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 4

27 de março de 2023

Sumário

Modelos de regressão linear

1. Um dos modelos estatísticos mais usados na prática: **modelo de regressão**.
2. Exemplo mais simples: dados $(x_1, y_1), \ldots, (x_n, y_n)$ de duas variáveis contínuas X e Y num contexto em que sabemos a priori que a distribuição de probabilidades de Y pode depender de X, ou seja, X é a variável explicativa (ou preditora) e Y é a variável resposta.
3. Trata-se de um **modelo linear**, ou seja, os parâmetros aparecem no modelo de forma linear.
4. Trata-se, também, de um **modelo paramétrico**.

Regressão linear simples

- **Exemplo**: objetivo é avaliar como a distância com que indivíduos conseguem distinguir um determinado objeto (doravante indicada simplesmente como distância) varia com a idade.
- Figura 1: gráfico de dispersão: tendência decrescente da distância com idade.
- modelo: $y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n$
- α, β **parâmetros**
- modelo: regressão linear simples – RLS
- mais adequado: $y_i = \alpha + \beta(x_i - 18) + e_i, \quad i = 1, \ldots, n.$

Regressão linear simples

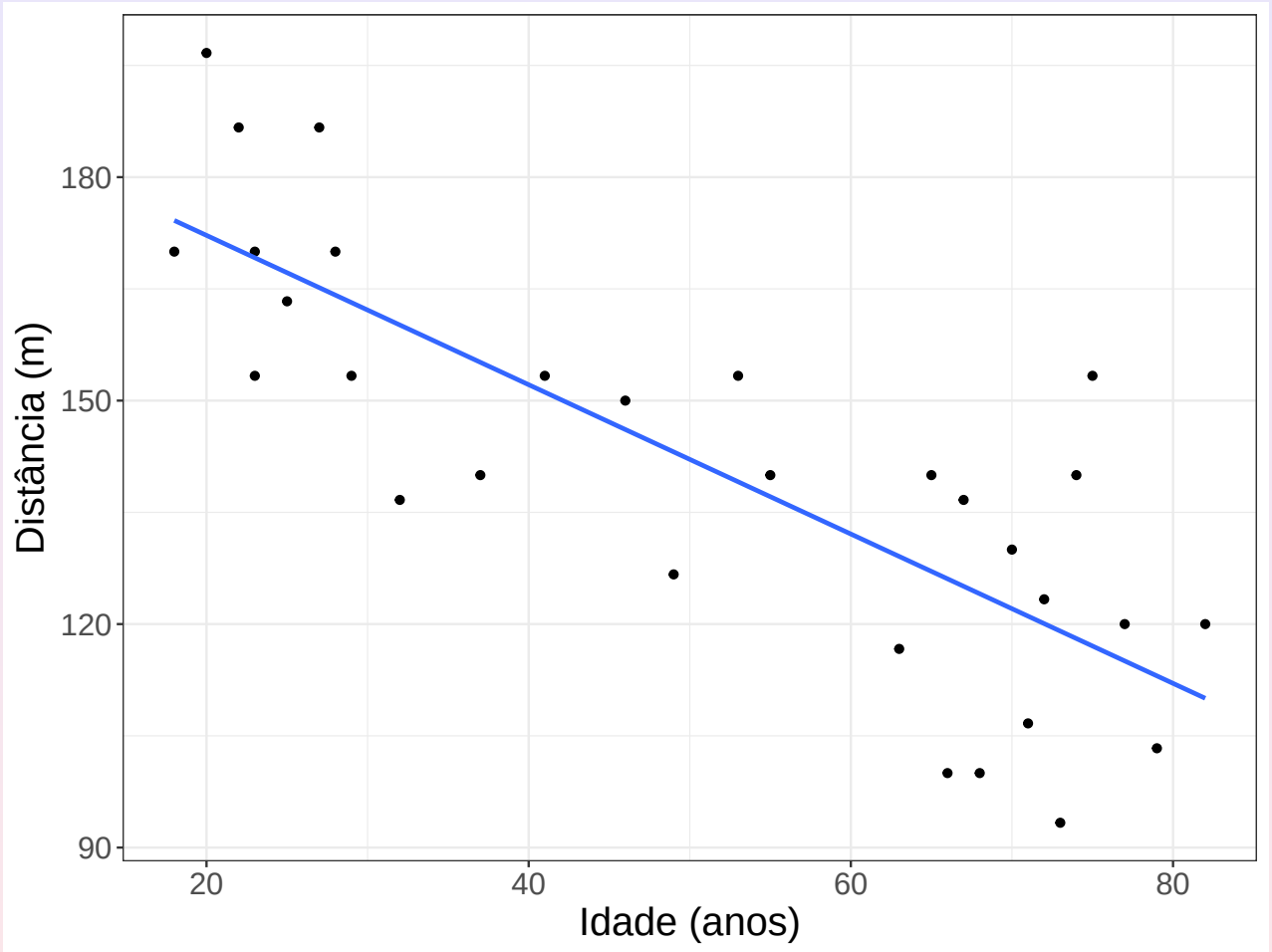

Figura 1: Gráfico de dispersão para os dados **distância**

Estimação do modelo RLS

- SQE: $Q(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$.
- Estimadores de mínimos quadrados (EMQ): minimizam a SQE.
- $$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{1}$$

  e

  $$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \tag{2}$$

  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 de σ^2 é

$$S^2 = \frac{1}{n-2} Q(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2, \tag{3}$$

em que onde $Q(\widehat{\alpha}, \widehat{\beta})$ é a **soma dos quadrados dos resíduos**, abreviadamente, *SQRes*.

Uso do R para o ajuste

- função lm() do pacote MASS
- modelo: $\text{distancia}_i = \alpha + \beta(\text{idade}_i - 18) + e_i, \quad i = 1, \ldots, n$
  $\hat{y}_i = 174.23 - 1.004(x_i - 18).$
- modelo: $\text{distancia}_i = \alpha + \beta\text{idade}_i + e_i, \quad i = 1, \ldots, n$
  $\hat{y}_i = 192.3 - 1.004x_i,$
- Residual standard error: 16.6 on 28 degrees of freedom
- Multiple R-squared: 0.6424, Adjusted R-squared: 0.6296

RLS-Avaliação do ajuste

- Uma vez ajustado o modelo, convém avaliar a qualidade do ajuste e um dos indicadores mais utilizados para essa finalidade é o **coeficiente de determinação** definido como

$$R^2 = \frac{SQTot - SQRes}{SQTot} = \frac{SQReg}{SQTot} = 1 - \frac{SQRes}{SQTot}$$

  em que a soma de quadrados total é $SQTot = \sum_{i=1}^{n}(y_i - \overline{y})^2$, a soma de quadrados dodos resíduos é $SQRes = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ e a soma de quadrados da regressão é $SQReg = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$.

- Esse coeficiente mede a porcentagem da variação total dos dados (em relação à sua média) explicada pelo modelo de regressão. O coeficiente de determinação deve ser acompanhado de outras ferramentas para a avaliação do ajuste, pois não está direcionado para identificar se todas as suposições do modelo são compatíveis com os dados sob investigação. Em particular, mencionamos os gráficos de resíduos, gráficos de Cook e gráficos de influência local.

R^2 ajustado

- Toda a vez que incluimos um preditor ao modelo, o R^2 aumenta; nunca decresce. Consequentemente, um modelo com mais parâmetros pode parecer que tenha o melhor ajuste (sobreajuste).
- O R^2 ajustado cresce somente quando o preditor incluído realmente aumenta o poder preditivo do modelo de regressão.
- O R^2 ajustado é calculado por

$$\overline{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right],$$

em que n é o tamanho da amostra e k o número de preditores do modelo.

RLS-Gráfico de resíduos

O gráfico de resíduos correspondente ao modelo ajustado aos dados **distancia** está apresentado na Figura 2.

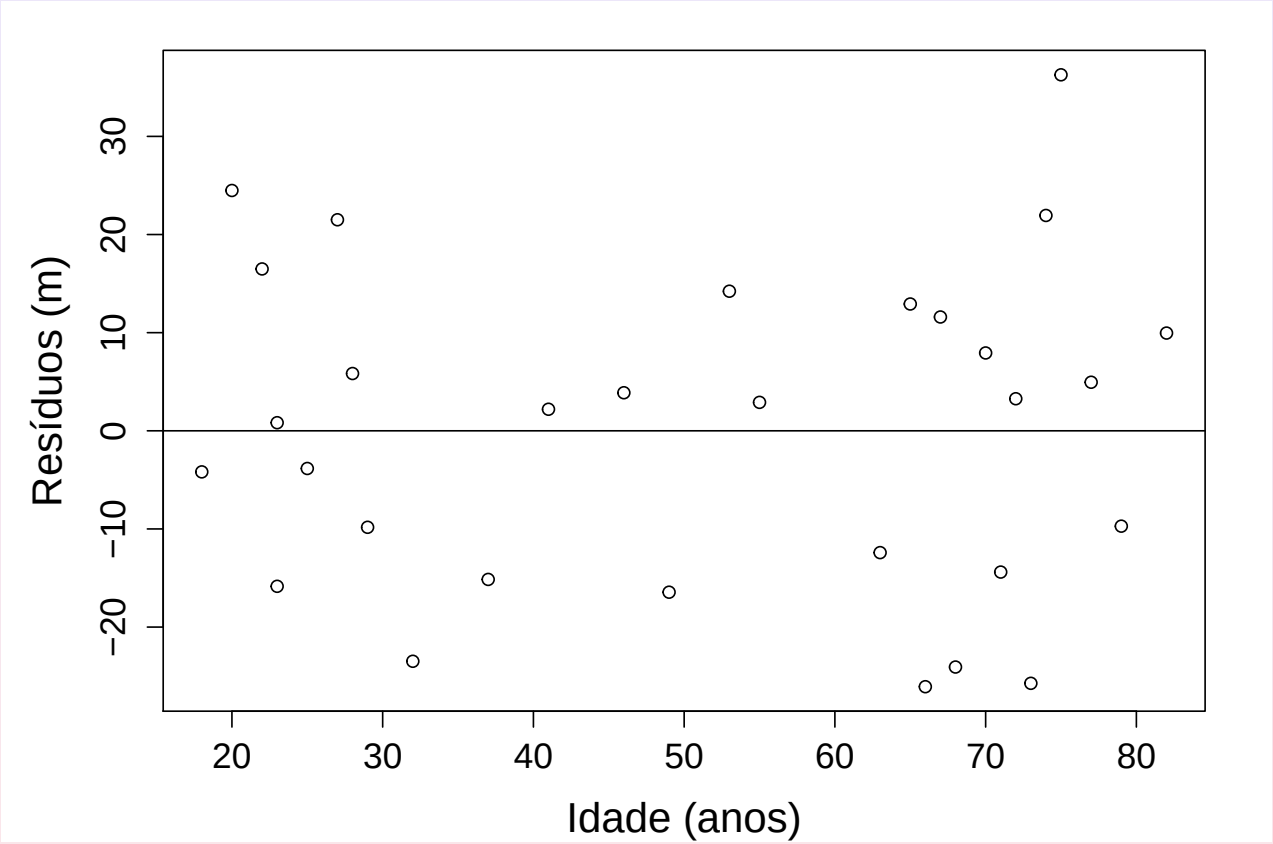

Figura 2: Gráfico de resíduos para o ajuste do modelo de regressão linear simples aos dados **distância**.

RLS-resíduos padronizados

- Para facilitar a visualização em relação à dispersão dos resíduos e para efeito de comparação entre ajustes de modelos em que as variáveis resposta têm unidades de medida diferentes, convém padronizá-los, *i.e.*, dividi-los pelo respectivo desvio padrão para que tenham variância igual a 1.
- Como os resíduos (ao contrário dos erros) são correlacionados, pode-se mostrar que

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

de forma que os **resíduos padronizados**, também chamados de **resíduos studentizados** são definidos por

$$\widehat{e}_i^* = \widehat{e}_i/(S\sqrt{1-h_{ii}}) \tag{4}$$

- Os resíduos padronizados são adimensionais e têm variância igual a 1, independentemente da variância da variável resposta. Além disso, para erros com distribuição Normal, cerca de 99% dos resíduos padronizados têm valor entre -3 e +3.
- h_{ii} : leverage (alavancagem) do i-ésimo preditor.

RLS- Resíduos padronizados

O gráfico de resíduos padronizados correspondente àquele da Figura 2 está apresentado na Figura 3.

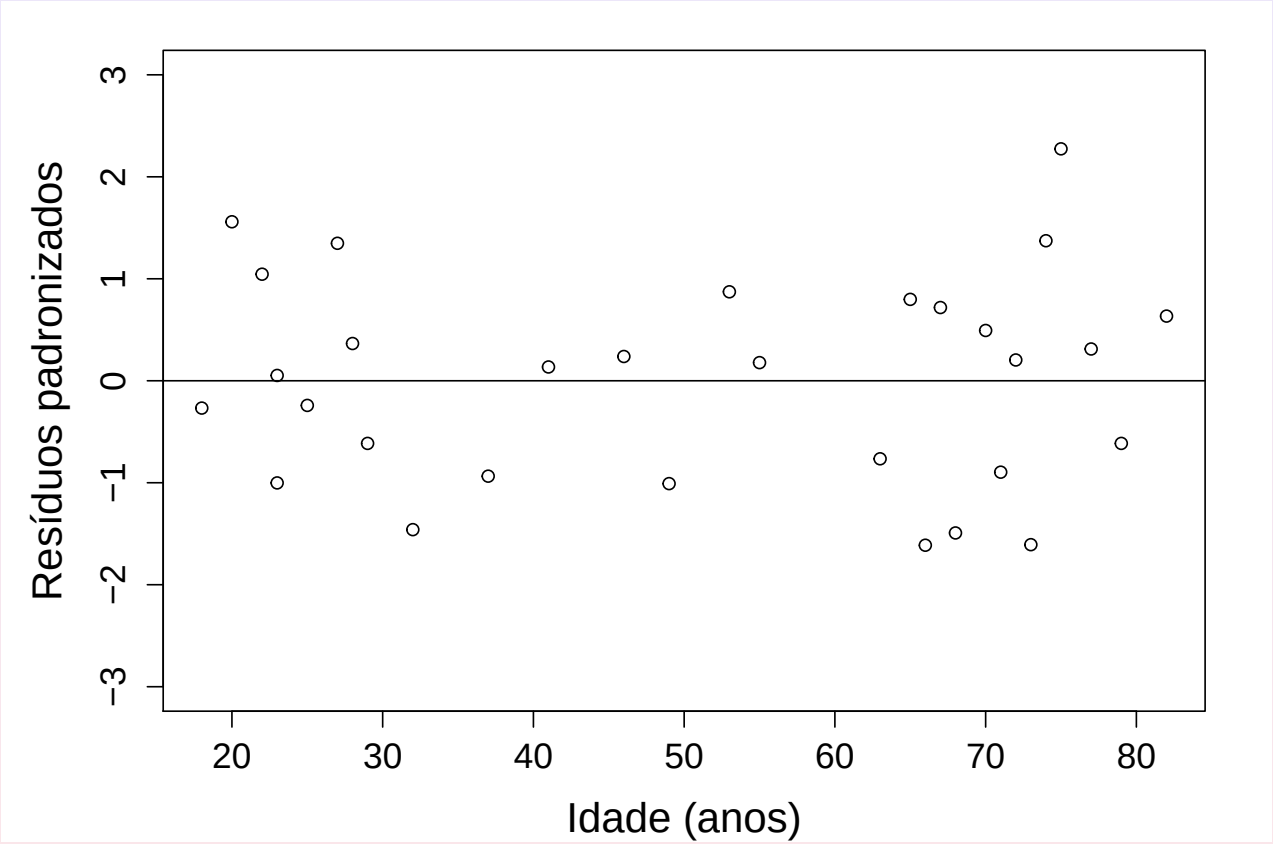

Figura 3: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados **distancia**.

RLS-Distância de Cook

Exemplo: Consideremos agora os dados (hipotéticos) dispostos na Tabela 1, aos quais ajustamos um modelo de regressão linear simples.

Tabela 1: 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

RLS-Distância de Cook

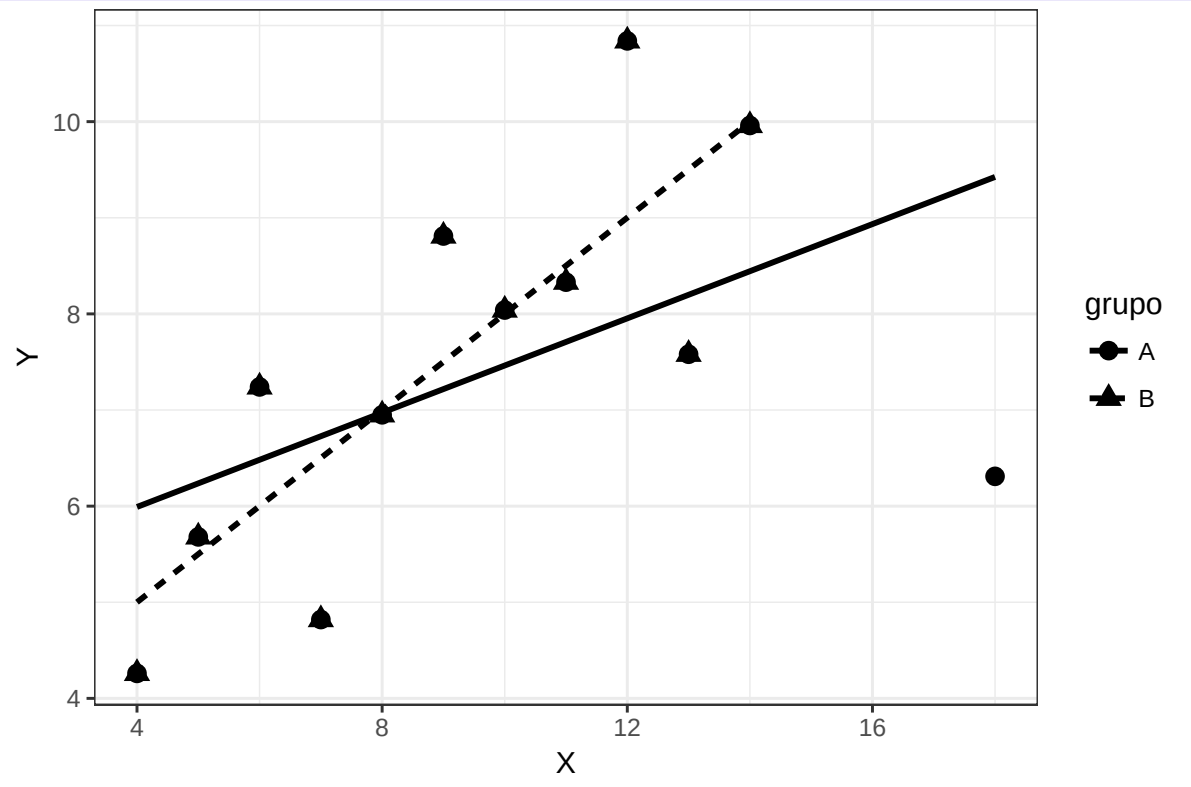

Figura 4: Gráfico de dispersão (com retas de regressão sobrepostas) para os dados da Tabela 1; curva sólida para dados completos e curva interrompida para dados com ponto influente eliminado.

RLS-Distância de Cook

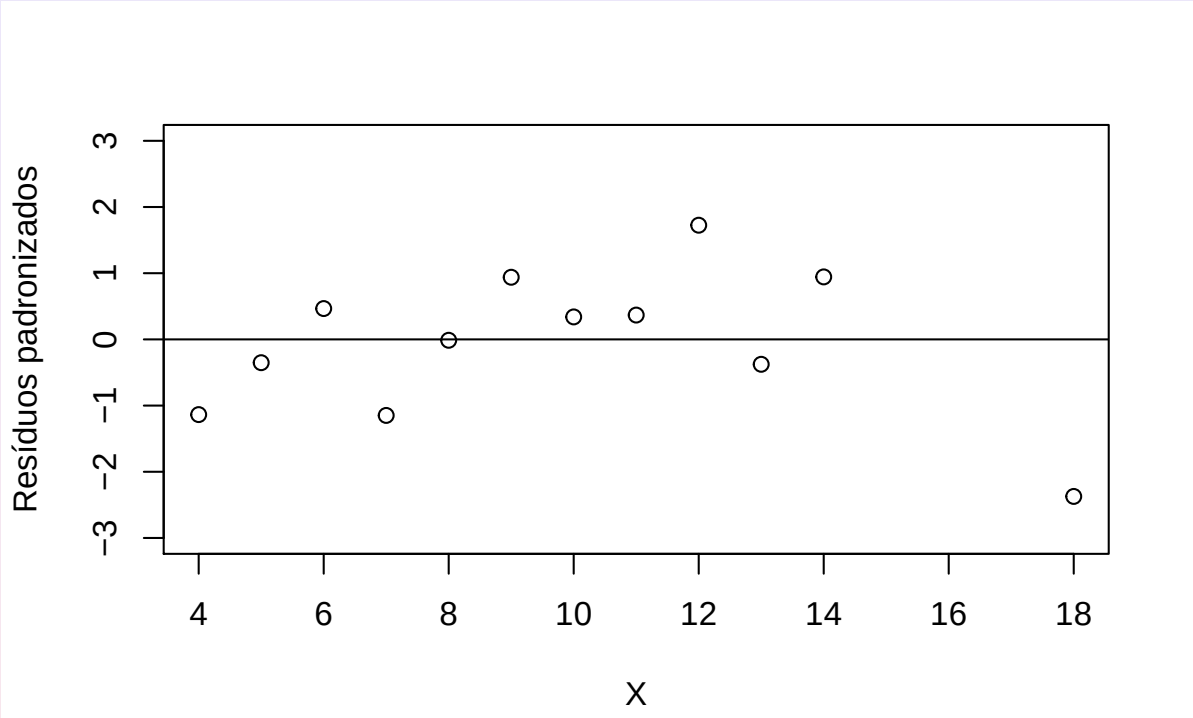

Figura 5: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados da Tabela 1.

RLS-Distância de Cook

- A distância de Cook é uma maneira de identificar pontos influentes (outliers) em um conjunto de preditores, que afetam o modelo. É uma combinação da alavancagem de cada observação e dos resíduos. Quanto maior a alavancagem, maior é a distância de Cook.
- Denotando por $\widehat{\mathbf{y}}$ o vetor (de dimensão n) com os valores preditos obtidos do ajuste do modelo baseado nas n observações e por $\widehat{\mathbf{y}}^{(-i)}$ o correspondente vetor com valores preditos (de dimensão n) obtido do ajuste do modelo baseado nas $n-1$ observações restantes após a eliminação da i-ésima, a **distância de Cook** é definida como

$$D_i = \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(-i)})^\top (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(-i)})}{(p+1)S}$$

em que p é o número de coeficientes de regressão e S é uma estimativa do desvio padrão.

- Pode-se mostrar que a distância de Cook (D_i) pode ser calculada sem a necessidade de ajustar o modelo com a omissão da i-ésima observação por meio da expressão

$$D_i = \frac{1}{p+1}\widehat{e}_i^2 \frac{h_{ii}}{(1-h_{ii})^2},$$

lembrando que h_{ii} é a leverage.

RLS-Distância de Cook

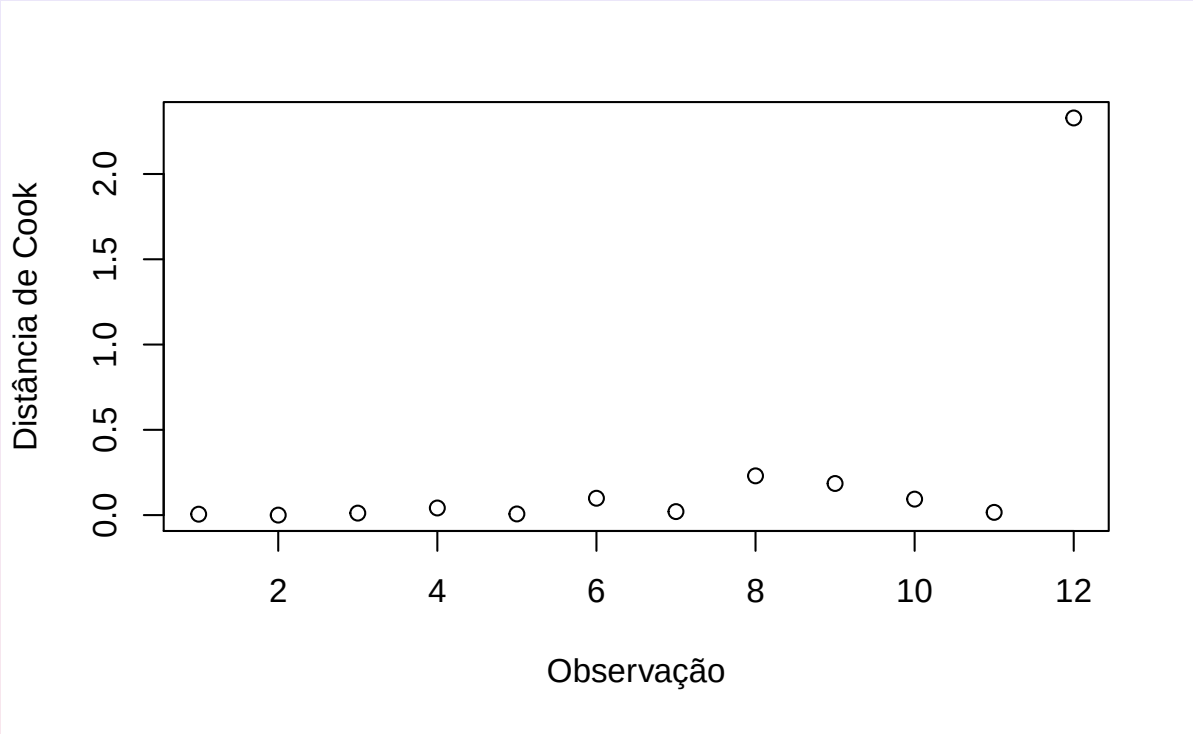

Figura 6: Gráfico de Cook correspondente ao ajuste do modelo de regressão linear aos dados da Tabela 1.

RLS-Gráficos QQ

- Nos casos em que se supõe que os erros têm distribuição Normal, pode-se utilizar gráficos QQ (quantis–quantis)com o objetivo de avaliar se os dados são compatíveis com essa suposição. É importante lembrar que esses gráficos QQ devem ser construídos com os quantis amostrais baseados nos resíduos e não com as observações da variável resposta, pois apesar de suas distribuições também serem normais, suas médias variam com os valores associados da variável explicativa, ou seja, a média da variável resposta correspondente a y_i é $\alpha + \beta x_i$.
- Convém observar que sob normalidade dos erros, os resíduos padronizados seguem uma distribuição t com $n - 2$ graus de liberdade e é dessa distribuição que se devem obter os quantis teóricos para a construção do gráfico QQ. Também deve-se lembrar que para valores de n maiores que 20 ou 30, os quantis da distribuição t se aproximam daqueles da distribuição Normal, tornando-as intercambiáveis para a construção do correspondente gráfico QQ.

RLS-Gráficos QQ

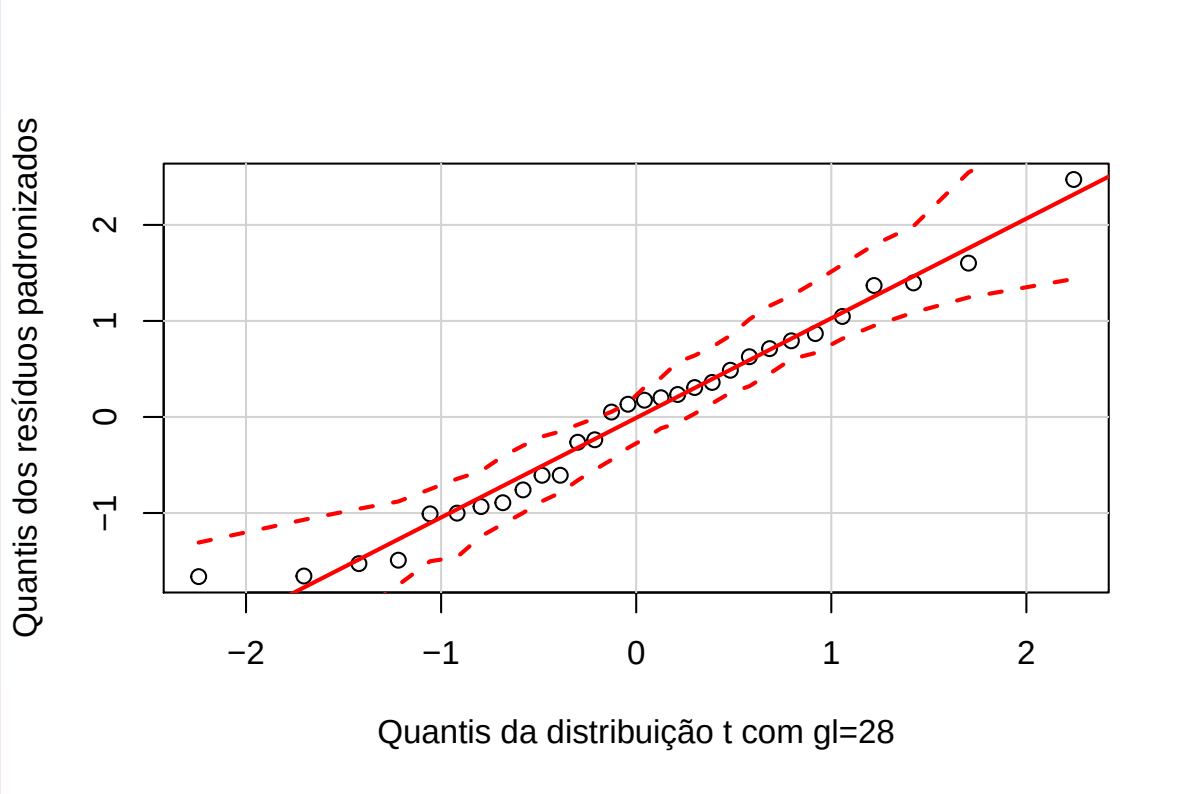

Figura 7: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados **distancia**.

RLS-Gráficos QQ

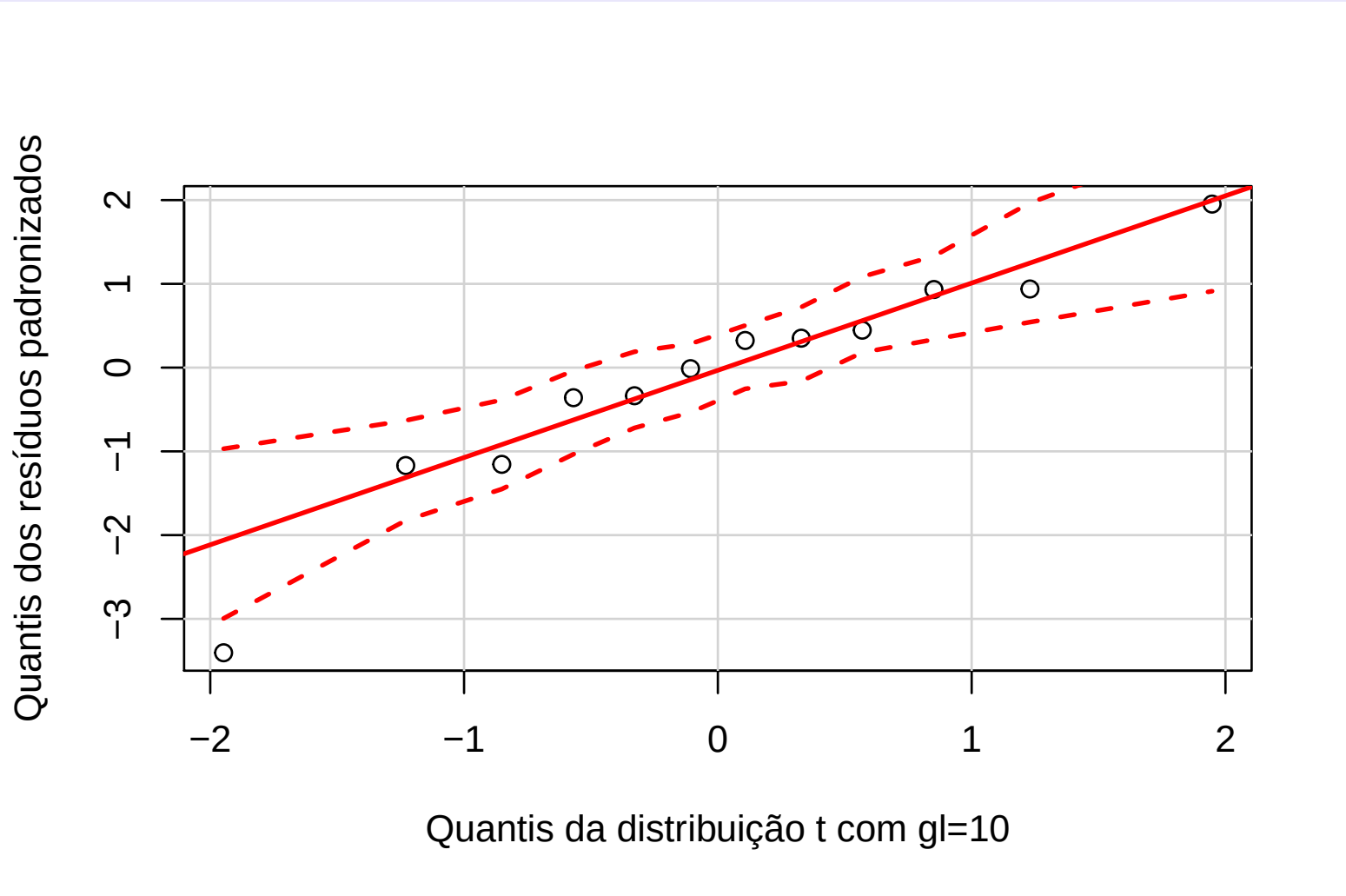

Figura 8: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 1 (com todas as observações).

RLS-Gráficos QQ

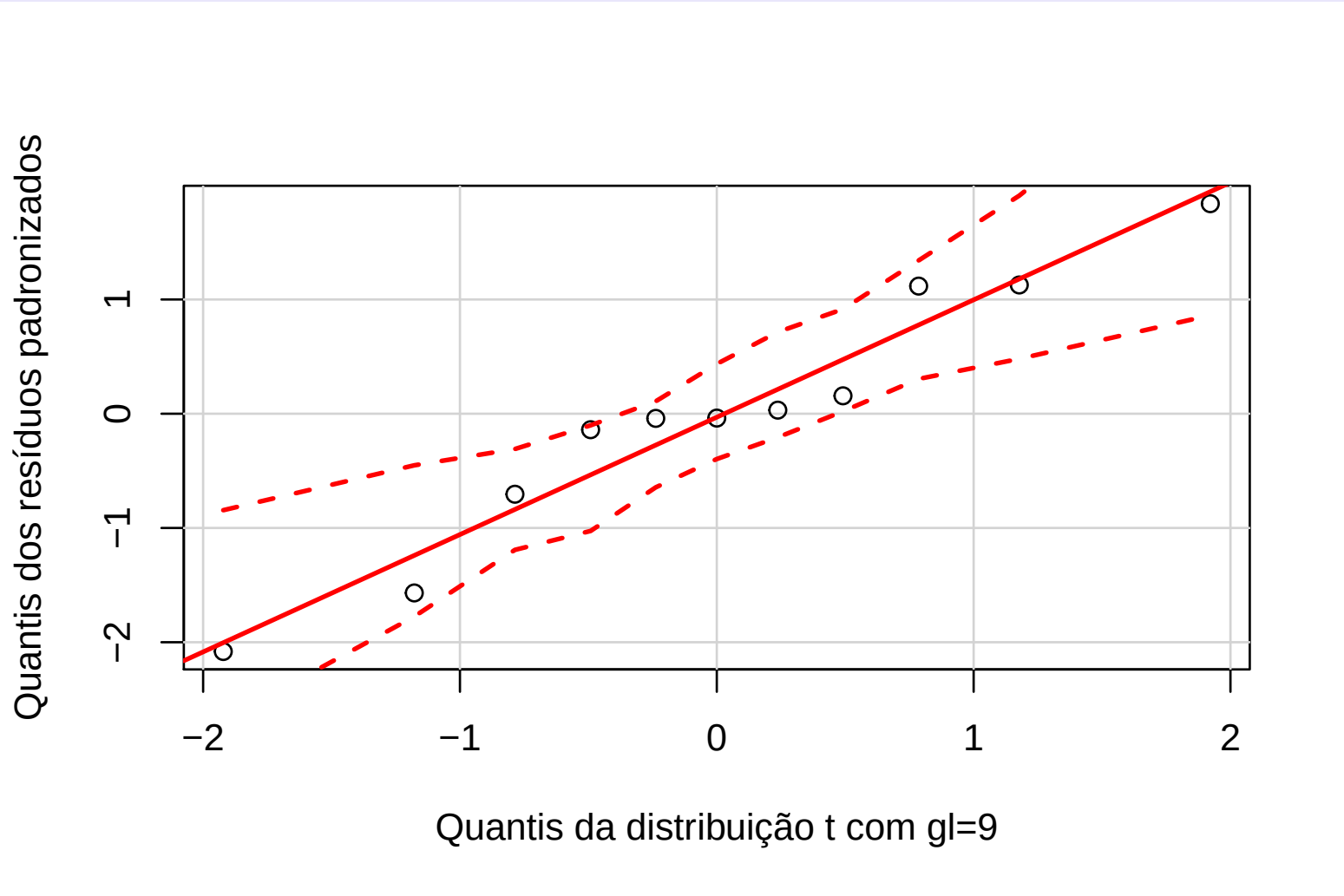

Figura 9: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 1 (sem a observação influente).

RLS-Dados correlacionados

Exemplo: Na Tabela 2 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 10.

Tabela 2: Peso (kg) de um bezerro nas primeiras 26 semanas após o nascimento

Semana	Peso
0	32,0
2	35,5
4	39,2
6	43,7
8	51,8
10	63,4
12	76,1
14	81,1
16	84,6
18	89,8
20	97,4
22	111,0
24	120,2
26	134,2

RLS-Dados correlacionados

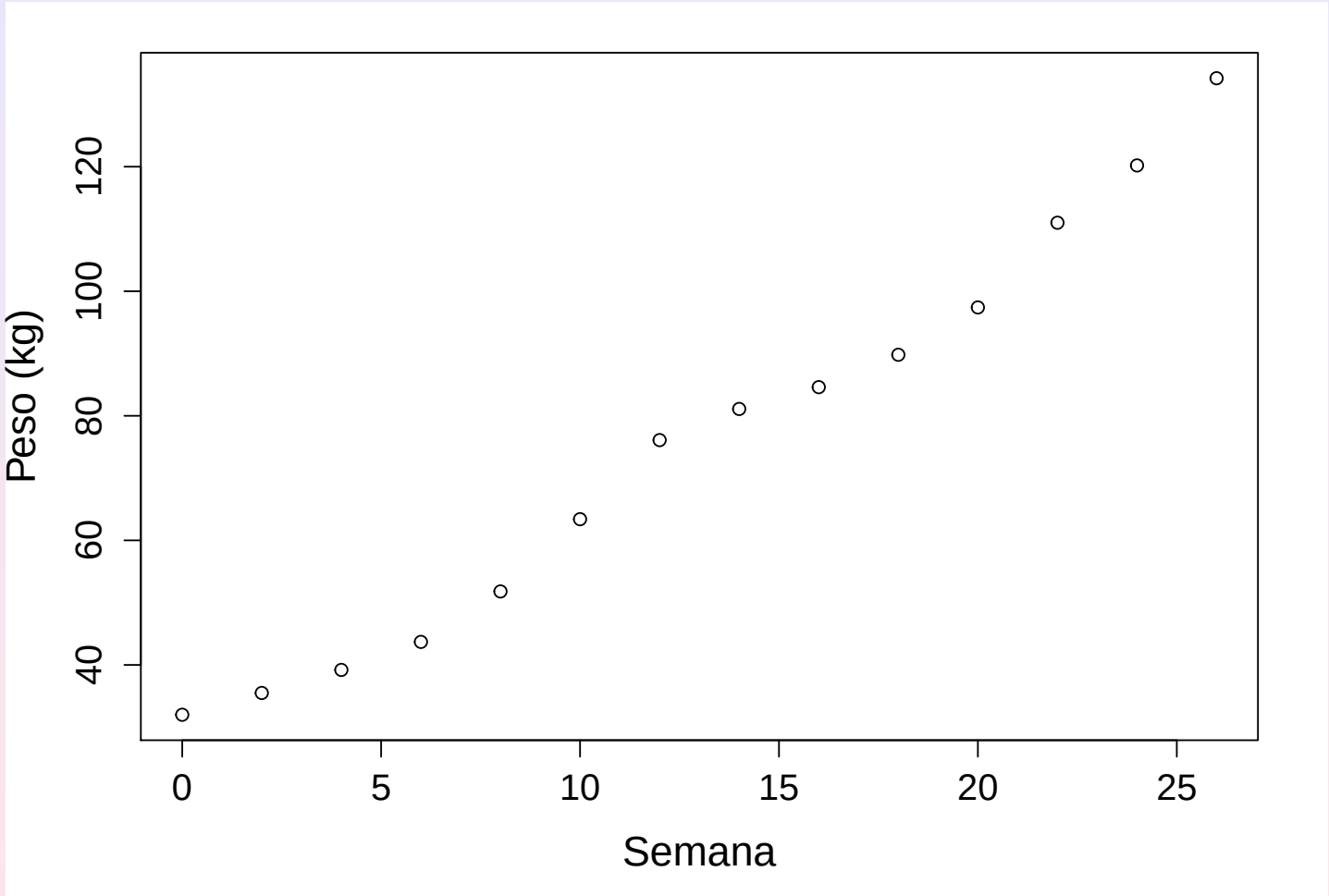

Figura 10: Gráfico de dispersão para os dados da Tabela 2

Dados correlacionados

- Tendo em vista o gráfico de dispersão, um possível modelo seria

$$y_t = \alpha + \beta t + \gamma t^2 + e_t, \tag{5}$$

  $i = 1, \ldots, 14$ em que y_t representa o peso do bezerro no instante t, α denota o valor esperado de seu peso ao nascer, β e γ representam os componentes linear e quadrático da curva que rege a variação temporal do peso no intervalo de tempo estudado e e_t denota um erro aleatório. 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 11 mostra sequências de resíduos positivos seguidas de sequências de resíduos negativos, sugerindo uma possível correlação positiva entre eles (autocorrelação).

RLS-Dados correlacionados

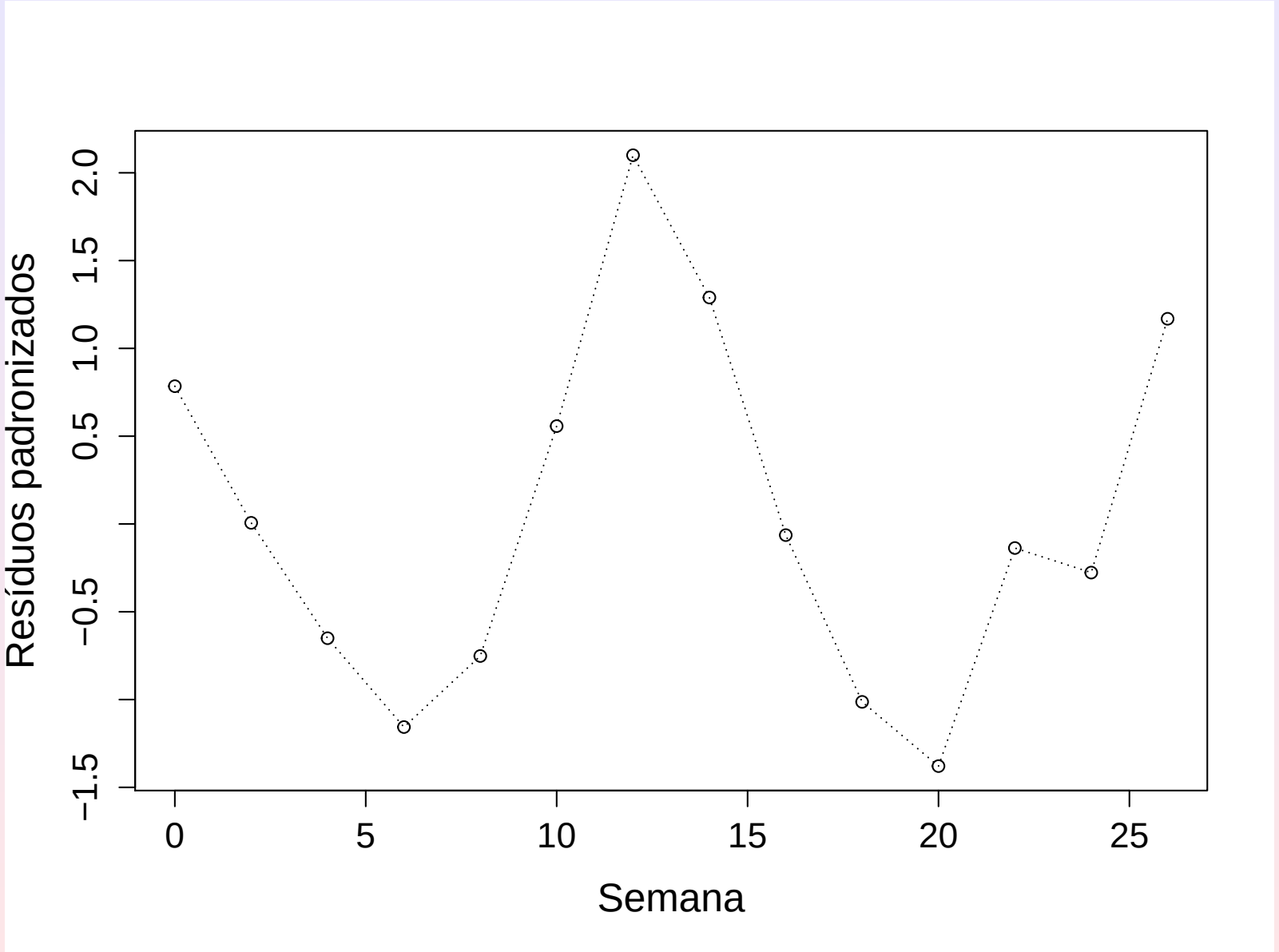

Figura 11: Resíduos studentizados obtidos do ajuste do modelo (5)

RLS-Dados correlacionados

- 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 (5) com

$$e_t = \rho e_{t-1} + u_t, \quad t = 1, \ldots, n \tag{6}$$

em que $u_t \sim N(0, \sigma^2)$, $t = 1, \ldots, n$, independentes e e_0 é uma constante (geralmente igual a zero). Essas suposições implicam que $\text{Var}(e_t) = \sigma^2/(1-\rho^2)$ e que $\text{Cov}(e_t, e_{t-s}) = \rho^s[\sigma^2/(1-\rho^2)]$.

- Para testar a hipótese de que os erros são não correlacionados pode-se utilizar a **estatística de Durbin-Watson**:

$$D = \sum_{t=2}^{n} (\widehat{e}_t - \widehat{e}_{t-1})^2 / \sum_{t=1}^{n} \widehat{e}_t^2, \tag{7}$$

em que $\widehat{e}_t$, $t = 1, \ldots, n$ são os resíduos obtidos do ajuste do modelo (5) por mínimos quadrados.

RLS-Dados correlacionados

- Expandindo (7) podemos verificar que

$$D \approx 2 - 2\frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=1}^{n} \widehat{e}_t^2}, \quad (8)$$

- Se os resíduos não forem correlacionados, então $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx 0$ e consequentemente, $D \approx 2$; se, por outro lado, os resíduos forem altamente correlacionados, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx \sum_{t=2}^{n} \widehat{e}_t^2$ e então $D \approx 0$; finalmente, se os resíduos tiverem uma grande correlação negativa, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx -\sum_{t=2}^{n} \widehat{e}_t^2$ e nesse caso, $D \approx 4$.
- Durbin and Watson (1950), Durbin and Watson (1951) e Durbin and Watson (1971) produziram tabelas da distribuição da estatística D que podem ser utilizados para avaliar a suposição de que os erros são não correlacionados.
- O valor da estatística de Durbin-Watson para os dados do Exemplo sob o modelo (5) 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 (5) - (6) poderá ser ajustado pelo **método de mínimos quadrados generalizados** ou por métodos de **Séries Temporais**.

RLS - Inferência

a) $E(\widehat{\alpha}) = \alpha$ e $E(\widehat{\beta}) = \beta$, ou seja, os EMQ 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) $\text{cov}(\widehat{\alpha}, \widehat{\beta}) = -\sigma^2 \overline{x} / \sum_{i=1}^{n} (x_i - \overline{x})^2$.

RLS - Inferência

Com a suposição adicional de normalidade, pode-se mostrar que

e) $y_i \sim N(\alpha + \beta x_i, \sigma^2)$

f) as estatísticas

$$t_{\widehat{\alpha}} = \frac{\widehat{\alpha} - \alpha}{S} \sqrt{\frac{n \sum (x_i - \overline{x})^2}{\sum x_i^2}}$$

e

$$t_{\widehat{\beta}} = \frac{\widehat{\beta} - \beta}{S} \sqrt{\sum (x_i - \overline{x})^2}$$

têm distribuição t de Student com $(n-2)$ graus de liberdade. Nesse contexto, os resíduos padronizados também seguem uma distribuição t de Student com $(n-2)$ graus de liberdade. Daí a denominação alternativa de resíduos studentizados

g) Com esses resultados é possível testar as hipóteses $H_0 : \alpha = 0$ e $H_0 : \beta = 0$, bem como construir intervalos de confiança para esses parâmetros.

RLS - Inferência

- **Teorema de Gauss-Markov**: EMQ têm variância mínima na classe dos estimadores não enviesados que sejam funções lineares das observações y_i (que não depende da suposição de normalidade dos erros).
- Quando os erros não seguem uma distribuição Normal, mas o tamanho da amostra é suficientemente grande, pode-se mostrar com o auxílio do **Teorema Limite Central** que sob certas condições de regularidade (usualmente satisfeitas na prática), os estimadores $\widehat{\alpha}$ e $\widehat{\beta}$ têm distribuições aproximadamente normais com variâncias que podem ser estimadas pelas expressões indicadas nos itens b) e c).

RLS - Previsão

- Um dos objetivos da análise de regressão é fazer previsões sobre a variável resposta com base em valores das variáveis explicativas.
- Uma estimativa para o valor esperado $E(Y|X = x_0)$ 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 anteriores pode-se mostrar que a variância de $\widehat{y}$ é

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

- 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

$$\widehat{y} \pm 1,96 S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

com S^2 denotando uma estimativa de σ^2. Podemos dizer que esse intervalo deve conter o verdadeiro valor esperado de $E(Y|X = x)$, *i.e.*, a média de Y para todas as observações em que $X = x_0$, com coeficiente de confiança de 95%.

RLS - Previsão

Isso não significa que esperamos que o intervalo contenha o verdadeiro valor de Y, digamos Y_0 para uma unidade de investigação para a qual $X = x_0$. Nesse caso precisamos levar em conta a variabilidade de $Y|X = x_0$ em torno de seu valor esperado $E(Y|X = x_0)$.
Como $Y_0 = \widehat{y} + e_0$ sua variância é

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

Então os limites superior e inferior de um **intervalo de previsão** (aproximado) para Y_0, com $\gamma = 95\%$, são

$$\widehat{y} \pm 1,96S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Note que se aumentarmos indefinidamente o tamanho da amostra, a amplitude do intervalo de confiança para o valor esperado tenderá para zero, porém a amplitude do intervalo de previsão correspondente a uma unidade específica tenderá para $2 \times 1,96 \times \sigma$.

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.

Apêndice: Quantis empíricos

- Suponha que a v.a. X tenha distribuição contínua, com f.d.a. F. Então, para $0 \leq p \leq 1$, o p - quantil de F é o valor Q_p satisfazendo $F(Q_p) = p$, ou seja,

$$F(Q_p) = P(X \leq Q_p) = p.$$

Se existir a inversa de F, então $Q_p = F^{-1}(p)$. No caso de X ser discreta, a definição tem que ser modificada: o p-quantil é o valor Q_p satisfazendo

$$\begin{aligned} P(X \leq Q_p) &\geq p, \\ P(X \geq Q_p) &\geq 1 - p. \end{aligned}$$

- Dado um conjunto de observações, podemos calcular os *quantis empíricos*. Uma maneira é considerar a *função de distribuição empírica* $\hat{F}_n$ como estimador de F, ou seja, dadas as observações $X_1, \ldots, X_n$ de X,

$$\hat{F}_n(x) = \frac{1}{n} \#\{i : \ 1 \leq i \leq n, \ X_i \leq x\}.$$

Então, o quantil Q_p é estimado pelo p-quantil de $\hat{F}_n$. Ou seja, o p-quantil estimado, q_p, seria definido por $\hat{F}_n(q_p) = p$. Contudo, usaremos um enfoque um pouco diferente.

Apêndice: Quantis empíricos

- Chamemos de $X_1, \ldots, X_T$ os valores observados e considere as estatísticas de ordem $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(T)}$. Um estimador consistente de Q_p é dado pelo p-quantil empírico, definido por

$$q_p = \begin{cases} X_{(i)}, & \text{se } p = p_i = (i - 0,5)/T,\ i = 1, \ldots, T \\ (1 - f_i)X_{(i)} + f_i X_{(i+1)}, & \text{se } p_i < p < p_{i+1} \\ X_{(1)}, & \text{se } 0 < p < p_1 \\ X_{(T)}, & \text{se } p_T < p < 1, \end{cases}$$

  onde $f_i = (p - p_i)/(p_{i+1} - p_i)$.

- Ou seja, ordenados os dados, q_p é uma das estatísticas de ordem, se p for da forma $p_i = (i - 0,5)/T$ e está na reta ligando os pontos $(p_i, X_{(i)})$ e $(p_{i+1}, X_{(i+1)})$, se p estiver entre p_i e p_{i+1}. Tomamos p_i da forma escolhida e não como i/T para que, por exemplo, a mediana calculada segundo esta definição coincida com a definição usual.

Apêndice: Quantis empíricos

Há dois tipos de gráficos $Q \times Q$: teóricos e empíricos.

- O primeiro tipo é usado para verificar se um conjunto de dados vem de determinada distribuição.
- O segundo tipo é usado para verificar se dois conjuntos de dados têm uma mesma distribuição.
- Para verificar se um conjunto de dados provém de uma distribuição especificada, consideramos o gráfico em que, no eixo horizontal, colocamos os quantis teóricos da distribuição hipotetizada para os dados, e no eixo vertical, os quantis empíricos dos dados, ambos calculados nos pontos p_i, acima. Se as observações realmente são provenientes da distribuição em questão, os pontos deverão estar distribuídos ao longo de uma reta.