MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 5

14 de março de 2024

Sumário

RLM-modelo

- Com p variáveis explicativas $X_1, \ldots, X_p$ e uma variável resposta Y, o **modelo de regressão linear múltipla** é expresso como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i, \quad i = 1, \ldots, n. \tag{1}$$

  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 (1) 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.

RLM-modelo

- Definindo $x_{i0} = 1$, $i = 1, \ldots, n$, podemos escrever (1) na forma

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i, \quad i = 1, \ldots, n.$$

Minimizando a soma dos quadrados do erros e_i, *i.e.*,

$$Q(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{p} \beta_j x_{ij}]^2,$$

em relação a $\beta_0, \ldots, \beta_p$ obtemos os **estimadores de mínimos quadrados**(EMQ) $\widehat{\beta}_j$, $j = 1, \ldots, p$, de modo que

$$\widehat{y}_i = \sum_{j=0}^{p} \widehat{\beta}_j x_{ij}, \quad i = 1, \ldots, n$$

são os **valores estimados** (sob o modelo).

- Os termos

$$\widehat{e}_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n \tag{2}$$

são os **resíduos**, cuja análise é fundamental para avaliar se modelos da forma (1) se ajustam bem aos dados.

RLM - o modelo

Para efeitos computacionais os dados correspondentes a problemas de regressão linear múltipla devem ser dispostos como indicado na Tabela 1.

Tabela 1: Matriz de dados

Y	X_1	X_2	$\cdots$	X_p
y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Em geral, a variável correspondente ao intercepto (que é constante e igual a um) não precisa ser incluída na matriz de dados; os pacotes computacionais incluem-na naturalmente no modelo a não ser que se indique o contrário.

RLM - o modelo

Para facilitar o desenvolvimento metodológico, convém expressar o modelo na forma matricial

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{3}$$

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**.

RLM - Exemplo

- Os dados **esteira** são provenientes de um estudo cujo objetivo é avaliar o efeito do índice de massa corpórea (IMC) e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO2) numa determinada fase do exercício.
- Para associar a distribuição do consumo de oxigênio (Y) com as informações sobre carga na esteira ergométrica (X_1) e IMC (X_2), consideramos o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \tag{4}$$

$i = 1, \ldots, 28$ com as suposições usuais sobre os erros (média zero, variância constante σ^2 e não correlacionados). Aqui, o parâmetro β_1 representa a variação no VO2 esperada por unidade carga para indivíduos com o mesmo IMC. O parâmetro β_2 tem interpretação semelhante com a substituição de carga na esteira por IMC e IMC por carga na esteira.

RLM - Exemplo

- Como não temos dados para indivíduos com IMC menor que 17,50 e carga menor que 32, o parâmetro β_0 deve ser interpretado como um fator de ajuste do plano que aproxima a verdadeira função que relaciona o valor esperado da variável resposta com as variáveis explicativas na região em que há dados disponíveis.
- Se substituíssemos X_1 por $X_1 - 32$ e X_2 por $X_2 - 17.5$, o termo β_0 corresponderia ao VO2 esperado para um indivíduo com IMC $= 17{,}50$ submetido a uma carga igual a 32 na esteira ergométrica.
  O modelo (4) pode ser expresso na forma matricial (3) 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.

RLM - Propriedades

- 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 (3) pode ser desenvolvido de maneira universal e corresponde à minimização da forma quadrática

$$Q(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2. \tag{5}$$

- Por meio da utilização de operações matriciais, obtém-se a seguinte expressão para os estimadores de mínimos quadrados

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{6}$$

- Sob a suposição de que $E(\mathbf{e}) = \mathbf{0}$ e $\text{var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$, em que $\mathbf{I}_n$ denota a matriz identidade de dimensão n, temos
  - i) $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,
  - ii) $\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

RLM - Propriedades

- Além disso, se adicionarmos a suposição de que os erros têm distribuição Normal, pode-se mostrar que o estimador (6) tem uma distribuição Normal multivariada, o que permite a construção de intervalos de confiança para ou testes de hipóteses sobre os elementos (ou combinações lineares deles) de $\boldsymbol{\beta}$ por meio de estatísticas com distribuições exatas. Mesmo sem a suposição de normalidade para os erros, um recurso ao **Teorema Limite Central** permite mostrar que a distribuição aproximada do estimador (6) é Normal, com média a $\boldsymbol{\beta}$ e matriz de covariâncias $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.
- Um estimador não enviesado de σ^2 é

$$\begin{aligned} s^2 &= [n-(p+1)]^{-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= [n-(p+1)]^{-1}\mathbf{y}^\top[\mathbf{I}_n-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{y}. \end{aligned}$$

- Com duas variáveis explicativas, o gráfico de dispersão precisa ser construído num espaço tridimensional, que ainda pode ser representado em duas dimensões; para mais que 2 variáveis explicativas, o gráfico de dispersão requer um espaço com mais do que três dimensões que não pode ser representado no plano. Por isso, uma alternativa é construir gráficos de dispersão entre a variável resposta e cada uma das variáveis explicativas.

RLM - Gráficos

Para os dados **esteira**, o gráfico de dispersão com três dimensões incluindo o plano correspondente ao modelo de regressão múltipla ajustado está disposto na Figura 1.

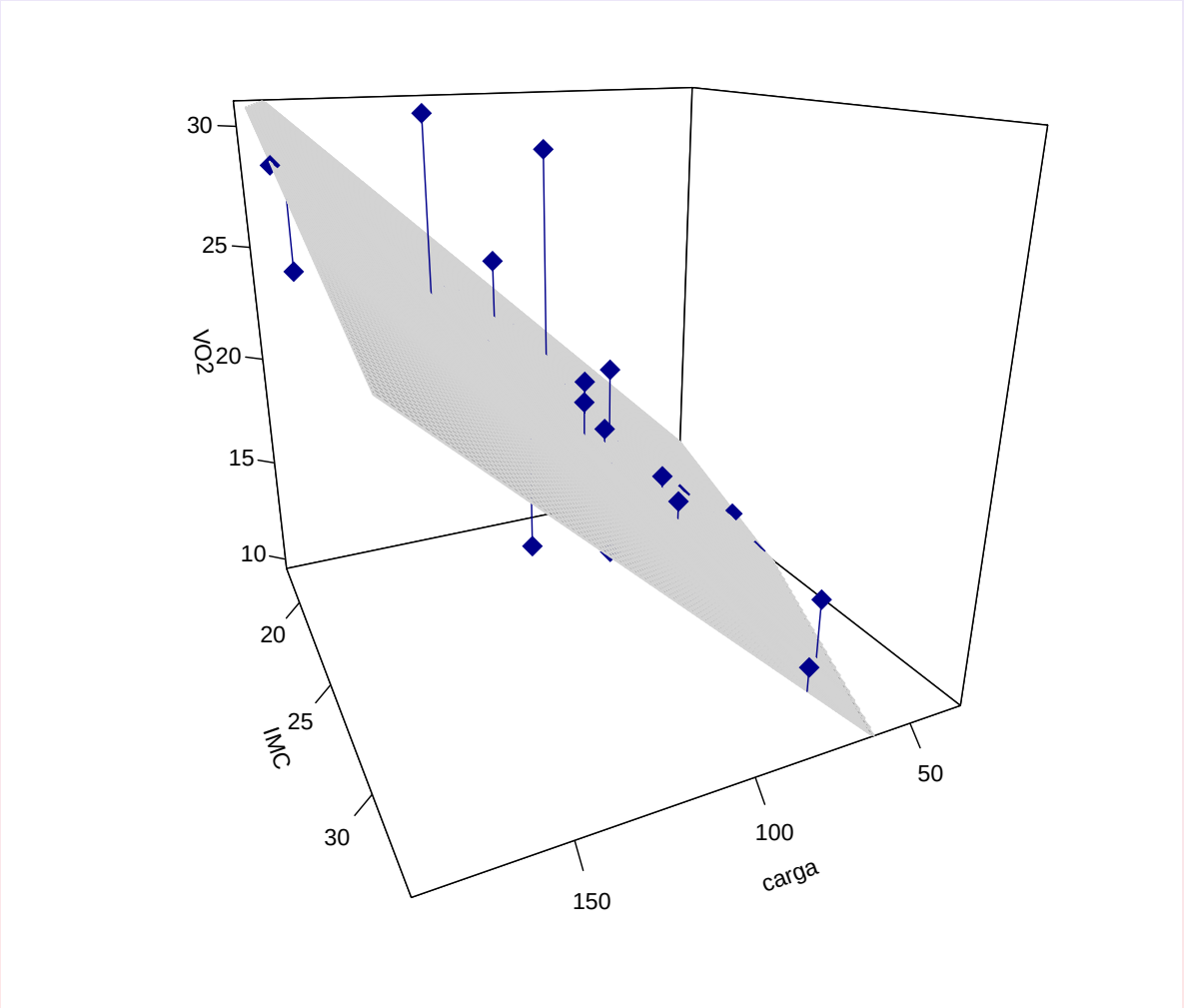

Figura 1: Gráficos de dispersão tridimensional para os dados **esteira**

RLM - Gráficos

Os gráficos de dispersão correspondentes a cada uma das duas variáveis explicativas estão dispostos na Figura 2 e indicam que a distribuição do VO2 varia positivamente com a carga na esteira e negativamente com o IMC.

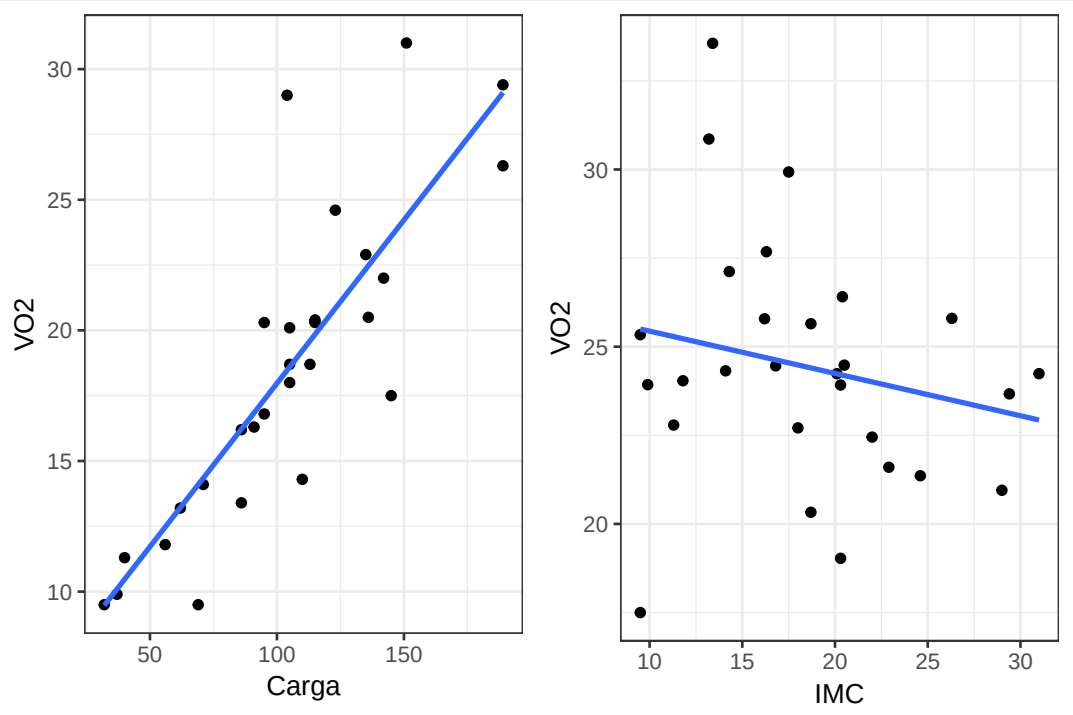

Figura 2: Gráficos de dispersão para os dados **esteira**.

RLM - Uso do R

O uso da função lm() conduz aos seguintes resultados.

Call:
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: $1.887e-08$

RLM - Uso do R

- Essa saída nos diz que os coeficientes (erro padrão) correspondentes ao ajuste do modelo (4) aos dados **esteira** são $\widehat{\beta_0} = 15,45$ $(4,45)$, $\widehat{\beta_1} = 0,13$ $(0,01)$ e $\widehat{\beta_2} = -0,41$ $(0,17)$. Então, segundo o modelo, o valor esperado do VO2 para um indivíduo (IMC fixado) aumenta de $0,13$ unidades para cada aumento de uma unidade da carga na esteira; similarmente, o valor esperado do VO2 para indivíduos submetidos à mesma carga na esteira diminui de $0,41$ unidades com o aumento de uma unidade no IMC.
- Embora o coeficiente de determinação $R^2 = 0,74$ sugira a adequação do modelo, convém avaliá-la por meio de outras ferramentas diagnósticas. No caso de regressão linear múltipla, gráficos de resíduos podem ter cada uma das variáveis explicativas ou os valores ajustados no eixo das abscissas. Para o exemplo, esses gráficos estão dispostos na Figura 3 juntamente com o gráfico contendo as distâncias de Cook.
- Os gráficos de resíduos padronizados não indicam um comprometimento da hipótese de homoscedasticidade embora seja possível suspeitar de dois ou três pontos discrepantes (correspondentes aos indivíduos com identificação 4, 8 e 28) que também são salientados no gráfico das distâncias de Cook. Veja também a Figura 1.

RLM - Uso do R

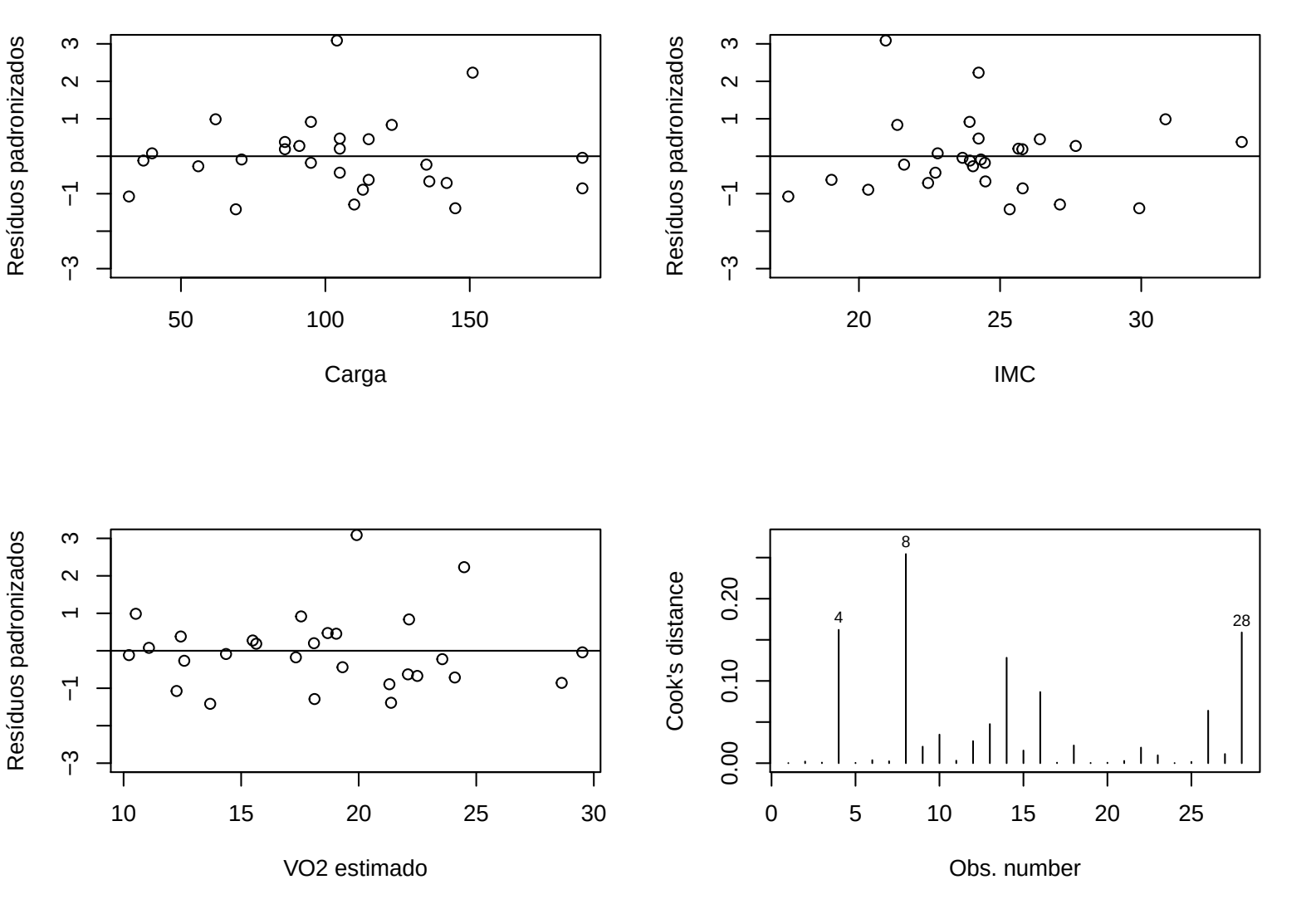

Figura 3: Gráficos de resíduos padronizados e distâncias de Cook para o ajuste do modelo (4) aos dados **esteira**.

frame

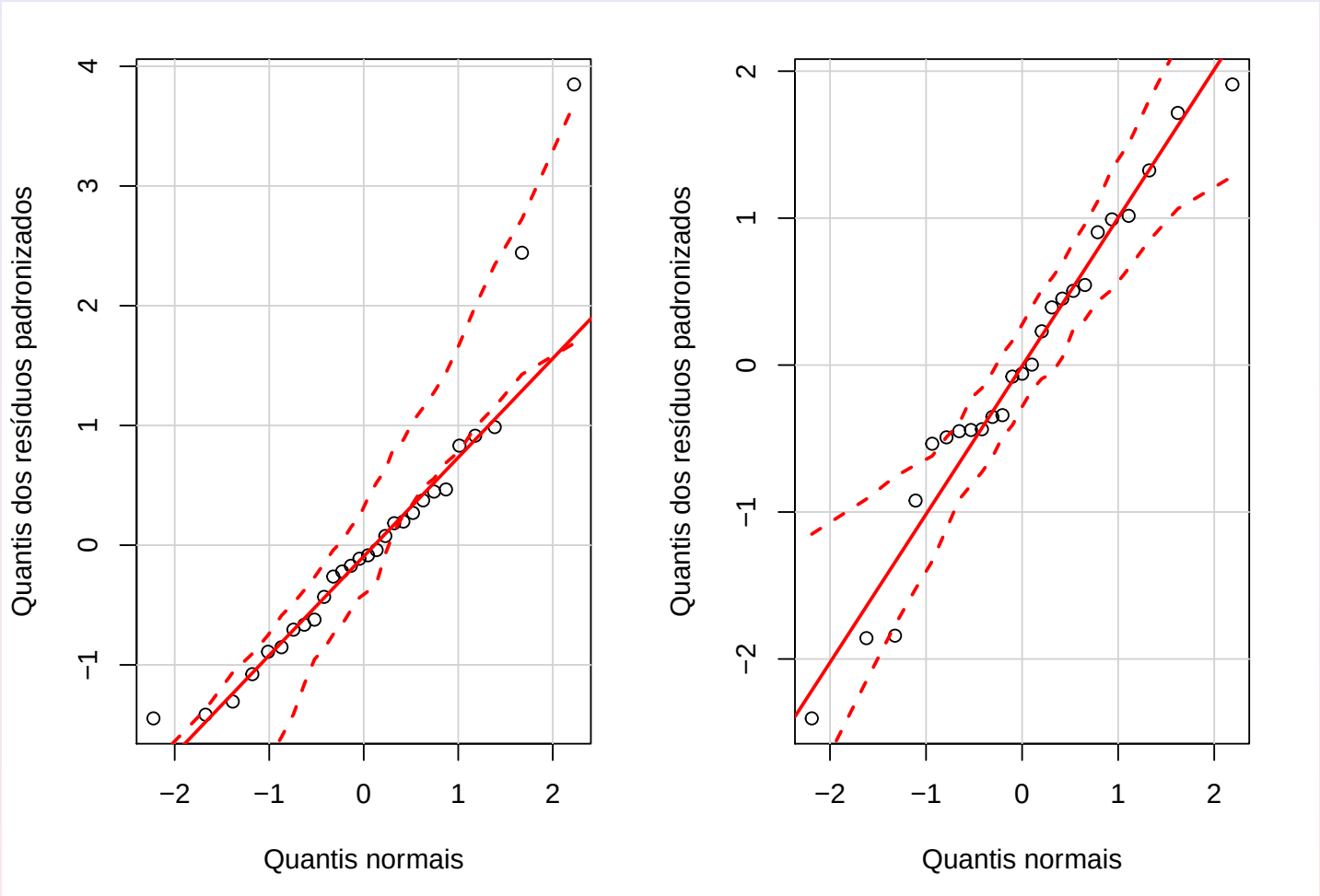

Figura 4: Gráficos QQ correspondentes ao ajuste do modelo (4) aos dados **esteira** com (painel esquerdo) e sem (painel direito) os pontos com identificação 4, 8 e 28.

RL - Regressão logística

- **Exemplo**. O conjunto de dados **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 (hormônio folículo estimulante) 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.
- A diferença entre esse problema e aqueles estudados nas seções anteriores está no fato de a variável resposta ser dicotômica e não contínua. Se definirmos a variável Y com valor igual a 1 no caso de resposta positiva e igual a zero no caso de resposta negativa, a resposta média será igual à proporção $p = E(Y)$ de pacientes com resposta positiva. Essencialmente, o objetivo da análise é modelar essa proporção como função da variável explicativa.

RL - o modelo

- Em vez de modelar a resposta média, convém modelar uma função dela, a saber o logaritmo da chance de resposta positiva para evitar estimativas de proporções com valores fora do intervalo (0, 1). O modelo correspondente pode ser escrito como

$$\log \frac{P(Y_i = 1|X = x)}{P(Y_i = 0|X = x)} = \alpha + \beta x_i, \ i = 1, \ldots, n. \tag{7}$$

- De forma equivalente,

$$P(Y_i = 1|X = x) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}, \ i = 1, \ldots, n. \tag{8}$$

RL - o modelo

- Neste contexto, o parâmetro α é interpretado como o logaritmo da chance de resposta positiva para pacientes com $x_i = 0$ (concentrações de inibina pré e pós estímulo iguais) e o parâmetro β corresponde ao logaritmo da razão entre a chance de resposta positiva para pacientes com diferença de uma unidade na variável explicativa.
- O ajuste desse modelo é realizado pelo método de máxima verossimilhança. A função de verossimilhança a ser maximizada é

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

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

RL - EMV

- A maximização da verossimilhança 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**

$$\sum_{i=1}^{n} \left\{ y_i - \frac{\exp(\widehat{\alpha} + \widehat{\beta} x_i)}{1 + \exp(\widehat{\alpha} + \widehat{\beta} x_i)} \right\} = 0 \quad \text{e} \quad \sum_{i=1}^{n} x_i \left\{ y_i - \frac{\exp(\widehat{\alpha} + \widehat{\beta} x_i)}{1 + \exp(\widehat{\alpha} + \widehat{\beta} x_i)} \right\} = 0.$$

- Como esse sistema de equações não tem solução explícita, deve-se recorrer a métodos iterativos como o **método de Newton-Raphson**.

RL - Uso do R

O uso da função glm() produz os resultados a seguir:

Call:
glm(formula = resposta ~ difinib, family = binomial, data = dados)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9770	-0.5594	0.1890	0.5589	2.0631

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-2.310455	0.947438	-2.439	0.01474
inib	0.025965	0.008561	3.033	0.00242

(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

RL - Uso do R

- Deviance: generaliza a SQR, usada em modelos lineares, para o caso de modelos lineares generalizados (GLM).
- Null deviance=2(LL(Modelo saturado)-LL(modelo nulo)); df=df(Saturado)-df(Nulo)=n-1
- Residual deviance=2(LL(ModeloSaturado)-LL(Modelo Proposto)); df=df(Saturado)-df(Proposto)=n-(p+1)
- Modelo saturado: modelo com um parâmetro para cada observação (dados ajustados exatamente)
- Modelo nulo: um parâmetro para todos os dados (estima um parâmetro)
- Modelo proposto: explica seus dados com p parâmetros + intercepto (p+1 parâmetros)

RL - Uso do R

- As estimativas dos parâmetros (com erro padrão entre parênteses) α e β correspondentes ao modelo ajustado aos dados **inibina** são, respectivamente,

$$\widehat{\alpha} = -2,31\ (0,95), \quad \widehat{\beta} = 0,03\ (0,01)$$

.

- Consequentemente, a chance de resposta positiva para pacientes com mesmo nível de inibina B pré e pós estímulo hormonal é $\exp(\widehat{\alpha}) = 0,10$.
- Essa chance fica multiplicada por $\exp(\widehat{\beta}) = 1,03$ para cada aumento de uma unidade na diferença entre os níveis de inibina B pré e pós estímulo hormonal.
- Os erros padrões de $\exp(\widehat{\alpha})$ e $\exp(\widehat{\beta})$ são calculados por meio do **método Delta**. Ver Nota de Capítulo 6.

RL - Uso do R

A função predict() pode ser usada para estimar a probabilidade de que a resposta seja positiva, dados os valores da variável explicativa. Algumas dessas probabilidades estão indicadas abaixo:

1	2	3	4	5	6
0.1190483	0.7018691	0.9554275	0.9988353	0.5797138	0.9588247
7	8	9	10		
0.8045906	0.8362005	0.9534173	0.8997726		

Por exemplo, o valor 0,1190483 foi obtido calculando-se

$$P(Y = 1|X = 11,90) = \frac{\exp\{-2,310455 + (0,025965)(11,90)\}}{1 + \exp\{-2,310455 + (0,025965)(11,90)\}}. \quad (9)$$

RL - uso do R

- Para prever se a resposta vai ser positiva ou negativa, temos que converter essas probabilidades previstas em rótulos de classes, "positiva" / ou "negativa". Considerando respostas positivas como aquelas cuja probabilidade seja maior do que 0,7, digamos, podemos utilizar a função table() para obter a seguinte tabela:

	resposta	
pred	negativa	positiva
negativa	11	5
positiva	2	14

- Os elementos da diagonal dessa tabela indicam os números de observações corretamente classificadas. Ou seja, a proporção de respostas corretas será (11+14)/32= 78%. Esse valor depende do limiar fixado, $0,7$, no caso. Um *default* usualmente fixado é $0,5$, e nesse caso, a proporção de respostas corretas vai aumentar.
- A utilização de Regressão Logística nesse contexto de classificação será detalhada no Capítulo 10.

Algumas considerações

- O modelo de RLM tem dois aspectos importantes: aditividade e linearidade.
- aditividade significa que o efeito de mudanças em um preditor X_j sobre a resposta Y é independente dos valores dos demais preditores.
- linearidade significa que uma mudança em Y devida a uma mudança unitária em X_j é constante, independentemente do valor de X_j.
- Uma maneira de estender o modelo linear é incluir interações, por exemplo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e.$$

- Outra maneira: considerar regressão polinomial. Nesse caso, temos uma função não linear, mas o modelo continua linear!
- Para verificar se há necessidade de um modelo não linear, fazer o gráfico dos resíduos *versus* x_i, no caso de RLS e de resíduos *versus* $\hat{y}_i$, no de RLM.

Referências

Morettin, P. A. and Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.