar o tamanho da amostra, o denominador $n-1$ aumenta, mas o numerador, $\sum_{i=1}^{n}(X_i - \overline{X})^2$, também aumenta de forma que o quociente e também sua raiz quadrada, que é uma estimativa do desvio padrão de X, permanecem estáveis a menos de flutuações aleatórias. Uma estimativa da variância da variável $\overline{X}$ é S^2/n, e dada a estabilidade do numerador S^2, esse

quociente, e consequentemente, sua raiz quadrada, que é o erro padrão da média $\overline{X}$, diminuem com o aumento do denominador n. Detalhes podem ser obtidos em Bussab e Morettin (2017).

Uma avaliação desses resultados por meio de simulação está apresentada na Figura 3.28.

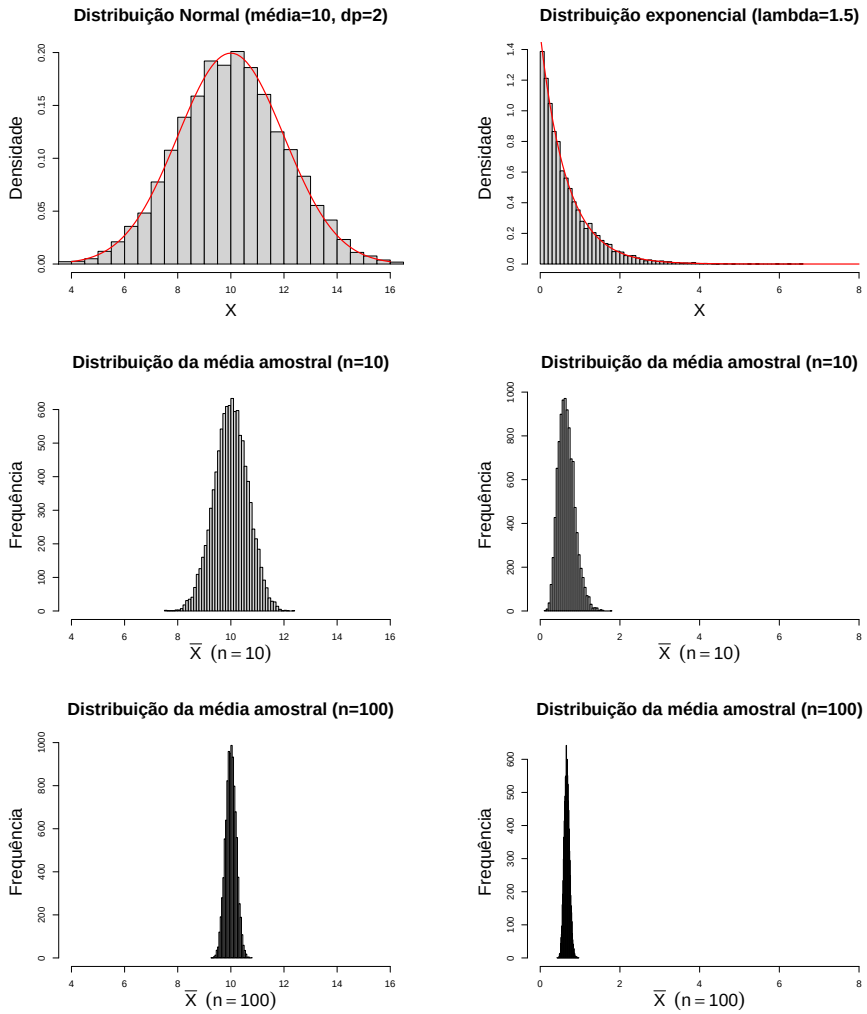

Figura 3.28: Efeito do tamanho da amostra na distribuição amostral da média.

Os histogramas apresentados na coluna da esquerda correspondem a dados simulados de uma distribuição normal com valor esperado $\mu = 10$ e desvio padrão $\sigma = 2$. O primeiro deles exibe um histograma obtido com 10000 dados e mimetiza a população. A média e o desvio padrão correspondentes são, respectivamente 10,009 e 2,011 que essencialmente representam os valores de μ e σ.

Geramos 10000 amostras de tamanhos $n = 10$ e 10000 amostras de tamanhos $n = 100$ dessa distribuição e para cada uma delas, calculamos as

médias ($\overline{X}$) e desvios padrões (S). Cada uma dessas 10000 médias $\overline{X}$ é uma estimativa do valor esperado populacional μ e cada um dos desvios padrões amostrais S é uma estimativa do desvio padrão populacional σ. Os outros dois gráficos exibidos na mesma coluna correspondem aos histogramas das médias (amostrais) das $M = 10000$ amostras para $n = 10$ e $n = 100$, e salientam o fato de que a dispersão das médias amostrais em torno de sua média, (que é essencialmente o valor esperado da população de onde foram extraídas) diminui com o aumento do tamanho das amostras. Médias e desvios padrões (erros padrões) dessas distribuições amostrais, além das médias dos desvios padrões amostrais estão indicados na Tabela 3.11.

Tabela 3.11: Medidas resumo de 10000 amostras de uma distribuição normal com média 10 e desvio padrão 2

Amostras	média($\overline{X}$)	ep($\overline{X}$)	média(S)
$n = 10$	9,997	$0{,}636 \approx 2/\sqrt{10}$	$1{,}945 \approx 2$
$n = 100$	9,999	$0{,}197 \approx 2/\sqrt{100}$	$1{,}993 \approx 2$

Os histogramas apresentados na coluna da direita da Figura 3.28 correspondem a dados simulados de uma distribuição exponencial com parâmetro $\lambda = 1{,}5$ para a qual o valor esperado é $\mu = 0{,}667 = 1/1{,}5$ e o desvio padrão também é $\sigma = 0{,}667 = 1/1{,}5$. O primeiro deles exibe um histograma obtido com 10000 dados e mimetiza a população. A média e o desvio padrão correspondentes são, respectivamente 0,657 e 0,659 que essencialmente são os valores de μ e σ.

Os outros dois gráficos exibidos na mesma coluna correspondem aos histogramas das médias (amostrais) de $M = 10000$ amostras de tamanhos $n = 10$ e $n = 100$, respectivamente da mesma distribuição exponencial, salientando o fato de que a dispersão das médias amostrais em torno de sua média (que é essencialmente o valor esperado da distribuição original) diminui. Além disso, esses histogramas mostram que a distribuição das médias amostrais é aproximadamente normal apesar de a distribuição de onde as amostras forma geradas ser bastante assimétrica. Médias e desvios padrões (erros padrões) dessas distribuições amostrais além das médias dos desvios padrões amostrais estão indicados na Tabela 3.12.

Tabela 3.12: Medidas resumo de 10000 amostras de uma distribuição exponencial com parâmetro $\lambda = 1{,}5$ (média=0,667 e desvio padrão=0,667)

Amostras	média($\overline{X}$)	ep($\overline{X}$)	média(S)
$n = 10$	0,669	$0{,}212 \approx 0{,}667/\sqrt{10}$	$0{,}617 \approx 0{,}667$
$n = 100$	0,666	$0{,}066 \approx 0{,}667/\sqrt{100}$	$0{,}659 \approx 0{,}667$

3.9 Intervalo de confiança e tamanho da amostra

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 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 o valor esperado 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 $me = kS/\sqrt{n}$ 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 supomos um modelo normal, a constante k para um intervalo com **coeficiente de confiança** de 95,4% é igual a 2. Esse valor corresponde à área sob a curva normal entre o valor esperado menos dois desvios padrões ($\mu - 2\sigma$) e o valor esperado mais dois desvios padrões ($\mu + 2\sigma$) como indicado na Figura 3.29.

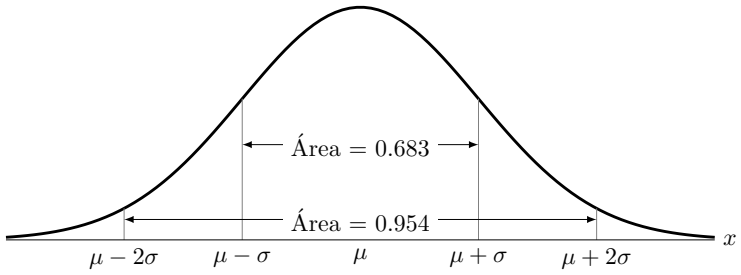

Figura 3.29: Áreas sob a curva normal com valor esperado μ e desvio padrão σ

Para um coeficiente de confiança igual a 95%, $k = 1{,}96$. Quando o tamanho da amostra é suficientemente grande, podemos utilizar esse valor, mesmo quando a distribuição de onde foi obtida a amostra não é normal. A margem de erro correspondente a um coeficiente de confiança de 95% é $me = 1{,}96S/\sqrt{n}$. Com base nessa margem de erro, podemos construir um **intervalo de confiança** para o valor esperado populacional da variável X. Os limites inferior e superior para esse intervalo são, respectivamente,

$$\overline{X} - 1{,}96S/\sqrt{n} \text{ e } \overline{X} + 1{,}96S/\sqrt{n}. \tag{3.21}$$

Se considerássemos um grande número de amostras dessa população sob as mesmas condições, o intervalo construído dessa maneira conteria o verdadeiro (mas desconhecido) valor esperado populacional (μ) em 95% dos casos. Dizemos então, que o intervalo de confiança tem confiança de 95%.

À guisa de exemplo, 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 $\widehat{p}(1-\widehat{p})/n$ (veja o 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,

$$\widehat{p} - 1{,}96\sqrt{\widehat{p}(1-\widehat{p})/n} \text{ e } \widehat{p} + 1{,}96\sqrt{\widehat{p}(1-\widehat{p})/n}. \tag{3.22}$$

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%.

Uma das perguntas mais frequentes com que o estatístico é defrontado é: *Qual é o tamanho da amostra necessário para que meus resultados sejam (estatisticamente) válidos?* Embora a pergunta faça sentido para quem a apresentou, ela não reflete exatamente o que se deseja e precisa ser reformulada para fazer algum sentido passível de análise estatística. Nesse contexto,

a pergunta apropriada seria: *Qual é o tamanho da amostra necessário para que meus resultados tenham uma determinada precisão?*

Especificamente no caso da estimação da média (populacional) μ de uma variável X, a pergunta seria *Qual é o tamanho da amostra necessário para que a estimativa $\overline{X}$ da média μ tenha uma precisão ε?* A resposta pode ser obtida da expressão (3.21), fazendo $\varepsilon = 1{,}96S/\sqrt{n}$ o que implica

$$n = (1{,}96S/\varepsilon)^2.$$

Então, para a determinação do tamanho da amostra, além da precisão desejada ε é preciso ter uma ideia sobre o desvio padrão S da variável sob investigação. Essa informação pode ser obtida por meio de uma **amostra piloto** ou por intermédio de outro estudo com características similares àquelas do estudo que está sendo planejado. Na falta dessas informações, uma estimativa grosseira do desvio padrão pode ser obtida como $S = [\max(X) - \min(X)]/4$ ou $S = [\max(X) - \min(X)]/6$ (veja a Figura 3.29).

Para estimação de proporções amostrais, o cálculo do tamanho da amostra pode ser baseado em (3.22). Nesse caso, $\varepsilon = 1{,}96\sqrt{\widehat{p}(1-\widehat{p})/n}$ o que implica

$$n = [1{,}96\sqrt{\widehat{p}(1-\widehat{p})}/\varepsilon]^2.$$

Observando que o valor máximo de $\widehat{p}(1-\widehat{p})$ é 0,25 e aproximando 1,96 por 2, o tamanho da amostra necessário para que a precisão aproximada da estimativa da proporção populacional seja de ε pontos percentuais é $n = 1/\varepsilon^2$. Portanto para que a precisão seja de 3 pontos percentuais como em pesquisas de intenção de votos, em geral, o tamanho da mostra deverá ser aproximadamente $n = 1/(0.03)^2 = 1111$.

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

3.10 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

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \log(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases} \tag{3.23}$$

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, *boxplots*) 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

$$d_p = \frac{\text{média} - \text{mediana}}{\text{medida de escala}}, \tag{3.24}$$

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

Exemplo 3.5. Consideremos a variável concentração de Fe em cascas de árvores da espécie *Tipuana tipu* disponível no arquivo `arvores`. Nas Figuras 3.30 e 3.31 apresentamos, respectivamente *boxplots* e histogramas para os valores originais da variável, assim como para seus valores transformados por (3.23) 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 de uma distribuição simétrica.

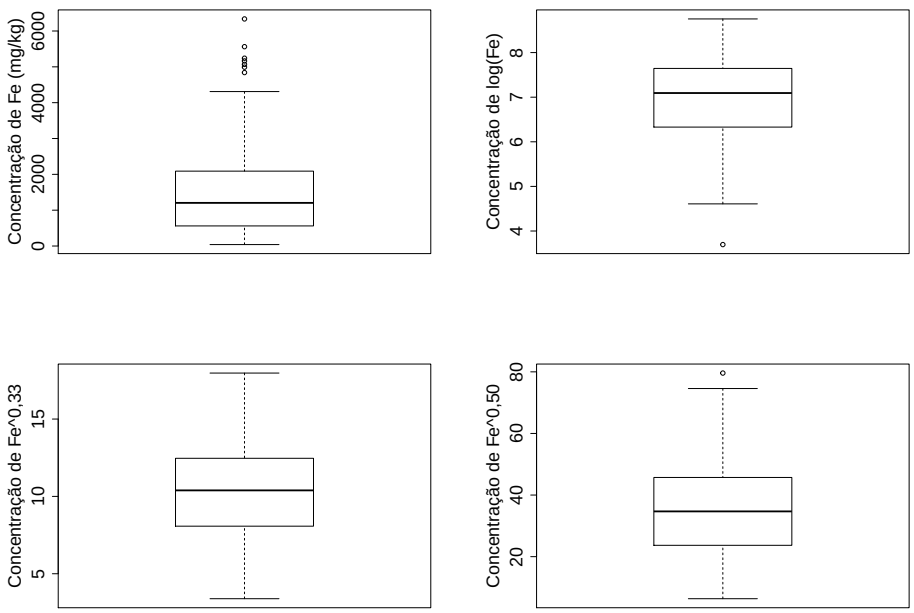

Figura 3.30: *Boxplots* com variável transformada.

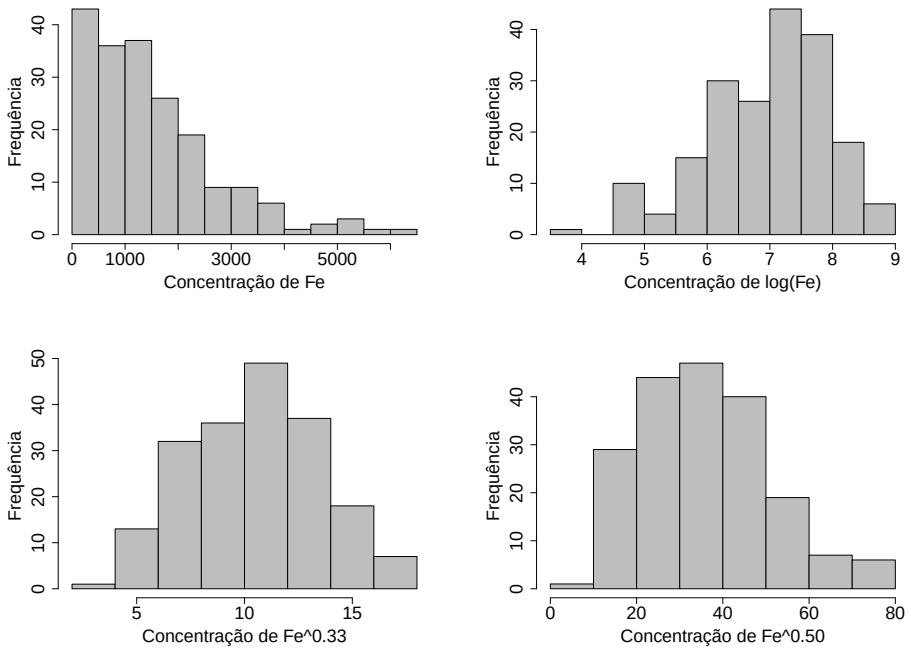

Figura 3.31: Histogramas com variável transformada.

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. 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 (3.23) 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 (3.23) é um caso particular das **transformações de Box-Cox** que são da forma

$$g(x) = \begin{cases} (x^p - 1)/p, & \text{se } p \neq 0 \\ \log(x), & \text{se } p = 0. \end{cases} \tag{3.25}$$

Veja Box e Cox (1964) para detalhes.

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

$$h = 1{,}349\widetilde{S}\left(\frac{\log n}{n}\right)^{1/3} \tag{3.26}$$

em que $\widetilde{S}$ é um estimador "robusto" do desvio padrão de X, minimiza o desvio máximo absoluto entre o histograma e a verdadeira densidade $f(x)$. Em (3.26). Veja a Nota de Capítulo 4 para detalhes sobre estimadores robustos do desvio padrão. O pacote `R` usa como *default* o valor de h sugerido por Sturges (1926), dado por

$$h = \frac{W}{1 + 3{,}322\log(n)}, \tag{3.27}$$

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

3) **Definição de histograma**

Consideremos um exemplo com K classes de amplitudes 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

$$\tilde{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} I(x - \tilde{x}_i; h/2), \tag{3.28}$$

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 3.32.

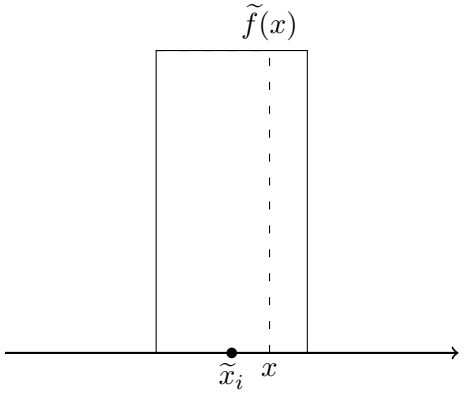

Figura 3.32: Detalhe para a construção de histogramas.

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

Pode-se verificar que, para uma distribuição normal e relação entre a distância interquartis d_Q e o desvio padrão σ satisfaz

$$d_Q = 1{,}349\sigma.$$

Logo, um estimador "robusto" para o desvio padrão populacional é

$$\widetilde{S} = d_Q/1{,}349.$$

Observe que substituindo $\widetilde{S}$ em (3.26), 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 30) 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

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

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

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

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

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

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 das variáveis `Peso`, `Altura`, `Coltot`, `HDL`, `LDL`, `Lesoes`, `Diabete` e `HA`. Com essa finalidade,

   a) Construa distribuições de frequências para as variáveis qualitativas.

   b) Construa histogramas, *boxplots* e gráficos de simetria para as variáveis contínuas.

   c) Construa uma tabela com medidas resumo para as variáveis contínuas.

   d) Avalie a compatibilidade de distribuições normais para as variáveis contínuas por meio de gráficos QQ.

2) Considere os dados do arquivo `antracose`.

   a) Construa uma tabela com as medidas de posição e dispersão estudadas para as variáveis desse arquivo.

   b) Construa histogramas e *boxplots* para essas variáveis e verifique que transformação é necessária para tornar mais simétricas aquelas em que a simetria pode ser questionada.

3) Considere as variáveis `Peso` e `Altura` de homens do conjunto de dados `rehabcardio`. Determine o número de classes para os histogramas correspondentes por meio de (3.26) e (3.27) e construa-os.

4) Considere o arquivo `vento`. Observe o valor atípico 61,1, que na realidade ocorreu devido a uma forte tempestade no dia 2 de dezembro. Calcule as medidas de posição e dispersão apresentadas na Seção 3.3. Quantifique o efeito do valor atípico indicado nessas medidas.

5) Construa gráficos ramo-e-folhas e *boxplot* para os dados do Exercício 4.

6) Transforme os dados do Exercício 4 por meio de (3.23) 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 (3.24).

7) Analise a variável `Temperatura` do arquivo `poluicao`.

8) Analise a variável `Salário de administradores`, disponível no arquivo `salarios`.

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

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

11) 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 distribuições simétricas. Calcule-a para os dados do Exercício 4.

12) Os dados disponíveis no arquivo `endometriose` são provenientes de um estudo em que o objetivo é verificar se existe diferença entre os grupos de doentes e controles quanto a algumas características observadas.

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

   b) Compare as distribuições das variáveis idade e concentração de PCR durante a menstruação (PCRa) para pacientes 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. Utilize um método gráfico para representá-la.

13) 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.

14) 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.

15) Os dados apresentados na Figura 3.33 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 3.33: Planilha com instantes de realização de chamados solicitando auxílio em estradas.

a) Construa um histograma para a distribuição de frequências dos instantes 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 e 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.

16) 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) Calcule 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 das notas desses dois grupos.

17) 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	85
Fraca	30	73,1	70	9,6	58	92	63	77

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

Indique se as seguintes afirmações estão corretas, justificando a sua respostas:

a) 5% e 50% dos estudantes com atividade física intensa e fraca, respectivamente, tiveram pulsação inferior a 70.

b) A proporção de estudantes com fraca atividade física com pulsação inferior a 63 é menor que a proporção de estudantes com atividade física intensa com pulsação inferior a 70.

c) A atividade física não tem efeito na média da pulsação dos estudantes.

d) Mais da metade dos estudantes com atividade física intensa têm pulsação maior que 82.

18) Considere os gráficos *boxplot* da Figura 3.34. Quais deles correspondem às pulsações dos estudantes submetidos a atividade física intensa e fraca?

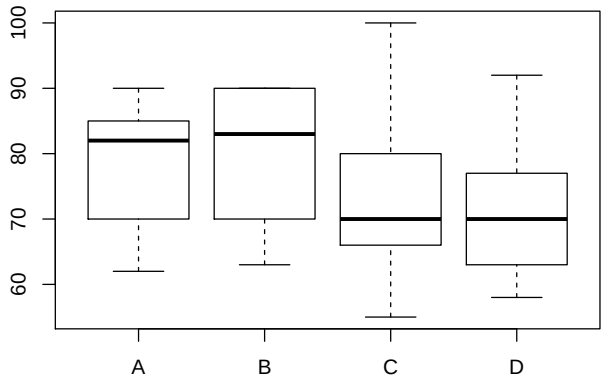

Figura 3.34: *Boxplots* para o Exercício 18.

a) A e B b) B e D c) A e C d) B e C

19) Os histogramas apresentados na Figura 3.35 mostram a distribuição das temperaturas (ºC) ao longo de vários dias de investigação para duas regiões (R1 e R2). Indique se as afirmações abaixo estão corretas, justificando as respostas:

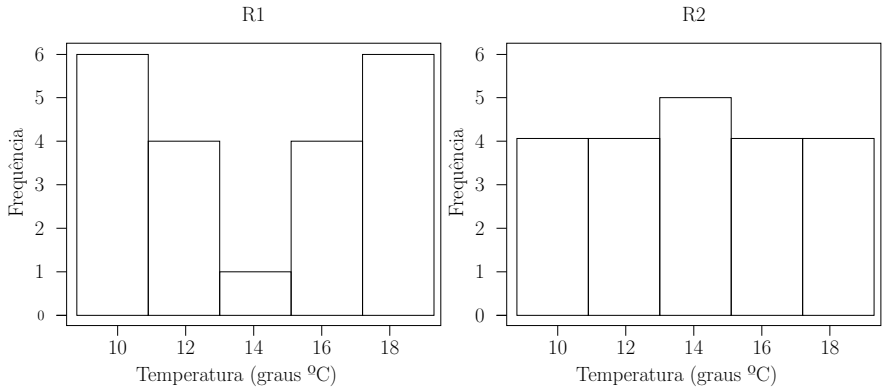

Figura 3.35: Histogramas para o Exercício 19

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

b) Não é possível comparar 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.

20) Na companhia A, a média dos salários é 10000 unidades e o 3º quartil é 5000. Responda as seguintes perguntas, justificando a sua respostas.

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?

21) 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.

22) 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.

23) A tabela abaixo representa a distribuição do número de dependentes por empregado de uma determinada empresa.

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

A mediana, média e moda cujos valores calculados por quatro estagiários, foram:

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

Indique qual deles está correto, justificando sua resposta.

24) Com relação ao Exercício 23, qual a porcentagem de empregados da empresa com 2 ou mais dependentes?

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

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

Tabela 3.13: 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.

26) 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.

27) 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 3.14: 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.

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

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

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

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

30) 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.

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

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 comparando-os com aqueles preconizados pela teoria.

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

> Life is complicated, but not uninteresting.
>
> Jerzy Neyman

4.1 Introdução

Neste capítulo trataremos da análise descritiva da **associação** entre duas variáveis. De maneira geral, dizemos que existe associação entre duas variáveis se o conhecimento do valor de uma delas nos dá alguma informação sobre alguma característica da distribuição (de frequências) da outra. Podemos estar interessados, por exemplo, na associação entre o grau de instrução e o salário de um conjunto de indivíduos. Diremos que existe associação entre essas duas variáveis se o salário de indivíduos com maior nível educacional for maior (ou menor) que os salários de indivíduos com menor nível educacional. 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 ou populacionais) 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 (ou **preditoras**) são aquelas cujas categorias ou valores são fixos, seja por planejamento, seja por condicionamento. **Variáveis respostas** são aquelas cujas categorias ou valores são aleatórios.

Num estudo em que se deseja avaliar o efeito do tipo 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 um de 4 tipos de aditivo. O consumo (em km/L) foi avaliado após um determinado período de tempo. Nesse contexto, a variável qualitativa "Tipo 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) observadas em 50 pacientes, ambas codificadas com os atributos 0=não tem e 1=tem. Nesse contexto, as duas variáveis são classificadas como respostas. A Tabela 4.1 contém a **distribuição de frequências conjunta** das duas variáveis.

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

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

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.

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

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

Os dados da Tabela 4.2 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.

Tabela 4.3: Porcentagens com totais nas colunas

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

Com base nessa tabela, podemos 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 planeja-**

mento. Nesse contexto, apenas as porcentagens calculadas como na Tabela 4.3 fariam sentido. Um enfoque análogo poderia ser adotado se fixássemos as frequências de "Insuficiência cardíaca" e considerássemos "Hipertensão arterial" como variável resposta. Nesse cenário, as porcentagens deveriam ser calculadas em relação ao total das linhas da tabela.

Num estudo idealizado para avaliar a preferência pela cor de um certo produto, selecionaram-se 200 mulheres e 100 homens aos quais foi perguntado se o preferiam nas cores vermelha ou preto. Os resultados estão dispostos na Tabela 4.4.

Tabela 4.4: Distribuição conjunta das variáveis Sexo e Cor de preferência

	Cor preferida		Total
Sexo	Vermelho	Preto	
Feminino	90	110	200
Masculino	30	70	100
Total	120	180	300

Neste caso, Sexo é uma variável explicativa, pois as frequências foram fixadas por planejamento e Cor preferida é a variável resposta. Só faz sentido calcular as frequências relativas tendo como base o total das linhas, como indicado na Tabela 4.5.

Tabela 4.5: Frequências relativas para a Tabela 4.4

	Cor preferida		Total
Sexo	Vermelho	Preto	
Feminino	45%	55%	100%
Masculino	30%	70%	100%
Total	40%	60%	100%

A frequência relativa de mulheres (67%) e de homens (33%) no conjunto de dados não reflete a distribuição de frequências de Sexo na população.

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 variável é considerada explicativa, por condicionamento 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**. Esse também é o cenário dos dados da Tabela 4.4, em que a variável Sexo é uma variável explicativa por planejamento e Cor preferida é a variável resposta.

Nas Tabelas 4.2 ou 4.3, por exemplo, há diferenças, que parecem não ser "muito grandes", o que nos leva a conjecturar que **na 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), ou seja, para a hipótese

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

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 a 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 (ou aproximadamente 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**, ou seja, admitindo que as frequências relativas de pacientes com ou sem hipertensão fossem, respectivamente, 64% e 36%, independentemente de terem ou não insuficiência cardíaca. Por exemplo, o valor 10,2 corresponde a 64% de 16, 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 4.6: Valores esperados das frequências na Tabela 4.3 sob H

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

Denotando 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,

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

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 $(r-1)(c-1)$ para tabelas $r \times c$ de forma que a decisão de rejeitar ou não a hipótese pode ser baseada nessa distribuição. Para o exemplo, o valor $\chi^2 = 1{,}3$ deve ser comparado com quantis da distribuição qui-quadrado com 1 grau de liberdade. 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, $T = 0{,}16$.

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[1]. 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%

Sem grande rigor, 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 os dados 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.

[1]Essa característica é conhecida como maldição (ou praga) da dimensionalidade.

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%

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), espera-se que elas sejam 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 4.10.

Tabela 4.10: 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. Uma estimativa do índice denominado κ de Cohen (1960), construído com esse propósito é

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

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 da estimativa do índice de Cohen é $\widehat{\kappa} = 1$. Quando a concordância entre as duas variáveis é menor do que a esperada pelo acaso, $\widehat{\kappa} < 0$. Uma estimativa da variância de $\widehat{\kappa}$ obtida por meio do método Delta é

$$\begin{aligned}\text{Var}(\widehat{\kappa}) &= n^{-1}\left\{\frac{\sum_i p_{ii}(1-\sum_i p_{ii})}{(1-\sum_i p_{ii})^2}\right. \\ &\quad + \frac{2(1-\sum_i p_{ii})[2\sum_i p_{ii}\sum_i p_{i+}p_{+i} - \sum_i p_{ii}(p_{i+}+p_{+i})]}{(1-\sum_i p_{i+}p_{+i})^3} \\ &\quad \left. + \frac{(1-\sum_i p_{ii})^2[\sum_i\sum_j p_{ij}(p_{i+}+p_{+j})^2 - 4\sum_i p_{i+}p_{+i}]}{(1-\sum_i p_{i+}p_{+i})^4}\right\}\end{aligned}$$

Para os dados da Tabela 4.10, em que há forte evidência de associação entre as duas variáveis ($\chi^2 = 444{,}02, gl = 4, p < 0{,}001$) temos $\widehat{\kappa} = 0{,}87$ sugerindo uma "boa" concordância entre as avaliações dos dois professores[2]. Na Tabela 4.11, por outro lado, temos $\kappa = -0{,}41$, sugerindo uma concordância muito fraca entre as duas variáveis, embora também haja forte evidência de associação ($\chi^2 = 652{,}44, gl = 4, p < 0{,}001$).

Tabela 4.11: Frequências de redações classificadas por dois professores

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

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 regras de classificação, como aquela proposta por Viera and Garrett (2005) e reproduzida na Tabela 4.12

Tabela 4.12: 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 con-

[2]Para uma interpretação ingênua sobre o significado do valor-p, consulte a Nota de Capítulo 8.

siderar o índice κ ponderado, cuja estimativa é

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

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, $w_{12} = w_{21} = w_{23} = w_{32} = 1 - 1/2 = 1/2$, $w_{13} = w_{31} = 1 - 2/2 = 0$.

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 (hipotéticos) 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 4.13. 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 4.13: 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 esquematizada como indicado na Tabela 4.14.

Tabela 4.14: 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 probabilidade[3] de que indivíduos que **sabemos** ser não fumantes contraírem câncer pulmonar; analogamente, π_1 corresponde à probabilidade de que indivíduos que **sabemos** ser fumantes contraírem câncer pulmonar.

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 (ou riscos) 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*) : $\omega = [\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$, que corresponde ao quociente entre as chances de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.[4]

No exemplo da Tabela 4.13 essas medidas de associação podem ser estimadas como

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

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

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

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

Em geral, embora 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

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

[3]O termo "frequência relativa" é substituído por "probabilidade" quando nos referimos às características populacionais (veja a Seção 3.5).

[4]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.

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 determina (*a posteriori*) 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 como variável explicativa e o hábito tabagista, como variável resposta. A Tabela 4.15 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 4.15: 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 4.16 representa as probabilidades pertinentes.

Tabela 4.16: 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 probabilidade de que indivíduos que **sabemos** não ter câncer pulmonar serem fumantes; analogamente, p_1 corresponde à probabilidade que indivíduos que **sabemos** ter câncer pulmonar serem não fumantes. 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 para fumantes e não fumantes, respectivamente. 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 4.15 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 estimativa da 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 4.15, não é possível estimar 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 4.17.

Tabela 4.17: Frequência de pacientes submetidos a um teste diagnóstico

Verdadeiro status	Resultado do teste		Total
	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 $[FP = P(T+|ND)]$ e pode ser estimada por $fp = n_{21}/n_{2+}$;

iv) **Falso negativo**: corresponde à probabilidade de resultado negativo para pacientes doentes $[FN = P(T-|D)]$ e pode ser estimada por $fn = n_{12}/n_{1+}$;

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 $ac = (n_{11}+n_{22})/n$.

Estimativas das variâncias dessas características estão apresentadas na Nota de Capítulo 7.

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 o 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 na população) 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 4.18 contém os dados (hipotéticos) da comunidade em que a doença é menos prevalente e a Tabela 4.19 contém os dados (hipotéticos) da comunidade em que a doença é mais prevalente.

Tabela 4.18: Frequência de pacientes submetidos a um teste diagnóstico (prevalência da doença = 15%)

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

Tabela 4.19: 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 4.20

Tabela 4.20: Características do teste aplicado aos dados das Tabelas 4.18 e 4.19

	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 4.21, 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 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.

Tabela 4.21: 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

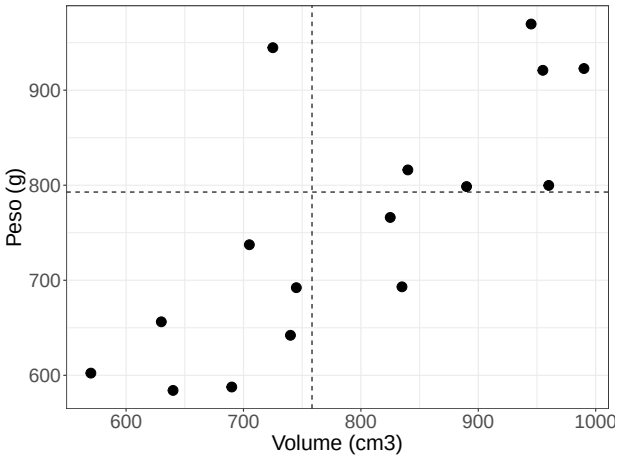

Figura 4.1: Gráfico de dispersão entre peso e volume do lobo direito de enxertos de fígado.

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**.[5] Mais especifica-

[5]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 $x_1 = 4$,

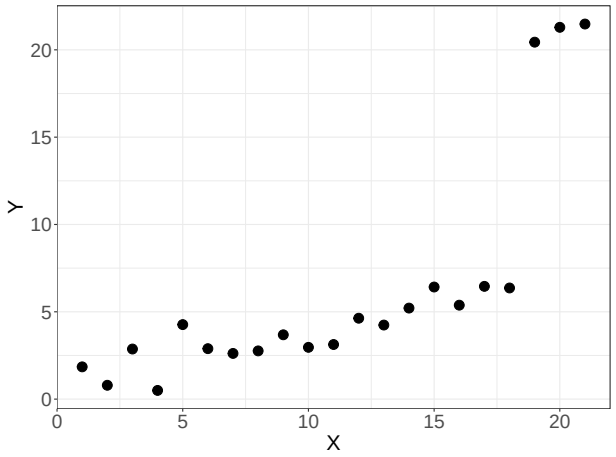

Figura 4.2: Gráfico de dispersão entre valores de duas variáveis X e Y.

mente, 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

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

Os dados correspondentes à Figura 4.2 foram gerados a partir da expressão $y_i = 1 + 0{,}25x_i + e_i$ com e_i simulado a partir de uma distribuição normal padrão e com as três últimas observações acrescidas de 15. Para esses dados obtemos $r_P = 0.768$ e $r_S = 0.926$. Eliminando as três observações com valores discrepantes, os coeficientes de correlação correspondentes são $r_P = 0.881$ e $r_S = 0.882$, indicando que o coeficiente de Spearman é mais sensível a associações não lineares.

Em resumo, o coeficiente de correlação de Spearman 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 $y_i = \exp(0.4x_i)$, $i = 1, \ldots, 20$.

$x_2 = 7$, $x_3 = 5$, $x_4 = 13$, $x_5 = 6$, $x_6 = 5$, 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 $2{,}5 = (2+3)/2$.

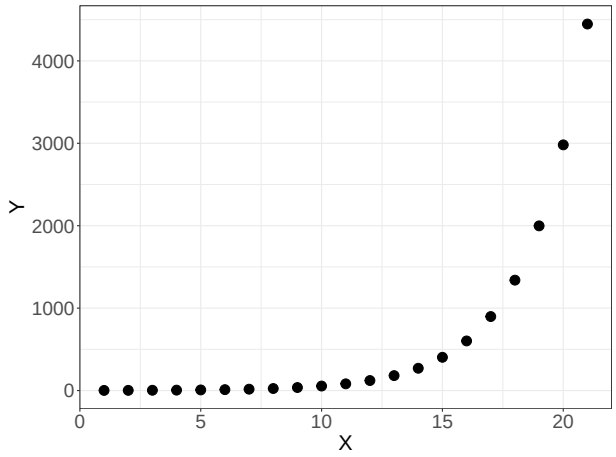

Figura 4.3: Gráfico de dispersão entre valores de duas variáveis X e Y.

Nesse caso, os valores dos coeficientes de correlação de Pearson e de Spearman são, respectivamente, $r_P = 0.742$ e $r_S = 1$ indicando que apenas este último é capaz de realçar a associação perfeita entre as duas variáveis.

Partição e janelamento

Um gráfico de dispersão para as variáveis concentração de monóxido de carbono (`CO`) e temperatura `temp` disponíveis no arquivo `poluicao` está apresentado na Figura 4.4 e não evidencia uma associação linear entre as duas variáveis. Para avaliar uma possível associação não linear é possível considerar diferentes medidas resumo de uma das variáveis (`temp`, por exemplo) para diferentes intervalos de valores da segunda variável, (`temp`, por exemplo) com o mesmo número de pontos, como indicado na Tabela 4.22.

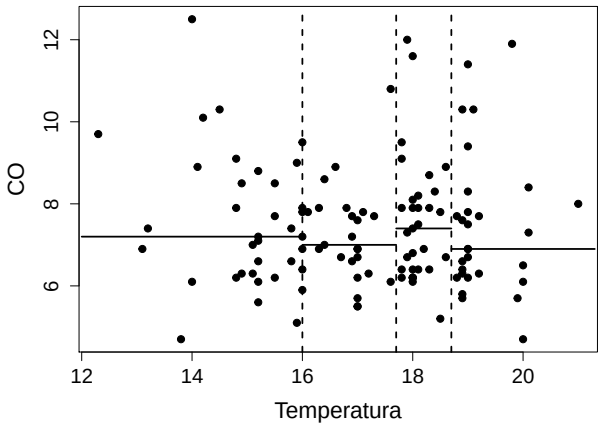

Figura 4.4: Gráfico de dispersão entre valores de duas variáveis X e Y.

Tabela 4.22: Medidas resumo para `CO` em diferentes intervalos de `temp`

Medida	Intervalo de temperatura			
resumo (CO)	12,0 - 16,0	16,0 - 17,7	17,7 - 18,7	18,7 - 21,0
n_i	29	31	31	29
Q_1	6,3	6,5	6,4	6,3
Mediana	7,2	7,0	7,4	6,9
Q_3	8,8	7,8	8,2	8,3
Média	7,6	7,2	7,6	7,5

Outra alternativa é considerar *boxplots* para `CO` em cada intervalo de temperatura como na Figura 4.5.

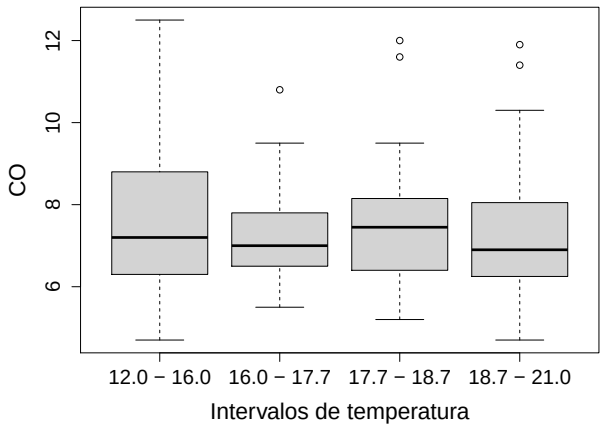

Figura 4.5: *Boxplots* para CO em diferentes intervalos de temperatura.

Embora a mediana de `CO` varie entre os diferentes intervalos de temperatura, não há uma indicação clara de que essa variação indique uma associação clara entre as variáveis. Os *boxplots* da Figura 4.5 sugerem que a variabilidade da concentração de CO é maior para o intervalo de temperaturas de 12 a 16 graus, atingindo aqui níveis maiores do que nos outros intervalos. Além disso, essa análise mostra que tanto o primeiro quartil quanta a mediana de `CO` são aproximadamente constantes ao longo dos intervalos, o que não acontece com o terceiro quartil.

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 segmen-

tos 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

```
> 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 4.6 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 perfil do atleta 9 é possivelmente atípico *outlier*. Na realidade, verificou-se que esse atleta era velocista e não fundista.

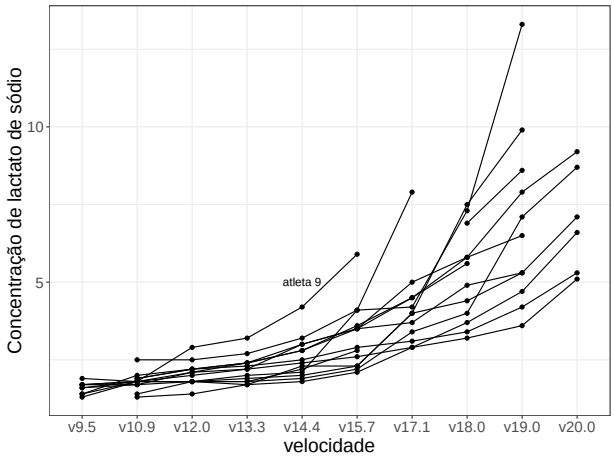

Figura 4.6: Gráfico de perfis individuais para os dados do Exemplo 4.4 (atletas fundistas).

Gráficos QQ para comparação de duas distribuições amostrais

Uma ferramenta adequada para comparar as distribuições de uma variável 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 forem iguais, 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, $y_{(1)} \leq \cdots \leq y_{(n)}$.

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,

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

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 4.7.

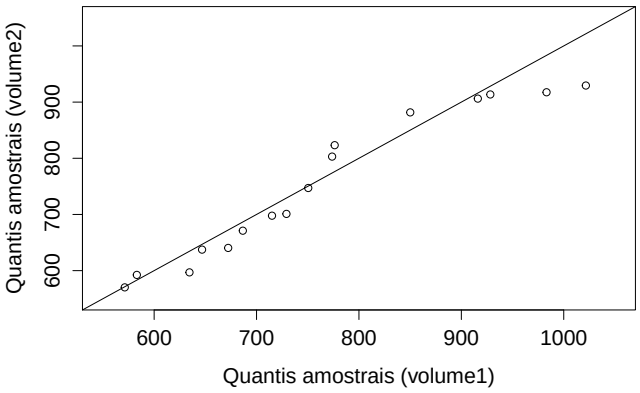

Figura 4.7: Gráfico QQ para avaliação da concordância de duas medidas ultrassonográficas do lobo direito do fígado.

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 de Cohen 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 variável.

Note que enquanto gráficos QQ são construídos a partir dos 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 do lobo direito do fígado medidos pelos dois observadores e indicados na Tabela 4.21 está apresentado na Figura 4.8.

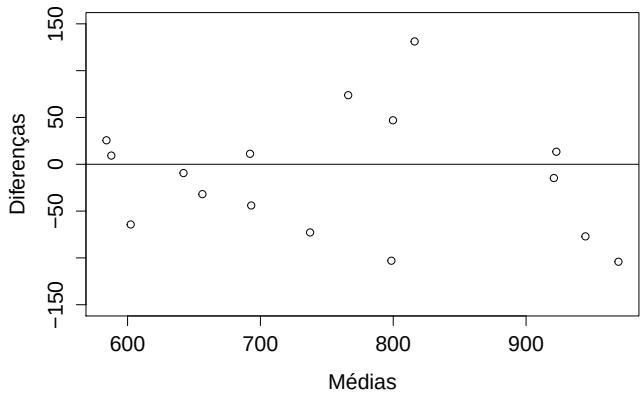

Figura 4.8: Gráfico de médias/diferenças de Tukey (Bland-Altman) para avaliação da concordância de duas medidas ultrassonográficas do lobo direito do fígado.

Os pontos no gráfico da Figura 4.8) 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 4.23 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 4.23: 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 4.23 estão apresentados nas Figuras 4.9 e 4.10.

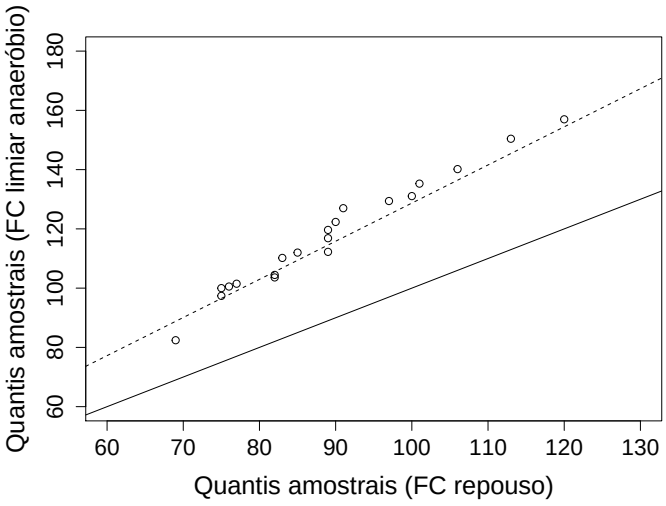

Figura 4.9: Gráfico QQ para comparação das distribuições de frequência cardíaca em repouso e no limiar anaeróbio.

Na Figura 4.9, 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 do 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 4.10.

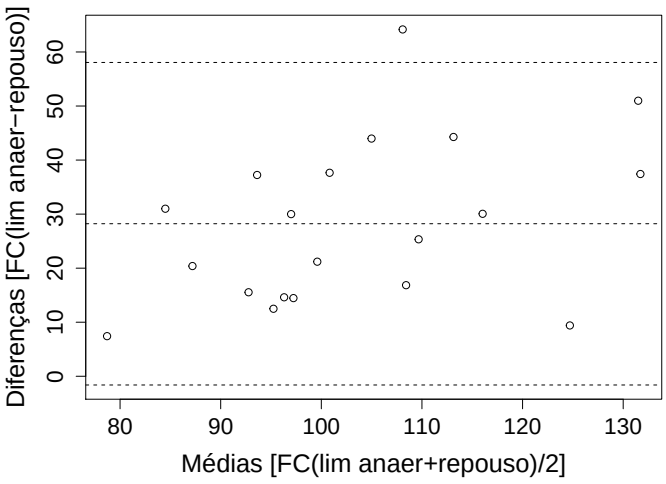

Figura 4.10: Gráfico de médias/diferenças de Tukey (Bland-Altman) para comparação das distribuições de frequência cardíaca em repouso e no limiar anaeróbio.

Exemplo 4.6. Considere o arquivo `temperaturas`, contendo dados de tem-

peratura para Ubatuba e Cananéia. O gráfico QQ correspondente está apresentado na Figura 4.11. 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.

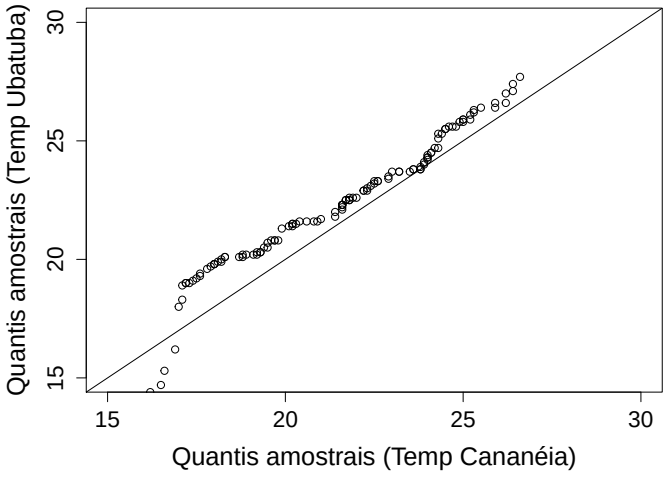

Figura 4.11: Gráfico QQ para comparação das distribuições de temperaturas de Ubatuba e Cananéia.

O gráfico de Bland-Altman correspondente, apresentado na Figura 4.12, sugere que acima de 17 graus, em média Ubatuba tende a ser 1 grau mais quente que Cananeia.

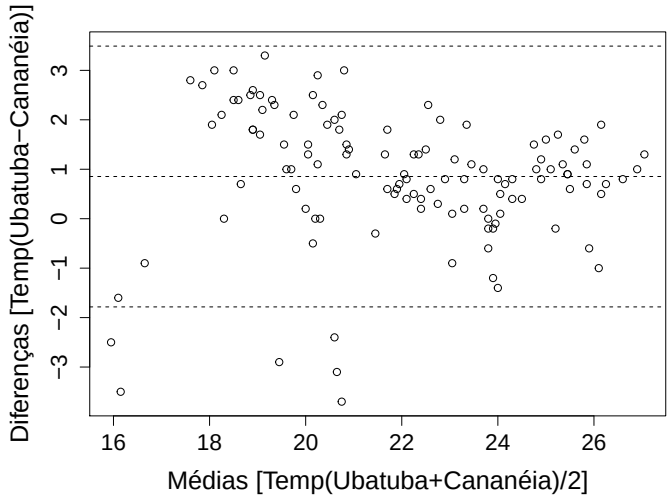

Figura 4.12: Gráfico de médias/diferenças de Tukey (Bland-Altman) para comparação das distribuições de temperaturas de Ubatuba e Cananéia.

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) entre outros elementos 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 4.24.

Tabela 4.24: 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 4.24 mostram que tanto as concentrações 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 4.13 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.

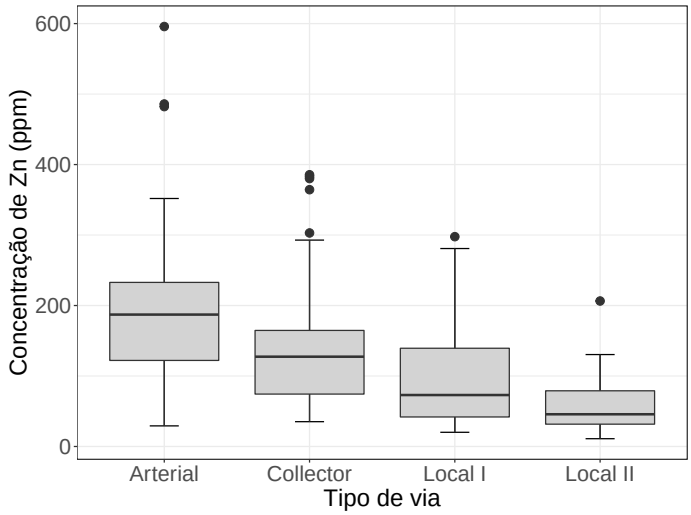

Figura 4.13: *Boxplots* para comparação das distribuições da concentração de Zn nas cascas de *tipuanas*.

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 - para detalhes, veja a Nota de Capítulo 6) 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 4.14 e reflete as mesmas conclusões obtidas com as análises anteriores.

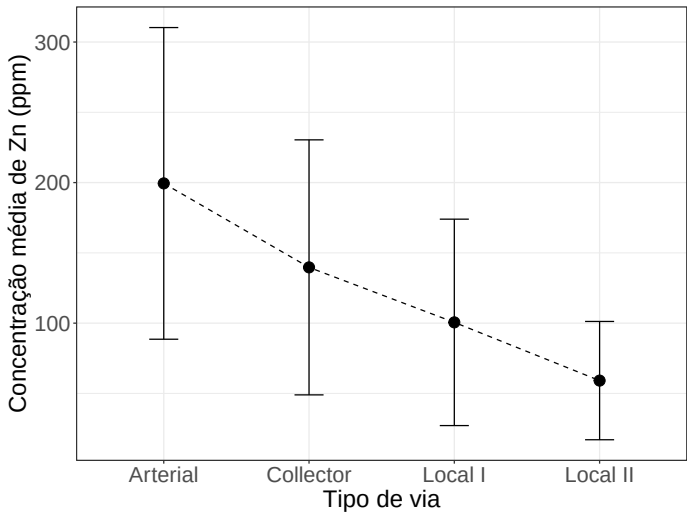

Figura 4.14: Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de *tipuanas*.

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.

Os segmentos de reta (linhas pontilhadas) que unem os pontos representando as médias não têm interpretação e servem apenas para salientar possíveis tendências de variação dessas médias.

Para propósitos inferenciais, uma técnica apropriada para a análise de dados com essa natureza é a **Análise de Variância** (com um fator), comumente cognominada ANOVA (*ANalysis Of VAriance*). O objetivo desse tipo de análise é avaliar diferenças entre as respostas esperadas das unidades de investigação na população da qual se supõe que os dados correspondem a uma amostra.

Um modelo bastante empregado para representar as distribuições da variável resposta das unidades de investigação submetidas aos diferentes tratamentos é

$$y_{ij} = \mu_i + e_{ij},\ i = 1, \ldots, a,\ j = 1, \ldots, n_i \tag{4.8}$$

em que y_{ij} representa a resposta da j-ésima unidade de investigação submetida ao i-ésimo tratamento, μ_i denota o valor esperado correspondente e

os e_{ij} representam erros aleatórios independentes para os quais se supõem distribuições normais com valores esperados iguais a zero e variância σ^2, constante, mas desconhecida. Uma representação gráfica desse modelo está disposta na Figura 4.15.

A hipótese a ser avaliada por meio da ANOVA é que os valores esperados das respostas associados aos a tratamentos são iguais, ou seja

$$H : \mu_1 = \ldots = \mu_a.$$

Se a ANOVA indicar que não existem evidências contrárias a essa hipótese, dizemos que não há **efeito de tratamentos**. Em caso contrário, dizemos que os dados sugerem que pelo menos uma das médias μ_i é diferente das demais.

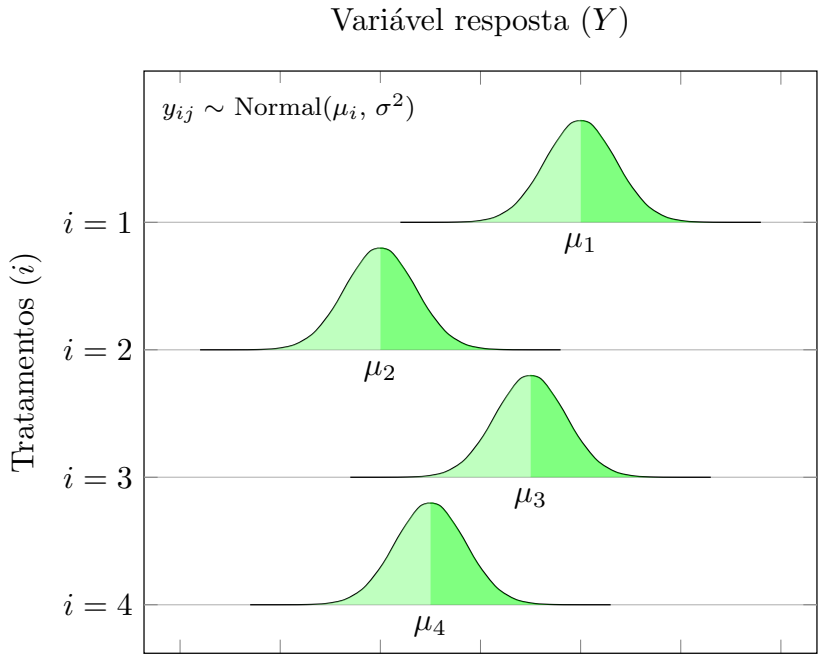

Figura 4.15: Representação de um modelo para ANOVA com um fator.

A concretização da ANOVA para a comparação dos valores esperados da concentração de Zn referentes aos diferentes tipos de via pode ser realizada por meio da função `aov()` com os comandos

```
> tipovia <- as.factor(tipuana$tipovia)
> anovaZn <- aov(Zn ~ tipovia, data=tipuana)
> summary(anovaZn)
```

O resultado, disposto na forma de uma tabela de ANOVA é

```
             Df  Sum Sq Mean Sq F value   Pr(>F)
tipovia       3  498525  166175   21.74 3.84e-12 ***
Residuals   189 1444384    7642
```

e sugere uma diferença altamente significativa ($p < 0{,}001$) entre os correspondentes valores esperados, ou seja, que pelo menos um dos valores esperados é diferente dos demais. O prosseguimento da análise envolve alguma técnica de **comparações múltiplas** para identificar se as concentrações esperadas de Zn correspondentes aos diferentes tipos de via são todas diferentes entre si ou se existem algumas que podem ser consideradas iguais. Para detalhes sobre esse tópico, o leitor pode consultar o excelente texto de Kutner et al. (2004).

Uma análise similar para os 76 *alfeneiros* está resumida na Tabela 4.25, e Figuras 4.16 e 4.17.

Tabela 4.25: 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

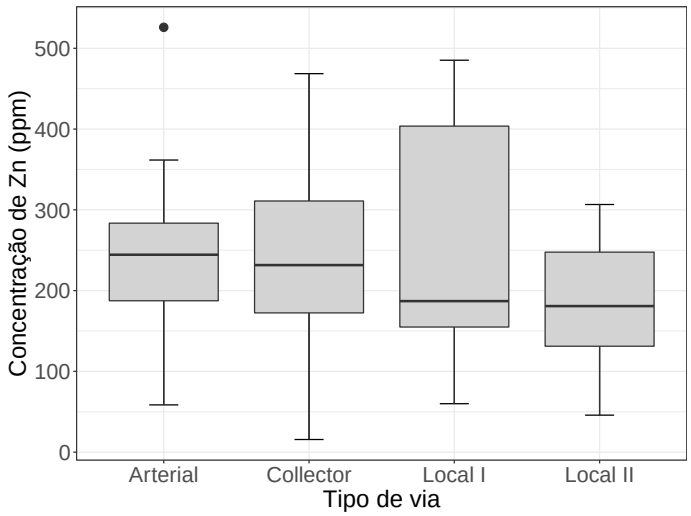

Figura 4.16: *Boxplots* para comparação das distribuições da concentração de Zn nas cascas de *alfeneiros*.

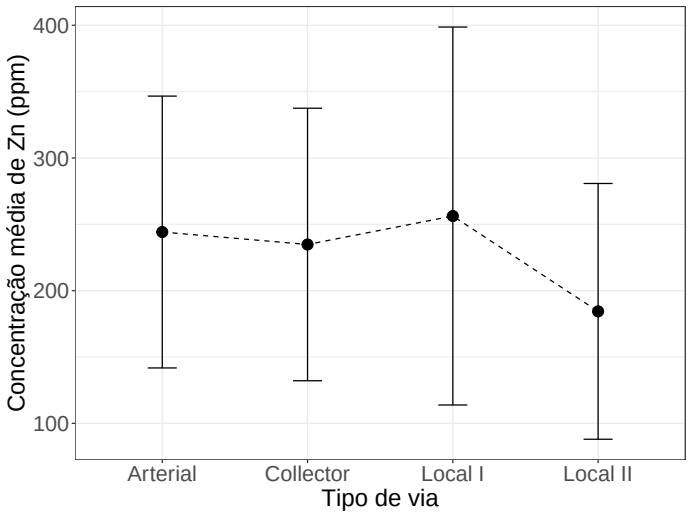

Figura 4.17: Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de *alfeneiros*.

Os valores dispostos na Tabela 4.25 e as Figuras 4.16 e 4.17 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.

A tabela de ANOVA correspondente é

```
            Df Sum Sq Mean Sq F value Pr(>F)
tipovia      3  27482    9161   0.712  0.548
Residuals   72 925949   12860
```

e não sugere que as concentrações esperadas de Zn nas cascas de *alfeneiros* sejam diferentes para árvores dessa espécie localizadas nas cercanias dos diferentes tipos de via ($p < 0{,}548$).

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 4.26.

Tabela 4.26: 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 qualitativa 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 4.26 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,

$$\overline{\text{Var}(S)} = \frac{\sum_{i=1}^{k} n_i \text{Var}_i(S)}{\sum_{i=1}^{k} n_i}, \tag{4.9}$$

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 $\overline{\text{Var}(S)} \leq \text{Var}(S)$, em que $\text{Var}(S)$ denota a variância da variável `Salário` obtida sem levar em conta `Grau de instrução`. Então podemos definir o

grau de associação entre as duas variáveis como o ganho relativo na variância obtido pela introdução da variável qualitativa. Explicitamente,

$$R^2 = \frac{\text{Var}(S) - \overline{\text{Var}(S)}}{\text{Var}(S)} = 1 - \frac{\overline{\text{Var}(S)}}{\text{Var}(S)}. \tag{4.10}$$

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 de Análise de Variância, cuja finalidade é comparar médias (populacionais) de distribuições normais independentes com mesma variância. 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 que $P(D|E)$ denota a probabilidade da ocorrência da moléstia para indivíduos expostos ao fator de risco, com os demais termos dessa expressão tendo interpretação similar.

Em **estudos retrospectivos** ou **caso-controle**, o planejamento envolve a escolha de amostras de tamanhos n_{+1} e n_{+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:

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

com $P(E|D)$ denotando a probabilidade de indivíduos com a moléstia terem sido expostos ao fator de risco e com interpretação similar para os demais termos da expressão. 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,

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

(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

$$\rho = \frac{\text{Cov}(X,Y)}{DP(X)DP(Y)} \tag{4.11}$$

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 28) que o mínimo de d ocorre quando

$$a = \frac{\sigma_X}{\sigma_Y}\rho(X,Y), \qquad b = E(X) - aE(Y), \tag{4.12}$$

e é dado por

$$\min d = \sigma_X^2[1 - \rho(X,Y)^2]. \tag{4.13}$$

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 (4.13) temos

$$\rho(X,Y) = \frac{\sigma_X^2 - \min_{a,b} E(|X - aY - b|^2)}{\sigma_X^2}, \tag{4.14}$$

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

$$r(\alpha) = \frac{S_{\tilde{x}+\tilde{y}}^2(\alpha) - S_{\tilde{x}-\tilde{y}}^2(\alpha)}{S_{\tilde{x}+\tilde{y}}^2(\alpha) + S_{\tilde{x}-\tilde{y}}^2(\alpha)}, \tag{4.15}$$

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 4.27 e dispostos num diagrama de dispersão na Figura 4.18.

Tabela 4.27: 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		

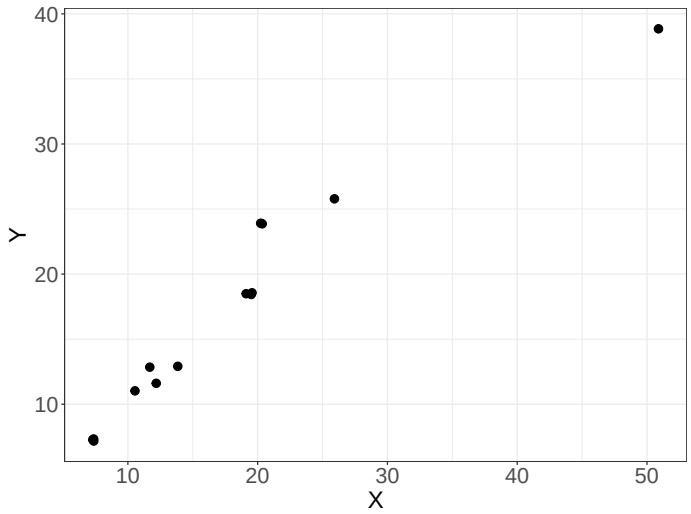

Figura 4.18: Gráfico de dispersão para os dados do Exemplo 4.9.

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 (4.15) obtemos $r(\alpha) = 0{,}973$, o que indica uma alta correlação entre as duas variáveis.

5) **Gráficos PP**

Na Figura 4.19, 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).

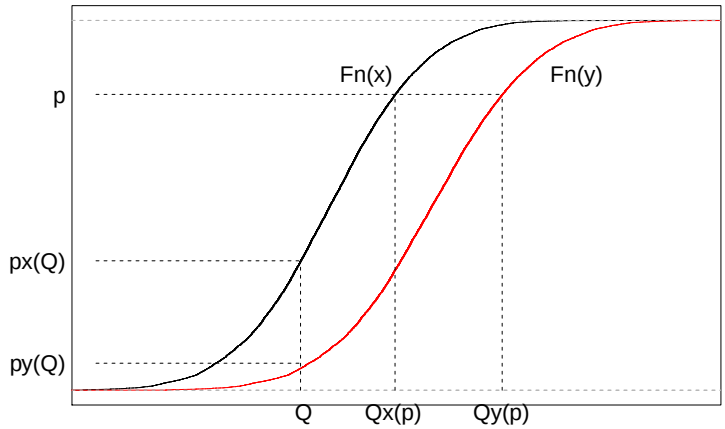

Figura 4.19: Quantis e probabilidades associados a duas distribuições.

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 4.20 e salientam a maior capacidade dos últimos para detectar assimetrias em distribuições de frequência.

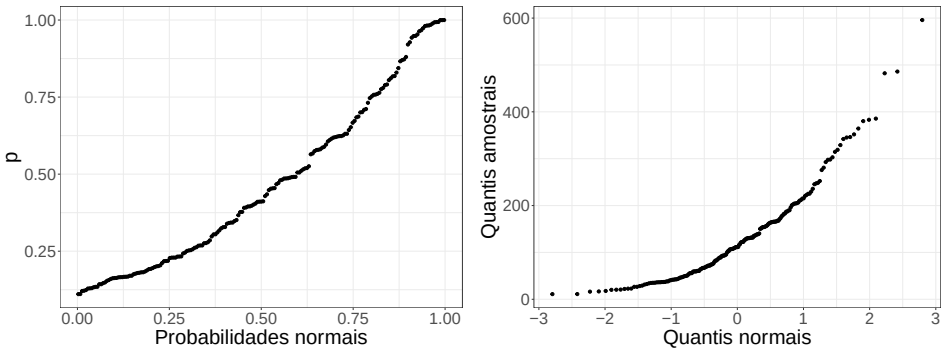

Figura 4.20: Gráficos PP e QQ para concentração de Zn em cascas de árvores da espécie *Tipuana*.

6) **Diferenças significativas**

Consideremos a distribuição de uma variável X (pressão arterial, por exemplo) em duas populações, A e B e admitamos que os valores esperados de X sejam μ_A e μ_B (desconhecidos), 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.

De forma geral, afirmar que uma diferença entre duas médias amostrais é significativa é dizer que as médias das populações de onde as amostras foram 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. A escolha do valor 0,05 para a decisão sobre a significância ou não da diferença é arbitrária embora seja muito utilizada na prática.

7) **Intervalos de confiança para o risco relativo e razão de chances**

Consideremos a seguinte tabela 2×2

Tabela 4.28: 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 (populacionais) 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 estimar a variância associada à estimativa do risco relativo, $rr = p_1/p_2$, é mais conveniente estimar a variância de $\log(rr)$. Com essa finalidade, recorremos ao **método Delta**[6], 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 intervalo de confiança com coeficiente de confiança aproximado de 95% para o logaritmo do risco relativo (populacional) RR são obtidos de

$$\log(p_1/p_2) \pm 1.96\sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}}. \tag{4.16}$$

Os limites do intervalo de confiança correspondente para o risco relativo (populacional) podem ser obtidos exponenciando-se os limites indicados em 4.16.

A razão de chances RC de doença entre indivíduos expostos e não expostos ao fator de risco é estimada por $rc = p_1(1-p_2)/p_2(1-p_1)$. 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 intervalo de confiança com coeficiente de confiança aproximado de 95% para o logaritmo do razão de chances (populacional) RC são obtidos de

$$\log[p_1(1-p_2)/p_2(1-p_1)] \pm 1.96\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{4.17}$$

[6] O método Delta é utilizado para estimar 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).

Assim como no caso do risco relativo, os limites do intervalo de confiança aproximado correspondente para a razão de chances pode ser obtido por meio da exponenciação dos limites indicados em (4.17).

Tendo em conta a forma da estimativa da sensibilidade de um teste diagnóstico, a saber, $\widehat{S} = n_{11}/n_{1+}$, uma estimativa sua variância, obtida por meio da aproximação da distribuição binomial por uma distribuição normal é $\widehat{\text{Var}(x)} = S(1-S)/n_{1+}$ e os limites inferior e superior de um intervalo de confiança com coeficiente de confiança aproximado de 95% para a sensibilidade populacional são, respectivamente,

$$S - 1{,}96\sqrt{S(1-S)/n_{1+}} \quad \text{e} \quad s + 1{,}96\sqrt{S(1-S)/n_{1+}} \tag{4.18}$$

Intervalos de confiança aproximados para as demais características de testes diagnósticos podem ser construídos de maneira análoga, substituindo em (4.18), $\widehat{S}$ e o denominador n_{1+} pelos valores correspondentes nas definições de especificidade, falsos positivos etc.

8) **Uma interpretação ingênua sobre o valor-p**

Embora haja controvérsias, uma medida de plausibilidade de uma hipótese (nula) estatística é o valor-p.

Considere a hipótese de que a probabilidade de cara, digamos θ, num lançamento de uma moeda seja igual a 0,5 contra uma hipótese alternativa de que essa probabilidade é maior do que 0,5 e admitamos que o objetivo seja decidir se $\theta = 0{,}5$ ou $\theta < 0{,}5$ com base em 10 lançamentos dessa moeda. Suponhamos que 10 caras tenham sido observadas nesses 10 lançamentos. A probabilidade de que isso ocorra para moedas com $\theta = 0{,}5$ pode ser calculada por meio da distribuição binomial e é igual a $1/1024 \approx 0{,}001$; esse é o valor-p associado ao resultado (10 caras em 10 lançamentos) e indica que embora esse resultado seja possível, ele é pouco provável. Se, com base nesse resultado, decidirmos que θ deve ser menor do 0,5, a probabilidade de termos decidido erroneamente é esse valor-p.

Se, tivermos observado 8 em vez de 10 caras nos 10 lançamentos da moeda, o valor-p correspondente é a probabilidade de que 8 ou mais caras sejam observadas que também pode ser obtido por meio da distribuição binomial e é igual a $56/1024 \approx 0{,}055$. Neste caso, se optarmos pela decisão de afirmar que θ deve ser menor do 0,5, a probabilidade de essa decisão esteja errada é $0{,}055 = 56/1024$.

A tomada da decisão depende das consequências de um possível erro mas é um problema extra estatístico. A conclusão estatística limita-se ao cálculo da probabilidade de uma decisão errada. Se decidirmos que a maior probabilidade de tomar a decisão errada for de 5% ou seja se adotarmos um **nível de significância** $\alpha = 0{,}05$, optaremos por dizer que $\theta < 0{,}5$ no primeiro caso (10 caras em 10 lançamentos da

moeda), mas não o faremos no segundo caso (8 ou mais caras em 10 lançamentos da moeda).

Para detalhes técnicos e generalizações, o leitor pode consultar Bussab e Morettin (2017), por exemplo.

4.6 Exercícios

1) Considere o conjunto de dados disponível no arquivo `empresa`. Compare as distribuições de frequências das variáveis `Estado civil`, `Grau de Instrução` e `Salário` para indivíduos com diferentes procedências.

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) Compare as distribuições das variáveis `Salário de professor secundário` e `Salário de administrador` por meio de um gráfico QQ e interprete os resultados.

   b) Calcule o coeficiente de correlação de Pearson e o coeficiente de correlação robusto (4.15) com $\alpha = 0{,}10$ entre essas duas variáveis.

4) 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`. Avalie a associação entre essas duas variáveis.

5) Analise a variável `Preço de veículos` segundo as categorias N (nacional) e I (importado) para o conjunto de dados disponível no arquivo `veiculos`.

6) Considere o conjunto de dados disponível no arquivo `coronarias`.

   a) Construa gráficos QQ para comparar a distribuição da variável `COL` de pacientes masculinos (=1) com aquela de femininos (=0). Repita a análise para a variável `IMC` 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 `TABAG4` e `ARTER` e calcule a intensidade de associação entre elas utilizando a estatística de Pearson, o coeficiente de contingência de Pearson e o coeficiente de Tschuprov.

7) Considere os dados do arquivo `endometriose`. Construa um gráfico QQ para comparar as distribuições da variável `Idade` de pacientes dos grupos Controle e Doente.

8) 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.

9) 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 `FC` correspondentes a pacientes nos diferentes níveis da variável `NYHA`. Comente os resultados.

10) Os dados da Tabela 4.29 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 4.29: 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

11) 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%. Considerando uma aproximação de 2 asas decimais, indique qual é a razão de chances (rc) de desistência entre as cidades A e B, justificando sua resposta

a) rc = 2,33 b) rc = 2,55 c) rc = 8,00 d) rc = 1,75 e) Nenhuma das respostas anteriores está correta.

12) De uma tabela construída para avaliar a associação entre tratamento (com níveis ativo e placebo) e cura (sim ou não) de uma certa moléstia obteve-se uma razão de chances igual a 2,0. Mostre 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.

13) Considere os dados do arquivo `esquistossomose`. Calcule a sensibilidade, especificidade, taxas de falsos positivos e falsos negativos, valores preditivos positivos e negativos e acurácia correspondentes aos cinco testes empregados para diagnóstico de esquistossomose.

14) 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.

15) 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 na veia porta e na via biliar tendo os resultados intraoperatórios como padrão ouro.

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 4.30

Tabela 4.30: 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

Indique qual a afirmação abaixo sobre a relação entre as variáveis X e Y é correta, justificando sua resposta.

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 4.31 e 4.32 e Figura 4.21

Tabela 4.31: VO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	1845	1707	795
Cardiopatas	57	1065	984	434
DPOC	46	889	820	381

Tabela 4.32: 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?

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?

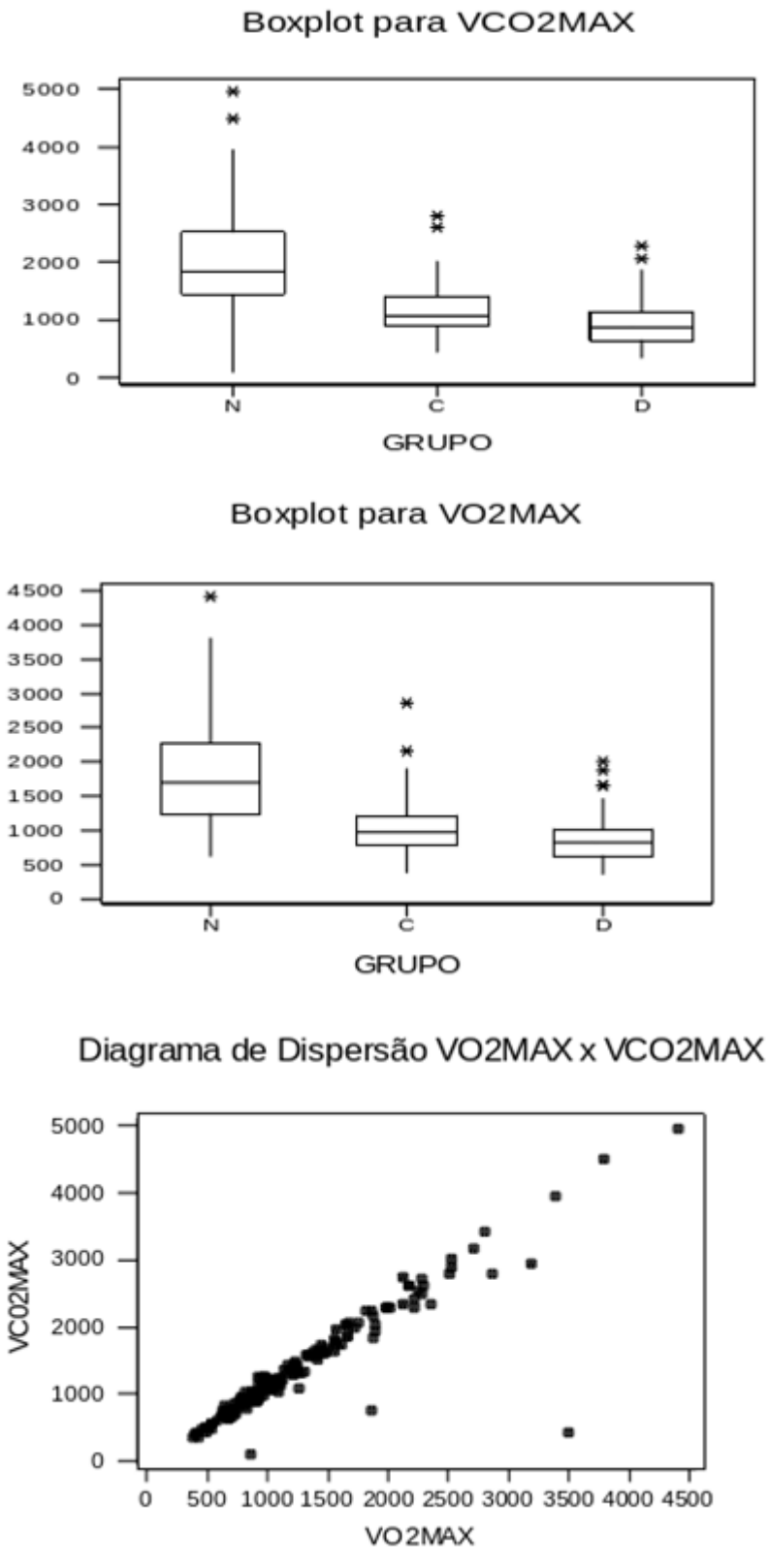

Figura 4.21: Gráficos para o Exercício 18.

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 de frequências

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 valores p em cada caso?

Detalhes podem ser obtidos em Afiune (2000).

20) 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 4.22 e 4.23.

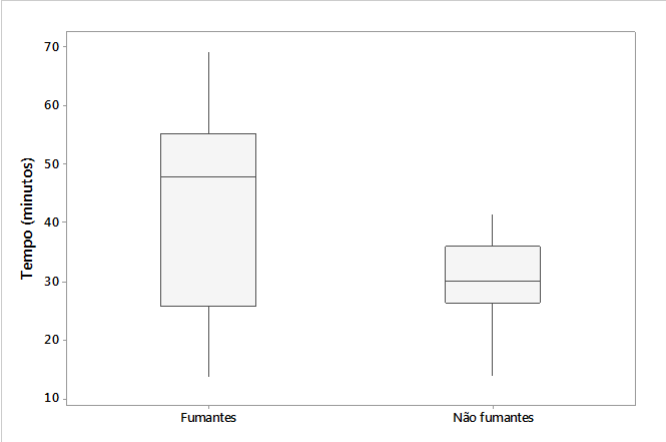

Figura 4.22: *Boxplots* do tempo até dormir nas populações Fumantes e Não Fumantes.

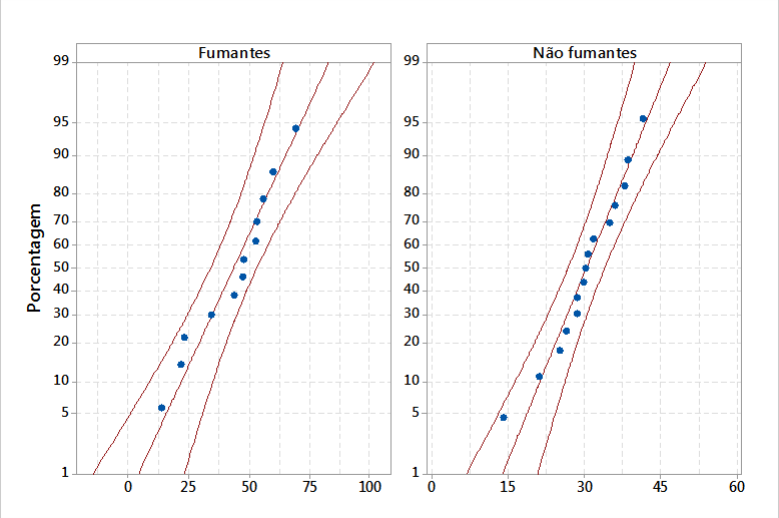

Figura 4.23: Gráfico QQ para as populações Fumantes e Não Fumantes.

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.

21) Em um estudo comparativo de duas drogas para hipertensão os resultados indicados nas Tabelas 4.33, 4.34 e 4.35 e Figura 4.24 foram usados para descrever a eficácia e a tolerabilidade das drogas ao longo de 5 meses de tratamento.

Tabela 4.33: 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 4.34: 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 4.35: 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

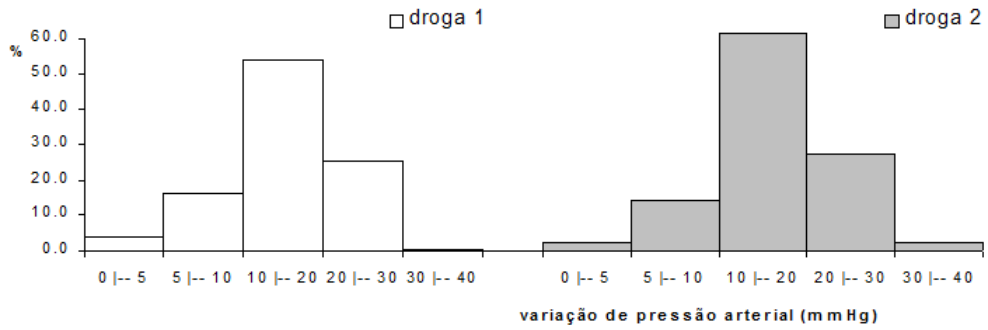

Figura 4.24: Histogramas para a variação de pressão arterial.

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 4.35).

22) Considere duas amostras de uma variável X com n unidades amostrais cada. Utilize a definição (4.9) para mostrar que $\overline{\text{Var}(X)} = \text{Var}(X)$ quando as médias das duas amostras são iguais.

23) Utilize o método Delta para calcular uma estimativa da variância da razão de chances (ver Nota de Capítulo 7).

24) Utilizando a definição da Nota de Capítulo 4, prove que se $\alpha = 0$, então $r(\alpha) = r$.

25) Mostre que para a hipótese de inexistência de associação numa tabela $r \times s$, a estatística (4.1) pode ser escrita como

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n},$$

em que n_{ij} é a frequência absoluta observada na linha i e coluna j e n_{i+} e n_{+j} são, respectivamente, os totais das linhas e colunas.

26) Prove que a expressão da estatística de Pearson do Exercício 10 pode ser escrita como

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}},$$

em que (f_{ij} e f^*_{ij} representam, respectivamente, as frequências relativas observada e esperada (sob a hipótese de inexistência de associação) correspondentes à cela i,j

27) Prove que (4.4) e (4.5) são equivalentes.

28) Prove as relações (4.12)-(4.14).

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, 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 monóxido de carbono (`CO`) e ozônio (`O3`), além de temperatura (`temp`) e umidade do ar (`umid`) observadas ao longo de 120 dias.

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, pois os modelos probabilísticos apropriados para esse tipo de dados envolvem distribuições conjuntas para as p variáveis, digamos $X_1, \ldots, X_p$, sob investigação. No caso discreto, eles são especificados por funções de probabilidade $P(X_1 = x_1, \ldots, X_p = x_p)$, e no caso contínuo, por funções densidade de probabilidade, $f(x_1, \ldots, x_p)$, que levam em consideração a provável correlação entre as variáveis observadas na mesma unidade amostral. 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 em unidades amostrais diferentes.

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.

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).

Muitas análises de dados multivariados *i.e.*, dados de várias variáveis, 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 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 13.

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 painéis com gráficos de dispersão para cada par de variáveis. Muitas vezes incluem-se coeficientes de correlação entre os diferentes pares de variáveis nos painéis situados acima ou abaixo da diagonal.

Exemplo 5.1. O gráfico do desenhista para as variáveis `preco`, `comp` e `motor` do arquivo `veiculos`, apresentado na Figura 5.1 pode ser gerado por meio dos comandos

```
> pairs(~preco + comp + motor, data=dados, upper.panel=panel.cor,
        cex=1.2, pch=19, cex.labels=3, cex.axis = 1.5)
```

Os coeficientes de correlação indicados nos painéis superiores podem ser calculados com a utilização da função

```
> panel.cor <- function(x, y, digits = 2, cex.cor, ...){
usr <- par("usr"); on.exit(par(usr))
par(usr = c(0, 1, 0, 1))
# correlation coefficient
r <- cor(x, y)
txt <- format(c(r, 0.123456789), digits = digits)[1]
txt <- paste("r= ", txt, sep = "")
text(0.5, 0.5, txt, cex=1.8)}
```

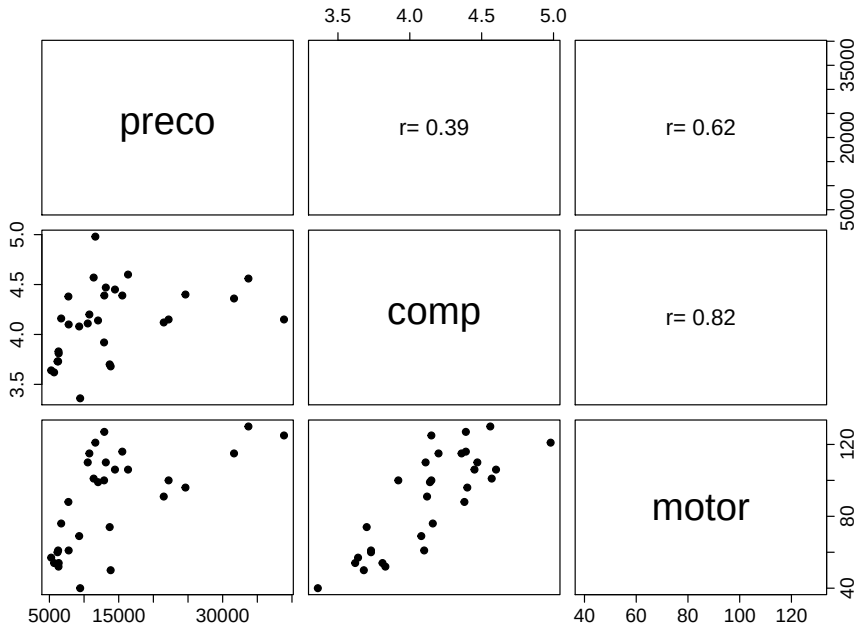

Figura 5.1: Gráfico do desenhista para os dados do arquivo `veiculos`.

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 é tão aparente quando consideramos as variáveis preço e comprimento: veículos com preços até 15000 apresentam comprimentos variando entre 3,5m e 4,0m, enquanto veículos com preços maiores que 15000, têm comprimento em torno de 4,5m.

A Figura 5.2 contém o mesmo tipo de gráfico para as variáveis CO, O3 e temp do arquivo `poluicao` e não mostra evidências considerável de associação entre cada par dessas variáveis.

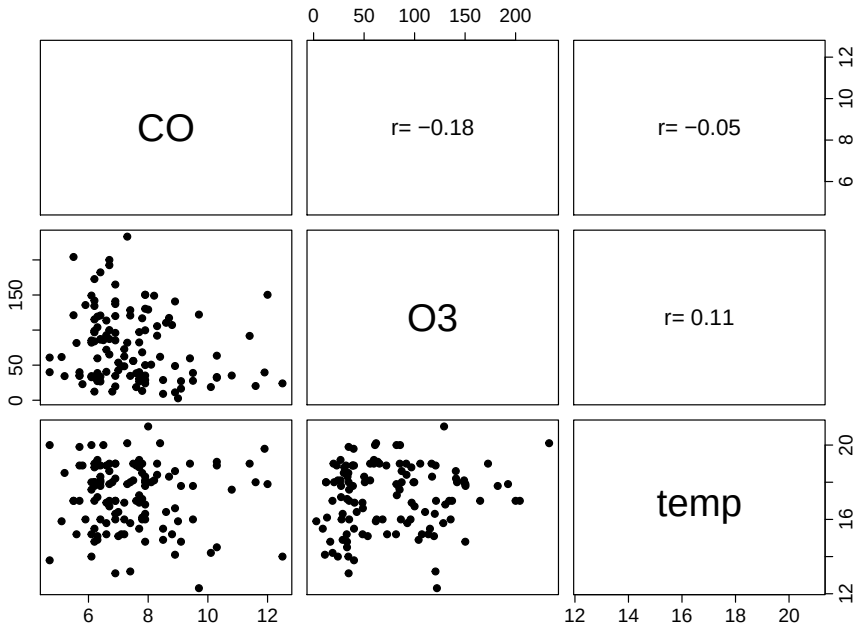

Figura 5.2: Gráfico do desenhista para os dados do arquivo `poluicao`.

Gráfico de dispersão simbólico

Gráficos de dispersão simbólicos ou **estéticos** (*aesthetic*) são essencialmente 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 `preco` e `comp` e em uma terceira variável qualitativa, `proc` (categorias `nacional` ou `importado`). Para cada par (`preco`, `comp`) usamos o símbolo △, para representar a categoria `nacional` e o símbolo ∘, para indicar a categoria `importado`. 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, pode ser construído por meio dos comandos

```
> g1 <- ggplot(veiculos, aes(comp, preco)) +
  geom_point(aes(shape=proc, color=proc), size=3) + theme_bw() +
  scale_color_manual(values=c("red", "blue")) +
  theme(axis.title = element_text(size=23)) +
  theme(legend.position="bottom", legend.direction="horizontal",
       legend.text=element_text(size=20))
  theme(axis.text.x = element_text(face="plain", size=13),
       axis.text.y = element_text(face="plain", size=13))
```

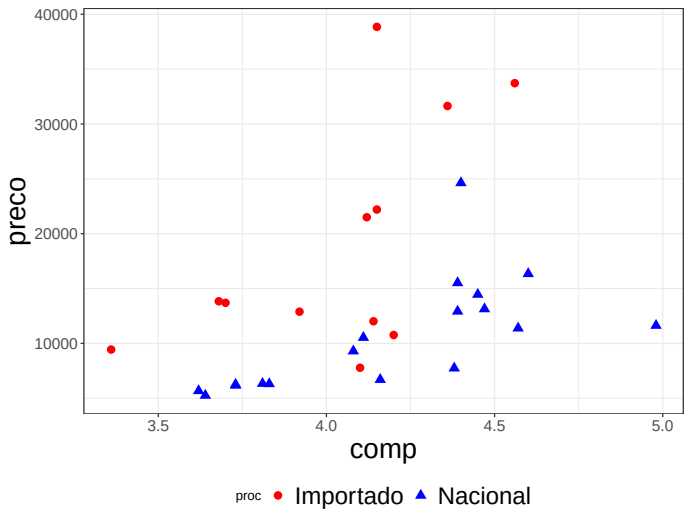

Figura 5.3: Gráfico de dispersão simbólico para as variáveis `preco`, `comp` e `proc` (Exemplo 5.2).

Uma alternativa para a representação gráfica das associações entre três variáveis quantitativas desse conjunto de dados consiste de um gráfico de dispersão com símbolos de diferentes tamanhos para representar uma delas. Por exemplo, na Figura 5.4, apresentamos o gráfico de dispersão de `preco` *versus* `comp`, com a variável `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.

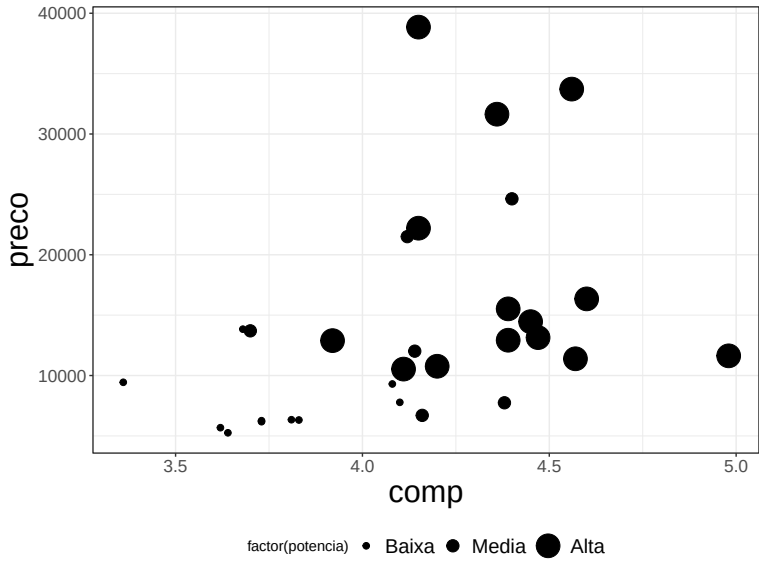

Figura 5.4: Gráfico de dispersão simbólico para as variáveis `preco`, `comp` e `motor` (Exemplo 5.2).

Os comandos do pacote ggplot2 utilizados para a construção do gráfico disposto na Figura 5.4 são

```
> 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 americanos, 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). Um gráfico de dispersão para as variáveis hwy *versus* displ, categorizado pela variável drv pode ser construído por meio dos comandos

```
> g1 <- ggplot(data=mpg) +
geom_point(mapping=aes(x=displ, y=hwy, color=drv)) +
geom_smooth(mapping=aes(x=displ, y=hwy, color=drv), se=FALSE) +
theme_bw() +
theme(legend.position="bottom", legend.direction="horizontal",
legend.text=element_text(size=20)) +
theme(axis.text.x = element_text(face="plain", size=18),
    axis.text.y = element_text(face="plain", size=18)) +
theme(axis.title = element_text(size=23))
```

Incluímos a opção geom_smooth para ajustar curvas suaves aos dados (usando o procedimento de suavização ***lowess***) de cada conjunto de pontos da variável drv, 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.

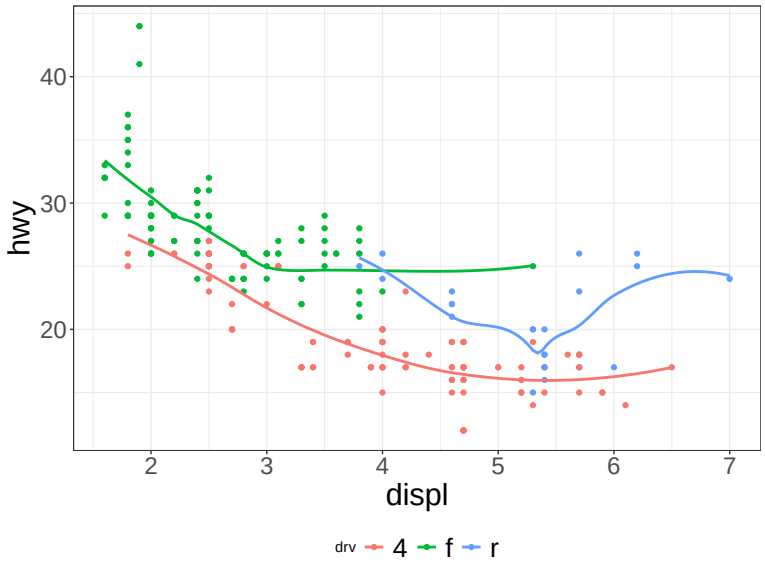

Figura 5.5: Gráfico de dispersão simbólico das variáveis `hwy` versus `displ`, categorizado pela variável `drv` com pontos e curvas *lowess*.

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 `motor` e `comp` de acordo com a faixa de preço (baixo, entre 5000 e 7000, médio, entre 5000 e 7000 ou alto, entre 12000 e 40000), como na Figura 5.6.

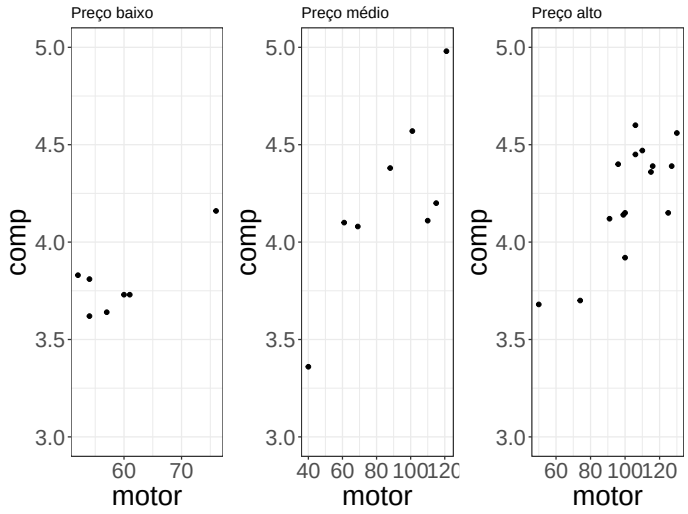

Figura 5.6: Janelamento para as variáveis preço *versus* comprimento, categorizado pela variável Potência do motor.

Esse gráfico sugere uma associação positiva entre comprimento e (potência) motor, independentemente da faixa de preço, com menos intensidade para veículos mais baratos.

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 os dados do arquivo `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

```
> 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 de Mg nas *tipua-*

nas 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 é similar para *tipuanas* e *alfeneiros* localizadas nas proximidades dos quatro tipos de vias considerados. As concentrações médias de Mg em *sibipirunas*, por outro lado, seguem um padrão diferente.

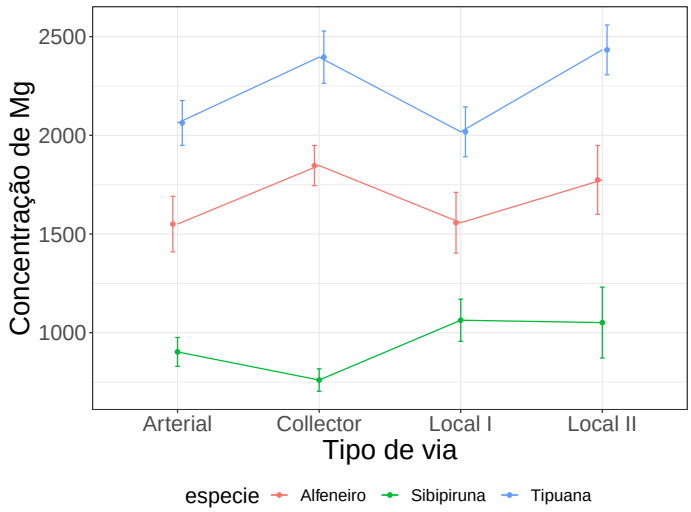

Figura 5.7: Gráfico de perfis médios para a concentração de Mg em cascas de árvores (as barras correspondem a erros padrões).

Quando 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** (com dois fatores). Nesse tipo de estudo, o objetivo inferencial é avaliar o "**efeito**" de cada fator e da "**interação**" entre eles em alguma característica da distribuição de uma variável resposta quantitativa contínua. Os termos "efeito" e "interação" estão entre aspas porque precisam ser definidos.

Com o objetivo de definir os "efeitos" dos fatores e de 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.*, sob 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 Tabela 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}

Os "efeitos" 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).

Para casos em que o fator A tem a níveis e o fator B tem b níveis, 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 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 submetidas ao nível j do fator B (faixa etária)** pode ser definido como a diferença $\mu_{1j} - \mu_{2j}$, que, no exemplo, corresponde à diferença entre o valor esperado da pressão diastólica de indivíduos com faixa etária j submetidos à droga 1 (ativa) e o valor esperado da pressão diastólica de indivíduos com essa mesma faixa etária submetidos à droga 2 (placebo). Analogamente, o **efeito do fator B (faixa etária) para indivíduos submetidos ao 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 indivíduos submetidos ao nível 1 do fator B e o efeito do fator A para indivíduos submetidos ao 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 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 submetidas ao nível 1 do fator A e o valor esperado da resposta para unidades 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, tanto 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.

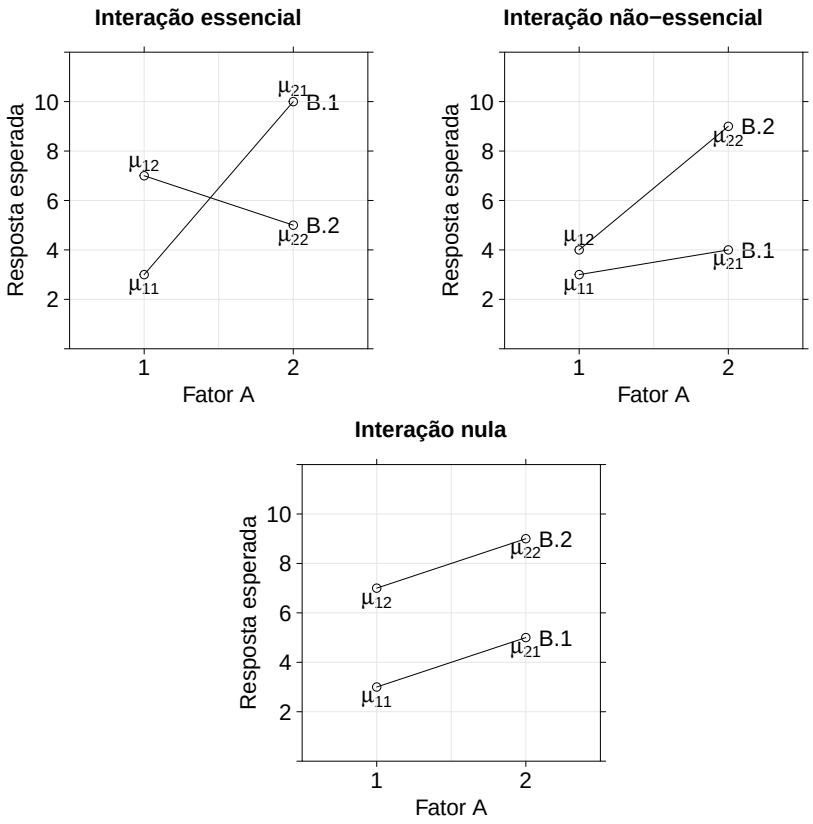

Figura 5.8: Gráfico de perfis médios (populacionais) com diferentes tipos de interação.

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 (variável resposta) de corpos de prova odontológicos . 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 adesivo 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.

Tabela 5.3: Medidas resumo 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

O gráfico de perfis médios correspondente está apresentado na Figura 5.9.

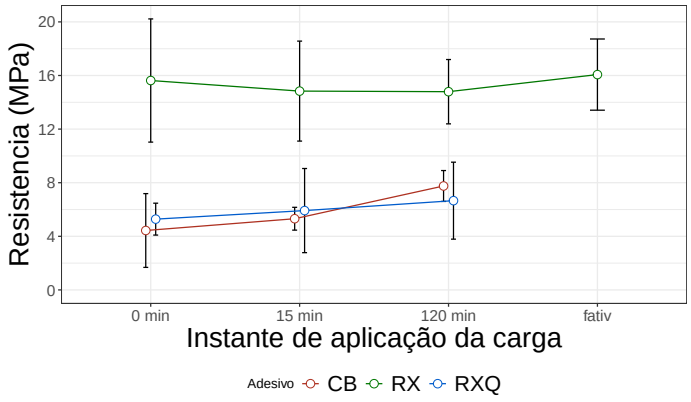

Figura 5.9: Gráfico de perfis de médias (com barras de desvios padrões) para os dados da Tabela 5.2.

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 entre os valores esperados populacionais. Nesse contexto, podemos esperar um efeito principal do fator Adesivo, segundo o qual, os adesivos CB e RXQ têm respostas esperadas iguais, mas menores que a resposta esperada 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 geram a tabela ANOVA apresentada em seguida.

```
> 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.

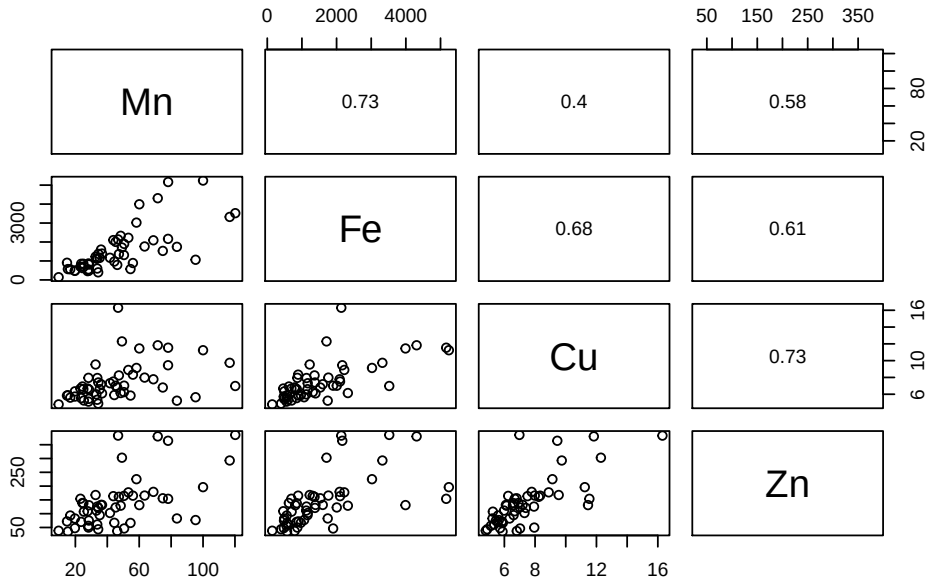

Figura 5.10: Gráfico do desenhista para os dados da concentração de elementos químicos em cascas de árvores.

Outros tipos de gráficos podem ser encontrados em Cleveland (1979) e Chambers et al. (1983), por exemplo.

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 dis-

postas 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 (4.11) 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 muito pequenas, exceto para O3 e Umid.

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 5.7. A tabela de frequências para a distribuição conjunta das variáveis `dismenorreia`, `esterilidade` e `endometriose` apresentadas no arquivo `endometriose2` 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, esterilidade, endometriose)
                           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 respostas, as frequências relativas calculadas em relação ao total de pacientes são obtidas por meio do comandos

```
> tabprop <- prop.table(tab)
> tabprop <- round(tabprop, 2)
```

cujo resultado é

```
                           endometriose  nao  sim
dismenorreia  esterilidade
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 podem ser estudadas com técnicas de **análise de dados categorizados**, por meio de **modelos log-lineares**. Detalhes podem ser obtidos em Paulino e Singer (2006), por exemplo.

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)
```

cujo resultado é

```
                          endometriose  nao  sim
dismenorreia  esterilidade
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 `esterilidade` e `endometriose` podem ser obtidas para cada nível de `dismenorreia` por meio das tabelas marginais; para `dismenorreia`=não, os comandos do pacote `vcd` são

```
> nao <- subset(endomet1, dismenorreia =="nao", na.rm=TRUE)
> attach(nao)
> tab1 <- ftable(esterilidade, endometriose)
```

com o seguinte resultado

```
             endometriose nao sim
esterilidade
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 `esterilidade` e `endometriose` podem ser obtidas para cada nível da variável `dismenorreia` por meio dos seguintes comandos do pacote `epiDisplay`

```
> endomet1 %$% mhor(esterilidade, 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** sugere 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 encontram-se 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,

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}.$$

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 principal 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

$$r_{uv} = \frac{S_{uv}}{\sqrt{S_u S_v}}, \tag{5.7}$$

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 (X_1) e número de clientes (X_2) 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

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 resposta é uma função linear ou quadrática de uma variável explicativa (preditora). 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), $j = 1, \ldots, n$ dos elementos 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 5.11 (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.

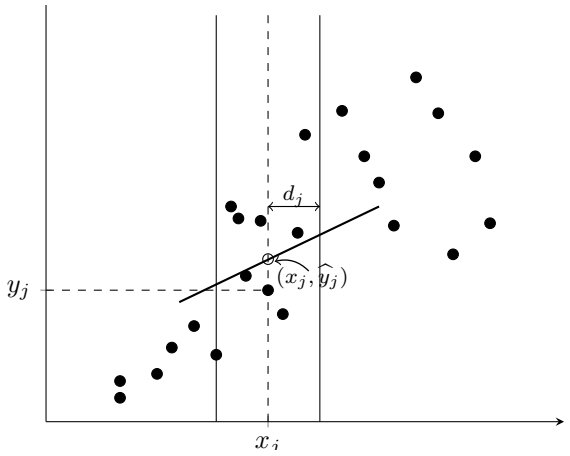

Figura 5.11: Faixa centrada em (x_j, y_j) para suavização por lowess.

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, por exemplo, 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 5.12.

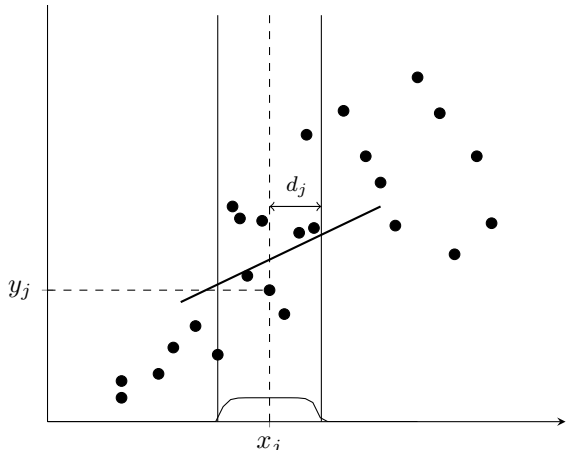

Figura 5.12: Atribuição de pesos para suavização por *lowess*.

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 atípicos (*outliers*). Quando existirem, refaça os cálculos, atribuindo pesos menores aos maiores resíduos, por exemplo, 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 valor absoluto do 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 5.13.

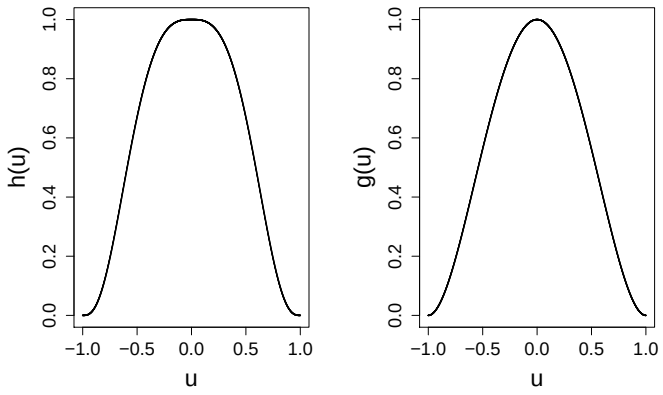

Figura 5.13: Gráficos das funções tricúbica $[h(u)]$ e biquadrática $[g(u)]$.

Para mais detalhes sobre o método *lowess* bem como sobre outros métodos de suavização o leitor poderá consultar Morettin e Tolói (2018), por exemplo.

O gráfico da Figura 5.14 contém curvas *lowess* (com dois níveis de suavização) ajustadas aos pontos do conjunto de dados `rotarod` 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. Os comandos utilizados para a construção da figura são

```
> 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)
```

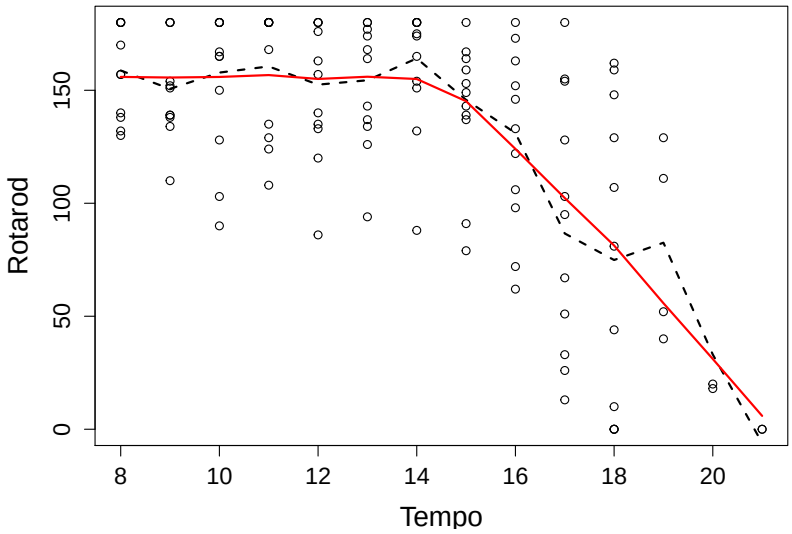

Figura 5.14: Curvas *lowess* com diferentes parâmetros de suavização ajustadas a um conjunto de dados.

3) **Parametrização de desvios médios**

Com a finalidade de explicitar efeitos principais e interação no modelo em que se deseja avaliar o efeito de dois fatores no valor esperado de uma variável resposta, 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 incorretamente 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 é **inidentificável**[1] e seus parâmetros não têm interpretação nem são estimáveis. Para torná-los interpretáveis e estimáveis, é preciso acrescentar **restrições de identificabilidade**, dentre as quais as mais frequentemente utilizadas são aquelas correspondentes à **parametrização de desvios de médias** e à **parametrização de cela de referência**, respectiva-

[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 (2015), entre outros, para detalhes.

mente definidas por

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \alpha\beta_{ij} = \sum_{j=1}^{b} \alpha\beta_{ij} = 0 \tag{5.11}$$

e

$$\alpha_1 = \beta_1 = \alpha\beta_{11} = \ldots = \alpha\beta_{1b} = \alpha\beta_{21} = \ldots = \alpha\beta_{a1} = 0 \tag{5.12}$$

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

$$\alpha\beta_{ij} = \mu_{ij} - b^{-1} \sum_{j=1}^{b} \mu_{ij} - a^{-1} \sum_{i=1}^{a} \mu_{ij}.$$

É nesse contexto que o parâmetro μ pode ser interpretado como **média geral** e representa a média dos valores esperados da resposta sob as diversas combinações dos níveis dos fatores A e B. O parâmetro α_i, chamado de **efeito do nível** i do fator A corresponde à diferença entre o valor esperado da resposta sob o nível i do do fator A e a média geral μ. O parâmetro β_j tem ma interpretação análoga e o parâmetro $\alpha\beta_{ij}$ corresponde à interação entre entre os níveis i do do fator A e j do fator B e pode ser interpretado como a diferença entre o valor esperado da resposta sob a combinação desses níveis dos fatores A e B e aquela que seria esperada quando não existe interação entre os dois fatores.

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

$$\alpha\beta_{ij} = \mu_{ij} - (\mu_{11} + \alpha_i + \beta_j), \;\; i = 2, \ldots, a, \; j = 2, \ldots, b,$$

de forma que o parâmetro μ é interpretado como referência, os parâmetros α_i, $i = 2, \ldots, a$ são interpretados como diferenças entre as respostas esperadas das unidades submetidas ao nível i do fator A relativamente àquelas obtidas por unidades submetidas ao tratamento associado ao nível 1 do mesmo fator, mantido fixo o nível correspondente ao fator B. Analogamente, os parâmetros β_j, $j = 2, \ldots, b$ podem ser interpretados como diferenças entre as respostas esperadas das unidades submetidas ao nível j do fator B relativamente àquelas obtidas por unidades submetidas ao tratamento associado ao nível 1 do mesmo fator, 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 submetidas ao tratamento correspondente à cela (i,j) e aquela esperada sob um modelo sem interação.

Em resumo, as definições do **efeito de um fator** e da **interação entre dois fatores** dependem da parametrização utilizada e são importantes para a interpretação dos resultados da análise, embora a conclusão seja a mesma qualquer que seja a alternativa adotada.

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, em que consideramos apenas dois estratos para efeito didático.

Tabela 5.5: Frequência de pacientes

	Fator	Status do paciente		
Estrato	de risco	doente	são	Total
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

$$rc_h = \frac{n_{h11}n_{h22}}{n_{h12}n_{h21}}.$$

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 ($H = 2$ no exemplo) estratos com pesos

$$w_h = \frac{n_{h12}n_{h21}}{n_{h++}} \Big/ \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}},$$

ou seja

$$\begin{aligned} rc_{MH} &= \sum_{h=1}^{H} w_h rc_h = \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}} \times \frac{n_{h11}n_{h22}}{n_{h12}n_{h21}} \Big/ \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}} \\ &= \sum_{h=1}^{H} \frac{n_{h11}n_{h22}}{n_{h++}} \Big/ \sum_{h=1}^{H} \frac{n_{h12}n_{h21}}{n_{h++}} \end{aligned}$$

Consideremos, por exemplo, os dados hipotéticos 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

	Fator	Doença			Razão
Clínica	de risco	sim	não	Total	de chances
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 de 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 e é

$$rc_{MH} = \frac{(5 \times 10)/22 + (3 \times 6)/19 + (3 \times 8)/16}{(7 \times 0)/22 + (9 \times 1)/19 + (4 \times 2)/16} = 6{,}56.$$

Outra vantagem da estatística de Mantel-Haenszel é que ela não é afetada pelo **Paradoxo de Simpson**, que ilustramos por meio de um exemplo. Com o objetivo de avaliar a associação entre a divulgação de propaganda e a intenção de compra de um certo produto, uma pesquisa foi conduzida em duas regiões. Os dados (hipotéticos) estão dispostos na Tabela 5.8.

Tabela 5.8: Frequências relacionadas a intenção de compra de um certo produto

		Intenção de compra			Razão
Região	Propaganda	sim	não	Total	de chances
1	não	50	950	1000	0,47
	sim	1000	9000	10000	
	Total	1050	9950	10000	
2	não	5000	5000	10000	0,05
	sim	95	5	100	
	Total	5095	5005	10100	

Segundo a Tabela 5.8, em ambas as regiões, a chance de intenção de compra com a divulgação de propaganda é pelo menos o dobro da chance de intenção de compra sem divulgação de propaganda. 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

	Intenção de compra			Razão
Propaganda	sim	não	Total	de chances
não	5050	5950	11000	6,98
sim	1095	9005	10100	
Total	6145	9950	21100	

A razão de chances obtida com os dados agrupados indica que a chance de intenção de compra quando não há divulgação de propaganda é cerca de 7 vezes aquela em que há divulgação de propaganda, invertendo a direção da associação encontrada nas duas regiões. Essa aparente incongruência é conhecida como o Paradoxo de Simpson e pode ser explicada por uma forte associação (com $rc = 0{,}001$) entre as variáveis Região e Divulgação de propaganda como indicado na Tabela 5.10.

Tabela 5.10: Frequências de divulgação de propaganda

	Região			Razão
Propaganda	1	2	Total	de chances
não	1000	10000	11000	0,001
sim	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 obtida nas duas regiões, em que a divulgação de propaganda está positivamente associada com a intenção de compra. Detalhes sobre o Paradoxo de Simpson podem ser encontrados em Paulino e Singer (2006), por exemplo.

5.7 Exercícios

1) Considere os dados do arquivo `tipofacial`.

   a) Construa um gráfico de dispersão simbólico para avaliar a relação entre as variáveis `altfac`, `proffac` e `grupo` e comente os resultados.

   b) Construa um gráfico do desenhista para avaliar a relação entre as variáveis `nsba`, `ns` e `sba` e comente os resultados.

2) Considere os dados do arquivo `antracose`.

   a) Categorize a variável `antracose` em três níveis (baixo, médio e alto).

   b) Construa um gráfico de dispersão simbólico para avaliar a relação entre as variáveis `htransp`, `ses` e `antracose` utilizando símbolos de tamanhos diferentes para os três níveis de `antracose` obtidos no item a). Comente os resultados.

3) Os dados dispostos na Figura 5.15 foram extraídos do arquivo `esforco` com a finalidade de avaliar a associação entre as variáveis `vo2rep` e `vo2pico` para pacientes chagásicos (CH) e controles (C). Comente as seguintes afirmações:

   a) Há pontos atípicos em ambos os casos.

   b) As variáveis são positivamente correlacionadas nos dois grupos de pacientes.

   c) Cada uma das variáveis têm dispersão semelhante nos dois grupos de pacientes.

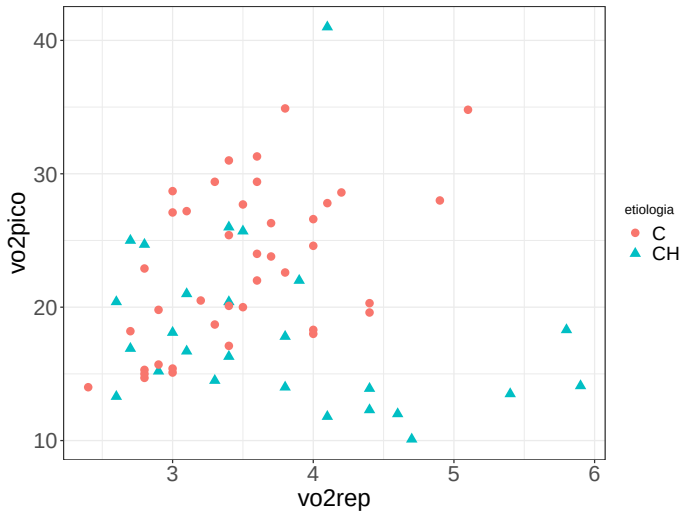

Figura 5.15: Gráfico de Pressão sistólica *versus* Idade para imigrantes.

4) 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 5.11.

Tabela 5.11: 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) Compare, descritivamente, 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 numa análise de variância? Justifique sua resposta.

5) 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 res-

posta observada foi o tempo para o desaparecimento dos sintomas (em dias). Os resultados obtidos estão dispostos na Tabela 5.12

Tabela 5.12: 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 esperado 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.

6) 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 ao material particulado fino e presença de diabetes ou hipertensão. As médias da capacidade vital forçada para cada combinação das categorias de diabetes ou hipertensão e exposição ao MP2,5 estão representadas na Figura 5.16.

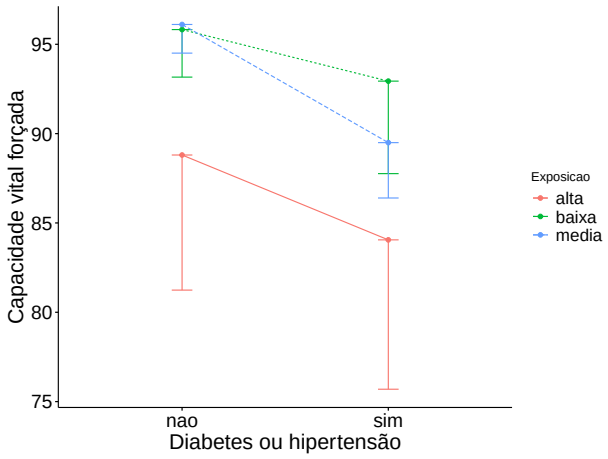

Figura 5.16: Capacidade vital forçada (% do predito).

a) Comente descritivamente os resultados obtidos, discutindo a possível interação entre diabetes/hipertensão e exposição ao material particulado.

b) Que comparações você faria para explicar essa possível interação? Justifique sua resposta.

7) 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 ºC, 21 ºC e 50 ºC. 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 .

   Com base nesse gráfico comente as seguintes afirmações

   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 ºC;

   d) há interação entre Temperatura e Tempo de vida.

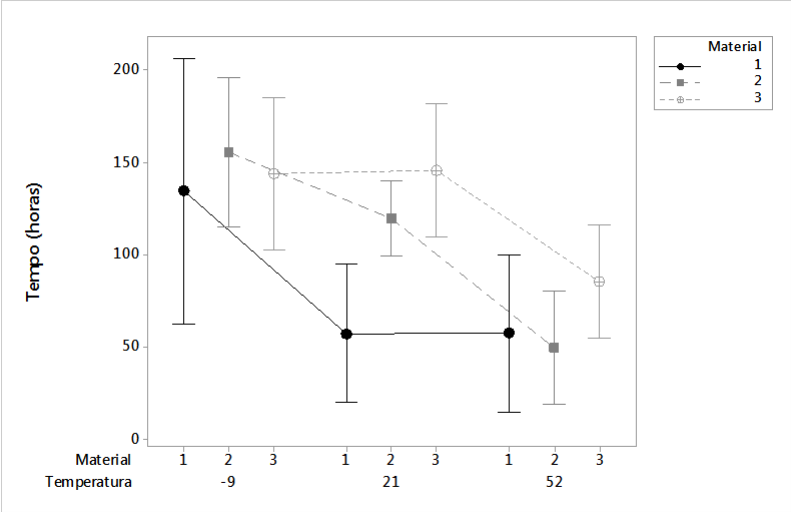

Figura 5.17: Gráfico das médias observadas do tempo de vida (h) e intervalos de confiança de 95% para as médias populacionais em cada combinação de temperatura e material.

8) Considere os dados do arquivo `esforco`.

   a) Construa gráficos do desenhista (*draftman's plots*) separadamente para cada etiologia (CH, ID e IS) com a finalidade de avaliar a associação entre os consumos de oxigênio (VO2) medidos nos três momentos de exercício (LAN, PCR e Pico) e indique os coeficientes de correlação de Pearson e de Spearman correspondentes.

   b) Para cada etiologia (CH, ID e IS), construa um gráfico de dispersão simbólico para representar a relação entre carga e VO2 no no momento PCR do exercício, e sobreponha curvas *lowess*. Que tipo de função você utilizaria para representar a relação entre as duas variáveis?

   c) Para cada um dos quatro momentos de exercício (Repouso, LAN, PCR e Pico), construa gráficos de perfis médios da frequência cardíaca para as diferentes combinações dos níveis de etiologia (CH, ID e IS) e gravidade da doença avaliada pelo critério NYHA. Em cada caso, avalie descritivamente as evidências de efeitos dos fatores Etiologia e Gravidade da doença e de sua interação.

   d) Utilize ANOVA para avaliar se as conclusões descritivas podem ser extrapoladas para a população de onde a amostra foi obtida.

9) 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 para cada combinação dos níveis de espécie e tipo de via. De uma forma geral, qual a relação entre os vetores de médias e matrizes de covariâncias para os diferentes níveis de espécie e tipo de via?

10) Considere os dados do arquivo `arvores`. Construa gráficos de perfis médios (com barras de desvios padrões) para avaliar o efeito de espécie de árvores e tipo de via na concentração de Fe. Utilize uma ANOVA com dois fatores para avaliar a possível interação e efeitos dos fatores na variável resposta. Traduza os resultados sem utilizar o jargão estatístico.

11) 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 sob os diferentes níveis do fator interunidades amostrais.

12) 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.

Curso	Professor	Desempenho	
		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.

13) Com base nos dados do arquivo `coronarias`, construa uma tabela de contingência 2×2×2×2 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 avaliamos, 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 (preditora) e Y é a variável resposta.

Exemplo 6.1[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.

[1] Exemplo adaptado de https://courses.lumenlearning.com/wmopen-concepts-statistics

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

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.

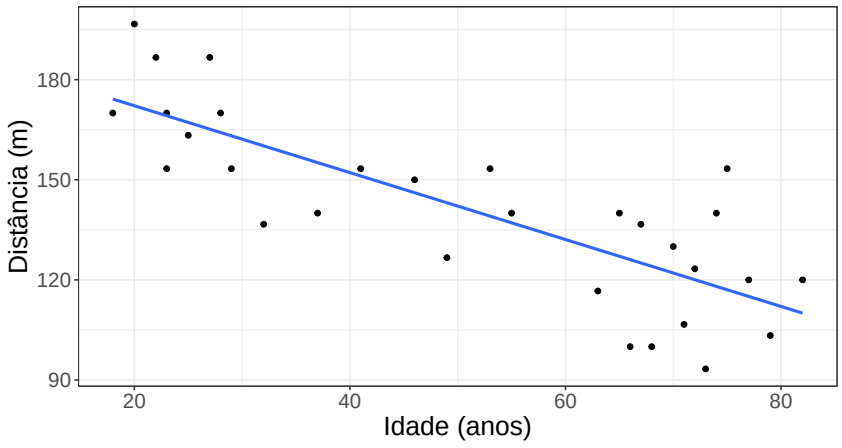

Figura 6.1: Gráfico de dispersão para os dados da Tabela 6.1 com reta de mínimos quadrados.

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 aleatórios que representam desvios entre as observações da variável resposta y_i e os pontos $\alpha+\beta x_i$, que correspondem aos valores esperados sob o modelo (6.1) para valores da variável explicativa iguais a x_i.[2] Em geral, supõe-se que a média (ou valor esperado) dos erros é nula, o que significa, de modo genérico, que existe uma compensação entre erros positivos e negativos e que, consequentemente, o objetivo da análise é modelar o valor esperado da variável resposta

$$\mathrm{E}(y_i) = \alpha + \beta x_i.$$

Uma representação gráfica desse modelo está apresentada na Figura 6.2.

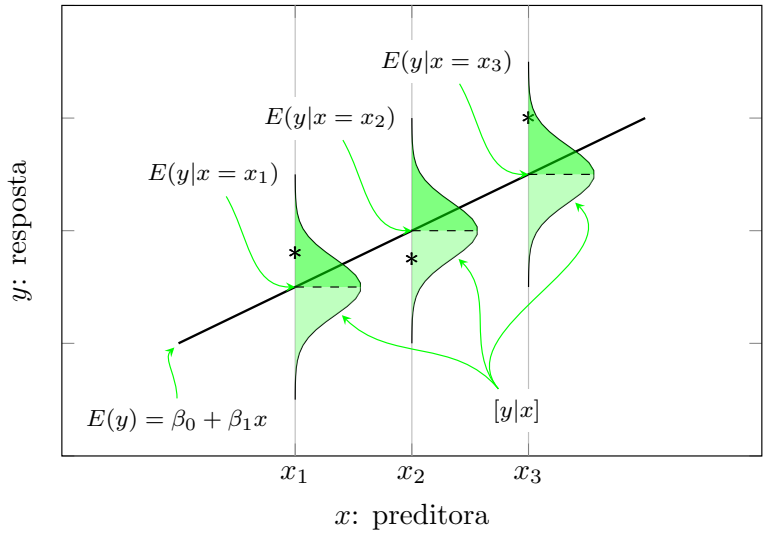

Figura 6.2: Representação gráfica do modelo (6.1).

No contexto do Exemplo 6.1, 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

$$y_i = \alpha + \beta(x_i - 18) + e_i, \quad i = 1, \ldots, n. \tag{6.2}$$

[2]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 x_i da variável explicativa X" deixa claro que o interesse da análise está centrado na distribuição de Y e não naquela de X.

Para esse modelo, o 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 de os parâmetros α e β serem 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. No entanto, 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()`. Na Seção 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 explicativas) 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 consiste em determinar $\widehat{\alpha}$ e $\widehat{\beta}$ que minimizem a distância entre cada observação o valor esperado definido 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 igualando as expressões resultantes a zero, obtemos as **equações de estimação**. 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 $Q(\widehat{\alpha},\widehat{\beta}) = \sum_{i=1}^{n}\widehat{e}_i^2$ é a **soma dos quadrados dos resíduos**, abreviadamente, $SQRes$. Os **resíduos** são definidos como

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta}x_i),\ i = 1, \ldots, n.$$

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.

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).

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 da distância esperada 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 dos 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, veja a Nota de Capítulo 3. Em essência, esse coeficiente mede a porcentagem da variação total dos valores da variável resposta (y_i) em relação à sua média ($\overline{y}$) 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 ***
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 da distância esperada 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 $R^2 = 0{,}63$. Detalhes sobre o coeficiente de determinação ajustado serão apresentados na Nota de Capítulo 3. 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 6.3.

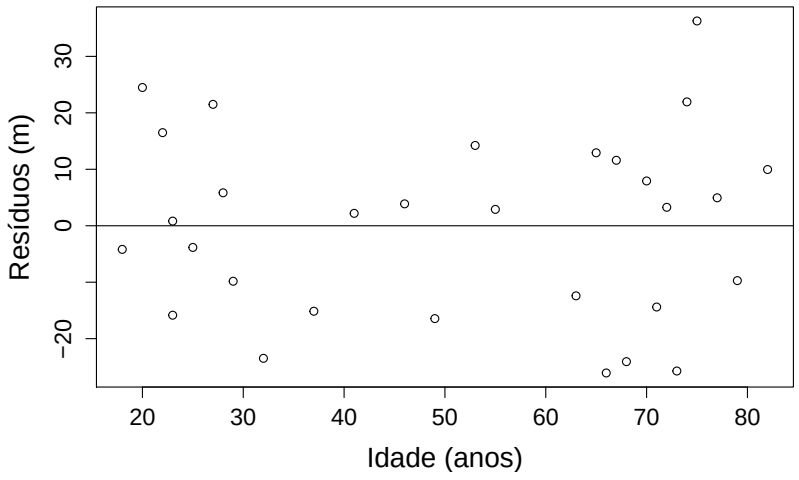

Figura 6.3: Gráfico de resíduos para o ajuste do modelo de regressão linear simples aos dados da Tabela 6.1.

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 seu desvio padrão é

$$DP(\widehat{e}_i) = \sigma\sqrt{1-h_{ii}} \;\text{ com }\; h_{ii} = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2},$$

de forma que os **resíduos padronizados**, também chamados de **resíduos estudentizados**, 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 dos erros (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 6.3 está apresentado na Figura 6.4.

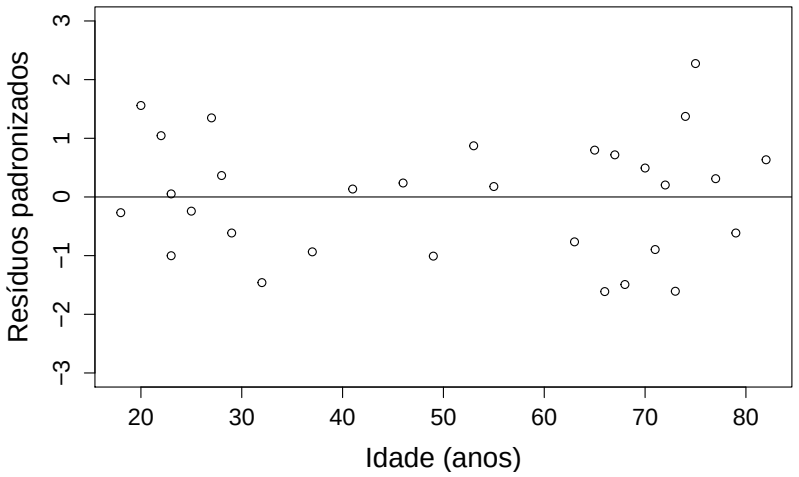

Figura 6.4: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados da Tabela 6.1.

Na Figura 6.4, 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: O ajuste de um modelo de regressão linear simples $CO_i = \alpha + \beta\ tempo_i + e_i, \quad i = 1, \ldots, n$ em que CO representa a concentração atmosférica de monóxido de carbono no dia (tempo) i contado a partir de 1 de janeiro de 1991 aos dados do arquivo `poluicao`) pode ser obtido por meio da função `lm()`:

```
> lm(formula = CO ~ tempo, data = dados)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.264608   0.254847  24.582  < 2e-16 ***
tempo       0.019827   0.003656   5.424 3.15e-07 ***
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 linear simples explica apenas uma pequena parcela da variabilidade dos dados.

Os gráficos de dispersão e de resíduos padronizados correspondentes ao ajuste desse modelo estão apresentados nas Figuras 6.5 e 6.6. 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.

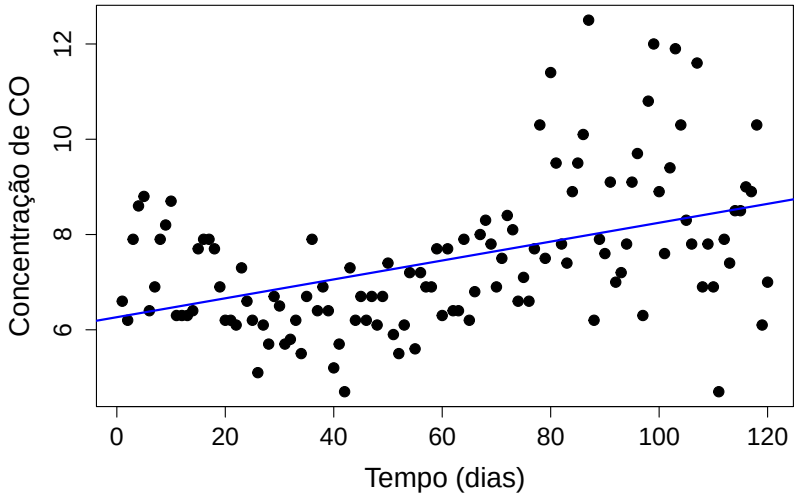

Figura 6.5: Gráfico de dispersão para os dados de monóxido de carbono.

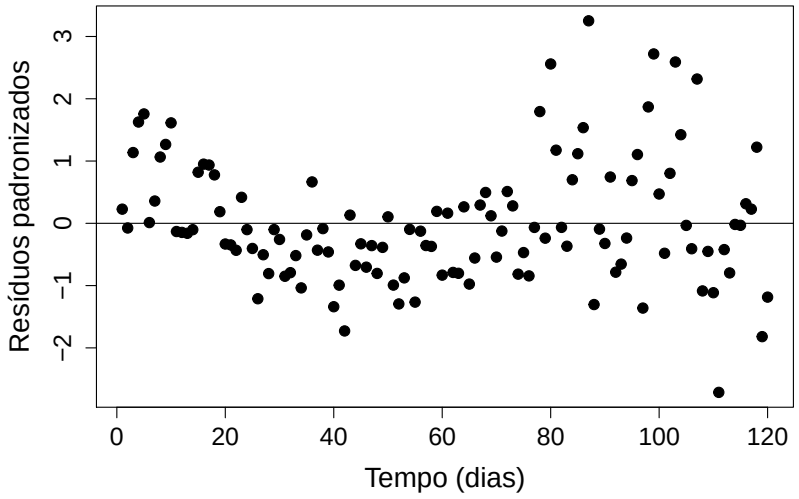

Figura 6.6: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados da concentração de CO.

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, quanto pelo gráfico de resíduos padronizados disposto na Figura 6.7. Detalhes sobre o

ajuste de modelos de regressão polinomial como esse, serão apresentados na Seção 6.3. Ainda assim, esse modelo polinomial não é o mais adequado em virtude da presença de **heteroscedasticidade**, 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.

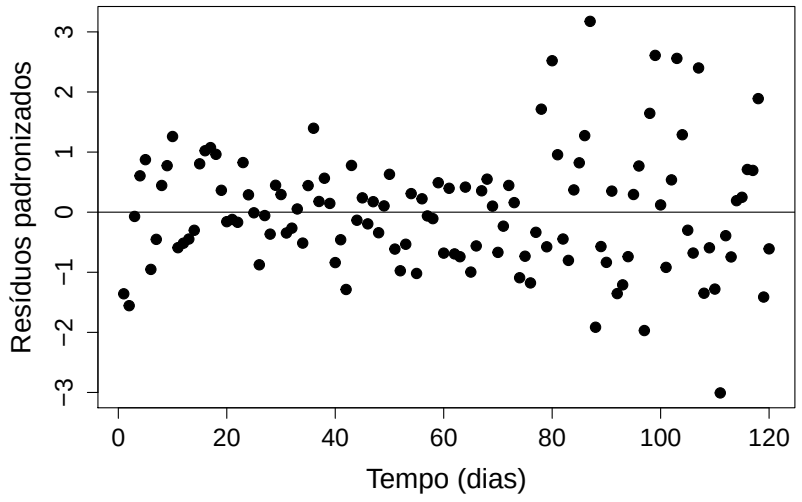

Figura 6.7: Gráfico de resíduos padronizados para o ajuste do modelo de regressão polinomial aos dados da concentração de CO.

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)

t	v_t	t	v_t
1	22,2	9	20,4
2	61,1	10	20,4
3	13,0	11	20,4
4	27,8	12	11,1
5	22,2	13	13,0
6	7,4	14	7,4
7	7,4	15	14,8
8	7,4		

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 6.8 e 6.9.

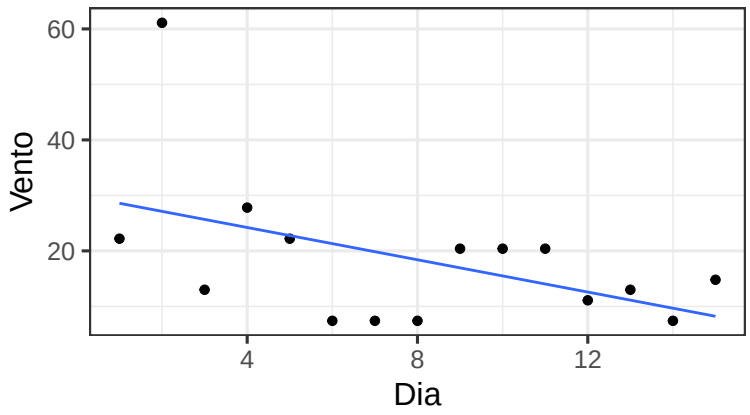

Figura 6.8: Gráfico de dispersão para os dados da Tabela 6.2 com reta de mínimos quadrados sobreposta.

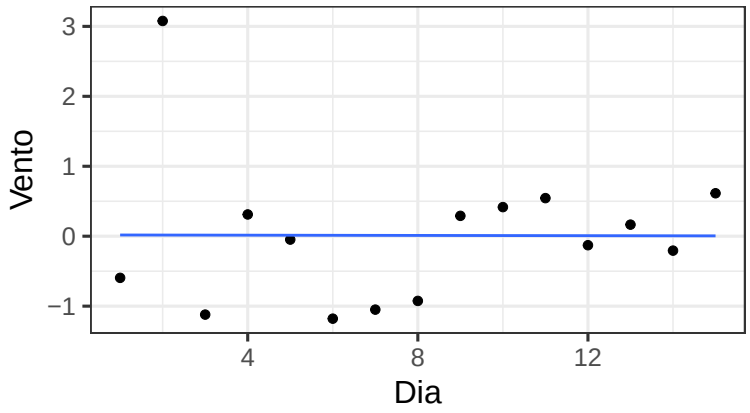

Figura 6.9: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados da Tabela 6.2.

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 atípica** (*outlier*). Na Nota de Capítulo 8, apresentamos um modelo alternativo com a finalidade de obtenção de estimativas **resistentes** (também chamadas de **robustas**) a pontos desse tipo.

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 6.10 e 6.11.

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

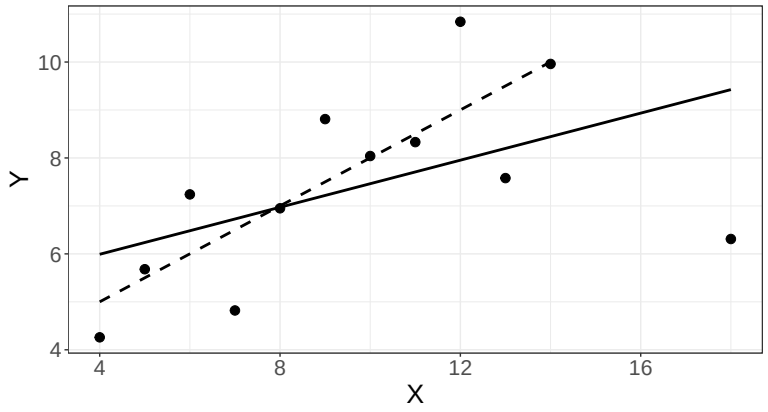

Figura 6.10: Gráfico de dispersão (com retas de regressão sobrepostas) para os dados da Tabela 6.3; curva sólida para dados completos e curva interrompida para dados com ponto influente eliminado.

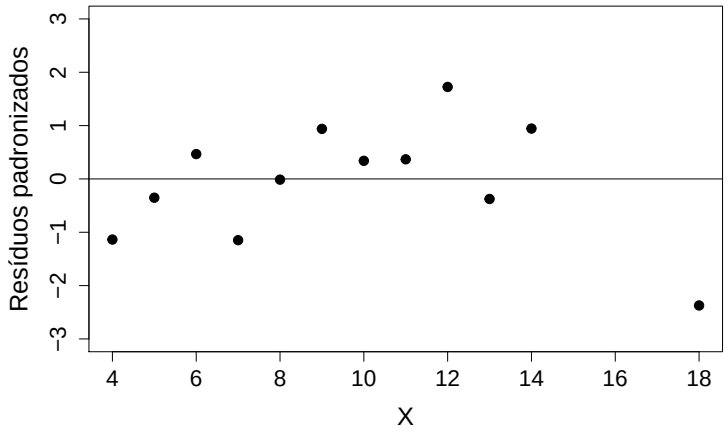

Figura 6.11: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados da Tabela 6.3.

Os dois gráficos contêm indicações de que o ponto associado aos valores $(X = 18, Y = 6.31)$ pode ser um ponto atípico. Isso fica mais evidente quando consideramos outra ferramenta diagnóstica conhecida como **gráfico de Cook** apresentado na Figura 6.12.

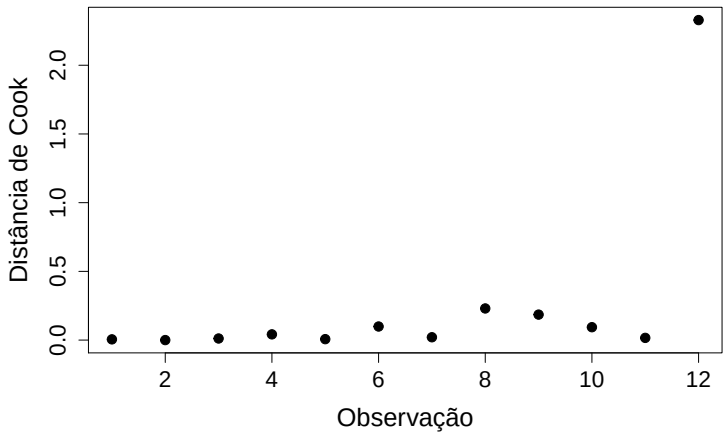

Figura 6.12: Gráfico de Cook correspondente ao ajuste do modelo de regressão linear simples aos dados da Tabela 6.3.

Esse gráfico é baseado na chamada **distância de Cook** (veja a 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 respectivas abscissas afastadas daquelas dos demais pontos do conjunto de dados. Para detalhes, consulte a Nota de Capítulo 5.

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 influente ($X = 18, Y = 6.31$) é eliminado do conjunto de dados está representada na Figura 6.10 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 x_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 deveriam obter os quantis teóricos para a construção do gráfico QQ. No entanto, 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. Na prática,

mesmo com valores de n menores, é comum construir esses gráficos baseados na distribuição normal.

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 6.13, 6.14 e 6.15.

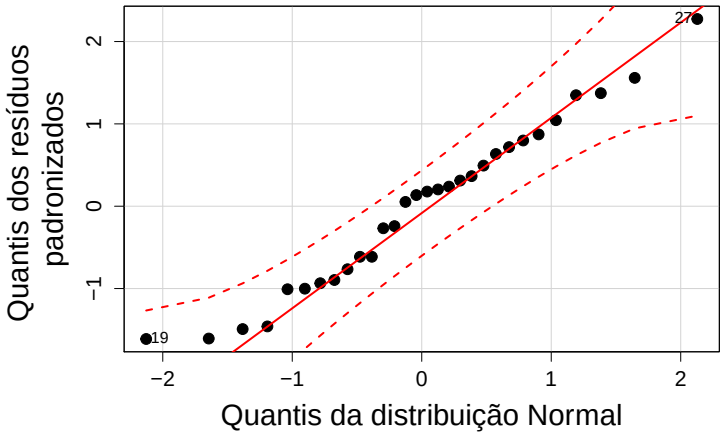

Figura 6.13: Gráfico QQ correspondente ajuste do modelo de regressão linear simples aos dados da Tabela 6.1.

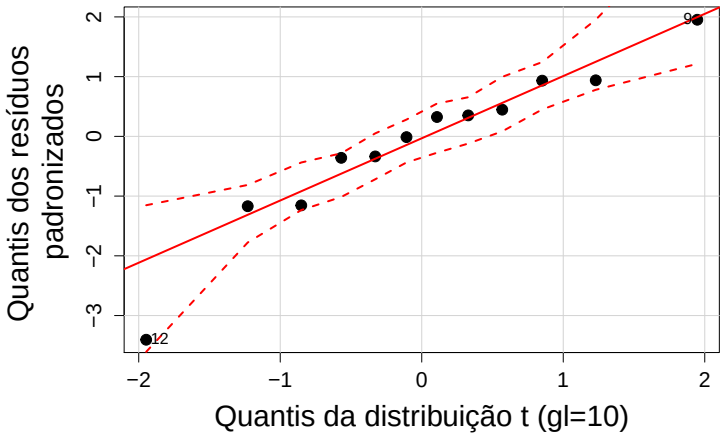

Figura 6.14: Gráfico QQ correspondente ajuste do modelo de regressão linear simples aos dados da Tabela 6.3 (com todas as observações).

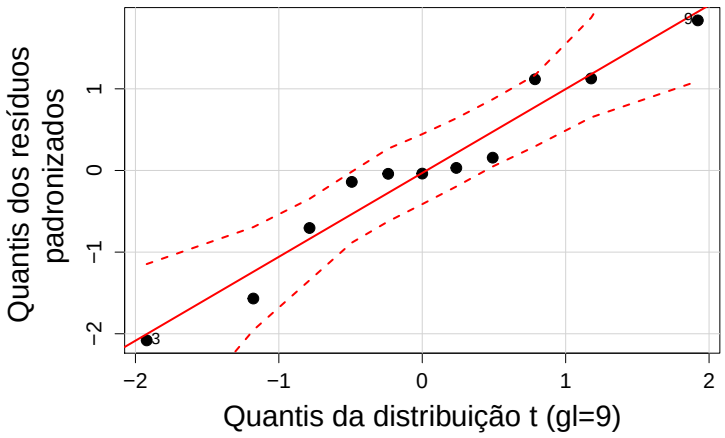

Figura 6.15: Gráfico QQ correspondente ajuste do modelo de regressão linear simples aos dados da Tabela 6.3 (sem a observação influente).

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 6.14). 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 (β), ou seja, avaliar a taxa com que a resposta esperada muda por unidade de variação da variável explicativa, essa suposição de normalidade da variável resposta tem efeito marginal na distribuição do estimador de mínimos quadrados desse parâmetro ($\widehat{\beta}$). 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 temporais**; 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 (veja a 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 6.16.

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

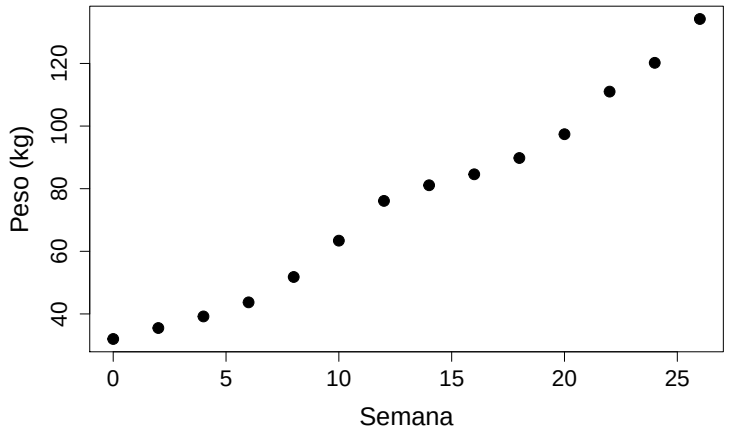

Figura 6.16: Gráfico de dispersão para os dados da Tabela 6.4

Tendo em vista o gráfico apresentado na Figura 6.16, um possível modelo para representar o valor esperado do peso em função do tempo é

$$y_t = \alpha + \beta t + \gamma t^2 + e_t, \tag{6.10}$$

$t = 1, \ldots, 14$ 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 6.17 mostra sequências de resíduos positivos seguidas de sequências de resíduos negativos, sugerindo uma possível correlação positiva entre eles (**autocorrelação**).

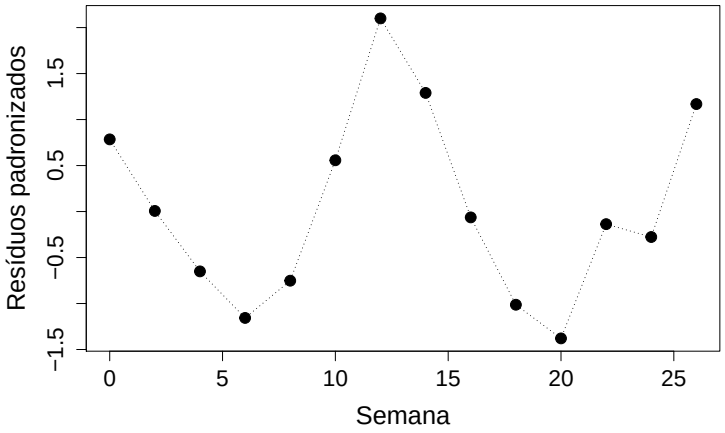

Figura 6.17: Resíduos estudentizados obtidos do ajuste do modelo (6.10)

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 $u_t \sim N(0, \sigma^2)$, $t = 1, \ldots, n$ são variáveis aleatórias 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 $\text{Cov}(e_t, e_{t-s}) = \rho^s[\sigma^2/(1-\rho^2)]$. O termo ρ é conhecido como **coeficiente de autocorrelação**.

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 tiverem uma forte correlação positiva, 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 e Watson (1950, 1951, 1971) produziram tabelas da distribuição da estatística D que

podem ser utilizadas 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 **método de mínimos quadrados generalizados** ou por métodos de **Séries Temporais**. Para detalhes sobre essas técnicas o leitor pode consultar Kutner et al. (2004) ou Morettin e Tolói (2018), respectivamente.

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 e 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, porque a análise dos dados completos 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

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 6.18 (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).

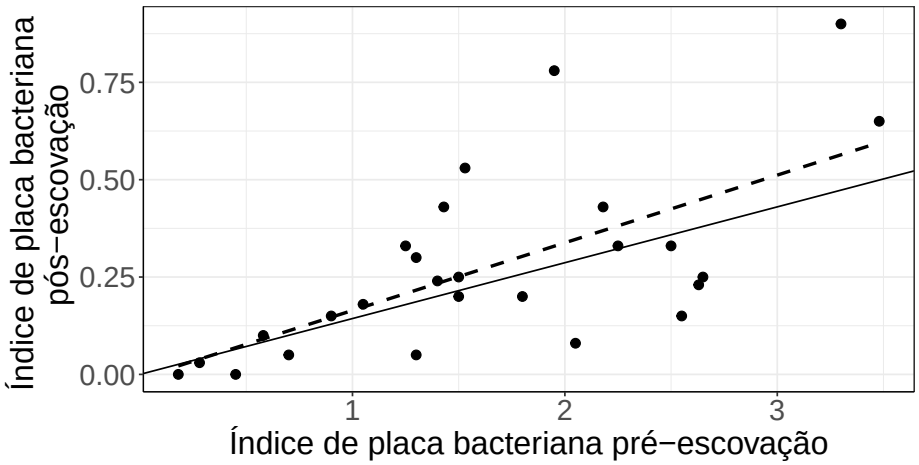

Figura 6.18: Gráfico de dispersão para os dados da Tabela 6.5; curva sólida para o modelo de regressão linear simples sem intercepto e curva interrompida para o modelo linearizado (6.20).

Singer e 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 $(0 < \beta < 1)$, maior será a redução do índice de placa.

Na Figura 6.19 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.

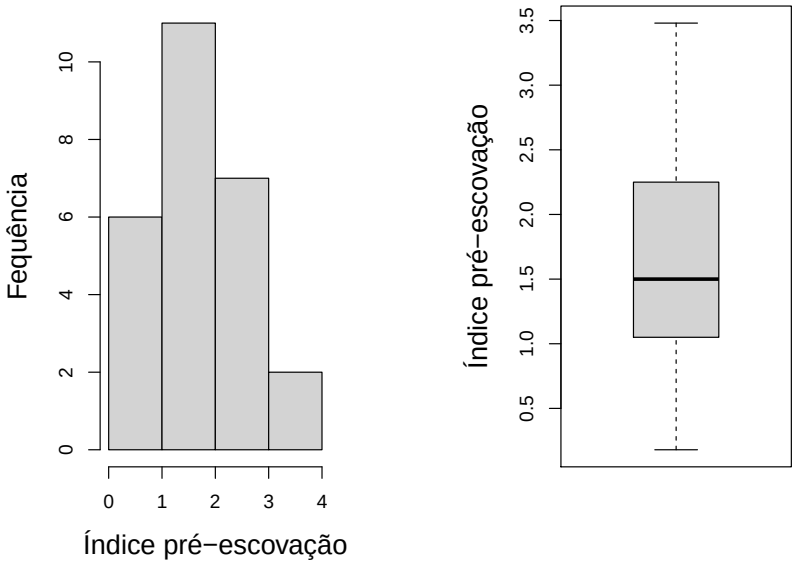

Figura 6.19: Histograma e *boxplot* para o índice de placa bacteriana pré escovação.

Esses resultados sugerem que uma transformação da forma $z^{(\theta)}$, com $0 \leq \theta < 1$, pode ser adequada para tornar os dados mais simétricos e estabilizar a variância (consulte a Seção 3.10 para detalhes sobre transformações de variáveis). Poderíamos 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 **linearização** do modelo, deixando-o no formato de um modelo de regressão linear simples para o qual dispomos de técnicas de ajuste amplamente conhecidas. Esse modelo, no entanto, exige que eliminemos os dois pares de

casos para os quais $Y = 0$, reduzindo para 24 o número de observações. O modelo resultante obtido com a transformação logarítmica é

$$y_i^* = \beta^* + \gamma x_i^* + e_i^*, \quad i = 1, \ldots, 24, \tag{6.15}$$

em que $y_i^* = \log(y_i)$, $x_i^* = \log(x_i)$, $\beta^* = \log \beta$ e $e_i^* = \log(e_i)$ são erros, que supomos ter média 0 e variância σ^2. Se, adicionalmente, supusermos que e_i^* tem distribuição normal, os erros originais, e_i terão distribuição log-normal, definida apenas para valores positivos, o que é compatível com as suposições adotadas para o modelo (6.14).

Na Figura 6.20 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.

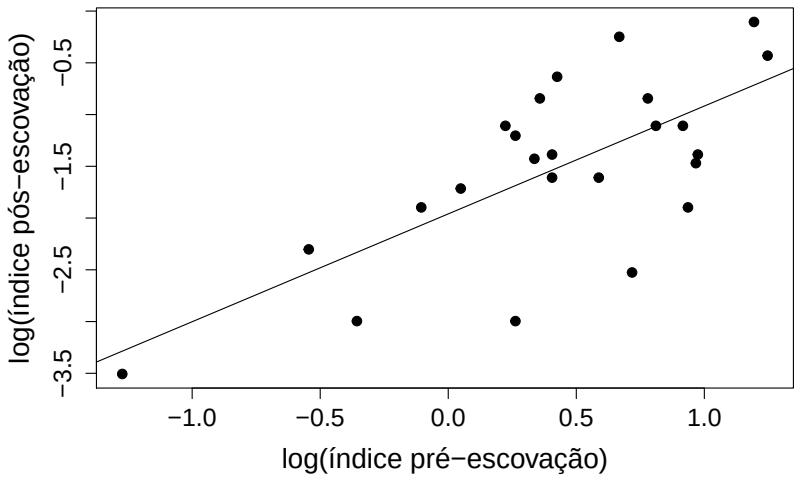

Figura 6.20: Gráfico de dispersão para os dados da Tabela 6.5 sob transformação logarítimica.

Usando o método de mínimos quadrados, a reta ajustada é

$$\widehat{y}_i^* = -1{,}960 + 1{,}042 x_i^*. \tag{6.16}$$

que corresponde a

$$\widehat{y}_i = 0{,}141 x_i^{1{,}042} \tag{6.17}$$

na concepção original, já que $\widehat{\beta} = \exp(\widehat{\beta}^*) = \exp(-1{,}960) = 0.141$. 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, por exemplo). O valor-p correspondente ao valor observado da estatística t é

$$p = P(\widehat{\gamma} > 1{,}042) = P\left[t_{22} > \frac{1{,}042 - 1}{S}\sqrt{\sum(x_i^* - \overline{x^*})^2}\right].$$

Como $\sum_i x_i^* = 10{,}246$, $\overline{x^*} = 0{,}427$, $\sum_i y_i^* = -36{,}361$, $\overline{y^*} = -1{,}515$, $\sum_i (x_i^* - \overline{x^*})^2 = \sum_i (x_i^*)^2 - 24(\overline{x^*})^2 = 12{,}149 - 24 \times 0{,}182) = 7{,}773$, obtemos $S^2 = 7{,}773/22 = 0{,}353$ e $S = 0{,}594$. Então

$$p = P[t_{22} > 0{,}42 \times 2{,}788/0{,}594] = P[t_{22} > 1{,}971] \approx 0{,}06$$

indicando que não há evidências fortes para rejeitar H_0. Como consequência, podemos dizer que a taxa esperada de redução da placa,

$$\mathrm{E}(y_i)/x_i = \beta \mathrm{E}(e_i),$$

é constante. Como concluímos que $\gamma = 1$, o modelo linear nos logaritmos das variáveis (6.15) fica reduzido a

$$y_i^* = \beta^* + x_i^* + e_i^*, \tag{6.18}$$

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

$$\widehat{y_i^*} = -1{,}942 + x_i^* \tag{6.19}$$

que em termos das variáveis originais corresponde a

$$\widehat{y}_i = 0{,}1434 x_i, \quad i = 1, \ldots, 24. \tag{6.20}$$

Esse modelo corresponde a uma reta que passa pela origem e está representada por meio de uma linha sólida na Figura 6.18.

Os resíduos dos modelos (6.14) e (6.15) estão representados nas Figuras 6.21 e 6.22, respectivamente. Esses gráficos sugerem que os resíduos dos dois modelos estão aleatoriamente distribuídos em torno de zero, mas não são totalmente compatíveis com a distribuição adotada, pois sua variabilidade não é constante. Para uma análise do conjunto de dados do qual este exemplo foi extraído e em que a suposição de heterocedasticidade é levada em conta, o leitor deve consultar Singer e Andrade (1997).

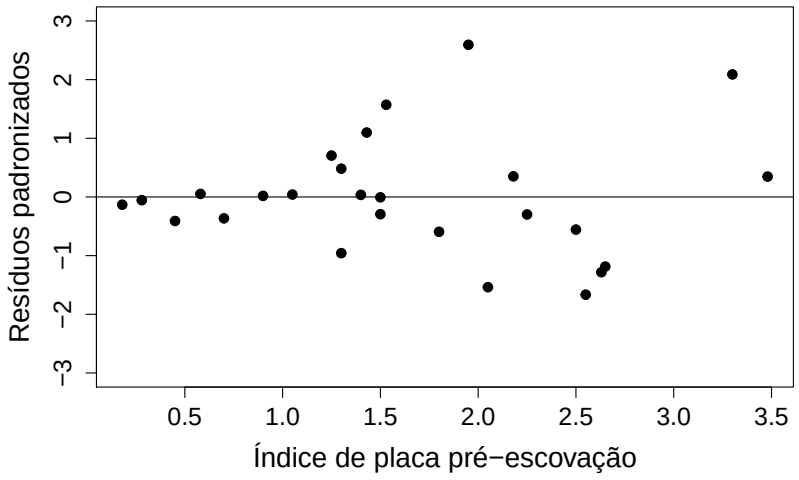

Figura 6.21: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados da Tabela 6.5.

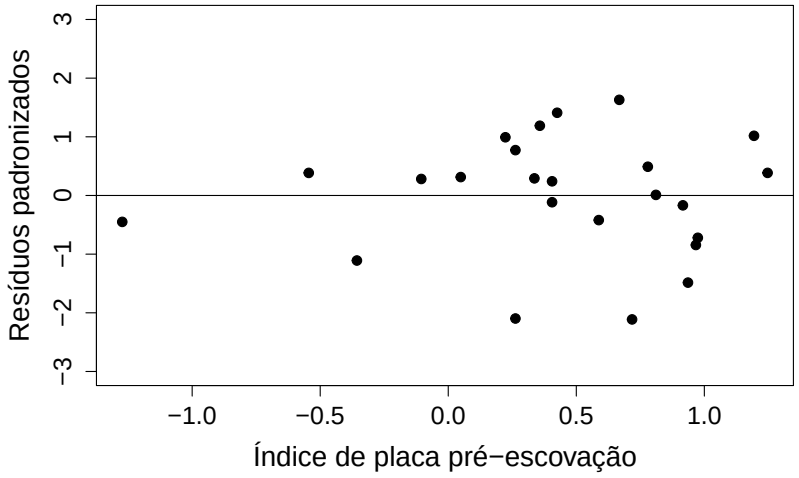

Figura 6.22: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados logaritmizados da Tabela 6.5.

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** $\widehat{\beta}_j$, $j = 1, \ldots, p$, 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 apropriada para ajuste de modelos de regressão

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`) foram extraídos de um estudo em que um dos objetivos era 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.

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

Para associar a distribuição do consumo de oxigênio (Y) com as informações sobre IMC (X_1) e carga na esteira ergométrica (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 variação esperada no VO2 por unidade do IMC para indivíduos com a mesma carga na esteira. O parâmetro β_2 tem interpretação semelhante com a substituição de IMC por carga na esteira e carga na esteira por IMC. 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 $X_1 - 17{,}50$ e X_2 por $X_2 - 32$, 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, pode-se demonstrar que

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 mormal 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** (veja a Nota de Capítulo 1) permite mostrar que a distribuição **aproximada** do estimador (6.26) é normal, com média $\boldsymbol{\beta}$ 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 do 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, e também dos valores ajustados.

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 6.23. Os gráficos de dispersão correspondentes a cada uma das duas variáveis explicativas estão dispostos na Figura 6.24 e indicam que a distribuição do VO2 varia positivamente com a carga na esteira e negativamente com o IMC.

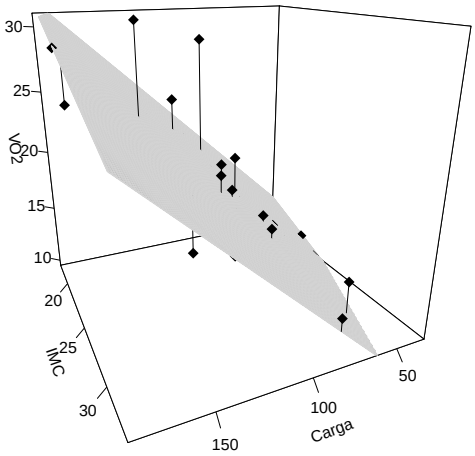

Figura 6.23: Gráficos de dispersão tridimensional para os dados da Tabela 6.7.

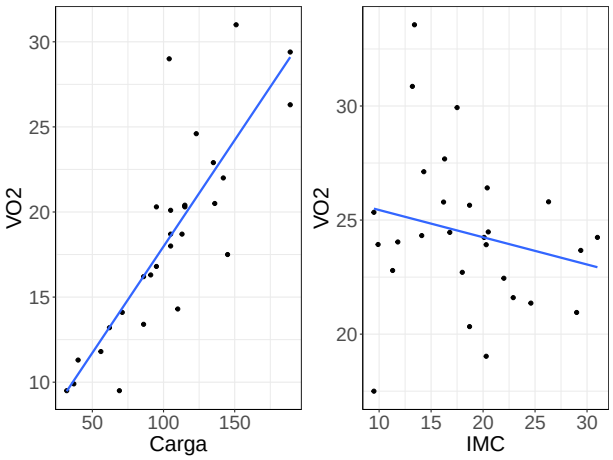

Figura 6.24: Gráficos de dispersão para os dados da Tabela 6.7.

Por meio da função `lm()`, podemos ajustar o modelo, obtendo os seguintes resultados

```
> lm(formula = VO2 ~ IMC + carga, data = esteira)
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 ***
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
```

Esses resultados indicam que os coeficientes e erros padrões (entre parênteses) 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 indivíduos com o mesmo IMC 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 ajustado, $R^2_{aj} = 0{,}74$, 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 6.25 juntamente com o gráfico contendo as distâncias de Cook.

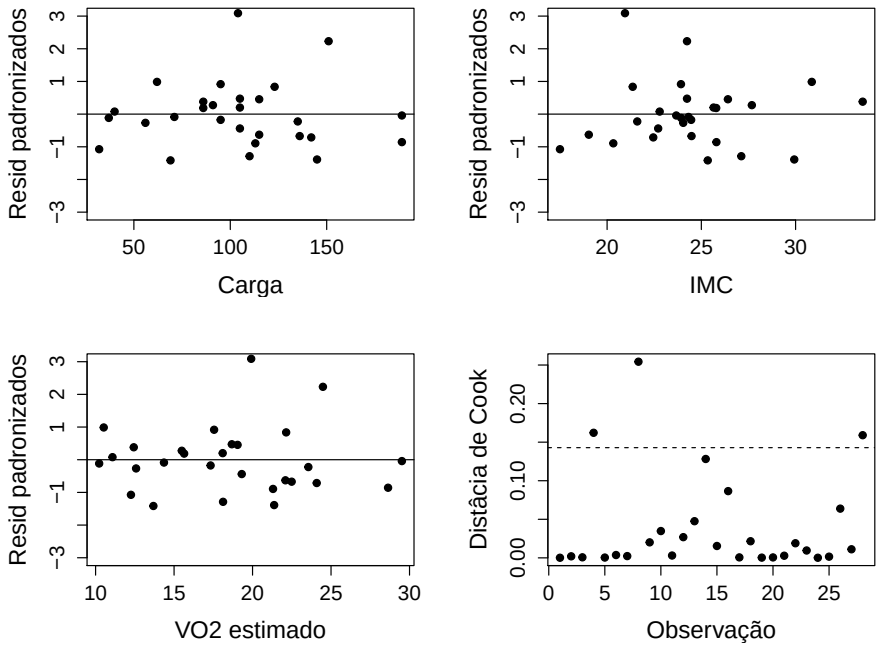

Figura 6.25: Gráficos de resíduos padronizados e distâncias de Cook para o ajuste do modelo (6.24) aos dados da Tabela 6.7.

Os gráficos de resíduos padronizados não indicam um comprometimento da hipótese de homocedasticidade embora seja possível suspeitar de dois ou três pontos atípicos (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 na literatura (não sem controvérsias), em que resíduos associados a distâncias de Cook maiores que $4/n$ [ou $4/(n-p)$] são considerados resíduos associados a **pontos influentes**. Em todo o caso, convém lembrar que o propósito dessas ferramentas é essencialmente exploratório e que as decisões sobre a exclusão de pontos atípicos 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 6.26.

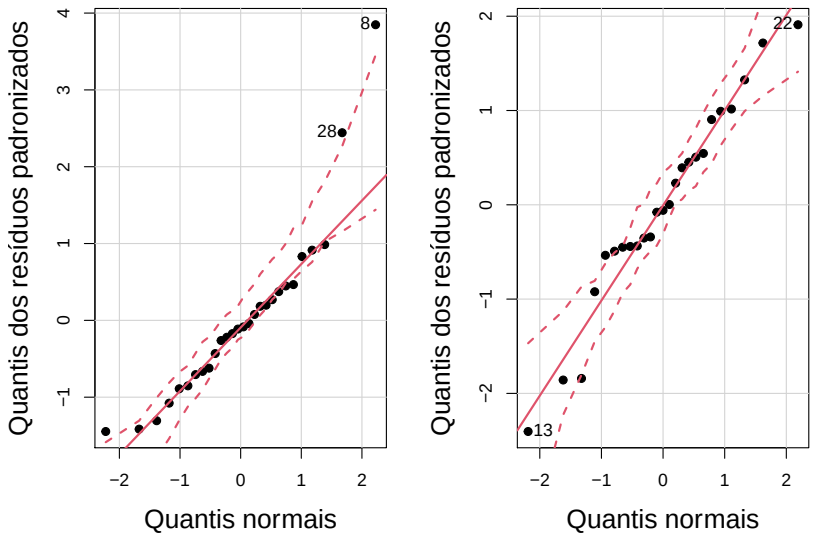

Figura 6.26: Gráficos QQ correspondentes ao ajuste do modelo (6.24) aos dados da Tabela 6.7 com (painel esquerdo) e sem (painel direito) os pontos com rótulos 4, 8 e 28.

O ajuste do modelo aos 25 dados obtidos com a exclusão dos pontos rotulados 4, 8 e 28 pode ser realizado por meio dos comandos

```
> esteirasem <- subset(dados, (dados$ident!=4 & dados$ident!=8 &
                               dados$ident!=28))
> esteirasem.fit <- lm(VO2 ~ IMC + carga, data = esteirasem)
> summary(esteirasem.fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.89307    3.47071   4.291 0.000296 ***
IMC         -0.35631    0.12606  -2.827 0.009823 **
carga        0.11304    0.01052  10.743 3.23e-10 ***
Residual standard error: 1.987 on 22 degrees of freedom
Multiple R-squared:  0.8581,Adjusted R-squared:  0.8452
```

O coeficiente de determinação ajustado é $R^2_{aj} = 0{,}85$. Os gráficos de dispersão, resíduos padronizados e de distâncias de Cook correspondentes estão dispostos na Figura 6.27 e também sugerem um ajuste melhor.

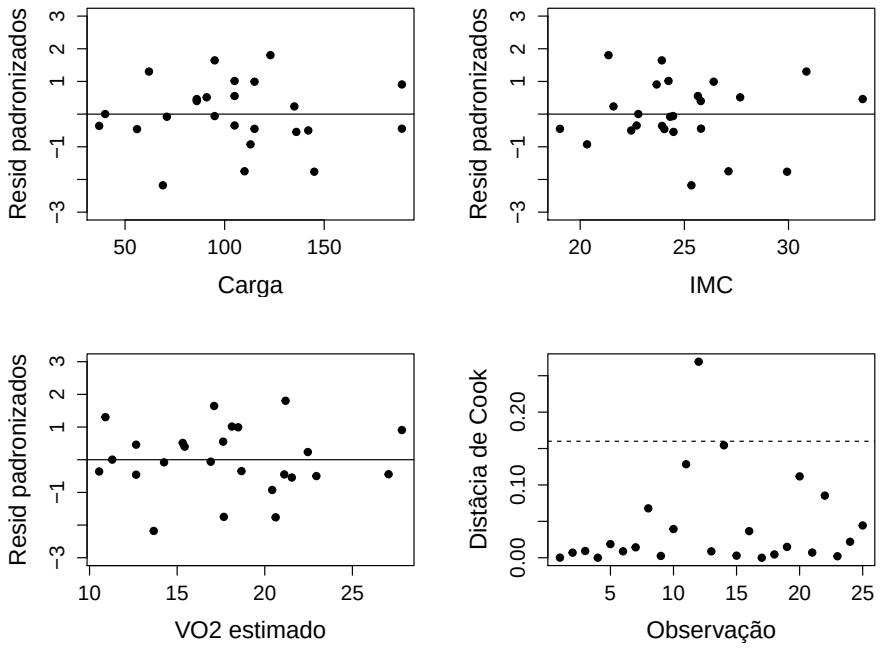

Figura 6.27: Gráficos de dispersão, resíduos padronizados e de distância de Cook correspondentes ao ajuste do modelo (6.24) aos dados da Tabela 6.7 sem os pontos com rótulos 4, 8 e 28.

Sob o modelo ajustado aos dados com as três observações excluídas, uma estimativa do valor esperado do VO2 **para indivíduos** com IMC = 25 submetido a uma carga na esteira igual a 100 e o correspondente intervalo de confiança podem ser obtidos por meio dos comandos

```
> beta0 <- esteirasem.fit$coefficients[1]
> beta1 <- esteirasem.fit$coefficients[2]
> beta2 <- esteirasem.fit$coefficients[3]
> yestim <- beta0 + beta1*25 + beta2*100
> round(yestim, 2)
     17.29
> s <- summary(esteirasem.fit)$sigma
> xtxinv <-summary(esteirasem.fit)$cov.unscaled
> s
[1] 1.987053
> xtxinv
             (Intercept)           IMC         carga
(Intercept)  3.050826884 -1.044085e-01 -3.970913e-03
IMC         -0.104408474  4.024411e-03  4.174432e-05
carga       -0.003970913  4.174432e-05  2.804206e-05
> xval <- c(1, 25, 100)
> varyestim <-s^2*xval%*%xtxinv%*%xval
> liminfic95 <- yestim - 1.96**sqrt(varyestim)
```

```
> limsupic95 <- yestim + 1.96**sqrt(varyestim)
> round(liminfic95, 2)
     15.98
> round(limsupic95, 2)
    18.6
```

A previsão para o valor do VO2 para um **um indivíduo genérico** é a mesma que aquela obtida para a estimação do correspondente valor esperado, ou seja 17,29. No entanto ao intervalo de previsão associado, deve ser acrescentada a variabilidade da resposta relativamente à sua média (veja a Nota de Capítulo 2). Esse intervalo é obtido com os comandos

```
> liminfip95 <- yestim - 1.96*sqrt(varyestim + s^2)
> limsupip95 <- yestim + 1.96*sqrt(varyestim + s^2)
> round(liminfip95, 2)
      [,1]
[1,] 13.32
> round(limsupip95, 2)
      [,1]
[1,] 21.26
```

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 esperada varia em função da potência instalada e da área construída. Os gráficos de dispersão entre a variável resposta (produção) e cada uma das variáveis explicativas estão dispostos na Figura 6.28 e sugerem que essas duas variáveis são linear e positivamente 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

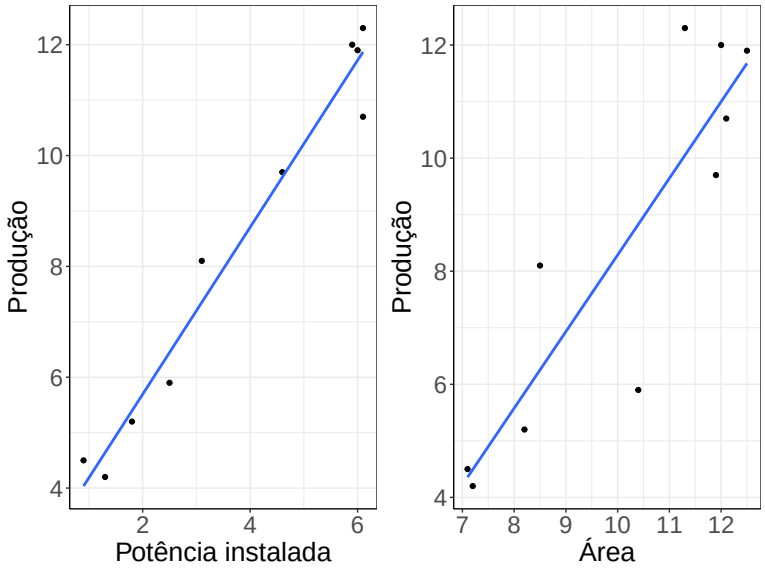

Figura 6.28: Gráficos de dispersão correspondentes aos dados da Tabela 6.8.

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 6.28 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 esperada na presença da outra. O valor-p associado ao teste da hipótese de que $\beta_2 = 0$ é $p = 0{,}34$ e sugere que esse coeficiente pode ser considerado nulo. Em resumo, 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 6.29 e corroboram a conclusão de que apenas a variável potência instalada é suficiente para a explicação da variação da produção esperada.

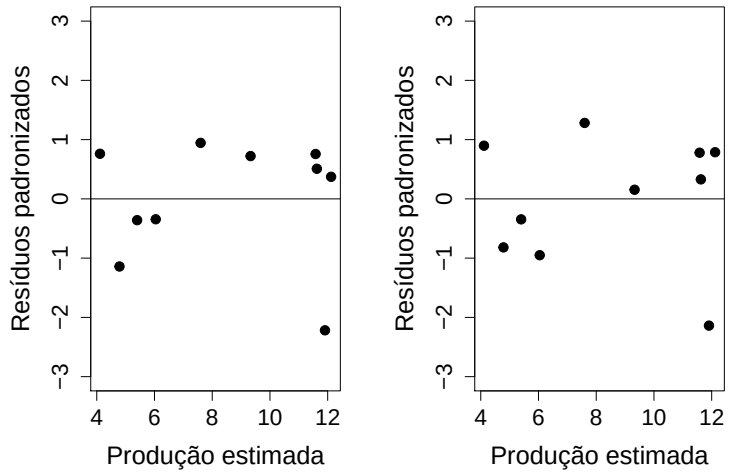

Figura 6.29: Gráficos de resíduos padronizados correspondentes aos modelos ajustados aos dados da Tabela 6.8 com duas (painel esquerdo) ou uma (painel direito) variável explicativa (potência).

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 (veja a Nota de Capítulo 3 para detalhes).

Uma outra ferramenta útil para avaliar a importância marginal de uma variável explicativa 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 construído 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 com duas variáveis.

Para o exemplo da Tabela 6.8 o gráfico da variável adicionada está apresentado na Figura 6.30. O coeficiente da reta ajustada ($-0{,}26$) não é significativo, sugerindo que a variável X_2 não precisa ser utilizada para explicar a variação esperada da variável resposta. Compare a inclinação (negativa) da reta representada nessa figura com aquela (positiva) da reta representada no painel direito da Figura 6.28.

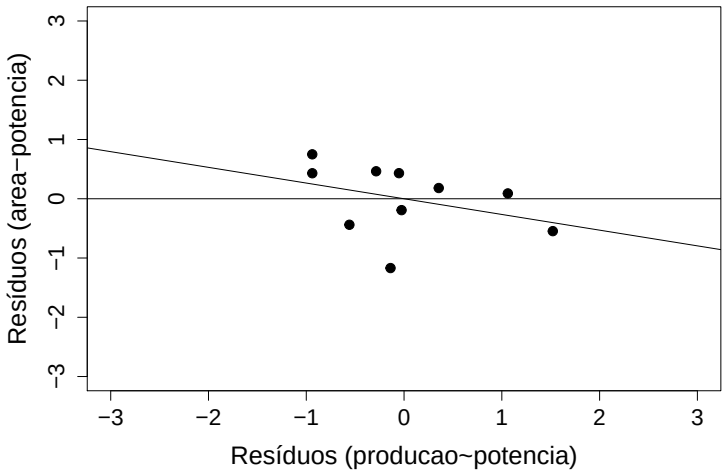

Figura 6.30: Gráfico da variável adicionada correspondente ao modelo ajustado aos dados da Tabela 6.8.

6.4 Regressão para dados longitudinais

Consideremos os dados do Exemplo 2.3 (disponíveis no arquivo `bezerros` e dispostos na Tabela 2.2) correspondentes a um estudo cujo objetivo é avaliar a variação de peso de bezerros entre a $12a$ e a $26a$ 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 6.31.

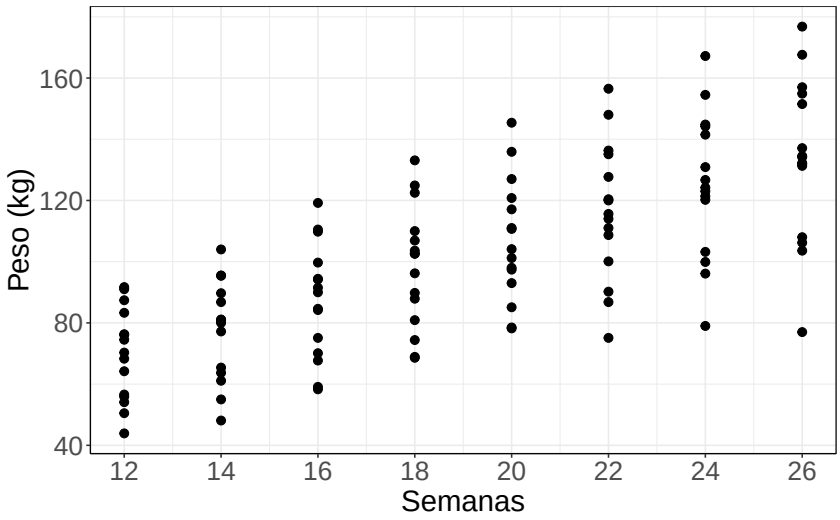

Figura 6.31: Gráfico de dispersão para os dados da Tabela 2.2.

Esse gráfico sugere que o crescimento dos animais poderia 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 os erros 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 comportamento, convém construir um **gráfico de perfis** em que as observações realizadas num mesmo animal são ligadas por segmentos de reta, como indicado na Figura 6.32. A correlação entre as observações realizadas no mesmo animal fica evidenciada no gráfico do desenhista disposto na Figura 6.33.

Pode-se notar que no gráfico de perfis, a variabilidade da resposta é similar em todos os instantes de observação e que os perfis individuais têm aproximadamente as mesmas inclinações. 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 incorpora essas características é um **modelo linear misto** expresso por

$$y_{ij} = \alpha + \beta(x_j - 12) + a_i + e_{ij}, \tag{6.28}$$

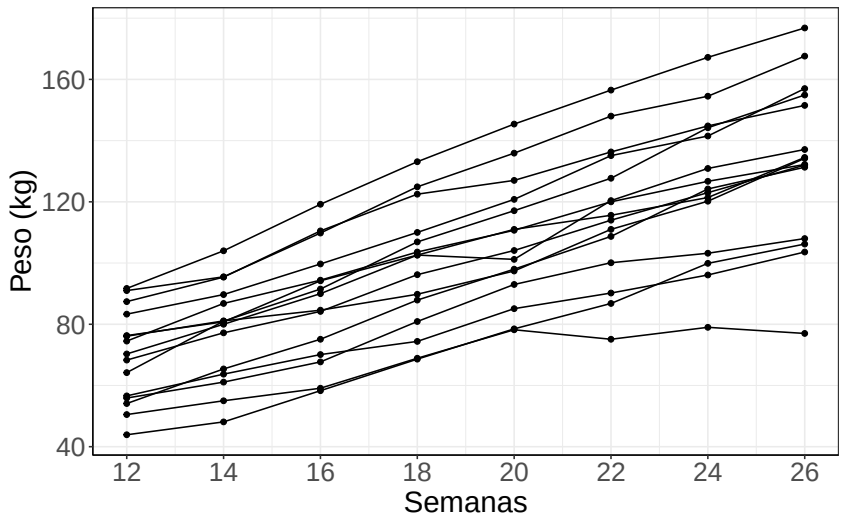

Figura 6.32: Gráfico de perfis para os dados da Tabela 2.2.

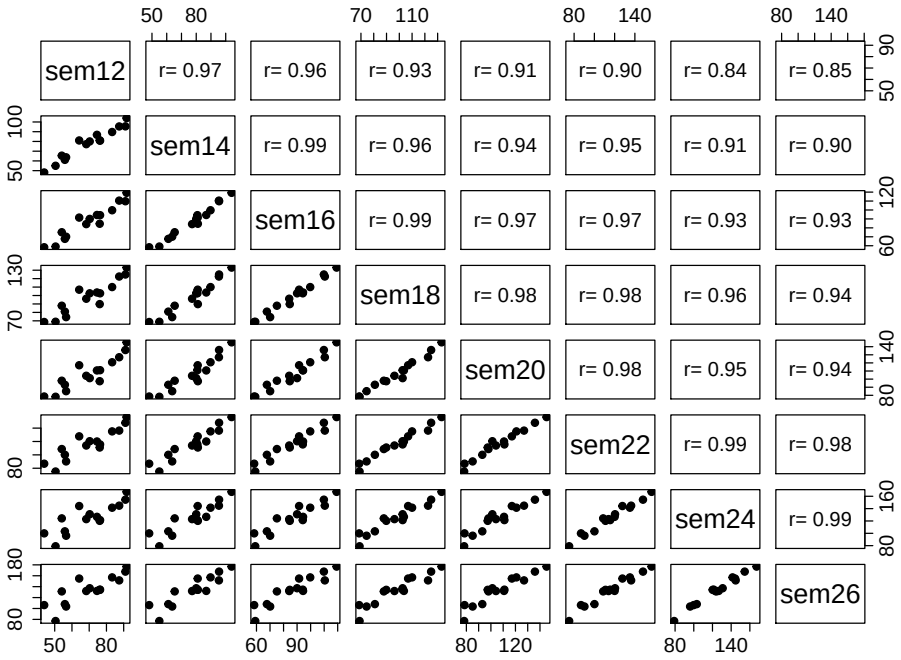

Figura 6.33: Gráfico do desenhista para os dados da Tabela 2.2.

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 homocedástico, 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 (a_i) porque admitimos que 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. Se o foco da análise se restringisse apenas aos 15 animais observados, um modelo de regressão linear simples expresso como

$$y_{ij} = \alpha_i + \beta(x_j - 12) + e_{ij}$$

com as suposições usuais para os erros e_{ij} seria adequado. Nesse caso, o parâmetro α_i corresponderia ao valor esperado do peso do i-ésimo animal na semana 12 e não haveria elementos para se fazer inferência sobre o correspondente valor esperado populacional.

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, nomeadamente, **resíduos condicionais** (associados à variabilidade intraunidades amostrais), **resíduos dos efeitos aleatórios** (associados à variabilidade interunidades amostrais), e **resíduos marginais** (associados à variabilidade de cada observação em relação aos valores populacionais), 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).

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

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 resposta Y com valor igual a 1 no caso de resposta positiva e igual a zero no caso de resposta negativa, a resposta esperada será igual à probabilidade $p = \mathrm{E}(Y)$ de que pacientes tenham resposta positiva. Assim como no caso de modelos de regressão linear, o objetivo da análise é modelar a resposta esperada, que neste caso é uma probabilidade, como função da variável explicativa. Por razões técnicas e de interpretação, em vez de modelar essa resposta esperada, convém modelar uma função dela, a saber, o logaritmo da chance de resposta positiva (consulte a Seção 4.2) para evitar estimativas de probabilidades com valores fora do intervalo (0, 1). O modelo correspondente pode ser escrito como

$$\log \frac{P(Y_i = 1|X = x_i)}{P(Y_i = 0|X = x_i)} = \alpha + \beta x_i,\ i = 1, \ldots, n. \tag{6.29}$$

ou equivalentemente (veja o Exercício 20), como

$$P(Y_i = 1|X = x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)},\ i = 1, \ldots, n. \tag{6.30}$$

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 as chances de resposta positiva para pacientes em que a variável explicativa difere por uma unidade (veja o Exercício 21).

O ajuste desse modelo é realizado pelo **método de máxima verossimilhança**[3].

Dada uma amostra $\{\mathbf{x}, \mathbf{y}\} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, a função de verossimilhança a ser maximizada é

$$\ell(\alpha, \beta | \{\mathbf{x}, \mathbf{y}\}) = \prod_{i=1}^{n} [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}$$

com

$$p(x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

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 os métodos **Newton-Raphson** ou **Fisher scoring**. Para detalhes, o leitor poderá consultar Paulino e Singer (2006), por exemplo.

O ajuste do modelo de regressão logística aos dados do Exemplo 6.9 por meio da função `glm()` produz os resultados a seguir:

```
> modelo1 <- glm(formula = resposta ~ difinib, family = binomial,
                 data = dados)
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 **
(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
```

[3]Para detalhes sobre o método de máxima verossimilhança, o leitor poderá consultar Bussab e Morettin (2017) ou Bickel e Doksum (2015), entre outros.

As estimativas dos parâmetros α e β (com erro padrão entre parênteses) correspondentes ao modelo ajustado aos dados da Tabela 6.9 são, respectivamente, $\widehat{\alpha} = -2{,}310$ (0,947) e $\widehat{\beta} = 0{,}026$ (0,009). Consequentemente, uma estimativa da chance de resposta positiva para pacientes com mesmo nível de inibina B pré e pós estímulo hormonal é $\exp(\widehat{\alpha}) = 0{,}099$ (0,094). Essa chance fica multiplicada por $\exp(\widehat{\beta}) = 1{,}026$ (0,009) para cada aumento de uma unidade na diferença entre os níveis de inibina B pré e pós estímulo hormonal.[4]

Dada a natureza não linear do modelo adotado, intervalos de confiança para essa chance e para essa razão de chances devem ser calculados a partir da exponenciação dos limites dos intervalos de confiança associados aos parâmetros α e β. Para os dados do exemplo em análise, esses intervalos de confiança podem ser obtidos por intermédio dos comandos

```
> ic95chance0 <- exp(modelo1$coefficients[1] +
    c(-1, 1) * 1.96*sqrt(summary(modelo1)$cov.scaled[1,1]))
> ic95chance0
[1] 0.01549197 0.63541545
> ic95razaochances <- exp(modelo1$coefficients[2] +
    c(-1, 1) * 1.96*sqrt(summary(modelo1)$cov.scaled[2,2]))
> ic95razaochances
[1] 1.009228 1.043671
```

A função `predict.glm()` pode ser usada para estimar a probabilidade de resposta positiva, dado algum valor da variável explicativa. Os dados da Tabela 6.10 indicam diferentes níveis da diferença inibina B pré e pós estímulo hormonal e as correspondentes probabilidades de resposta positiva previstas pelo modelo.

Tabela 6.10: Diferenças entre os níveis de inibina B pré e pós estímulo hormonal e probabilidades de resposta positiva previstas

Difinib	10	50	100	200	300	400	500
Prob	0,11	0,27	0,57	0,95	0,99	1,00	1,00

Por exemplo, o valor 0,57 correspondente a uma diferença inibina B pré e pós igual a 100 foi obtido calculando-se

$$\widehat{P}(X = 1|X = 100) = \frac{\exp\{-2{,}310455 + (0{,}025965)(100)\}}{1 + \exp\{-2{,}310455 + (0{,}025965)(100)\}}. \qquad (6.31)$$

Para classificar a resposta como positiva ou negativa, é preciso 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:

[4] Os erros padrões de $\exp(\widehat{\alpha})$ e $\exp(\widehat{\beta})$ são calculados por meio do **método Delta**. Veja a Nota de Capítulo 6.

```
          resposta
glm.pred negativa positiva
negativa    11        5
positiva     2       14
```

Os elementos da diagonal dessa tabela indicam os números de observações corretamente classificadas pelo modelo. Ou seja, a proporção de respostas corretas será (11+14)/32= 78%. Esse valor depende do limiar fixado, igual a 0,7, no caso. Um *default* usualmente fixado é 0,5, e nesse cenário, a proporção de respostas corretas pode diminuir. A utilização de Regressão Logística nesse contexto de classificação será detalhada no Capítulo 9.

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 (veja a 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** [para detalhes, consulte Paulino e Singer (2006), por exemplo].

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 6.11.

Tabela 6.11: 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.

a) 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).

b) O parâmetro β corresponde ao logaritmo da razão entre a chance de lesão coronariana $\geq 50\%$ para homens não hipertensos com menos de 55 anos e a chance correspondente para mulheres com as mesmas características (de idade e de hipertensão).

c) 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 (de sexo e de hipertensão) e menos de 55 anos.

d) 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.

Com o recurso ao pacote `ACD` e à função `loglinWLS()`, as estimativas dos parâmetros obtidas pelo método de mínimos quadrados generalizados (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 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 6.12.

Tabela 6.12: Estimativas (e intervalos de confiança de 95%) para a chance e razões de chances associadas aos dados da Tabela 6.11

	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

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 não possa ser estimada (como em estudos retrospectivos), 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 6.13.

Tabela 6.13: Estimativas das chances de lesão coronariana para $\geq 50\%$ para pacientes com diferentes níveis dos fatores de risco obtidas com os dados da Tabela 6.11

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

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 $2{,}94 (= 1{,}95 \times 1{,}51)$ vezes 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.

O modelo de regressão logística pode ser generalizado para o caso em que a variável resposta tem mais do que dois possíveis valores. Por exemplo, num estudo em que se quer avaliar a associação entre textura, cor, gosto de um alimento e o seu destino comercial, a variável resposta pode ter as categorias "descarte", "varejo nacional" ou "exportação". Nesse caso, o modelo é conhecido como **regressão logística politômica** ou **regressão logística multinomial** e não será abordado neste texto. O leitor poderá consultar Paulino e Singer (2006) para detalhes. Para efeito de classificação, técnicas alternativas e mais empregadas na prática serão abordadas no Capítulo 9.

6.6 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. 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 estimadores de mínimos quadrados são não enviesados.

b) $\text{Var}(\widehat{\alpha}) = \sigma^2 \sum_{i=1}^n x_i^2/[n\sum_{i=1}^n (x_i - \overline{x})^2]$.

c) $\text{Var}(\widehat{\beta}) = \sigma^2/\sum_{i=1}^n (x_i - \overline{x})^2$.

d) $\text{Cov}(\widehat{\alpha},\widehat{\beta}) = -\sigma^2\overline{x}/\sum_{i=1}^n (x_i - \overline{x})^2$.

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 cenário, 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 estudentizados**.

Com esses resultados é possível testar as hipóteses $H_0 : \alpha = \alpha_0$ e $H_0 : \beta = \beta_0$, em que α_0 e β_0 são constantes conhecidas (usualmente iguais a zero), bem como construir intervalos de confiança para esses parâmetros.

Um teorema importante conhecido como **Teorema de Gauss-Markov** (e que não inclui a suposição de normalidade dos erros) afirma que os os estimadores de mínimos quadrados têm variância mínima na classe dos estimadores não enviesados que sejam funções lineares das observações y_i.

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 de mínimos quadrados, $\widehat{\alpha}$ e $\widehat{\beta}$, têm distribuições aproximadamente normais com variâncias que podem ser estimadas pelas expressões apresentadas nos itens b) e c) indicados anteriormente. 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 $\text{E}(Y|X = x_0)$ 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

$$\text{Var}(\widehat{y}) = \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\Big].$$

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. Sem muito rigor, podemos dizer que esse intervalo deve conter o verdadeiro valor esperado de $\text{E}(Y|X = x_0)$, *i.e.*, o valor esperado 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 associado a uma determinada 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 $\text{E}(Y|X = x_0)$.

Como $Y_0 = \widehat{y} + e_0$, sua variância é

$$\text{Var}(Y_0) = \text{Var}(\widehat{y}) + \text{Var}(e_0) = \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x - \overline{x})^2}\Big] + \sigma^2.$$

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 de investigação 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 $S^2 = \sum_{i=1}^n (y_i - \overline{y})^2/(n-1)$. Para relacionar esse resultado com aquele obtido por meio de um modelo de regressão linear para os mesmos dados, podemos escrever

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}). \tag{6.33}$$

Uma representação gráfica dessa relação está apresentada na Figura 6.34.

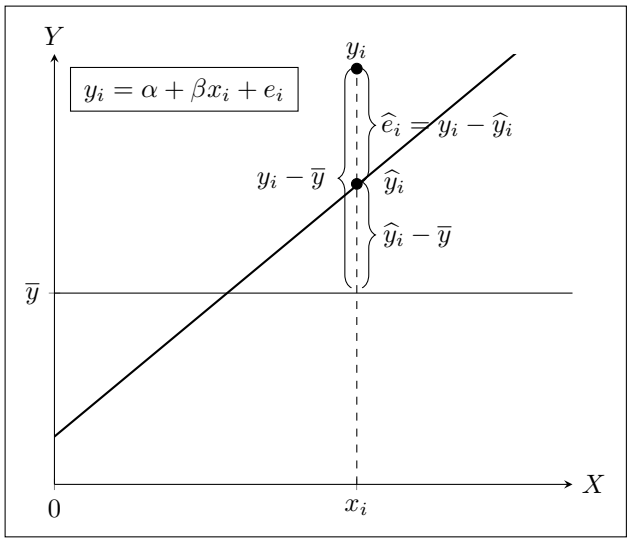

Figura 6.34: Representação gráfica da decomposição (6.33).

Pode-se mostrar (veja o Exercício 17) 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 escrita como a soma de um termo correspondente à variabilidade dos resíduos ($SQRes$) com outro correspondente à variabilidade explicada pela 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), pode-se calcular a soma de quadrados devida à regressão por meio de

$$SQReg = \widehat{\beta}^2 \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

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 (de dimensão n) com valores preditos 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. É possível mostrar que a distância de Cook 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 e alavancagem**.

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

$$\Delta\widehat{\beta}(y_j) = \frac{(x_j - \overline{x})\Delta y_j}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{6.34}$$

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.

Lembrando que a estimativa do intercepto é

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} = \frac{\sum_i y_i}{n} - \widehat{\beta}\overline{x},$$

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

$$\Delta\widehat{\alpha}(y_j) = \Big[\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2}\Big]\Delta y_j. \tag{6.35}$$

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 (6.34) e (6.35) 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.

Alavancagem (*leverage*) mede o efeito de uma variação na ordenada de um ponto particular (x_j, y_j) sobre o valor ajustado $\widehat{y}_j$. Observe que

$$\widehat{y}_j - \overline{y} = \widehat{\alpha} + \widehat{\beta}x_j - (\widehat{\alpha} + \widehat{\beta}\overline{x}) = \widehat{\beta}(x_j - \overline{x})$$

e portanto

$$\widehat{y}_j = \frac{\sum_i y_i}{n} + \frac{\sum_i (x_i - \overline{x})y_i}{\sum_i (x_i - \overline{x})^2}(x_j - \overline{x}),$$

e quando y_j é alterado para $y_j + \Delta y_j$, temos

$$\widehat{y}_j + \Delta\widehat{y}_j = \frac{\sum_i y_i + \Delta y_j}{n} + \frac{\sum_i (x_i - \overline{x})y_i + (x_j - \overline{x})\Delta y_j}{\sum_i (x_i - \overline{x})^2}(x_j - \overline{x})$$

e, então,

$$\Delta\widehat{y}_j = \left[\frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right] \Delta y_j = h_{jj}\Delta y_j.$$

O fator h_{jj} é chamado **repercussão** e depende, basicamente, da distância entre x_j e $\overline{x}$. Se x_j for muito menor ou muito maior que $\overline{x}$, um acréscimo no valor de $\widehat{y}_j$ vai "empurrar" a ordenada do ponto, y_j, correspondente para baixo ou para cima, respectivamente.

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 função $g(\widehat{\beta})$. Por meio de uma expansão de Taylor, pode-se mostrar que sob certas condições usualmente válidas na prática,

$$\text{Var}[g(\widehat{\beta})] \approx [g'(\widehat{\beta})]^2 \sigma^2_{\widehat{\beta}},$$

em que $g'(\widehat{\beta})$ denota a primeira derivada de g calculada no ponto $\widehat{\beta}$.

No caso multivariado, em que $\widehat{\boldsymbol{\beta}}$ é um estimador de dimensão $p \times 1$ com matriz de covariâncias $\text{Var}(\widehat{\boldsymbol{\beta}}) = \mathbf{V}(\widehat{\boldsymbol{\beta}})$ sob certas condições usualmente satisfeitas na prática, a variância de uma função $g(\widehat{\boldsymbol{\beta}})$ pode ser aproximada por

$$\text{Var}[g(\widehat{\boldsymbol{\beta}})] \approx [\partial g(\widehat{\boldsymbol{\beta}})/\partial\widehat{\boldsymbol{\beta}}]^\top \mathbf{V}(\widehat{\boldsymbol{\beta}})[\partial g(\widehat{\boldsymbol{\beta}})/\partial\widehat{\boldsymbol{\beta}}]$$

em que $[\partial g(\widehat{\boldsymbol{\beta}})/\partial\widehat{\boldsymbol{\beta}}] = [\partial g(\widehat{\boldsymbol{\beta}})/\partial\widehat{\beta}_1, \ldots, \partial g(\widehat{\boldsymbol{\beta}})/\partial\widehat{\beta}_p]^\top$.

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.

Para contornar esse problema, consideremos um caso com p variáveis explicativas $\mathbf{x} = (x_1, \ldots, x_p)^\top$ e seja 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 $\widetilde{y}_j$ o número de respostas $Y = 1$ observadas entre as m_j unidades amostrais com o mesmo valor $\mathbf{x}_j$. A frequência esperada correspondente à frequência observada $\widetilde{y}_j$ é

$$m_j\widehat{p}_j = m_j\frac{\exp(\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}})},$$

em que $\widehat{\boldsymbol{\beta}}$ é o estimador do vetor de parâmetros de modelo.

O **resíduo de Pearson** é definido como

$$\widehat{e}_j = \frac{\widetilde{y}_j - m_j\widehat{p}_j}{\sqrt{m_j\widehat{p}_j(1-\widehat{p}_j)}}$$

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 aproximada χ^2 com $M-(p+1)$ graus de liberdade quando o modelo está bem ajustado.

Para evitar problemas com a distribuição aproximada de Q_P quando $M \approx n$, convém agrupar os dados de alguma forma. Hosmer e Lemeshow (1980, 2013) 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 estimadas (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 está 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 atípicas (*outliers*). Como alternativa, podemos considerar modelos de **regressão resistente**, em que os estimadores são baseados em medianas.

Para o ajuste de um desses modelos, inicialmente, dividimos o conjunto de n pontos em três grupos de tamanhos aproximadamente iguais. Chamemos esses grupos de E, C e D (de esquerdo, central e direito). 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 respectivamente, por

$$b_0 = \frac{y_D - y_E}{x_D - x_E},$$

e

$$a_0 = \frac{1}{3}\left[(y_E - b_0 x_E) + (y_C - b_0 x_C) + (y_D - b_0 x_D)\right].$$

Convém notar as diferenças entre b_0 e (6.6) e entre a_0 e (6.7). A correspondente reta resistente ajustada é

$$\widetilde{y}_i = a_0 + b_0 x_i, \quad i = 1, \ldots, n.$$

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 6.8 quanto o gráfico de resíduos (Figura 6.9) 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 6.35.

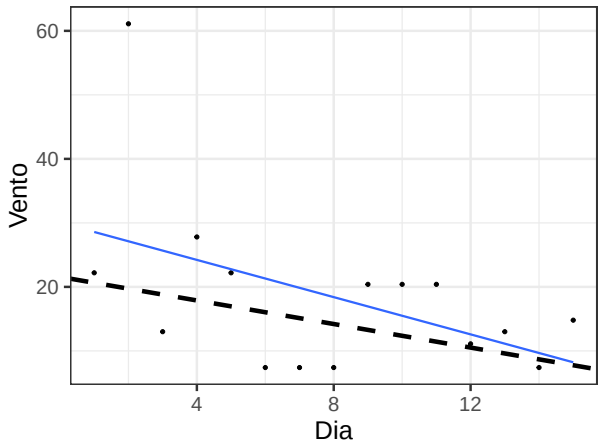

Figura 6.35: Gráfico de dispersão com retas de mínimos quadrados (linha cheia) e resistente (linha tracejada) correspondentes ao ajuste do modelo de regressão linear simples aos dados da Tabela 6.2.

Como nesse caso $n = 3 \times 5$, 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

$$\widetilde{v}_t = 21{,}56 - 0{,}92t.$$

Esta reta não é tão afetada pelo ponto discrepante (que não foi eliminado da análise).

9) **Formulação geral do modelo de regressão**

O modelo de regressão múltipla (6.21) pode ser escrito na forma

$$y_i = f(\mathbf{x}_i,\boldsymbol{\beta}) + e_i,\ i = 1,\ldots,n, \tag{6.36}$$

em que

$$f(\mathbf{x}_i,\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

com $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^\top$ e $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_p)^\top$. Esse modelo pode ser generalizado como

$$f(\mathbf{x},\boldsymbol{\beta}) = \sum_{j=0}^{M-1} \beta_j \phi_j(\mathbf{x}),$$

em que $\phi_j(\cdot)$, $j = 0,\ldots,M-1$ com $\phi_0(x) = 1$ são funções pertencentes a uma base de funções. Essa formulação é útil no contexto das **redes neurais** apresentadas no Capítulo 14.

Em notação matricial, temos

$$f(\mathbf{x},\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\mathbf{x}),$$

com $\boldsymbol{\phi}(\mathbf{x}) = [\phi_0(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x})]^\top$.

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 $[\phi_j(x) = \exp\{[-(x - \mu_j)/(2s)]^2\}]$, com μ_j denotando parâmetros de posição e s denotando o parâmetro de dispersão;

d) sigmoide $[\phi_j(x) = \sigma\{(x - \mu_j)/s\}]$ em que, $\sigma(a)$ pode ser qualquer uma das funções a)-d) da Seção 10.5.

e) Fourier, em que $[\phi_j(x)]$ é uma cossenoide [ver Morettin (2014)].

f) ondaletas, em que $[\phi_j(x)]$ é uma ondaleta [ver Morettin (2014)].

Tanto a técnica de mínimos quadrados quanto as técnicas de regularização apresentadas no Capítulo 8 podem ser aplicadas a essa formulação mais geral.

6.7 Exercícios

1) 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 6.14.

   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 6.14: Peso real e volume 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

2) 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, $i = 1, \ldots, 18$.

3) Os dados da Tabela 6.15 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 6.15: Área e Preço 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 $1000m^2$.

4) 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.

5) A Tabela 6.16 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 que não saiba implementar suas sugestões).

Tabela 6.16: 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

6) 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. Observe que o modelo proposto tem respaldo estatístico para valores do número de cadeiras variando entre 105 e 172; inferência para valores fora desse intervalo dependem da suposição de que o modelo também é válido nesse caso.
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?

7) 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.

8) 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 do 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.

9) Considere a seguinte reta de regressão ajustada a um conjunto de dados em que se pretendia estimar o volume de certos recipientes (V)a partir de seus diâmetros (D): $\mathrm{E}(V) = 7{,}68 + 0{,}185D$

Verifique se as afirmações abaixo estão corretas, justificando sua resposta:

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) Os dados abaixo são provenientes de uma pesquisa cujo objetivo é avaliar o efeito da dosagem de uma certa droga (X) na redução de pressão arterial (Y) de pacientes hipertensos.

Homens		**Mulheres**	
Dose	Redução de pressão	Dose	Redução de pressão
1	3	2	4
3	5	3	7
4	9	5	11
6	15	6	14
		6	13

O pesquisador sugeriu o seguinte modelo para a análise dos dados

$$y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}) + e_{ij}$$

$i = 1{,}2,\ j = 1, \ldots, n_i$ em que os erros e_{ij} seguem distribuições $N(0{,}\sigma^2)$ independentes e $\bar{x}$ denota a dose média empregada no estudo.

a) Interprete os parâmetros do modelo.
b) Escreva o modelo na forma matricial.

11) Para avaliar o efeito da dose (D) de uma certa droga na redução da pressão arterial ($RedPA$) controlando o sexo (S) do paciente, o seguinte modelo de regressão foi ajustado a um conjunto de dados:

$$\mathrm{E}(RedPA) = 2 + 0{,}3S + 1{,}2(D - 10)$$

em que $S = 0$ (Masculino) e $S = 1$ (Feminino). Verifique se as afirmações abaixo estão corretas, justificando sua resposta:

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

12) O gráfico QQ da Figura 6.36 corresponde ao ajuste de um modelo de regressão linear múltipla.

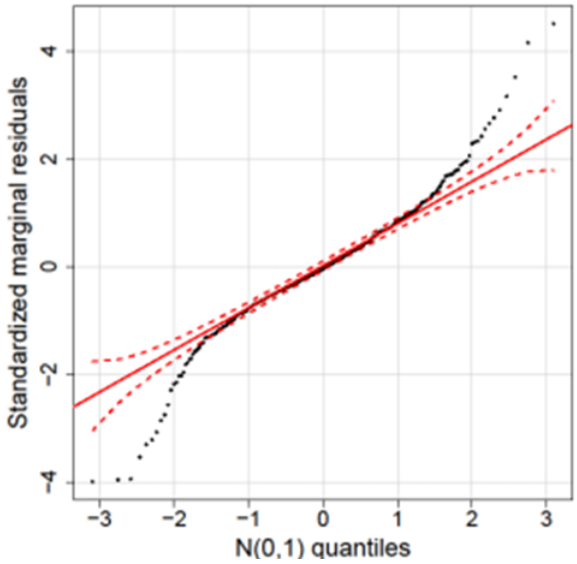

Figura 6.36: Gráfico QQ correspondente ajuste de um modelo de regressão linear múltipla.

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) Os dados da tabela abaixo foram obtidos de um estudo cujo objetivo era avaliar a relação entre a quantidade de um certo aditivo (X) e o tempo de vida (Y) de um determinado alimento. Os valores substituídos por ? ficaram ilegíveis depois que o responsável pelo estudo derramou café sobre eles.

X (g/kg)	5	10	15	20	30
Y (dias)	3.2	?	?	?	?

Um modelo de regressão linear simples (com a suposição de normalidade e independência dos erros) foi ajustado aos dados gerando os seguintes resultados:

Tabela de ANOVA

Fonte de variação	*gl*	*SQ*	*QM*	*F*	*Valor p*
Regressão	1	42,30	42,30	156,53	0,001
Resíduo	3	0,81	0,27		
Total	4	43,11			

Intervalos de confiança (95%)

Parâmetro	*Limite inferior*	*Limite superior*
Intercepto	-0,19	2,93
X	0,25	0,42

Resíduos

Observação	*Resíduos*
1	0,14
2	-0,55
3	0,66
4	-0,23
5	-0,01

$$(\mathbf{X}^t\mathbf{X})^{-1} = \begin{pmatrix} 0.892 & -0.043 \\ -0.043 & 0.003 \end{pmatrix}$$

a) Escreva o modelo na forma matricial e interprete seus parâmetros.
b) Construa um intervalo de confiança para o valor esperado do tempo de vida do produto quando a quantidade de aditivo utilizada é de 25 g/kg.
c) Construa um intervalo de previsão para o valor do tempo de vida do produto quando a quantidade de aditivo utilizada é de 25 g/kg.
d) Reconstrua a tabela dos dados, *i.e.*, calcule os valores de Y substituídos por ?.

Observação: O quantil de ordem 97,5% da distribuição t com 3 graus de liberdade é 3,182.

14) Considere o modelo ajustado para o Exemplo 6.3 e avalie a qualidade do ajuste utilizando todas as técnicas de diagnóstico discutidas neste capítulo.

15) 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, $0 < \gamma < 1$.

16) Considere o modelo especificado no Exercício 14 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 $\mathrm{E}(y_i|x_i + 1)$.

17) Mostre que $SQTot = SQRes + SQReg$.

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) A Tabela 6.17 contém 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 de regressão logística para avaliar a intensidade dessa relação.

Sugestão: Considere o ponto médio de cada intervalo como valor da variável explicativa.

Tabela 6.17: 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 do arquivo `endometriose2`. Com objetivo inferencial, ajuste um modelo de regressão logística, tendo `endometriose` como variável resposta e `idade`, `dormenstrual`, `dismenorreia` e `tipoesteril` como variáveis explicativas. Interprete os coeficientes do modelo em termos de chances e razões de chances.

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 ou mais eventos, 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 domínio, duas características são importantes: as definições 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 e Wang (2003), entre outros, para uma exposição mais detalhada.

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** (à direita). No entanto, sabe-se que o tempo é maior que um determinado valor chamado de **tempo de censura** (veja a 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.)

Por esse motivo, em estudos de sobrevivência, a variável resposta é definida pelo par (T, δ) em que T é o tempo associado à unidade amostral, que pode corresponder a uma falha se o indicador de censura δ tiver valor 1, ou a uma censura se o indicador δ tiver valor 0. Um exemplo de organização de dados dessa natureza está apresentado na Tabela 7.1.

[1] Apesar dessa terminologia, falha pode ter tanto uma conotação negativa, como a morte de um paciente, quanto positiva, como a sua cura.

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

Na Tabela 7.1, há indicação de que a unidade amostral A falhou no tempo 5,0 e de que a unidade amostral B não falhou até o tempo 12,0 e que foi censurada nesse instante.

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.

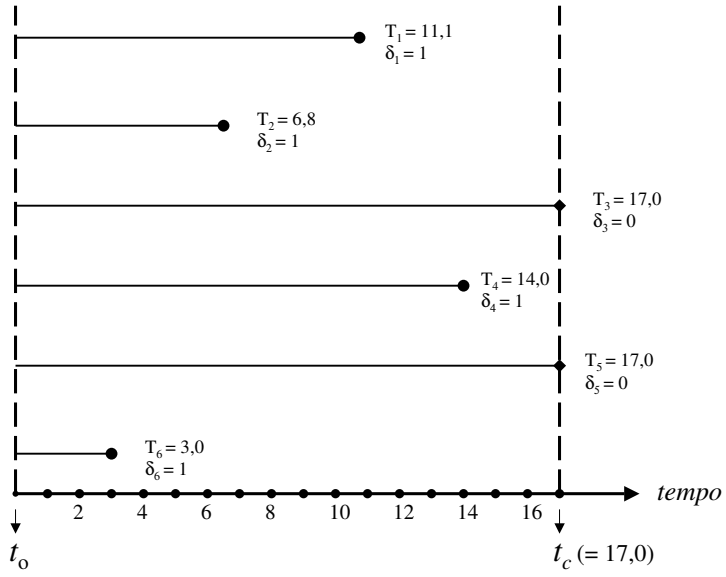

Figura 7.1: Representação esquemática de dados de sobrevivência.

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.

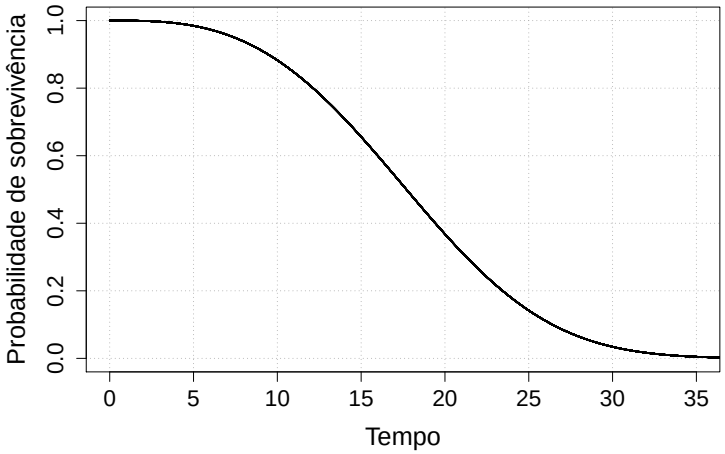

Figura 7.2: Função de sobrevivência teórica.

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. Os "saltos" ocorrem nos instantes em que há falhas. Em muitos casos, as censuras também são representadas nesse tipo de gráfico, como veremos adiante.

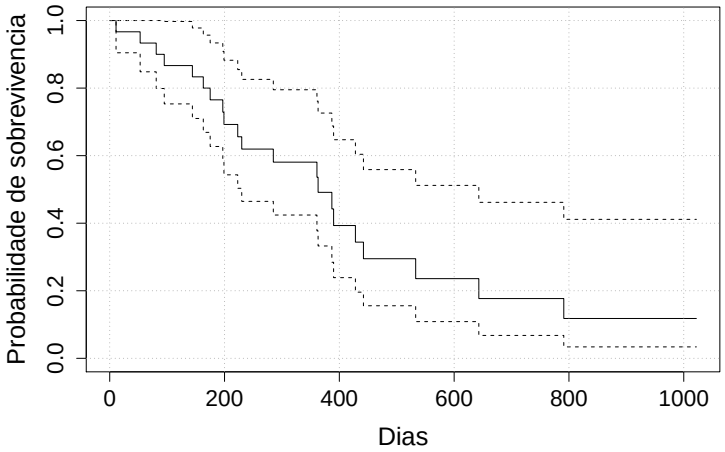

Figura 7.3: Função de sobrevivência observada.

Outra função de interesse na análise de dados de sobrevivência é a **função de risco** (*hazard function*) também conhecida como **função de taxa de falhas**, definida como

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \approx \frac{P(T = t)}{P(T > t)}.$$

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. 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 decresce com o tempo e usualmente é empregado para representar riscos pós cirúrgicos; um modelo probabilístico adequado nesse caso é um modelo **modelo Weibull**. No painel c), 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 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**.

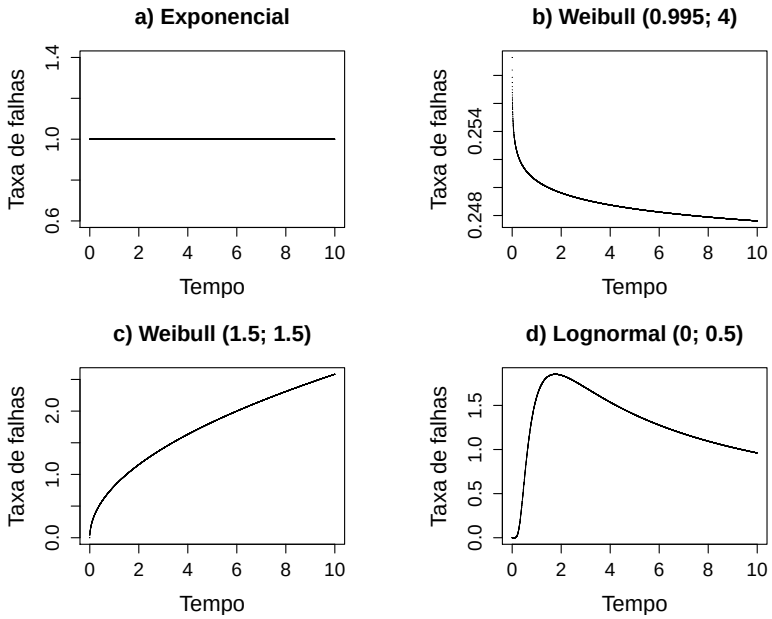

Figura 7.4: Exemplos de funções de risco

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

$$h(t) = -\frac{S'(t)}{S(t)} \quad \text{e} \quad S(t) = \exp[-\int_0^t h(s)ds]$$

em que $S'(t)$ indica a derivada de S calculada no instante t.

A **função de risco acumulado** (ou de **taxa de falhas acumuladas**) é definida como

$$H(t) = \int_0^t h(s)ds.$$

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

$$\widehat{F}(t) = \frac{\text{número de observações} \leq t}{\text{número de observações}}$$

e consequentemente, um estimador da função de sobrevivência é $\widehat{S}(t) = 1 - \widehat{F}(t)$. Para dados censurados, o **estimador de Kaplan-Meier** também conhecido como **estimador do limite de produtos** (*product limit estimator*) é o mais utilizado na prática e é baseado na representação da sobrevivência num instante t como um produto de probabilidades condicionais 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

$$\begin{aligned} S(10) &= P(T > 10) = P(T > 10 \cap T > 8) = P(T > 10|T > 8)P(T > 8) \\ &= P(T > 10|T > 8)P(T > 8|T > 5)P(T > 5) \\ &= P(T > 10|T > 8)P(T > 8|T > 5)P(T > 5|T > 2)P(T > 2). \end{aligned}$$

Considerando os instantes de falha ordenados, $t_{(0)} \leq t_{(1)} \leq \ldots \leq t_{(n)}$, definindo $t_{(0)} = t_0 = 0$ como o 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 ordenados $t_{(j)}$ de interesse são aqueles em que ocorreram falhas ou censuras. Definindo $R[t_{(i)}]$ como o número de unidades em risco no instante $t_{(i)}$, *i.e.*, para as quais o evento de interesse não ocorreu ou que não foram censuradas até o instante imediatamente anterior a $t_{(i)}$ 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 para a curva de sobrevivência é definido como

$$\widehat{S}(t) = 1 \text{ se } t < t_{(1)}$$

e

$$\widehat{S}(t) = \prod_{t_{(i)} < t} \{1 - M_i/R[t_{(i)}]\} \text{ se } t_{(i)} < t.$$

A variância desse estimador pode ser estimada pela **fórmula de Greenwood**

$$\widehat{\text{Var}}[\widehat{S}(t)] = [\widehat{S}(t)]^2 \sum_{t_{(i)}<t} \frac{M_i}{R[t_{(i)}]\{R[t_{(i)}] - M_i\}}.$$

Para detalhes, consulte Lee e Wang (2003), entre outros. O **tempo médio de acompanhamento limitado à duração do estudo (T)** é definido como

$$\mu_T = \int_0^T S(t)dt$$

e pode ser estimado pela área sob a curva baseada no estimador de Kaplan-Meier,

$$\widehat{\mu}_T = \sum_{t_{(k)}\leq T} \widehat{S}(t_{(k-1)})[t_{(k)} - t_{(k-1)}].$$

O tempo médio de acompanhamento, $\widehat{\mu}_T$, corresponde à soma das áreas dos retângulos cujas bases são os *plateaux* definidos pelas falhas consecutivas no gráfico de $\widehat{S}(t)$ (veja a Nota de Capítulo 2 e a Figura 7.3). Um estimador da variância de $\widehat{\mu}_T$ é

$$\widehat{\text{Var}}(\widehat{\mu}_T) = \sum_{j=1}^{D} \left\{ \sum_{i=j}^{D} [\widehat{S}(t_{(i)})[t_{(i+1)} - t_{(i)}]^2 \frac{M_j}{R_{(j)}[R_{(j)} - M_j]} \right\}$$

em que $R[t_{(j)}]$ representa o número de unidades em risco, M_j é o número de falhas ocorridas exatamente nesse instante e D representa o número de instantes distintos em que ocorreram eventos.

Além disso, um estimador do **tempo mediano de sobrevivência** é

$$\widehat{t}_{med} = \{\inf t : \widehat{S}(t) \leq 0.5\}$$

ou seja, é o menor valor de t para o qual o valor da função de sobrevivência $\widehat{S}(t)$ 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 $\widehat{t}_p = \{\inf t : \widehat{S}(t) \leq 1 - p\}$. Expressões para as correspondentes variâncias são bastante complicadas e não são indicadas neste texto. Em geral, os pacotes computacionais calculam intervalos de confiança para os quantis.

Exemplo 7.1: 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+. O formato usual para apresentação desses dados é aquele apresentado na Tabela 7.2

Tabela 7.2: Formato usual para apresentação dos dados do Exemplo 7.1

ident	tempo	evento	ident	tempo	evento
1	6	1	12	17	0
2	6	1	13	19	0
3	6	1	14	20	0
4	6	0	15	22	1
5	7	1	16	23	1
6	9	0	17	25	0
7	10	1	18	32	0
8	10	0	19	32	0
9	11	0	20	34	0
10	13	1	21	35	0
11	16	1			

Para efeito do cálculo do estimador de Kaplan-Meier, convém dispor os dados com o formato das quatro primeiras colunas da Tabela 7.3.

Tabela 7.3: Formato apropriado para cálculo do estimador de Kaplan-Meier (Exemplo 7.1)

j	Tempo $t_{(j)}$	Falhas em $t_{(j)}$	Censuras em $(t_{(j-1)}, t_{(j)})$	Unidades em risco $(R[t_{(j)}])$	$\widehat{S}[t_{(j)}]$
0	0	0	0	21	1
1	6	3	0	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	1	15 = 17-(1+1)	$0{,}81 \times 14/15 = 0{,}75$
4	13	1	2	12 = 15-(1+2)	$0{,}75 \times 11/12 = 0{,}69$
5	16	1	0	11 = 12-(1+0)	$0{,}69 \times 10/11 = 0{,}63$
6	22	1	3	7 = 11-(1+3)	$0{,}63 \times 6/7 = 0{,}54$
7	23	1	0	6 = 7-(1+0)	$0{,}54 \times 5/6 = 0{,}45$

Para os dados da Tabela 7.3, o 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.

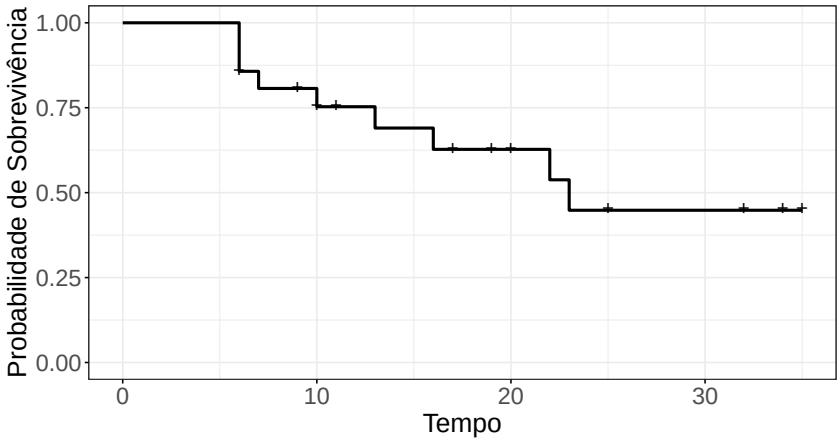

Figura 7.5: Curva de sobrevivência estimada para o Exemplo 7.1.

Esse gráfico, bem com medidas resumo para os dados podem ser obtido por meio das funções `Surv()`, `survfit()` e `ggsurvplot()` dos pacotes `survival` e `survminer` conforme os comandos.

```
> surv_object <- Surv(time = dados$tempo, event = dados$evento)
> mod1 <- survfit(surv_object ~ 1, data = dados)
> print(mod1, rmean = 35)
         n     events     *rmean *se(rmean)     median
     21.00       9.00      23.29       2.83      23.00
    * restricted mean with upper limit =  35
summary(mod1)
time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

> km <- ggsurvplot(mod1, data = dados, xlab = "Tempo",
      ylab = "Probabilidade de Sobrevivência",
      legend = 'none' , conf.int = F, palette = 'black',
      ggtheme = theme_bw() + theme(aspect.ratio = 0.5),
  font.x = c(18), font.y = c(18), font.tickslab = c(16))
```

Os resultados indicam que o tempo médio de sobrevivência (limitado a tempo = 35) é 23,3 (com erro padrão 2,8) e o tempo mediano de sobrevivência é 23,0.

Exemplo 7.2: 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 até 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. A censura ocorreu para os animais que não encontraram a plataforma em até 2 minutos. Os dados estão disponíveis no arquivo `piscina`.

Estimativas para as curvas de sobrevivência referentes ao Exemplo 7.2 estão dispostas na Figura 7.6 e estatísticas daí decorrentes, na Tabela 7.4. Os comandos utilizados para a obtenção desses resultados são

```
> surv_object <- Surv(time = ratos$Tempo, event = ratos$Delta)
> mod1 <- survfit(surv_object ~ ratos$Grupo, data = ratos)
> summary(mod1)
> print(survfit(surv_object ~ ratos$Grupo, data = ratos), print.rmean=TRUE)
> quant <- quantile(mod1, probs = c(0.25, 0.5, 0.75))
> km <- ggsurvplot(mod1, data = ratos, xlab = "Segundos",
           ylab = "Probabilidade de Sobrevivência",
   conf.int = F, palette = 'colors', legend = c(0.7, 0.7),
   legend.title = "Grupo",
       legend.labs = c("Giro denteado dorsal", "Giro denteado ventral",
       "Controle"), break.x.by = 10,
   ggtheme = theme_bw() + theme(aspect.ratio = 0.6),
   font.x = c(20), font.y = c(20), font.tickslab = c(18),
   font.legend = c(18))
```

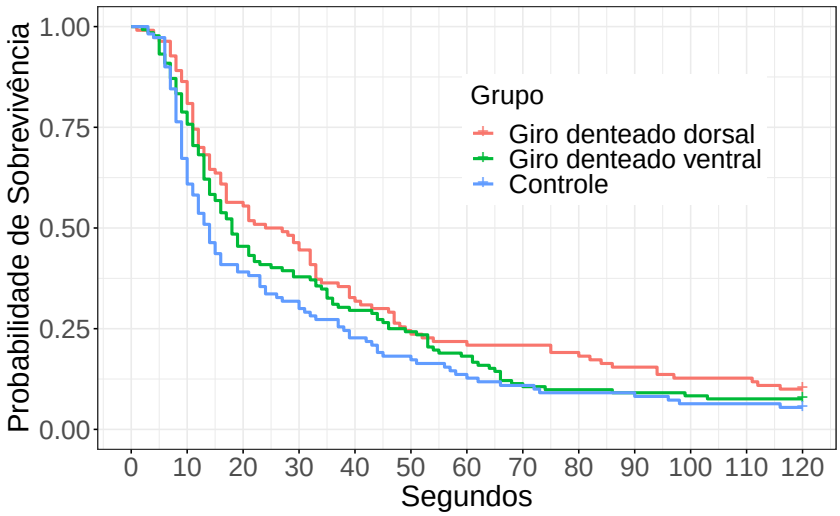

Figura 7.6: Curvas de sobrevivências estimadas para os dados do Exemplo 7.2.

Tabela 7.4: Medidas resumo com erros padrões ou intervalos de confiança aproximados (95%) entre parênteses (Exemplo 7.2)

Tratamento	Censuras	Tempo Médio	Primeiro Quartil	Tempo Mediano	Terceiro Quartil
Grupo 1	10,0%	39,5 (3,5)	11	26 (17-33)	49
Grupo 2	7,6%	33,3 (2.8)	11	18 (15-25)	48
Grupo 3	5,5%	28,6 (3,0)	9	14 (11-19)	38

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 7.5.

Tabela 7.5: 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 (óbito) e `delta` = 0, uma censura. A primeira dificuldade é ler as datas no formato indicado utilizando alguma função do `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 exibidas no formato **dd/mm/aa** no número de dias desde 01 de janeiro de 1970, com datas anteriores assumindo valores negativos, deixando os dados no formato indicado na Tabela 7.1. Consequentemente, o intervalo de tempo entre as datas de início do estudo e aquela de ocorrência do evento ou de censura pode ser calculado por diferença. 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.

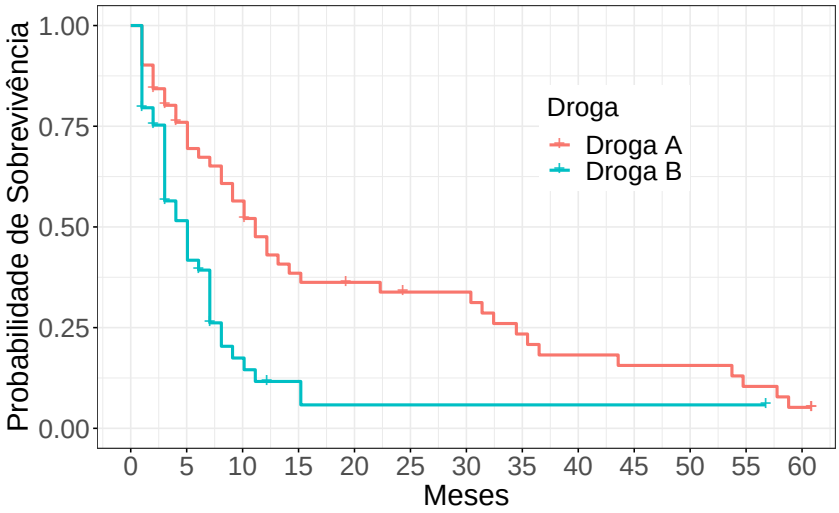

Figura 7.7: Curvas de sobrevivência estimadas para o Exemplo 7.3.

Tabela 7.6: Estatísticas descritivas com erros padrões ou intervalos de confiança entre parênteses para os dados do Exemplo 7.3

Tratamento	Censuras	Tempo Médio	Primeiro Quartil	Tempo Mediano	Terceiro Quartil
Droga A	17,6%	20,3 (2,9)	5,1	11,1 (8,1 - 30,4)	34,5
Droga B	22,4%	8,4 (2,6)	3,0	5,1 (3,0 - 7,1)	8,1

7.3 Comparação de curvas de sobrevivência

Um dos problemas oriundos de estudos como aquele descrito nos Exemplos 7.2 e 7.3 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 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 ordenados 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.

A comparação das curvas de sobrevivência correspondentes aos efeitos das duas drogas consideradas no Exemplo 7.3 pode ser concretizada por meio do comando

```
> survdiff(Surv(hiv$tempomeses,hiv$delta) ~ hiv$droga)
```

cujo resultado é

```
              N Observed Expected (O-E)^2/E (O-E)^2/V
hiv$droga=0 51       42     54.9      3.02      11.9
hiv$droga=1 49       38     25.1      6.60      11.9
 Chisq= 11.9  on 1 degrees of freedom, p= 6e-04
```

sugerindo uma diferença significativa ($p < 0{,}001$) entre as curvas (populacionais) associadas.

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. Para o Exemplo 7.2, a estatística de teste obtida com essa generalização tem valor 6,4 que comparado com uma distribuição qui-quadrado com 2 graus de liberdade sugere que as curvas de sobrevivência (populacionais) associadas aos três grupos são diferentes ($p = 0{,}041$).

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. 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 x, 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 7.7.

Tabela 7.7: Relação entre algumas distribuições para T e $\log(T)$

Distribuição de	
T	$\log(T)$
exponencial	Valores extremos
Weibull	Valores extremos
log logística	logística
log normal	normal

Esses modelos podem ser ajustados por meio do método da máxima verossimilhança e a função `survreg()` pode ser utilizada com esse propósito. Detalhes podem ser obtidos nas referências citadas no início do capítulo.

Os comandos e o resultado do ajuste de um modelo de tempo de falha acelerado Weibull aos dados do Exemplo 7.2 são

```
> survreg(formula = Surv(ratos$Tempo, ratos$Delta) ~ factor(ratos$Grupo),
                  data = ratos, dist = "weibull", scale = 1)
                    Value Std. Error     z      p
(Intercept)           3.783      0.101 37.64 <2e-16
```

```
factor(ratos$Grupo)2 -0.199      0.135 -1.47 0.1407
factor(ratos$Grupo)3 -0.373      0.140 -2.66 0.0078
Scale fixed at 1
Weibull distribution
Loglik(model)= -1491.2   Loglik(intercept only)= -1494.7
Chisq= 7.1 on 2 degrees of freedom, p= 0.029
Number of Newton-Raphson Iterations: 4
```

Esses resultados sugerem que animais do Grupo 2 têm o tempo de falha retardado por um fator $\exp(-0.199) = 0{,}82$ (IC95%: 0,79 - 0,85) relativamente aos animais do Grupo 1; para os animais do Grupo 3, esse fator de desaceleração é $= \exp(-0.373) =$ 0,69 (IC95%: 0,66 - 0,72) relativamente aos animais do Grupo 1.

Uma alternativa são os **modelos semipraramétricos** em que se destaca o **modelo de riscos proporcionais** (*proportional hazards model*) também conhecido como **modelo de regressão de Cox** e expresso como

$$h(t|\mathbf{X} = \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x}),$$

em que $h_0(t)$ representa uma **função de risco basal**, arbitrária, mas não negativa, *i.e.*, é o risco para unidades com $\mathbf{X} = \mathbf{0}$ e $\exp(\boldsymbol{\beta}^\top \mathbf{x})$ é a **função de risco relativo** com parâmetro $\boldsymbol{\beta}$ 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 (veja a Nota de Capítulo 3). 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.

O ajuste de modelos de riscos proporcionais pode ser realizado por meio das funções `coxph()` e `cox.zph()`. Para os dados do Exemplo 7.3, esses comandos e resultados correspondentes são

```
> coxmod1 <- coxph(formula = Surv(hiv$tempomeses, hiv$delta) ~ hiv$droga)
             coef exp(coef) se(coef)     z       p
hiv$droga 0.8309    2.2953   0.2418 3.436 0.00059
Likelihood ratio test=11.6  on 1 df, p=0.0006593
n= 100, number of events= 80
> summary(coxmod1)
```

```
          exp(coef) exp(-coef) lower .95 upper .95
hiv$droga     2.295     0.4357     1.429     3.687
cox.zph(coxmod1)
          chisq df    p
hiv$droga 0.555  1 0.46
GLOBAL    0.555  1 0.46
```

e sugerem que a hipótese de riscos proporcionais é aceitável ($p = 0{,}46$). Além disso, o fator droga é signifcativamente ($p < 0{,}001$) importante para o padrão de sobrevivência dos pacientes, indicando que o risco de óbito para pacientes tratados com a Droga B é 2,30 (IC95%: 1,43 - 3,68) vezes o risco de óbito para pacientes tratados com a Droga A.

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) **Tempo médio de sobrevivência**

Para cálculo do tempo médio de sobrevivência seria necessário acompanhar todas as unidades investigadas até que todas falhassem. Como em geral esse não é o caso, calcula-se o tempo médio de sobrevivência restrito à duração do estudo em questão, digamos τ. Então

$$\mu_\tau = \int_0^\tau t f(t) dt$$

em que f é a função densidade da variável de interesse, T (tempo). Fazendo $t = u$ e $f(t)dt = dv$ de forma que $v = F(t)$, a função distribuição de T, e integrando essa expressão por partes, temos

$$\mu_\tau = tF(t)\Big|_0^\tau - \int_0^\tau F(t)dt = \tau - \int_0^\tau F(t)dt = \int_0^\tau [1 - F(t)]dt = \int_0^\tau S(t)dt.$$

3) **Riscos não proporcionais**

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) Determinação dos instantes de tempo em que ocorrem mudanças no padrão da sobrevivência.

b) Ajuste de modelos diferentes para intervalos de tempo distintos.

c) Refinamento do modelo com a inclusão de variáveis explicativas dependentes do tempo.

d) Introdução de estratos.

7.6 Exercícios

1) Suponha que 6 ratos foram expostos a um material cancerígeno. O tempo até o desenvolvimento de um tumor com certas características foi registrado para cada um dos animais. Os ratos A, B e C desenvolvem o tumor em 10, 15 e 25 semanas, respectivamente. O rato D morreu sem tumor na vigésima semana de observação. O estudo terminou após 30 semanas, sem que os ratos E e F apresentassem o tumor.

   a) Defina cuidadosamente a resposta do estudo.

   b) Identifique o tipo de resposta (falha ou censura) para cada um dos ratos do estudo.

   c) Construa uma planilha de dados adequada para a análise por meio de alguma função do pacote `R`.

2) Os dados do arquivo `freios` são tempos de vida (em quilômetros) de pastilhas de freios de uma amostra de 40 carros do mesmo modelo selecionada aleatoriamente. Cada veículo foi monitorado para avaliar a influência dos seguintes fatores na duração das pastilhas: Ano do modelo (1 ou 2), Região de uso do carro (N: Norte ou S: Sul, Condições de dirigibilidade (A: predominantemente na cidade, B: predominantemente em estrada ou C: Uso misto).

   a) Construa as curvas de sobrevivência de Kaplan-Meier considerando cada um dos fatores separadamente.

   b) Teste a hipótese de que os níveis de cada fator (individualmente) não alteram a durabilidade das pastilhas (use o teste log-rank ou outro teste mais conveniente).

   c) Considere apenas os fatores significativamente associados ao tempo de falha, construa as curvas de Kaplan-Meier correspondentes e calcule os tempos médios e medianos de sobrevivência das pastilhas de freios.

3) 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 7.8.

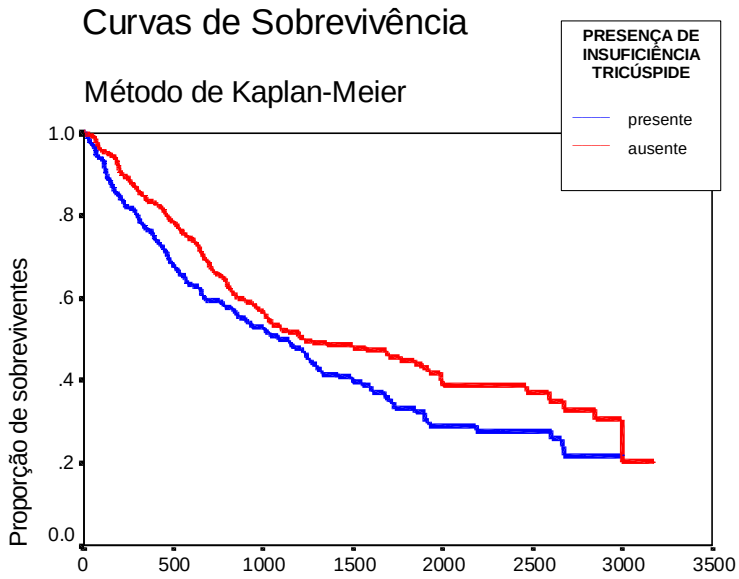

Figura 7.8: Curva de sobrevivência estimada para o estudo de transplante cardíaco

Tabela 7.8: 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 7.8.

d) Qual a razão para se incluir um intervalo de confiança na análise?

4) Os dados da Tabela 7.9 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 7.9: 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 7.10.

Tabela 7.10: 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.

5) 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) 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.

6) O arquivo `rehabcardio` contém dados de um estudo cujo objetivo era avaliar a sobrevivência de pacientes infartados submetidos ou a tratamento clínico ou a um programa de exercícios. Com essa finalidade, considere que

   i) o evento de interesse é definido como revascularização miocárdica, ocorrência de um segundo infarto ou óbito pós admissão no estudo;

   ii) os pacientes foram acompanhados até 31/12/2000;

   iii) datas incompatíveis com a data de nascimento devem ser descartadas.

   Construa curvas de Kaplan-Meier para comparar o padrão de sobrevivência dos pacientes submetidos a cada grupo e com base nelas, estime o primeiro quartil e o tempo médio 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.

Capítulo 8

Regularização e Modelos Aditivos Generalizados

> It is better to solve the right problem approximately than to solve the wrong problem exactly.
>
> John Tukey

8.1 Introdução

O objetivo das técnicas abordadas neste capítulo é selecionar modelos de regressão que, segundo algum critério, sejam eficientes para prever valores de variáveis respostas contínuas ou discretas. Neste contexto, a estratégia do aprendizado estatístico consiste em ajustar vários modelos a um conjunto de dados de treinamento e escolher aquele que gere as melhores previsões com dados de um conjunto de dados de validação. Em geral, esse processo é concretizado por meio de **validação cruzada** (veja a Nota de Capítulo 1).

Como alternativas aos modelos de regressão linear múltipla empregados com objetivo de previsão, abordaremos **modelos de regularização** e **modelos aditivos generalizados**.

Se a associação entre as variáveis preditoras e a variável resposta for linear, os modelos de regressão ajustados por mínimos quadrados gerarão previsores não enviesados mas com variância grande a não ser que o número de dados (n) seja bem maior que o número de variáveis preditoras (p). Nos casos em que n não é muito maior que p, os modelos de regularização podem gerar previsões com menor variância sem alterar o viés consideravelmente. Quando há variáveis preditoras que não são associadas à variável resposta, os modelos de regressão ajustados por mínimos quadrados poderão ser mais complexos do que o desejado, pois os coeficientes associados a essas variáveis não serão anulados. Modelos de regularização *Lasso* por outro lado, têm a característica de anular os coeficientes das variáveis preditoras cuja associação com a variável resposta seja irrelevante. Quando o objetivo é inferencial, a eliminação de variáveis preditoras deve ser avaliada com cautela, pois mesmo que a associação entre elas e a variável resposta não seja evidenciada pelos dados, é importante mantê-las no modelo por razões subjacentes ao problema investigado. Esse é o caso da variável "Idade" em problemas da área médica, pois em muitos casos, sabe-se que independentemente de seu coeficiente não ser significativo, ela pode estar associada com a resposta.

Os modelos aditivos generalizados, por sua vez, permitem avaliar associações não lineares com forma não especificada entre as variáveis preditoras e a variável resposta e podem servir como sugestão para a inclusão de termos não lineares em modelos de regressão ajustados por mínimos quadrados.

Os critérios mais usados para a escolha do melhor modelo são o erro quadrático médio (MSE), sua raiz quadrada ($RMSE$), o coeficiente de determinação (R^2) cujas definições estão nas Seções 1.6 e 6.2 para variáveis respostas contínuas ou a taxa de erros no caso de variáveis respostas discretas.

8.2 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 8.1 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 valores observados 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** (*overfitting*).

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. Em termos gerais, essas técnicas servem para ajustar modelos de regressão baseados em uma função de perda que contém um termo de penalização. Esse termo tem a finalidade de reduzir a influência de coeficientes responsáveis por flutuações excessivas.

Embora haja várias técnicas de regularização, consideraremos apenas três: 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 regularização ***Elastic Net***.

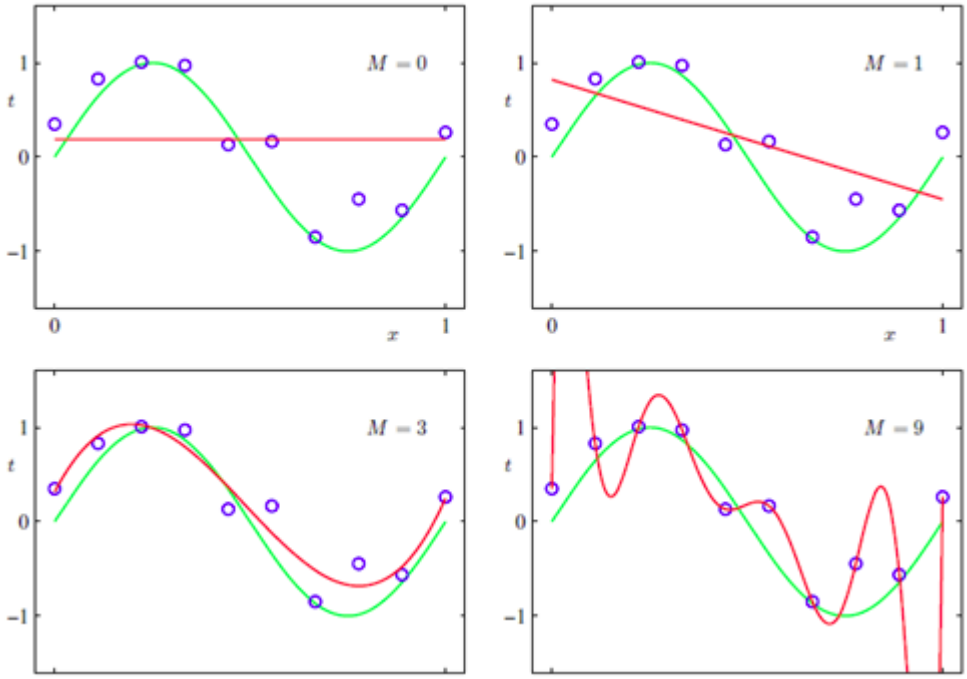

Figura 8.1: Ajuste de modelos polinomiais a um conjunto de dados hipotéticos (M indica o grau do polinômio ajustado).

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 que têm contribuição limitada para efeito de previsão. Isso pode implicar uma **solução esparsa**.[1] Na regularização *Ridge*, por outro lado, o termo de regularização usa uma soma de quadrados dos parâmetros e um termo 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 seguimos as ideias apresentadas em Hastie et al. (2017), James et al. (2017) e Medeiros (2019).

8.2.1 Regularização L_2 (*Ridge*)

A técnica de regressão *Ridge* foi introduzida por Hoerl e Kennard (1970) para tratar do problema da multicolinearidade mas também pode ser utilizada para corrigir problemas ligados ao sobreajuste. Consideremos o modelo de regressão

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_p x_{pt} + e_t, \quad t = 1,2,\ldots,n, \tag{8.1}$$

[1] 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, e_t indicando erros de média zero e $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_p)^\top$ denotando os parâmetros a serem estimados. Os **estimadores de mínimos quadrados penalizados** correspondem à solução de

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \tag{8.2}$$

em que $\lambda \geq 0$ é o **coeficiente de regularização**, que controla a importância relativa entre a minimização da soma de quadrados dos erros [o primeiro termo do segundo membro de (8.2)] e o termo de penalização, $\lambda \sum_{j=1}^{p} \beta_j^2$. Se $\lambda = \infty$, não há variáveis a serem incluídas no modelo e se $\lambda = 0$, obtemos os estimadores de mínimos quadrados. A escolha de λ deve ser um dos componentes da estratégia para a determinação de estimadores regularizados. Convém lembrar que o intercepto não é considerado no termo de penalização, dado que o interesse está na associação entre as variáveis preditoras e a variável resposta.

Algumas propriedades dessa classe de estimadores são:

i) O estimador *Ridge* não é consistente; no entanto, pode-se mostrar que é assintoticamente consistente sob condições sobre λ, p e n.

ii) O estimador *Ridge* é enviesado para os parâmetros não nulos.

iii) A técnica de regularização *Ridge* não serve para a seleção de modelos.

iv) 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 3.

Obter o mínimo em (8.2) é equivalente a minimizar a soma de quadrados não regularizada $[\sum_{t=1}^{n}(y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2]$ sujeita à restrição

$$\sum_{j=1}^{p} \beta_j^2 \leq m, \tag{8.3}$$

para algum valor apropriado m, ou seja, é um problema de otimização com **multiplicadores de Lagrange**. O **estimador *Ridge*** pode ser expresso como

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{8.4}$$

em que $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$ é a matriz de especificação do modelo e $\mathbf{y} = (y_1, \ldots, y_n)^\top$ é o vetor de respostas.

8.2.2 Regularização L_1 (*Lasso*)

Consideremos, agora, o **estimador *Lasso***, obtido de

$$\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{8.5}$$

Neste caso, a restrição (8.3) é substituída por

$$\sum_{j=1}^{p} |\beta_j| \leq m. \tag{8.6}$$

Algumas propriedades estatísticas do estimador *Lasso* são:

i) Parâmetros que correspondem a preditores redundantes são encolhidos para zero.

ii) O estimador *Lasso* é enviesado para parâmetros não nulos.

iii) Sob certas condições, o estimador *Lasso* descarta as variáveis irrelevantes do modelo atribuindo pesos nulos aos respectivos coeficientes.

iv) 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,

$$\widehat{\beta}_j(\lambda) = \text{sinal}(Z_j)\,(|Z_j| - \lambda/2)_+\,, \tag{8.7}$$

em que $(x)_+ = \max\{x,0\}$.

Para outras propriedades, veja Medeiros (2019) e Bühlmann e van de Geer (2011).

A característica de geração de soluções esparsas pode ser esquematizada por meio da Figura 8.2.

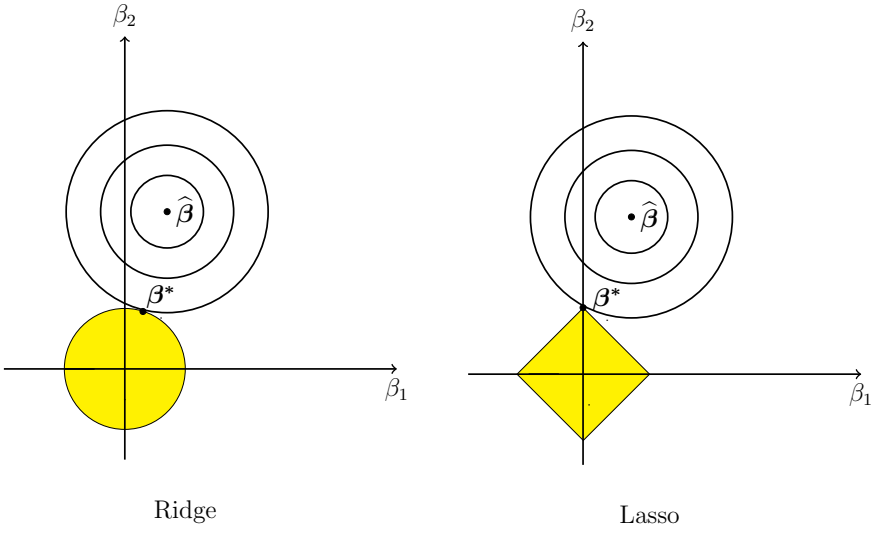

Figura 8.2: Esparsidade dos modelos *Ridge* e *Lasso*.

No painel esquerdo, representamos o estimador de mínimos quadrados $\widehat{\boldsymbol{\beta}}$ e o estimador $\widehat{\boldsymbol{\beta}}_{Ridge}$ (representado por $\boldsymbol{\beta}^*$). Os círculos concêntricos representam curvas cujos pontos correspondem a somas de quadrados de resíduos $[\sum_{i=1}^n (y_i - \widehat{y}_i)^2]$ constantes. Quanto mais afastados esses círculos estiverem de $\widehat{\boldsymbol{\beta}}$, maior será o valor da soma de quadrados de resíduos correspondente. A região colorida representa a restrição $\sum_{j=1}^p \beta_j^2 \leq m$ em que m depende do coeficiente de penalização λ. O estimador $\widehat{\boldsymbol{\beta}}_{Ridge}$ corresponde ao ponto em que um dos círculos tangencia a região de restrição. Note que ambos os componentes de $\boldsymbol{\beta}^*$ são diferentes de zero. No painel direito, a região colorida corresponde à restrição $\sum_{j=1}^p |\beta_j| \leq m$ e nesse caso, a regularização *Lasso* pode gerar uma solução esparsa, ou seja, com algum elemento do vetor $\widehat{\boldsymbol{\beta}}_{Lasso}$ (representado por $\boldsymbol{\beta}^*$) igual a zero. Na figura, o componente β_1 de $\boldsymbol{\beta}^*$ é nulo.

8.2.3 Outras propostas

O estimador *Elastic Net* (*EN*) é definido por

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda_1,\lambda_2) = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda_1 \sum_{i=1}^{p} \beta_i^2 + \lambda_2 \sum_{i=1}^{p} |\beta_i| \right], \tag{8.8}$$

em que $\lambda_1 \geq 0$ e $\lambda_2 \geq 0$ são coeficientes de regularização. Para este estimador, a restrição de penalização (8.3) é substituída por

$$\lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \leq m. \tag{8.9}$$

Pode-se mostrar que sob determinadas condições, o estimador *Elastic Net* é consistente.

Na Figura 8.3 apresentamos esquematicamente uma região delimitada pela restrição

$$J(\boldsymbol{\beta}) = (1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \leq m$$

para algum m, com $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, além daquelas delimitadas pelas restrições *Ridge* e *Lasso*.

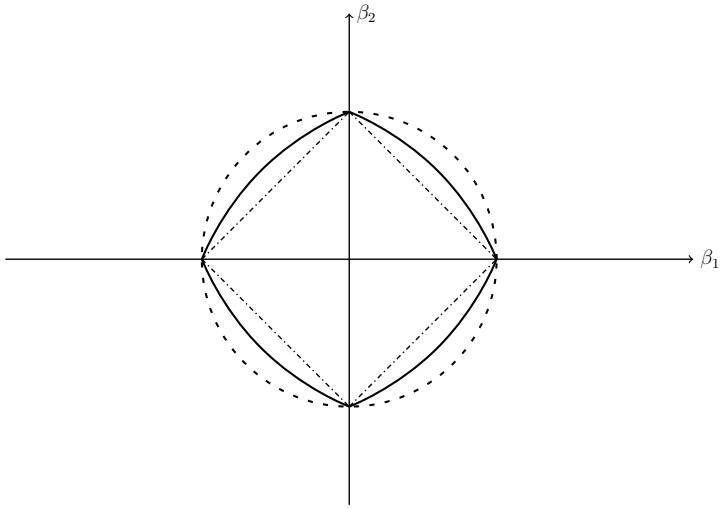

Figura 8.3: Geometria das restrições *Elastic Net* (curva contínua), *Ridge* (curva tracejada) e *Lasso* (curva pontilhada)

O estimador ***Lasso* adaptativo** (*AL*) é definido por

$$\widehat{\boldsymbol{\beta}}_{AL}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{i=1}^{p} w_i |\widetilde{\beta}_i| \right], \tag{8.10}$$

em que $w_1, \ldots, w_p$ são pesos não negativos pré definidos e $\widetilde{\beta}_i$, $i = 1, \ldots, p$ são estimadores iniciais (por exemplo, estimadores *Lasso*). Usualmente, toma-se $w_j = |\widetilde{\beta}_j|^{-\tau}$, para $0 < \tau \leq 1$.

O estimador ***Lasso* adaptativo** é consistente sob condições não muito fortes.

A função `adalasso()` do pacote `parcor` pode ser usada para calcular esse estimador. O pacote `glmnet` 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 8.1: Consideremos os dados do arquivo `antracose2`, extraídos de um estudo cuja finalidade era avaliar o efeito da idade (`idade`), tempo vivendo em São Paulo (`tmunic`), horas diárias em trânsito (`htransp`), carga tabágica (`cargatabag`), classificação sócio-econômica (`ses`), densidade de tráfego na região onde habitou (`densid`) e distância mínima entre a residência a vias com alta intensidade de tráfego (`distmin`) num índice de antracose (`antracose`), que é uma medida de fuligem (*black carbon*) depositada no pulmão. Detalhes sobre esse estudo podem ser obtidos em Takano et al. (2019).

Como o índice de antracose índice varia entre 0 e 1, consideramos

$$logrc = \log[\text{índice de antracose}/(1 - \text{índice de antracose})]$$

como variável resposta.

O conjunto de dados para a análise pode ser preparado por meio dos comandos

```
> pulmao0 <- read.xls("/home/julio/Desktop/antracose2.xls",
sheet='dados', method="tab")
> pulmao1 <- na.omit(pulmao0)
> pulmao <- pulmao1[which(pulmao1$antracose != 0),]
> pulmao$logrc <- log(pulmao$antracose/(1-pulmao$antracose))
```

Os estimadores de mínimos quadrados dos coeficientes de um modelo linear podem ser obtidos por meio dos comandos

```
> pulmao_lm <- lm(logrc ~ idade + tmunic + htransp + cargatabag +
                          ses + densid + distmin, data=pulmao)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.977e+00  2.459e-01 -16.169  < 2e-16 ***
idade        2.554e-02  2.979e-03   8.574  < 2e-16 ***
tmunic       2.436e-04  2.191e-03   0.111 0.911485
htransp      7.505e-02  1.634e-02   4.592 5.35e-06 ***
cargatabag   6.464e-03  1.055e-03   6.128 1.61e-09 ***
ses         -4.120e-01  1.238e-01  -3.329 0.000926 ***
densid       7.570e+00  6.349e+00   1.192 0.233582
distmin      3.014e-05  2.396e-04   0.126 0.899950
Residual standard error: 1.014 on 598 degrees of freedom
Multiple R-squared:  0.1965,Adjusted R-squared:  0.1871
F-statistic: 20.89 on 7 and 598 DF,  p-value: < 2.2e-16
```

Utilizando o pacote `glmnet`, ajustamos um modelo de regressão *Ridge* por meio de validação cruzada com os comandos

```
> y <- pulmao$logrc
> X <- pulmao[,-c(5,9)]
> X <- data.matrix(X)
> regridgecv = cv.glmnet(X, y, alpha = 0)
> plot(regridgecv)
```

O parâmetro $\alpha = 0$ indica regressão *Ridge* e o comando `plot` gera o gráfico da Figura 8.4. Nesse gráfico, a curva em vermelho representa a variação do erro quadrático médio (MSE) em função do logaritmo do coeficiente de regularização λ, juntamente com a indicação dos desvios padrões correspondentes. Uma das linhas verticais pontilhadas indica o valor de $\log(\lambda)$ correspondente ao menor erro quadrático médio obtido no ajuste e a outra correspondente ao valor de $\log(\lambda)$ associado ao modelo com a maior regularização em que o erro quadrático médio obtido por validação cruzada não seja maior do que o correspondente erro mínimo mais um desvio padrão.

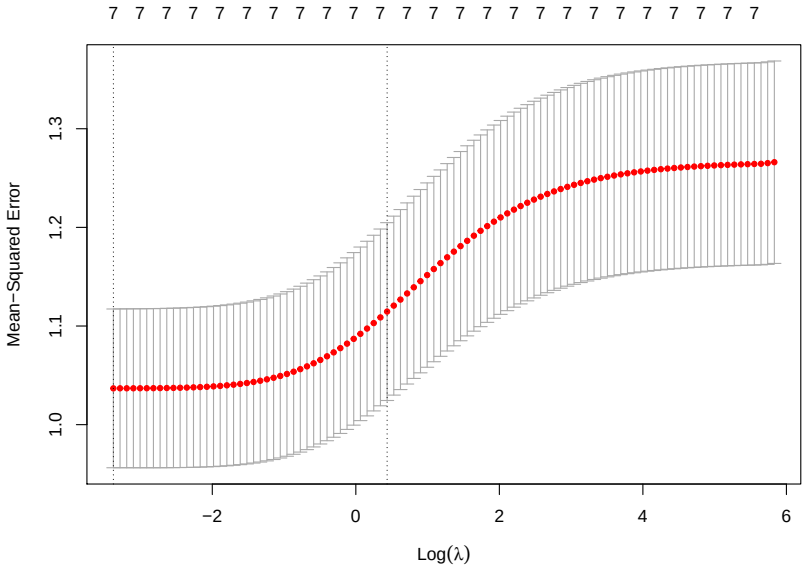

Figura 8.4: Gráfico para avaliação do efeito do coeficiente de regularização *Ridge* aos dados do Exemplo 8.1.

Os coeficientes do ajuste correspondente ao valor mínimo do coeficiente λ juntamente com esse valor são obtidos com os comandos

```
>
coef(regridgecv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
(Intercept) -3.905299e+00
idade        2.456715e-02
tmunic       4.905597e-04
htransp      7.251095e-02
cargatabag   6.265919e-03
ses         -3.953787e-01
densid       7.368120e+00
distmin      3.401372e-05
> regridgecv$lambda.min
[1] 0.03410028
```

Note que a abscissa da linha pontilhada situada à esquerda na Figura 8.4 corresponde a $\log(0{,}03410028) = -3{,}37845$. Segundo o modelo, todas as 7 variáveis são mantidas e algumas (`idade`, `htransp`, `cargatabag`, `ses` e `densid`) têm coeficientes "encolhidos" em direção a zero quando comparadas com aquelas obtidas por

mínimos quadrados. Os valores preditos para os elementos do conjunto de dados e a correspondente raiz quadrada do MSE ($RMSE$) são obtidos por meio dos comandos

```
> predict(regridgecv, X, s = "lambda.min")
> sqrt(regridgecv$cvm[regridgecv$lambda == regridgecv$lambda.min])
[1] 1.050218
```

Para o modelo de regressão *Ridge*, a $RMSE = 1{,}050218$ é ligeiramente maior do que aquela obtido por meio do modelo de regressão linear múltipla, $RMSE = 1{,}014$.

O ajuste do modelo de regressão *Lasso* ($\alpha = 1$) juntamente com o gráfico para a escolha do coeficiente λ, disposto na Figura 8.5, são obtidos com

```
> reglassocv = cv.glmnet(X, y, alpha = 1)
> plot(reglassocv)
```

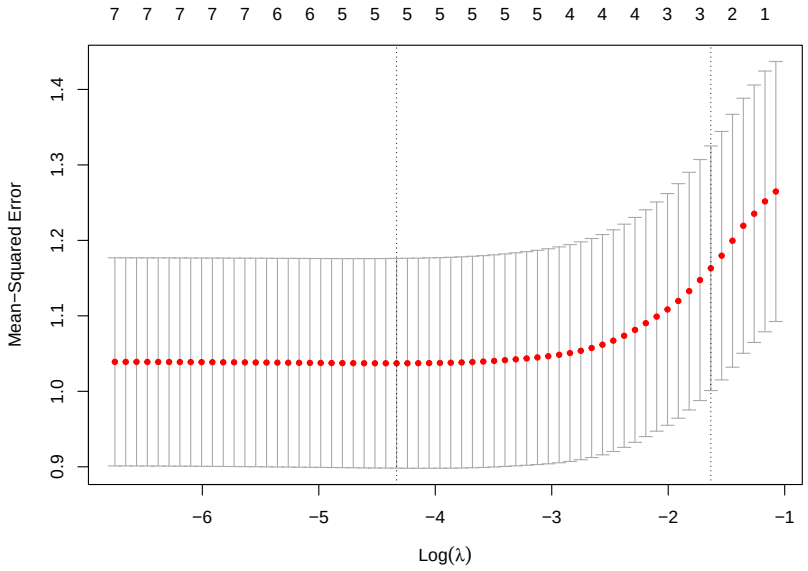

Figura 8.5: Gráfico para avaliação do efeito do coeficiente de regularização *Lasso* aos dados do Exemplo 8.1.

As duas linhas pontilhadas que indicam os limites para valores $\log(\lambda)$ sugerem que o modelo associado ao menor erro quadrático médio é obtido com 5 das 7 variáveis disponíveis. Os coeficientes correspondentes à regularização *Lasso*, o valor mínimo do coeficiente λ e o $RMSE$ são gerados por intermédio dos comandos

```
> coef(reglassocv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
(Intercept) -3.820975473
idade        0.024549358
tmunic       .
htransp      0.069750435
cargatabag   0.006177662
ses         -0.365713282
densid       5.166969594
distmin      .
```

```
> reglassocv$lambda.min
[1] 0.01314064
> sqrt(reglassocv$cvm[reglassocv$lambda == reglassocv$lambda.min])
[1] 1.018408
```

Neste caso, todos os coeficientes foram encolhidos em direção ao zero e aqueles correspondentes às variáveis `tmunic` e `distmin` foram anulados.

Para o modelo *Elastic Net* com $\alpha = 0{,}5$, os resultados são

```
> regelncv = cv.glmnet(X, y, alpha = 0.5)
> regelncv$lambda.min
[1] 0.02884367
> coef(regelncv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
(Intercept) -3.776354935
idade        0.024089256
tmunic       .
htransp      0.068289153
cargatabag   0.006070319
ses         -0.354080190
densid       4.889074555
distmin      .           .
> sqrt(regelncv$cvm[regelncv$lambda == regelncv$lambda.min])
[1] 1.0183
```

Quando avaliados com relação aos valores preditos para os elementos do conjunto de dados, os 3 procedimentos de regularização produzem erros quadráticos médios similares, com pequena vantagem para aquele obtido por meio de *Elastic Net*.

Exemplo 8.2: Consideremos os dados do arquivo `coronarias`, selecionando os 1032 elementos sem observações omissas para as seguintes variáveis: `LO3` (resposta) e `IDADE1`, `IMC`, `HA`, `PDR`, `PSR`, `COLS`, `TRIGS` e `GLICS` (preditoras). Um modelo de regressão logística pode ser ajustado com os seguintes comandos

```
> glm(formula = LO3 ~ IDADE1 + IMC + HA + PDR + PSR + COLS + TRIGS +
    GLICS, family = "binomial", data = coronarias)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0265  -1.2950   0.7234   0.8983   1.5563
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.1712877  0.8172749   0.210  0.83399
IDADE1       0.0362321  0.0070450   5.143  2.7e-07 ***
IMC         -0.0075693  0.0175103  -0.432  0.66554
HA           0.4767869  0.1545457   3.085  0.00203 **
PDR         -0.0016821  0.0808568  -0.021  0.98340
PSR         -0.0737184  0.0436001  -1.691  0.09088 .
COLS        -0.0019213  0.0010348  -1.857  0.06336 .
TRIGS        0.0009716  0.0008296   1.171  0.24153
GLICS       -0.0038084  0.0017947  -2.122  0.03384 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1295.0  on 1031  degrees of freedom
Residual deviance: 1242.6  on 1023  degrees of freedom
AIC: 1260.6
Number of Fisher Scoring iterations: 4
```

Para avaliar o desempenho do modelo relativamente à previsão para os elementos do conjunto de dados, podemos construir uma tabela de confusão (em que a classe predita é aquela com maior valor da probabilidade predita pelo modelo) por meio dos comandos

```
> yhat <- ifelse(modreglog$fitted.values > 0.5, 1, 0)
> table(coronarias$LO3, yhat)
   yhat
      0   1
  0  32 299
  1  30 671
```

A taxa de classificações erradas correspondente é de $31{,}9\% = [(299 + 30)/1032]$.

Utilizaremos técnicas de regularização para ajuste de um modelo com as mesmas variáveis, dividindo o conjunto de dados num conjunto de treinamento e num conjunto de validação. Com essa finalidade, consideramos os comandos

```
> y <- coronarias$LO3
> X <- coronarias[, c("IDADE1", "IMC", "HA", "PDR", "PSR", "COLS",
"TRIGS", "GLICS")]
> X <- data.matrix(X)
> cor.rows <- sample(1:1032, .67*1034)
> cor.train <- X[cor.rows, ]
> cor.valid <- X[-cor.rows, ]
> y.train <- y[cor.rows]
> y.valid <- y[-cor.rows]
```

Os comandos para o ajuste do modelo de regressão *Ridge* aos dados de treinamento com escolha do parâmetro λ por meio de validação cruzada e construção do gráfico correspondente juntamente com os resultados são

```
> regridgecv = cv.glmnet(cor.train, y.train, alpha = 0, family="binomial",
type.measure = "class")
> plot(regridgecv)
> coef(regridgecv, s = "lambda.min")
9 x 1 sparse Matrix of class "dgCMatrix"
(Intercept)  0.4293387897
IDADE1       0.0323811126
IMC         -0.0022071010
HA           0.4674088493
PDR         -0.0294993531
PSR         -0.0727593423
COLS        -0.0021478285
TRIGS        0.0007545016
GLICS       -0.0026381393
> regridgecv$lambda.min
[1] 0.01557885
```

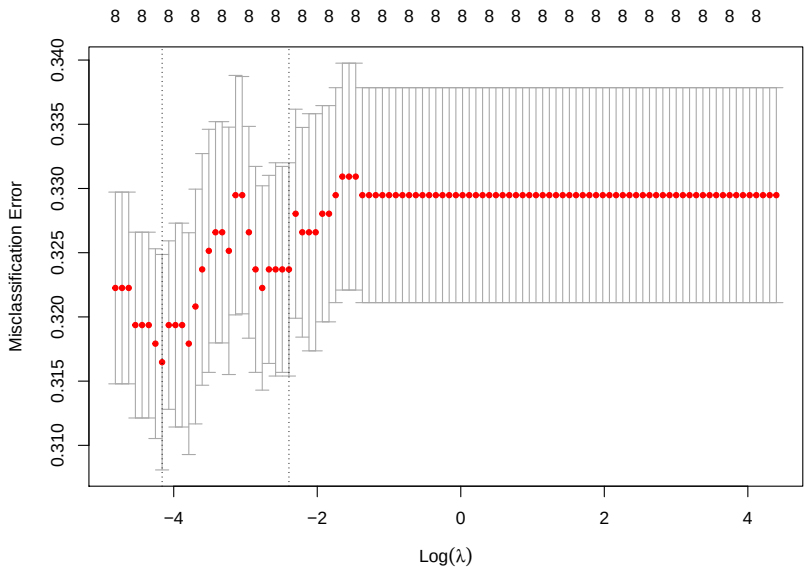

Figura 8.6: Gráfico para avaliação do efeito do coeficiente de regularização *Ridge* aos dados do Exemplo 8.2.

O comando para obtenção de medidas de desempenho do modelo *Ridge* para os dados de validação obtidos com o valor mínimo do parâmetro λ é

```
> assess.glmnet(regridgecv, newx = cor.valid, newy = y.valid,
type.measure="class")
```

Dentre as medidas de desempenho produzidas está a taxa de erros de classificação, que no caso é 0,3117647. Comandos similares podem ser considerados para produzir taxas de erros de classificação para os dados de treinamento (0,316474) e para os dados do conjunto completo (0,3149225).

Uma tabela de confusão para comparação entre as classes observadas e preditas pelo modelo *Ridge* no conjunto completo é obtida com o comando

```
> confusion.glmnet(regridgecv, newx = X, newy = y, s = "lambda.min",
type.measure="class")
        True
Predicted   0   1 Total
    0      35  29    64
    1     296 672   968
    Total 331 701  1032
 Percent Correct:  0.6851
```

Os comandos `glmnet` para uma análise similar envolvendo o ajuste de um modelo de regressão *Lasso*, juntamente com os resultados produzidos são

```
> reglassocv = cv.glmnet(cor.train, y.train, alpha = 1, family="binomial",
type.measure="class")
> coef(reglassocv, s = "lambda.min")
9 x 1 sparse Matrix of class "dgCMatrix"
(Intercept) -0.3550130628
IDADE1       0.0263456404
IMC          .
```

```
HA            0.2655187474
PDR           .
PSR          -0.0254148108
COLS         -0.0008159609
TRIGS         .
GLICS        -0.0011471304
> reglassocv$lambda.min
[1] 0.01670469
```

Aqui vale notar a natureza esparsa da solução, em que os coeficientes das variáveis `IMC`, `PDR` e `TRIGS` foram zerados. As taxas de erros de classificação obtidas por meio desse modelo para os dados dos conjuntos de validação, de treinamento e completo são, respectivamente, 0,3058824, 0,316474 e 0,3129845.

Os resultados do ajuste com regularização *Elastic Net* são

```
> regelncv = cv.glmnet(cor.train, y.train, alpha = 0.5, family="binomial",
type.measure="class")
> coef(regelncv, s = "lambda.min")
9 x 1 sparse Matrix of class "dgCMatrix"
(Intercept)  0.1006473735
IDADE1       0.0310848999
IMC          .
HA           0.4100168033
PDR         -0.0107214867
PSR         -0.0631703734
COLS        -0.0016021300
TRIGS        0.0003340047
GLICS       -0.0021747061
> regelncv$lambda.min
[1] 0.01200671
```

Neste caso, há um compromisso entre a regularização *Ridge*, em que nenhum dos coeficientes é zerado e a regularização *Lasso*, em que três coeficientes são zerados. As taxas de erros de classificação obtidas por meio do modelo de regularização *Elastic Net* para os dados dos conjuntos de validação, de treinamento e completo são, respectivamente, 0,3088235, 0,316474 e 0,3139535.

8.3 Modelos aditivos generalizados (*GAM*)

Modelos lineares têm um papel muito importante na análise de dados, tanto pela facilidade de ajuste quanto de interpretação. De uma forma geral, os modelos lineares podem ser expressos como

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) + e_i \tag{8.11}$$

$i = 1, \ldots, n$ em que as funções f_i são conhecidas. No modelo de regressão polinomial de segundo grau, por exemplo, $f_1(x_{i1}) = x_{i1}$ e $f_2(x_{i1}) = x_{i1}^2$. Em casos mais gerais, poderíamos ter $f_1(x_{i1}) = x_{i1}$ e $f_2(x_{i1}) = \exp(x_{i1})$. Em muitos problemas reais, no entanto, nem sempre é fácil especificar a forma das funções f_i e uma alternativa proposta por Hastie e Tibshirani (1996) envolve os chamados **Modelos Aditivos Generalizados** (*Generalized Additive Modelos - GAM*) que são expressos como (8.11) sem a especificação da forma das funções f_i.

Quando a distribuição da variável resposta y_i pertence à **família exponencial**, o modelo pode ser considerado como uma extensão dos **Modelos Lineares Generalizados** (*Generalized Linear Models - GLM*) e é expresso como

$$g(\mu_i) = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) \tag{8.12}$$

em que g é uma **função de ligação** e $\mu_i = \mathrm{E}(y_i)$ (veja a Nota de Capítulo 4).

Existem diversas propostas para a representação das funções f_i que incluem o uso de *splines* naturais, *splines* suavizados e regressões locais. A suavidade dessas funções é controlada por parâmetros de suavização, que devem ser determinados *a priori*. Curvas muito suaves podem ser muito restritivas, enquanto curvas muito rugosas podem causar sobreajuste.

O procedimento de ajuste dos modelos aditivos generalizados depende da forma escolhida para as funções f_i. A utilização de *splines* naturais, por exemplo, permite a aplicação direta do método de mínimos quadrados, graças à sua construção a partir de **funções base**. Para *splines* penalizados, o processo de estimação envolve algoritmos um pouco mais complexos, como aqueles conhecidos sob a denominação de **retroajustamento** (*backfitting*). Para detalhes sobre o ajuste dos modelos aditivos generalizados, consulte Hastie e Tibshirani (1990) e Hastie et al. (2008).

Para entender o conceito de *splines*, consideremos o seguinte modelo linear com apenas uma variável explicativa

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n. \tag{8.13}$$

A ideia subjacente aos modelos aditivos generalizados consiste na substituição do termo $\beta_1 x_i$ em (8.13) por um conjunto de transformações conhecidas, $b_1(x_i), \ldots, b_t(x_i)$, gerando o modelo

$$y_i = \alpha_0 + \alpha_1 b_1(x_i) + \ldots + \alpha_t b_t(x_i) + e_i, \quad i = 1, \ldots, n. \tag{8.14}$$

O modelo de regressão polinomial de grau t é um caso particular de (8.14) com $b_j(x_i) = x_i^j,\ j = 1, \ldots, t$.

Uma proposta para aumentar a flexibilidade da curva ajustada consiste em segmentar o domínio da variável preditora e ajustar diferentes polinômios de grau d aos dados de cada um dos intervalos gerados pela segmentação. Cada ponto de segmentação é chamado de **nó** e uma segmentação com k nós envolve $k + 1$ polinômios. Na Figura 8.7, apresentamos um exemplo com 5 polinômios de terceiro grau e 4 nós. Nesse exemplo, o modelo aditivo generalizado é expresso como

$$y_i = \begin{cases} \alpha_{01} + \alpha_{11} x_i + \alpha_{21} x_i^2 + \alpha_{31} x_i^3 + e_i, & \text{se } x_i \leq -0{,}5, \\ \alpha_{02} + \alpha_{12} x_i + \alpha_{22} x_i^2 + \alpha_{32} x_i^3 + e_i, & \text{se } -0{,}5 < x_i \leq 0, \\ \alpha_{02} + \alpha_{13} x_i + \alpha_{23} x_i^2 + \alpha_{33} x_i^3 + e_i, & \text{se } 0 < x_i \leq 0{,}5, \\ \alpha_{02} + \alpha_{14} x_i + \alpha_{24} x_i^2 + \alpha_{34} x_i^3 + e_i, & \text{se } 0{,}5 < x_i \leq 1, \\ \alpha_{05} + \alpha_{15} x_i + \alpha_{25} x_i^2 + \alpha_{35} x_i^3 + e_i, & \text{se } x_i > 1, \end{cases} \tag{8.15}$$

com as funções base, $b_1(x), b_2(x), ..., b_k(x)$, construídas com a ajuda de funções indicadoras. Esse modelo é conhecido como **modelo polinomial cúbico segmentado**.

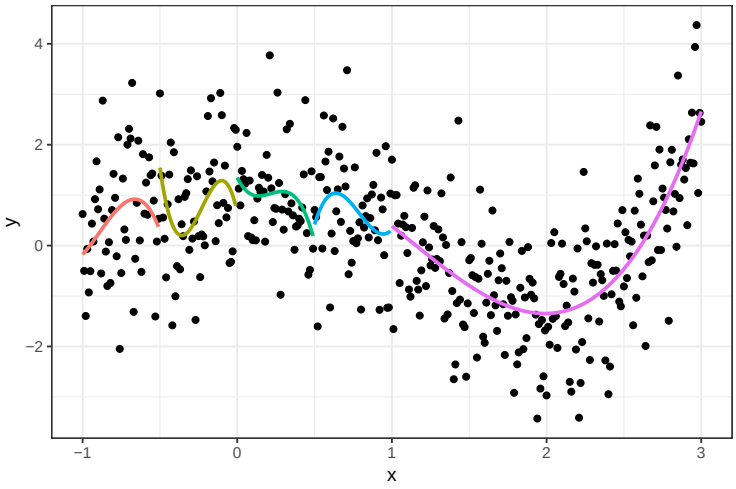

Figura 8.7: Polinômios de terceiro grau ajustados aos dados de cada região segmentada da variável X. Os nós são os pontos $x = -0{,}5$, $x = 0$, $x = 0{,}5$ e $x = 1$.

A curva formada pela junção de cada um dos polinômios na Figura 8.7 não é contínua, apresentando "saltos" nos nós. Essa característica não é desejável, pois essas descontinuidades não são interpretáveis. Para contornar esse problema, podemos definir um *spline* de grau d como um polinômio segmentado de grau d com as $d-1$ primeiras derivadas contínuas em cada nó. Essa restrição garante a continuidade e suavidade (ausência de vértices) da curva obtida.

Utilizando a representação por bases (8.14), um *spline* cúbico com k nós pode ser expresso como

$$y_i = \alpha_0 + \alpha_1 b_1(x_i) + \alpha_2 b_2(x_i) + \ldots + \alpha_{k+3} b_{k+3}(x_i) + e_i, \quad i = 1, ..., n, \qquad (8.16)$$

com as funções $b_1(x), b_2(x), \ldots, b_{k+3}(x)$ escolhidas apropriadamente. Usualmente, essas funções envolvem três termos polinomiais, a saber, x, x^2 e x^3 e k termos $h(x, c_1), \ldots, h(x, c_k)$ da forma

$$h(x, c_j) = (x - c_j)^3_+ = \begin{cases} (x - c_j)^3, & \text{se } x < c_j, \\ 0, & \text{em caso contrário,} \end{cases}$$

com $c_1, \ldots, c_k$ indicando os nós. Com a inclusão do termo α_0, o ajuste de um *spline* cúbico com k nós envolve a estimação de $k+4$ parâmetros e, portanto, utiliza $k+4$ graus de liberdade. Mais detalhes sobre a construção desses modelos podem ser encontrados em Hastie (2008) e James et al. (2017).

Além das restrições sobre as derivadas, podemos adicionar **restrições de fronteira**, exigindo que a função seja linear nas regiões em que $x < c_1$ e $x > c_k$. Essas restrições diminuem a variância dos valores extremos gerados pela variável preditora, produzindo estimativas mais estáveis. Um *spline* cúbico com restrições de fronteira é chamado de *spline* natural.

No ajuste de *splines* cúbicos ou naturais, o número de nós determina o grau de suavidade da curva e a sua escolha pode ser feita por validação cruzada conforme indicado em James et al. (2017). De uma forma geral, a maior parte dos nós é posicionada nas regiões em que há mais informação sobre a variável preditora, isto é, com mais observações. Por pragmatismo, para modelos com mais de uma variável preditora, costuma-se adotar o mesmo número de nós para todas.

Os *splines* suavizados constituem uma classe de funções suavizadoras que não utilizam a abordagem por funções bases. De maneira resumida, um *spline* suavizado é uma função f que minimiza

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int f''(u)^2 du \tag{8.17}$$

em que f'' corresponde à segunda derivada da função f e indica sua curvatura; quanto maior for a curvatura maior a penalização. O primeiro termo dessa expressão garante que f se ajustará bem aos dados, enquanto o segundo penaliza a sua variabilidade, isto é, controla a suavidade de f, que é regulada pelo parâmetro $\lambda \geq 0$. A função f se torna mais suave conforme λ aumenta. A escolha desse parâmetro é geralmente feita por validação cruzada.

Outro método bastante utilizado no ajuste funções não lineares para a relação entre a variável preditora X e a variável resposta Y é conhecido como **regressão local**. Esse método consiste em ajustar modelos de regressão simples em regiões em torno de cada observação x_0 da variável preditora X. Essas regiões são formadas pelos k pontos mais próximos de x_0, sendo que o parâmetro $s = k/n$ determina o quão suave ou rugosa será a curva ajustada. O ajuste é feito por meio de mínimos quadrados ponderados, com pesos inversamente proporcionais à distância entre cada ponto da região centrada em x_0 e x_0. Aos pontos dessas regiões mais afastados de x_0 são atribuídos pesos menores. Um exemplo é o *lowess*, discutido na Nota de Capítulo 2 do Capítulo 5.

Para uma excelente exposição sobre *splines* e penalização o leitor pode consultar Eilers e Marx (1996, 2021).

Modelos aditivos generalizados podem ser ajustados utilizando-se a função `gam()` do pacote `mgcv`. Essa função permite a utilização de *splines* como funções suavizadoras. Para regressão local, é necessário usar a função `gam()` do pacote `gam`. Também é possível utilizar o pacote `caret`, a partir da função `train()` e `method = "gam"`.

Exemplo 8.3: Consideremos os dados do arquivo `esforco` com o objetivo de prever os valores da variável `vo2fcpico` (VO2/FC no pico do exercício) a partir das variáveis `NYHA`, `idade`, `altura`, `peso`, `fcrep` (frequência cardíaca em repouso) e `vo2rep` (VO2 em repouso). Um modelo inicial de regressão linear múltipla também pode ser ajustado por meio dos seguintes comandos da função `gam()`

```
> mod0 <- gam(vo2fcpico ~ NYHA + idade + altura + peso + fcrep
                              + vo2rep, data=esforco)
```

Como não especificamos nem a distribuição da resposta, nem a função de ligação, a função `gam()` utiliza a distribuição normal com função de ligação identidade, conforme indica o resultado

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + idade + altura + peso + fcrep + vo2rep
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.80229    4.43061  -1.084 0.280642
NYHA1       -0.45757    0.50032  -0.915 0.362303
NYHA2       -1.78625    0.52629  -3.394 0.000941 ***
```

```
NYHA3       -2.64609    0.56128  -4.714 6.75e-06 ***
NYHA4       -2.43352    0.70532  -3.450 0.000780 ***
idade       -0.05670    0.01515  -3.742 0.000284 ***
altura       0.09794    0.02654   3.690 0.000342 ***
peso         0.08614    0.01739   4.953 2.48e-06 ***
fcrep       -0.07096    0.01318  -5.382 3.84e-07 ***
vo2rep       0.35564    0.24606   1.445 0.151033
---
R-sq.(adj) =  0.607   Deviance explained = 63.5%
GCV =  4.075  Scale est. = 3.7542    n = 127
```

Para avaliar a qualidade do ajuste, produzimos gráficos de dispersão entre os resíduos do ajuste do modelo e cada uma das variáveis preditoras. Esses gráficos estão dispostos na Figura 8.8 e sugerem relações possivelmente não lineares, pelo menos em alguns casos.

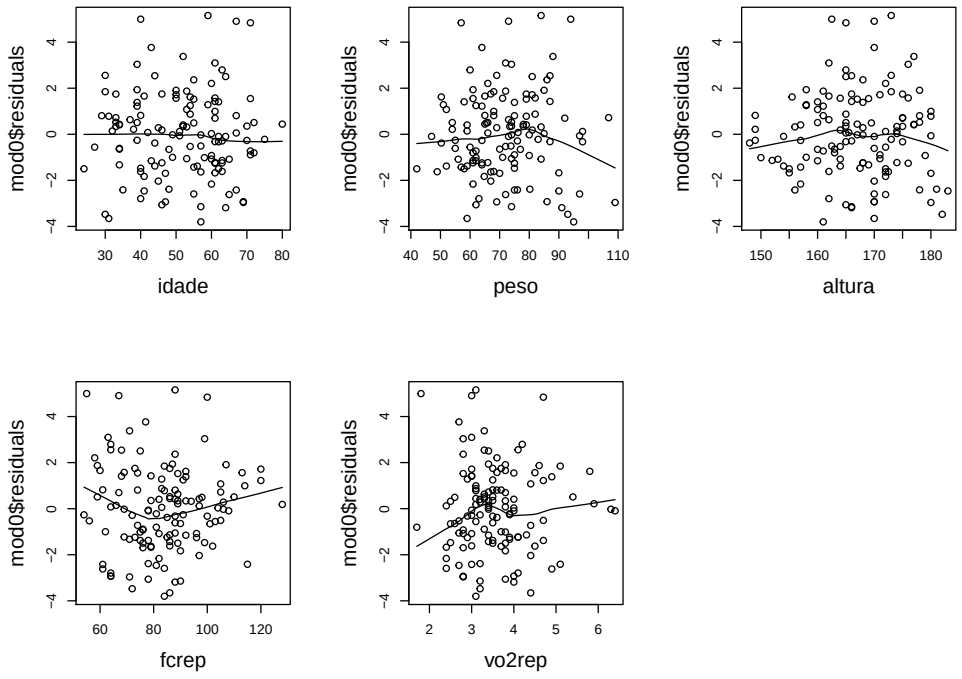

Figura 8.8: Gráficos de dispersão (com curva *lowess*) entre `vo2fcpico` e cada variável preditora contínua considerada no Exemplo 8.3.

Uma alternativa é considerar modelos *GAM* do tipo (8.11) em que as funções f_i são expressas em termos de *splines*. Em particular, um modelo *GAM* com *splines* cúbicos para todas as variáveis preditoras contínuas pode ser ajustado por meio do comando

```
> mod1 <- gam(vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                          s(fcrep) + s(vo2rep), data=esforco)
```

gerando os seguintes resultados:

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) + s(fcrep) +
    s(vo2rep)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2101     0.3207  31.841  < 2e-16 ***
NYHA1        -0.5498     0.4987  -1.103 0.272614
NYHA2        -1.8513     0.5181  -3.573 0.000522 ***
NYHA3        -2.8420     0.5664  -5.018 1.99e-06 ***
NYHA4        -2.5616     0.7031  -3.643 0.000410 ***
---
Approximate significance of smooth terms:
            edf Ref.df      F  p-value
s(idade)  1.000  1.000 15.860  0.00012 ***
s(altura) 5.362  6.476  3.751  0.00142 **
s(peso)   1.000  1.000 22.364 6.32e-06 ***
s(fcrep)  1.742  2.185 16.236 3.95e-07 ***
s(vo2rep) 1.344  1.615  0.906  0.47319
---
R-sq.(adj) =   0.64   Deviance explained = 68.2%
GCV = 3.9107  Scale est. = 3.435     n = 127
```

O painel superior contém estimativas dos componentes paramétricos do modelo e o painel inferior, os resultados referentes aos termos suavizados. Neste caso, apenas a variável categorizada `NYHA` não foi suavizada, dada sua natureza categorizada.

A coluna rotulada `edf` contém os graus de liberdade efetivos associados a cada variável preditora. Para cada variável preditora contínua não suavizada, perde-se um grau de liberdade; para as variáveis suavizadas a atribuição de graus de liberdade é mais complexa em virtude do número de funções base e do número de nós utilizados no processo de suavização. Variáveis com `edf`=1 têm efeito linear e poderiam ser incluídas sem suavização no modelo. A coluna rotulada `Ref.df` corresponde a graus de liberdade aproximados utilizados para o cálculo da estatística `F`. A suavização é irrelevante apenas para a variável `vo2rep` e dado que ela também não apresentou contribuição significativa no modelo de regressão linear múltipla, pode-se considerar um novo modelo *GAM* obtido com a sua eliminação.

Os gráficos dispostos na Figura 8.9, produzidos por meio do comando `plot(mod1, se=TRUE)` evidenciam esse fato; além disso mostram a natureza "mais não linear" da variável `altura` (com `edf`= 5.362).

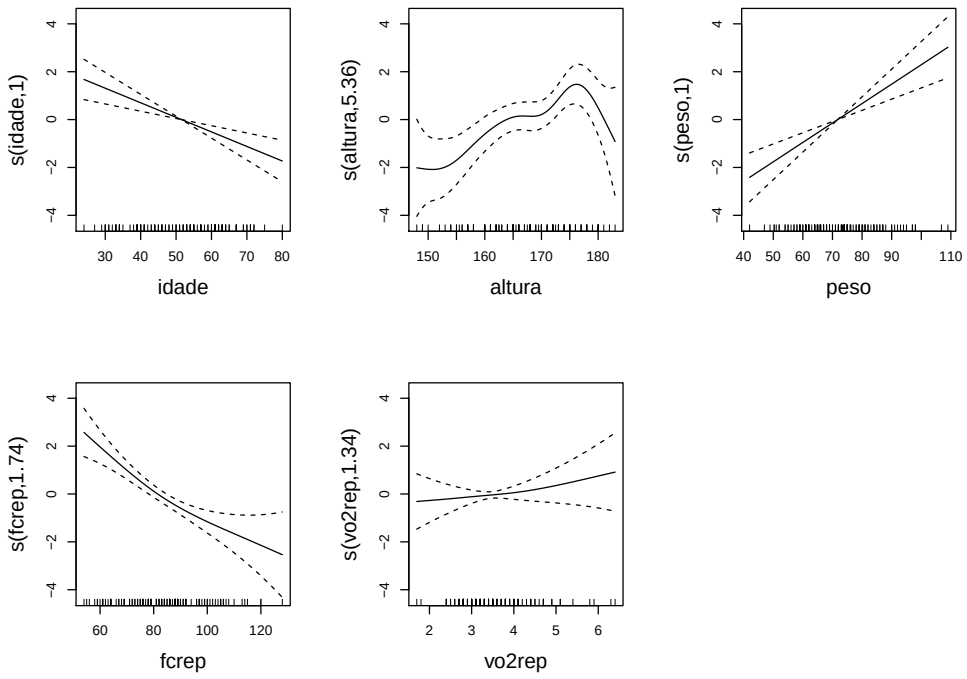

Figura 8.9: Funções suavizadas (com bandas de confiança) obtidas por meio do modelo *GAM* (mod1) para os dados do Exemplo 8.3.

Um modelo que incorpora essas conclusões pode ser ajustado por meio do comando

```
mod2 <- gam(vo2fcpico ~ NYHA + idade + s(altura) + peso + s(fcrep),
            data=esforco)
summary(mod2)
```

O resultado correspondente, apresentado a seguir, sugere que todas as variáveis preditoras contribuem significativamente para explicar sua relação com a variável resposta.

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + idade + s(altura) + peso + s(fcrep)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.90755    1.31397   6.018 2.24e-08 ***
NYHA1       -0.58181    0.49853  -1.167 0.245650
NYHA2       -1.83849    0.51605  -3.563 0.000539 ***
NYHA3       -2.96692    0.55123  -5.382 4.04e-07 ***
NYHA4       -2.48232    0.69802  -3.556 0.000551 ***
idade       -0.06152    0.01523  -4.040 9.79e-05 ***
peso         0.07656    0.01623   4.718 6.88e-06 ***
Approximate significance of smooth terms:
            edf Ref.df      F  p-value
```

```
s(altura) 5.311  6.426  3.857  0.00119 **
s(fcrep)  1.856  2.337 14.865 1.28e-06 ***
R-sq.(adj) =   0.64   Deviance explained = 67.8%
GCV = 3.8663  Scale est. = 3.435     n = 127
```

A qualidade do ajuste avaliada pelo coeficiente de determinação ajustado $R^2_{aj} = 0{,}64$ é ligeiramente superior ao valor obtido do ajuste do modelo de regressão linear múltipla, $R^2_{aj} = 0{,}61$ Uma avaliação adicional pode ser realizada por meio de uma análise de resíduos e de comparação dos valores observados e preditos. Para essa finalidade, o comando `gam.check(mod2)` gera os gráficos apresentados na Figura 8.10 que não evidenciam problemas no ajuste.

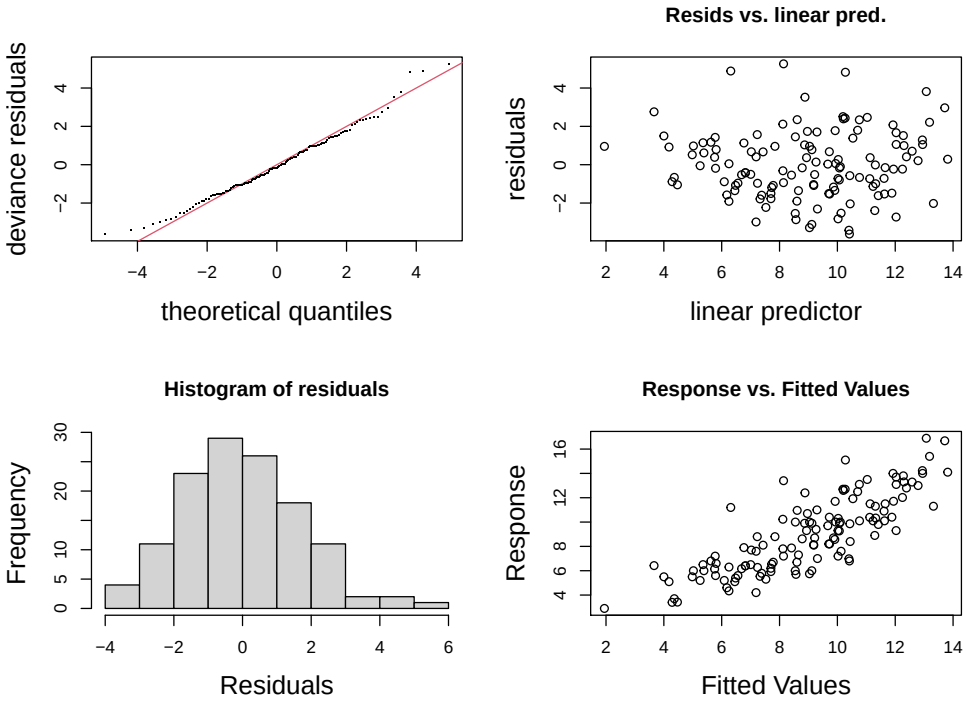

Figura 8.10: Gráficos diagnósticos para o ajuste do modelo *GAM* aos dados do Exemplo 8.3.

Além disso, é possível comparar os modelos por meio de uma **análise de desviância**, que pode ser obtida com o comando `anova(mod0, mod1, mod2, test= "F")`.

```
Analysis of Deviance Table
Model 1: vo2fcpico ~ NYHA + idade + altura + peso + fcrep + vo2rep
Model 2: vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) + s(fcrep) +
    s(vo2rep)
Model 3: vo2fcpico ~ NYHA + idade + s(altura) + peso + s(fcrep)
  Resid. Df Resid. Dev      Df Deviance      F  Pr(>F)
1    117.00     439.24
2    109.72     383.18  7.2766   56.052 2.2425 0.03404 *
3    111.24     387.58 -1.5129   -4.399 0.8465 0.40336
```

Esses resultados mostram que ambos os modelos *GAM* são essencialmente equivalentes ($p = 0.403$) mas significativamente diferentes ($p = 0.034$) do modelo de regressão linear múltipla.

A previsão para um novo conjunto dados em que apenas os valores das variáveis preditoras estão disponíveis pode ser obtida por meio do comando `predict(mod2, newdata=esforcoprev, se=TRUE, type="response")`. Consideremos, por exemplo, o seguinte conjunto com dados de 5 novos pacientes

```
idade  altura   peso   NYHA   fcrep   vo2rep
 66     159      50     2      86      3,4
 70     171      77     4     108      4,8
 64     167      56     2      91      2,5
 42     150      67     2      70      3,0
 54     175      89     2      91      2,9
```

O resultado da previsão com o modelo adotado é

```
$fit
     1         2         3         4         5
 4.632615  5.945157  5.928703  7.577097 10.273719
$se.fit
     1         2         3         4         5
0.6747203 0.7155702 0.6255449 0.7731991 0.5660150
```

Exemplo 8.4: Consideremos novamente os dados do arquivo `coronarias` analisados no Exemplo 8.2. O modelo de regressão logística ali considerado também pode ser ajustado por meio do comando

```
mod1 <- gam(LO3 ~  IDADE1 + IMC + HA + PDR + PSR + COLS + TRIGS +
           GLICS, data=coronarias, family = "binomial")
```

Sob esse modelo, todas as variáveis (com exceção de `HA`, que é caegorizada) têm um efeito linear no logaritmo da chance de $LO3$. Para avaliar essa suposição, podemos recorrer a um modelo aditivo generalizado ajustado por meio do comando

```
mod2 <- gam(LO3 ~  s(IDADE1) + s(IMC) + HA + s(PDR) + s(PSR) +
                   s(COLS) + s(TRIGS) + s(GLICS),
               data=coronarias, family = "binomial")
summary(mod2)
```

obtendo os seguintes resultados

```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6281     0.1179   5.329  9.9e-08 ***
HA            0.3714     0.1608   2.309   0.0209 *
Approximate significance of smooth terms:
            edf Ref.df Chi.sq  p-value
s(IDADE1) 5.817  6.909 32.053 5.17e-05 ***
s(IMC)    1.269  1.494  2.934 0.212418
s(PDR)    1.260  1.476  0.210 0.890799
s(PSR)    3.251  4.085  7.428 0.118488
s(COLS)   5.612  6.625 16.617 0.013485 *
s(TRIGS)  2.971  3.770  5.129 0.313417
```

```
s(GLICS)  3.288  3.988 21.673 0.000242 ***
R-sq.(adj) =  0.0853   Deviance explained = 9.09%
UBRE = 0.19011  Scale est. = 1         n = 1032
```

Segundo esse modelo, além de `HA`, as variáveis `IDADE1`, `COLS` e `GLICS` são significativamente importantes para previsão, as três de forma não linear. Um modelo reduzido à luz dessas conclusões, ajustado por meio do comando

```
mod3 <- gam(LO3 ~  HA + s(IDADE1) + s(COLS) + s(GLICS), data=coronarias,
            family = "binomial")
summary(mod3)
```

gerando os seguintes resultados

```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6755     0.1096   6.161 7.25e-10 ***
HA            0.2595     0.1428   1.817   0.0692 .
Approximate significance of smooth terms:
            edf Ref.df Chi.sq  p-value
s(IDADE1) 5.499  6.596  26.06 0.000298 ***
s(COLS)   3.412  4.172  14.71 0.006339 **
s(GLICS)  3.217  3.908  22.31 0.000196 ***
R-sq.(adj) =  0.0715   Deviance explained = 7.01%
UBRE = 0.19422  Scale est. = 1         n = 1032
```

Os gráficos apresentados na Figura 8.11 salientam o comportamento não linear das variáveis suavizadas.

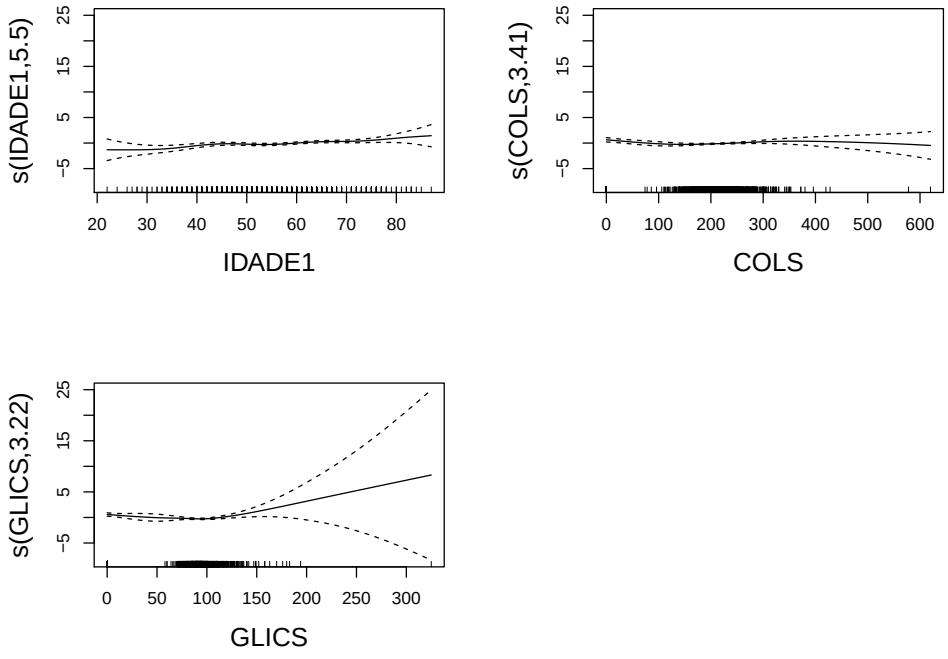

Figura 8.11: Funções suavizadas (com bandas de confiança) obtidas por meio do modelo *GAM* (mod3) para os dados do Exemplo 8.4.

Ferramentas de diagnóstico para dados com resposta dicotômica são mais complicadas do que aquelas destinadas a respostas contínuas. Uma alternativa é aquela descrita na Nota de Capítulo 7 do Capítulo 6.

Valores preditos pelo modelo para o conjunto original e a tabela de confusão associada são obtidos por meio de

```
> pred <- predict(mod3, newdata=coronarias)
> conf_gam <- table(pred>.5, coronarias$LO3)
> conf_gam
         0   1
  FALSE 159 191
  TRUE  172 510
```

e indicam um erro de previsão de 35,2% = (191 + 172)/1032.

8.4 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 utilizadas no ajuste do modelo.

Em modelos de regressão, o erro de previsão é definido como $EP = \mathrm{E}(y_0 - \widehat{y}_0)^2$ em que y_0 representa uma nova observação e $\widehat{y}_0$ é a previsão obtida pelo modelo. O **erro quadrático médio** (MSE) 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 validação**) para avaliar se ele tem um bom desempenho ou não.

O algoritmo proposto por Efron e Tibshirani (1993), conhecido por ***LOOCV*** (*Leave-One-Out Cross Validation*) e bastante utilizado nesse processo é o seguinte:

i) Dado um conjunto com n elementos, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, o modelo é ajustado n vezes, em cada uma delas eliminando um elemento; o valor previsto para o elemento eliminado, denotado por $\widehat{y}_{-i}$, é calculado com base no modelo ajustado aos demais $n-1$ elementos.

ii) O erro de previsão é estimado por

$$MSE_{(-i)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{-i})^2. \tag{8.18}$$

Como alternativa para (8.18) pode-se considerar

$$MSE_{(-i)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_{-i}}{1-h_i}\right)^2, \tag{8.19}$$

em que h_i é a **alavanca** (*leverage*), definida por

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

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 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 8.12 para o caso $k = 5$.

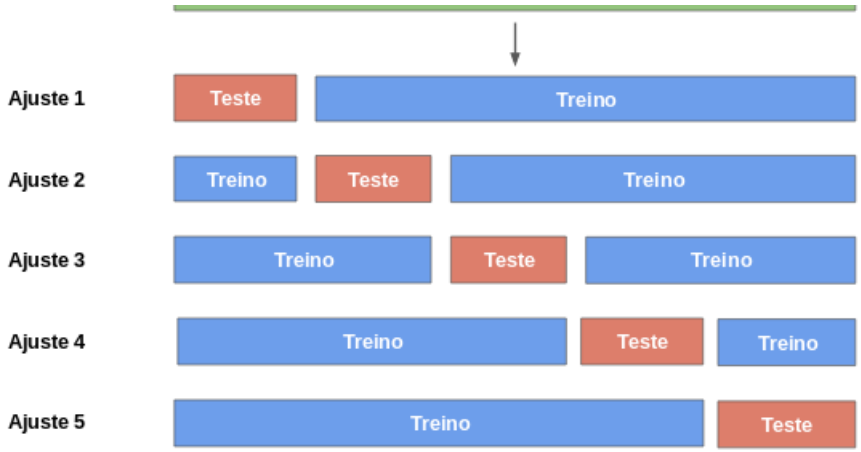

Figura 8.12: Representação esquemática da divisão dos dados para validação cruzada de ordem $k = 5$.

O correspondente erro de previsão é estimado por

$$MSE_{(k-fold)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i. \tag{8.20}$$

em que o erro quadrático médio (MSE) obtido no i-ésimo ajuste, $i = 1, \ldots, k$, é

$$MSE_i = \sum_{j=1}^{n_i}(y_{0j} - \widehat{y}_{0j})^2/n_i$$

com y_{0j}, $\widehat{y}_{0j}$ e n_i denotando, respectivamente, os valores observado e predito para o j-ésimo elemento e o número de elementos do i-ésimo conjunto de validação.

Nos casos em que o interesse é classificação, o MSE é substituído pela **taxa de erros** associada a um classificador $\widehat{g}$, obtido do ajuste do modelo aos dados de treinamento $\mathcal{T} = \{(\mathbf{x}_1^{(\mathcal{T})}, y_1^{(\mathcal{T})}), \ldots, (\mathbf{x}_t^{(\mathcal{T})}, y_t^{(\mathcal{T})})\}$. Essa taxa é definida como

$$TE = \frac{1}{v}\sum_{i=1}^{v} I[y_i^{(\mathcal{V})} \neq \widehat{y}_i^{(\mathcal{V})}] \tag{8.21}$$

em que $y_i^{(\mathcal{V})}$ denota a classe correspondente ao i-ésimo elemento do conjunto de dados de validação, $\mathcal{V} = \{(\mathbf{x}_1^{(\mathcal{V})}, y_1^{(\mathcal{V})}), \ldots, (\mathbf{x}_v^{(\mathcal{V})}, y_v^{(\mathcal{V})})\}$, $\widehat{y}_i^{(\mathcal{V})}$, o valor pre-

dito correspondente obtido por meio do classificador $\widehat{g}$ e a média é calculada em relação a todos os v dados desse conjunto.

2) **Viés da regularização *Ridge*.**

Fazendo $\mathbf{R} = \mathbf{X}^\top\mathbf{X}$, o estimador *Ridge* (8.2) pode ser expresso como

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = (\mathbf{I} + \lambda\mathbf{R}^{-1})^{-1}\widehat{\boldsymbol{\beta}}_{\text{MQ}}, \tag{8.22}$$

em que $\widehat{\boldsymbol{\beta}}_{\text{MQ}}$ denota o estimador de mínimos quadrados ordinários. A esperança condicional de (8.22), dada $\mathbf{X}$, é

$$\text{E}[\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda)] = (\mathbf{I} + \lambda\mathbf{R}^{-1})^{-1}\boldsymbol{\beta} \neq \boldsymbol{\beta} \tag{8.23}$$

indicando que o estimador *Ridge* é enviesado.

3) **Escolha do parâmetro de regularização λ**

A escolha do parâmetro de regularização λ pode ser baseada em validação cruzada ou em algum critério de informação.

Considerando a decomposição em valores singulares de $\mathbf{X}$, nomeadamente, $\mathbf{X} = \mathbf{UDV}^\top$, em que $\mathbf{U}$ denota 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 $\mathbf{X}^\top\mathbf{X}$), pode-se provar que

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \mathbf{V}\left[\text{diag}\left(\frac{d_1}{d_1^2+\lambda}, \frac{d_2}{d_2^2+\lambda}, \ldots, \frac{d_p}{d_p^2+\lambda}\right)\right]\mathbf{U}^\top\mathbf{y}.$$

Seja $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ uma grade de valores para λ. Para e escolha de um valor apropriado para λ, podemos usar um critério de informação do tipo

$$\widehat{\lambda} = \arg\min_{\lambda\in\Lambda}[-\log \text{verossimilhança} + \text{penalização}],$$

como

$$\begin{aligned} AIC &= \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{2}{n}, \\ BIC &= \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{\log n}{n}, \\ HQ &= \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{\log\log n}{n}, \end{aligned}$$

em que $\text{gl}(\lambda)$ é o número de graus de liberdade associado a λ, nomeadamente,

$$\text{gl}(\lambda) = \text{tr}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\right] = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda},$$

e

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n - \text{gl}(\lambda)}\sum_{t=1}^{n}[y_t - \widehat{\boldsymbol{\beta}}_{Ridge}(\lambda)^\top\mathbf{x}_t]^2.$$

4) **Modelos Lineares Generalizados**

Com a finalidade de avaliar a relação entre variáveis respostas com distribuição na classe da **família exponencial** e variáveis explicativas, Nelder e Wederburn (1972) propuseram os chamados **Modelos Lineares Generalizados** (*Generalized Linear Models*).

A função densidade (de probabilidade) de variáveis com distribuição na família exponencial pode ser expressa de uma forma geral como

$$f(y|\theta,\phi) = \exp\{[a(\phi)]^{-1}[y\theta - b(\theta)]\} + c(y,\theta), \tag{8.24}$$

em que θ e ϕ são parâmetros e a, b e c são funções conhecidas. Pode-se mostrar que

$$\mathrm{E}(y) = \mu = b'(\theta) = db(\theta)/d\theta$$

e que

$$\mathrm{Var}(y) = a(\phi)b''(\theta) = a(\phi)d^2b(\theta)/d\theta^2 = a(\phi)V(\mu),$$

com

$$V(\mu) = d^2b(\theta)/d\theta^2 = d\mu(\theta)/d\theta. \tag{8.25}$$

A expressão (8.25) é conhecida como **função de variância** e relaciona a variância com o valor esperado de y.

Muitas distribuições podem ser expressas na forma (8.24). Em particular, para mostrar que a distribuição normal com parâmetros μ e σ^2 pertence a essa família, basta fazer

$$a(\theta) = \sigma^2,\ \theta = \mu,\ b(\theta) = \theta^2/2,\ \text{e}\ c(y,\theta = -[y^2/\sigma^2 + \log(2\pi\sigma^2)]/2.$$

Para a distribuição binomial com parâmetros n e p , os termos de (8.24) são

$$a(\phi) = 1,\ \theta = \log[p/(1-p)],\ b(\theta) = -n\log(1-p)$$

e

$$c(y,\phi) = \log\{n!/[y!(n-y)!]\}.$$

O modelo linear generalizado para variáveis da família exponencial é definido como

$$g(\mu) = \mathbf{x}^\top\boldsymbol{\beta},$$

em que $\mathbf{x}$ é um vetor com os valores de variáveis explicativas, $\boldsymbol{\beta}$ é um vetor de coeficientes a serem estimados e g, conhecida como **função de ligação**, identifica a função do valor esperado (μ) que se pretende modelar linearmente. Para a distribuição normal, definindo $g(\mu)$ como a função identidade, o modelo se reduz ao modelo linear gaussiano padrão. Para a distribuição binomial, definindo $g(\mu)$ como $\log[\mu/(1-\mu)]$, obtemos o modelo de regressão logística.

A vantagem dessa formulação generalizada é que vários modelos podem ser ajustados por meio de um único algoritmo. Dada uma amostra aleatória $[(y_1,\mathbf{x}_1),\ldots,(y_1,\mathbf{x}_1)]^\top$ em que a distribuição de Y pertence à família exponencial, a função log-verossimilhança é

$$\ell[\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y}] = \sum_{i=1}^{n}\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\}, \tag{8.26}$$

com $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)^\top$ e $\mu_i = \mathrm{E}(y_i) = b'(\theta_i) = g^{-1}(\mathbf{x}_i^\top\boldsymbol{\beta})$.

Em geral, a maximização de (8.26) requer métodos iterativos como aqueles discutidos no Apêndice A. Nesse contexto, o algoritmo de Newton-Raphson pode ser implementado por meio do processo iterativo

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + [\mathbf{I}(\boldsymbol{\beta}^{(k)})]^{-1}\mathbf{u}(\boldsymbol{\beta}^{(k)}),\ k = 0{,}1.....$$

em que

$$\mathbf{u}(\boldsymbol{\beta}) = \frac{\partial \ell[\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y}]}{\partial \boldsymbol{\beta}} \ \text{ e } \ \mathbf{I}(\boldsymbol{\beta}) = -\mathrm{E}\left\{\frac{\partial^2 \ell[\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{y}]}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right\}$$

denotam, respectivamente, a **função escore** e a correspondente matriz de **informação de Fisher**.

8.5 Exercícios

1) Obtenha os estimadores *Ridge*, *Lasso* e *Elastic Net* para os dados do Exemplo 6.7 e comente os resultados.

2) Repita a análise do Exemplo 8.2. Justifique possíveis diferenças encontradas.

3) Ajuste um modelo *GAM* aos dados do Exemplo 6.9 e compare os resultados com aqueles obtidos por meio do modelo de regressão linear.

4) Reanalise os dados do Exemplo 8.3 adotando uma distribuição gama com função de ligação logarítmica para a variável resposta. Compare os resultados com aqueles obtidos sob a suposição de normalidade.

5) Considere o conjunto de dados `esforco`, centrando o interesse na predição da variável resposta Y: VO2 (consumo de oxigênio) com base nas variáveis preditoras X_1: Idade, X_2: Peso, X_3: Superfície corpórea e X_4: IMC (índice de massa corpórea) ($n = 126$).

   a) Ajuste um modelo de regressão linear aos dados, utilizando o método dos mínimos quadrados e analise os resultados.

   b) Ajuste o mesmo modelo por meio de regularização *Ridge*, obtenha λ, a raiz do erro quadrático médio e o coeficiente R^2.

   c) Repita o procedimento utilizando regularização *Lasso* e *Elastic Net*.

   d) Compare os resultados obtidos com os ajustes obtidos nos itens a), b) e c).

6) Considere o conjunto de dados `antracose`, selecionando aleatoriamente 70% das observações para treinamento e as restantes para validação.

   a) Ajuste modelos de regressão linear múltipla e de regularização *Ridge*, *Lasso* e *Elastic Net* ao conjunto de treinamento com o objetivo de previsão da variável `antracose` com base nas variáveis preditoras `idade`, `tmunic`, `htransp`, `cargatabag`, `ses`, `densid` e `distmin`.

   b) Compare o desempenho dos modelos no conjunto de validação.

   c) Repita os itens a) e b) com outra seleção aleatória dos conjuntos de treinamento e de validação.

   d) Construa uma tabela com as medidas de desempenho de todos os modelos ajustados e comente os resultados.

e) Utilize modelos *GAM* para avaliar se as variáveis preditoras estão linearmente associadas com a variável resposta e compare o resultado do modelo adotado sob esse enfoque com aqueles dos demais modelos.

7) Mostre que o estimador *Ridge* pode ser expresso como (8.22) e que seu valor esperado é (8.23).

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.

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical Statistics*. 2nd edition. Chapman and Hall.

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. (1969). *Probability*. Reading: Addison-Wesley Publishing Co.

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.

Broyden, C.G. (1965). A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, **19**, 577-593.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, **30**, 927-961.

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.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245-276.

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*, **49**, 327-335.

Chollet, F. (2018). *Deep Learning with R*. Shelter Island, NY: Manning Publications Co.

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.

Dantzig, G.B. (1963). *Linear Programming and Extensions*. Princeton: Princeton University Press.

Davidon, W.C. (1959). Variable metric for minimization. Report ANL-5990 Rev., Argonne National Laboratories, Argonne, Illinois.

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.

Dennis, J.E. and Moré, J.J. (1977). Quasi-Newton methods, motivation and theory. *SIAMM Review*, **19**, 46-89.

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.

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**, 89-121.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference.* New York: Cambridge University Press.

Eilers, P. H. C. and Marx, B. D. (2021). *Practical Smoothing: The Joys of P-splines.* Cambridge: 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. 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.

Faraway, J.J. and Augustin, N.H. (2018). When small data beats big data. *Statistics and Probability Letters*, **136**, 142-145.

Ferreira, D.F. (2011). *Análise Discriminante*. Encontro Mineiro de Estatística, São João Del-Rei, M.G.

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

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society, Series A, **222**, 309-368.

Fletcher, R. and Powell, M.J.D. (1963). A rapid convergent descent method for minimization. *Computer Journal*, **6**, 163-168.

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, 2nd 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.

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.

Guttman, L. (1954). Some necessary conditions for common factor analysis.*Psychometrika*, XIX, 149-161.

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.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, **30**, 179-185.

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*, 3rd edition. 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.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement,* **20**, 141-151.

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.

Lindstrom, M. (2016). *Small Data: The Tiny Clues that Uncover Huge Trends*. London: St. Martin's Press.

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. August 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.

Meng, X.L. (2014). A trio of inference problems that could win you a Nobel Prize in Statistics (if you help fund it). In: **Lin, X., et al. (Eds). Past, Present, and Future of Statistical Science.** Boca Raton: CRC Press.

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.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

Nelder at al. (1988).

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A,* **231**, 289-337.

Nocedal, J. and Wright S. (2006). *Numerical Optimization.* Segunda Edição. Springer.

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 in-*

suficiência cardíaca encaminhados para tratamento cirúrgico." São Paulo: Centro de Estatística Aplicada do IME/USP.

Pepe, M.S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the Receiver Operating Characteristic curve. *Biometrics*, **62**, 221-229.

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 cognitive 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.

Stigler, S. M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge: Harvard University Press.

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.

Takano, A.P.C., Justo, L.T., Santos, N.V., Marquezini, M.V., André, P.A., Rocha, F.M.M., Barrozo, L.V., Singer, J.M., André, C.D.S., Saldiva, P.H.N. and Veras, M.M., (2019). Pleural anthracosis as an indicator of lifetime exposure to urban air pollution: an autopsy-based study in São Paulo. *Environmental Research*, **173**, 23-32.

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 Remissivo

Accelerated failure time models, 263
Data Science, 1
Draftsman's display, 150
Lowess, 154, 170
Odds ratio, 107
Proportional hazards model, 264
Stem and leaf, 51

Acurácia, 110
Alavanca, 236, 296
Amostra, 26, 67
    aleatória simples, 26, 70, 79
    piloto, 81
Amplitude, 60
ANOVA, 126
Análise
    condicional, 204
    de agrupamentos, 13
    de componentes independentes, 14
    de componentes principais, 14, 150
    de confiabilidade, 251
    de desviância, 292
    de regressão, 112
    de séries de tempo, 201
    de séries temporais, 201
    de variância, 84, 126, 157
    exploratória de dados, 25
Aprendizado
    automático, 4
    com Estatística, 3
    não supervisionado, 13
Associação entre variáveis, 97
Autocorrelação, 202

Banco de dados, 26

Capa de Cohen, 104
Censura, 251
    intervalar, 265
    à direita, 265
    à esquerda, 265
Chance, 107, 225
Ciência de dados, 1
Classe modal, 57
Classificador
    de Bayes, 13
    KNN, 13
Classificação, 12
Coeficiente
    de autocorrelação, 203
    de confiança, 79
    de contingência, 102
    de correlação, 133
    de correlação de Pearson, 113
    de correlação de Spearman, 113
    de determinação, 191
    de determinação ajustado, 220, 235
    de penalização, 275
    de regularização, 276
    de Tschuprov, 102
Comparações múltiplas, 128
Concordância, 104, 118
Conjunto
    de dados, 17
Curtose, 64

Dado
de treinamento, 295
de validação, 295
estruturado, 10
longitudinal, 30, 116, 201, 221
não estruturado, 10
omisso, 28
Desvio
absoluto médio, 59
mediano absoluto, 60
médio, 59
padrão, 59
Diferença significativa, 136
Distribuição
conjunta, 98
de frequências, 48
de valores extremos, 263
Distância
de Cook, 199, 215
interquartis, 60

Efeito
aleatório, 224
de tratamento, 127
principal, 159
Ensaio clínico, 26
Equação de estimação, 191, 226
Erro
aleatório, 211
padrão, 76
propagação de, 38
quadrático médio, 11, 295
Especificidade, 109
Estatística, 7
de Durbin-Watson, 203
de Mantel-Haenszel, 176
de ordem, 56, 71
de Pearson, 101, 238
Estimador
de Kaplan Meier, 255
de mínimos quadrados, 190, 210
de mínimos quadrados penalizados, 276
do limite de produtos, 255
não enviesado, 59
resistente, 197
robusto, 197
Estrutura de dados, 9
Estudo
caso-controle, 132
observacional, 26
prospectivo, 106, 131
retrospectivo, 108, 132
Expansão de Taylor, 84

Falso
negativo, 110
positivo, 110
Família exponencial, 286, 298
Fator, 156
de risco, 106
efeito, 157
interação, 157
Forma quadrática, 212
Fronteira
de decisão linear, 271
Função
biquadrática, 172
de distribuição empírica, 255
de ligação, 298
de probabilidade, 68
de risco, 253
de risco acumulado, 254
de risco basal, 264
de sobrevivência, 252
de variância, 298
densidade de probabilidade, 68
discriminante de Fisher, 271
distribuição acumulada, 71
distribuição empírica, 71
escore, 299
tricúbica, 171

Graus de liberdade, 59
Gráfico
dotplot, 50
da variável adicionada, 220
de barras, 48
de Bland-Altman, 120
de Cook, 192, 198
de dispersão, 111
de dispersão simbólico, 152

de dispersão unidimensional, 50
de médias/diferenças, 120
de perfis, 222
de perfis individuais, 116
de perfis médios, 125, 156
de pizza, 48
de quantis, 62
de resíduos, 192
de simetria, 63
do desenhista, 150, 222
PP, 135
QQ, 72, 118, 199
ramo e folhas, 51
torta, 48

Heteroscedasticidade, 196
Hipótese
de homogeneidade, 101
de independência, 101
Homocedasticidade, 194

Inferência Estatística, 25, 68
Influência local, 192
Informação
sistemática, 10
Inteligência artificial, 4
Interação
essencial, 159
não essencial, 159
Intervalo
de confiança, 80, 233
de previsão, 233

Janelamento, 155

Limiar suave, 277
LOOCV, 295

Margem de erro, 79
Matlab, 17
Matriz
de correlações, 165, 169
de covariâncias, 165, 168
de dados, 10, 150, 164
Mediana, 56
Medida
de localização, 56
de tendência central, 56
resistente, 57
robusta, 57
Megadados, 1
Meia média, 56
Microdados, 2
Minitab, 17
Moda, 57
Modelo
de regressão de Cox, 264
de regressão linear múltipla, 209
de regressão linear simples, 190
de regressão polinomial, 190
de riscos proporcionais, 264
de tempo de falha acelerado, 263
exponencial, 254
hipergeométrico, 262
Linear Generalizado, 298
linear misto, 222
linearizável, 206
log normal, 254
não linear, 190
paramétrico, 263
probabilístico, 67
semiparamétrico, 264
Weibull, 254
Momento centrado, 61
Multiplicadores de Lagrange, 276
Média, 56
aparada, 56
Método
de Mantel-Haenszel, 168
de máxima verossimilhança, 226
de mínimos quadrados, 7, 212, 242
de mínimos quadrados generalizados, 204, 228
de Newton-Raphson, 226
Delta, 138, 227
Fisher scoring, 226

Observação atípica, 197

Paradoxo de Simpson, 177
Parametrização, 158

de cela de referência, 174
de desvios de médias, 174
de desvios médios, 174
de médias de celas, 158
Partição, 155
Parâmetro, 68, 189
de localização, 158
Percentil, 58
Ponto
alavanca, 199
influente, 199
Posto, 113
Prevalência, 110
Previsão, 10, 271
Probabilidades, 6
Processo estocástico, 70

Quantil, 57
empírico, 57
Quartil, 58

Razão de chances, 107, 137, 226, 228
Rede neural, 240
Redução de dimensionalidade, 14
Regressão
linear, 7
linear múltipla, 190, 209
linear simples, 190
logística multinomial, 231
logística politômica, 231
resistente, 239
segmentada, 173
Regularização, 274
Elastic Net, 274
Lasso, 274
Ridge, 274
Resíduo, 101, 191, 210
condicional, 224
da desviância, 238
de efeito aleatório, 224
de Pearson, 238
estudentizado, 193, 232
marginal, 224
padronizado, 193
Risco
atribuível, 107
relativo, 107, 137, 264

SAS, 17
Sensibilidade, 109
Sobreajuste, 7, 274
SPlus, 17
Suavização, 170
Série temporal, 149, 204

Tabela
atuarial, 268
de contingência, 100
de dupla entrada, 100
Teorema
de Gauss-Markov, 232
Limite Central, 212, 232
Teste
log rank, 262

Unidade amostral, 28

Validação cruzada, 6, 12, 273, 295, 297
de ordem k, 296
Valor
atípico, 65
discrepante, 65
esperado, 101
preditivo negativo, 110
preditivo positivo, 110
Variáveis
comonotônicas, 132
contramonotônicas, 132
Variável
bimodal, 57
contínua, 46
discreta, 46
explicativa, 98
nominal, 46
ordinal, 46
padronizada, 86
preditora, 98
qualitativa, 46
quantitativa, 46
resposta, 98
valor esperado, 68

Variância, 58
    aparada, 60, 133