Regressão Linear Múltipla

Gilberto A. Paula
Instituto de Matemática e Estatística - Universidade de São Paulo
e-mail: giapaula@ime.usp.br

Junho 2023

Resumo

O principal objetivo deste texto é apresentar uma síntese dos principais tópicos relacionados com regressão linear múltipla, tais como estimação por mínimos quadrados e máxima verossimilhança, procedimentos inferenciais e de teste de hipóteses, além de métodos de diagnóstico, conceito de interação, comparação de médias, regressão ponderada, multicolinearidade e seleção de modelos. Exemplos ilustrativos são apresentados ao longo do texto e vários exercícios teóricos e aplicados são propostos no final do texto. Uma abordagem mais completa pode ser encontrada, por exemplo, no livro de Montgomery, Peck e Vining (2021).

1 Introdução

Denote por $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ uma amostra aleatória de tamanho n de uma determinada população, em que $y_1, \ldots, y_n$ representam os valores observados da variável resposta (assumida contínua), enquanto $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ denotam valores observados de variáveis explicativas, para $i = 1, \ldots, n$. O principal objetivo da regressão linear múltipla é tentar explicar o valor esperado da variável resposta dados os valores das variáveis explicativas. A formulação mais usual é a seguinte:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. Muitas vezes tem-se um intercepto em (1), sendo nesse caso assumido que $x_{i1} = 1 \; \forall i$.

A suposição de normalidade para os erros pode ser relaxada para amostras grandes, contudo para amostras pequenas e moderadas essa suposição

é crucial para fazer inferência. De (1) segue que $Y_i|\mathbf{x}_i \overset{\text{ind}}{\sim} \text{N}(\mu_i, \sigma^2)$ com $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, para $i = 1, \ldots, n$.

Em forma matricial o modelo (1) fica expresso na forma

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X}$ é a matriz modelo de dimensão $n \times p$ dada por

$$\mathbf{X} = \begin{bmatrix} x_{i1} & x_{12} & ... & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix},$$

$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ com $\boldsymbol{\epsilon} \sim \text{N}_n(0, \sigma^2\mathbf{I}_n)$ e $\mathbf{I}_n$ a matriz identidade de ordem n.

2 Solução de Mínimos Quadrados

A estimativa de mínimos quadrados de $\boldsymbol{\beta}$ é obtida minimizando a função objetivo $S(\boldsymbol{\beta})$ que corresponde a minimizar a soma dos quadrados dos erros

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

A derivada parcial de $S(\boldsymbol{\beta})$ com relação a β_j fica dada por

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}),$$

para $j = 1, \ldots, p$. Assim, a derivada de $S(\boldsymbol{\beta})$ com relação a $\boldsymbol{\beta}$ é um vetor de dimensão $p \times 1$ expresso na forma

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

A estimativa de mínimos quadrados $\widehat{\boldsymbol{\beta}}$ é obtida igualando-se a primeira derivada a zero

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \Rightarrow -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Se $\mathbf{X}$ é uma matriz de posto coluna completo então tem-se uma solução única

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

Na Figura 1 é apresentada uma representação geométrica da solução de mínimos quadrados, em que $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$ corresponde à projeção ortogonal de $\mathbf{y}$ através do projetor linear $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, no subespaço gerado pelas colunas da matriz $\mathbf{X}$, denotado por $\mathrm{C}(\mathbf{X})$. Por outro lado, $r = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ definido como vetor de resíduos ordinários, corresponde à projeção ortogonal de $\mathbf{y}$ através do projetor linear $(\mathbf{I}_n - \mathbf{H})$, no subespaço complementar $\mathrm{C}^c(\mathbf{X})$, denominado ortocomplemento de $\mathrm{C}(\mathbf{X})$.

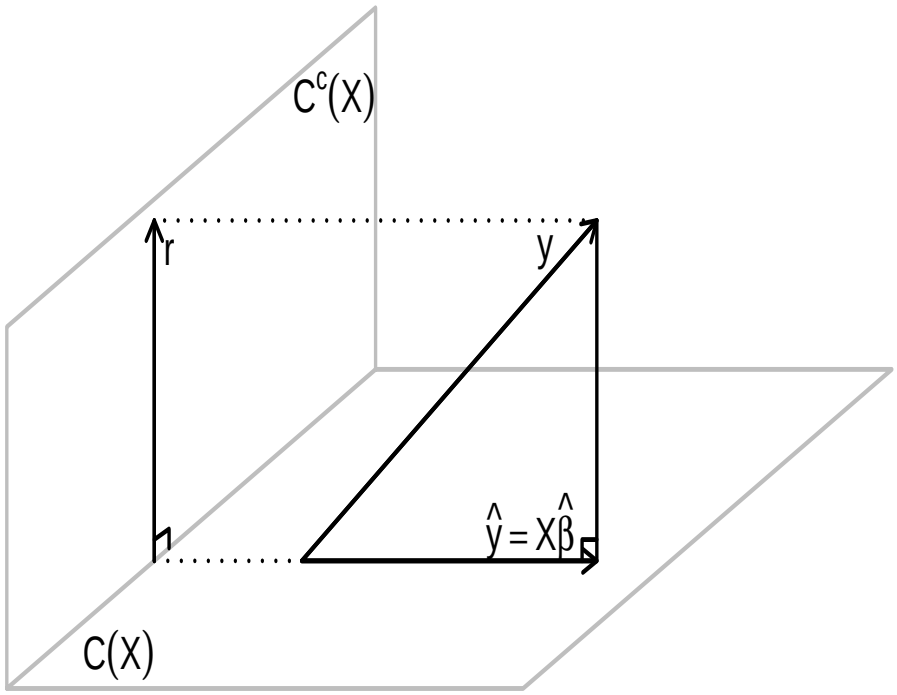

Figura 1: Representação geométrica da solução de mínimos quadrados referente ao modelo de regressão linear múltipla (2), em que $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$ é o resíduo ordinário e $\mathrm{C}(\mathbf{X})$ denota o subespaço gerado pelas colunas da matriz $\mathbf{X}$ e $\mathrm{C}^c(\mathbf{X})$ o ortocomplemento.

É preciso verificar se a raiz da primeira derivada é de fato um ponto de mínimo da superfície formada por $(S(\boldsymbol{\beta}), \boldsymbol{\beta}^\top)^\top$. Deriva-se então novamente $S(\boldsymbol{\beta})$ com relação a β_ℓ, obtendo-se

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_\ell} = 2\sum_{i=1}^{n} x_{ij}x_{i\ell},$$

para $j, \ell = 1, \ldots, p$. Assim, a matriz de segundas derivadas de $S(\boldsymbol{\beta})$ com relação a $\boldsymbol{\beta}$ tem dimensão $p \times p$ e fica expressa na forma

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = 2\mathbf{X}^\top\mathbf{X}.$$

Como é assumido que $\mathbf{X}$ tem posto coluna completo então $\mathbf{X}^\top\mathbf{X}$ é uma matriz positiva definida, logo $S(\boldsymbol{\beta})$ é uma superfície convexa e $\widehat{\boldsymbol{\beta}}$ é ponto de mínimo.

Resumindo, tem-se que $\mathbf{Y}|\mathbf{X} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ e como consequências $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ e $\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}_n$, em que $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$. Seguem as seguintes propriedades do estimador de mínimos quadrados:

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \mathrm{E}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}(\mathbf{Y}|\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Ou seja, $\widehat{\boldsymbol{\beta}}$ é um estimador não tendencioso de $\boldsymbol{\beta}$. A matriz de variância-covariância de $\widehat{\boldsymbol{\beta}}$ fica dada por

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= \mathrm{Var}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{Var}(\mathbf{Y}|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

Logo, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ e conforme mostrado em Montgomery et al. (2021, Apêndice C.4) $\widehat{\boldsymbol{\beta}}$ tem a menor variância entre todos os estimadores lineares não viesados de $\boldsymbol{\beta}$.

Pelo Teorema de Pitágoras aplicado ao triângulo retângulo da Figura 1, tem-se que

$$\begin{aligned}|| \mathbf{y} ||^2 &= || \hat{\mathbf{y}} ||^2 + || \mathbf{y} - \hat{\mathbf{y}} ||^2 \\ \sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,\end{aligned}$$

em que $|| \mathbf{v} || = \sqrt{v_1^2 + \cdots + v_n^2}$ denota norma ou comprimento do vetor $\mathbf{v} = (v_1, \ldots, v_n)^\top$. Se o modelo tem intercepto segue da solução de mínimos

quadrados $\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ que $\sum_{i=1}^n r_i = 0$. Logo, obtém-se a decomposição de somas de quadrados

$$\text{SQT} = \text{SQReg} + \text{SQRes},$$

em que $\text{SQT} = \sum_{i=1}^n (y_i - \bar{y})^2$ é a soma de quadrados total, $\text{SQReg} = \sum_{i=1}^n (y_i - \bar{y})^2$ é a soma de quadrados devido à regressão, enquanto $\text{SQRes} = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$ é a soma de quadrados de resíduos. Uma maneira de avaliar a qualidade do ajuste é comparar SQReg com SQT através do coeficiente de determinação

$$\text{R}^2 = \frac{\text{SQReg}}{\text{SQT}} = 1 - \frac{\text{SQRes}}{\text{SQT}},$$

em que $0 \leq \text{R}^2 \leq 1$. Quanto mais próximo R^2 está de 1 melhor a qualidade do ajuste. Contudo, como o coeficiente de determinação cresce à medida que o número p de parâmetros aumenta, recomenda-se a utilização do coeficiente de determinação ajustado

$$\overline{\text{R}}^2 = 1 - \frac{\text{QMRes}}{\text{QMT}},$$

em que $\text{QMRes} = \frac{\text{SQRes}}{n-p}$ e $\text{QMT} = \frac{\text{SQT}}{p-1}$ e $0 \leq \overline{\text{R}}^2 \leq 1$. É possível estabelecer a seguinte relação:

$$\overline{\text{R}}^2 = 1 - (1 - \text{R}^2)\frac{(n-1)}{(n-p)}.$$

Portanto, segue que $\overline{\text{R}}^2 \leq \text{R}^2$.

2.1 Regressão Linear Simples

Considere agora o modelo de regressão linear simples definido por

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, $x_1, \ldots, x_n$ são valores observados da variável explicativa X e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. A matriz modelo de dimensão $n \times 2$ fica dada por

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Assim, obtém-se

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix} \quad \text{e} \quad \mathbf{X}^\top\mathbf{y} = (n\bar{y}, \sum x_i y_i)^\top.$$

em que $\bar{x} = \frac{\sum x_i}{n}$ e $\bar{y} = \frac{\sum y_i}{n}$. Logo,

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{1}{n\mathrm{S}_{xx}} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix},$$

em que $\mathrm{S}_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$. O estimador de mínimos quadrados fica dado por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \bar{y} - \widehat{\beta}_2\bar{x} \\ \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}} \end{bmatrix}$$

com $\mathrm{S}_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. A matriz de variância-covariância assume a forma

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{\sigma^2}{n\mathrm{S}_{xx}} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}.$$

Daí segue que $\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n\mathrm{S}_{xx}}$, $\mathrm{Var}(\widehat{\beta}_2) = \frac{\sigma^2}{\mathrm{S}_{xx}}$ e $\mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = -\frac{\sigma^2\bar{x}}{\mathrm{S}_{xx}}$.

Supondo que X é uma variável quantitativa contínua, o coeficiente de correlação linear amostral de Pearson entre X e Y é expresso na forma

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2\}^{\frac{1}{2}}},$$

em que $-1 \leq r \leq 1$. Aternativamente, tem-se que

$$r_{xy} = \frac{\mathrm{S}_{xy}}{\{\mathrm{S}_{xx}\mathrm{SQT}\}^{\frac{1}{2}}} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}}\sqrt{\frac{\mathrm{S}_{xx}}{\mathrm{SQT}}} = \widehat{\beta}_2\sqrt{\frac{\mathrm{S}_{xx}}{\mathrm{SQT}}}.$$

Por outro lado, obtém-se

$$\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i = (\bar{y} - \bar{x}\widehat{\beta}_2) + \widehat{\beta}_2 x_i = \bar{y} + (x_i - \bar{x})\widehat{\beta}_2.$$

Logo $(\widehat{y}_i - \bar{y}) = (x_i - \bar{x})\widehat{\beta}_2$ e portanto $\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2 = \widehat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2$. Então, segue que $\mathrm{SQReg} = \widehat{\beta}_2^2\mathrm{S}_{xx}$. E desde que $\widehat{\beta}_2 = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}}$ obtém-se

$$\mathrm{SQReg} = \widehat{\beta}_2\mathrm{S}_{xy} \rightarrow \widehat{\beta}_2 = \frac{\mathrm{SQReg}}{\mathrm{S}_{xy}}.$$

Finalmente, segue a relação

$$r_{xy}^2 = \widehat{\beta}_2^2 \frac{\mathrm{S}_{xx}}{\mathrm{SQT}} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}} \frac{\mathrm{SQReg}}{\mathrm{S}_{xy}} \frac{\mathrm{S}_{xx}}{\mathrm{SQT}} = \frac{\mathrm{SQReg}}{\mathrm{SQT}} = \mathrm{R}^2.$$

Ou seja, o coeficiente de determinação R^2 coincide com o quadrado do coeficiente de correlação linear amostral de Pearson entre X e Y na regressão linear simples.

3 Teste de Hipóteses

Inicialmente, supor que o interesse é avaliar se os coeficientes da regressão são nulos, que corresponde a testar as hipóteses

$$\mathrm{H}_0 : \beta_2 = \cdots = \beta_p = 0 \;\; \text{contra} \;\; \mathrm{H}_1 : \beta_j \neq 0,$$

para pelo menos algum $j = 2, \ldots, p$. A estatística F fica expressa na forma

$$\mathrm{F} = \frac{\mathrm{SQReg}/(p-1)}{\mathrm{SQRes}/(n-p)} = \frac{\mathrm{QMReg}}{\mathrm{QMRes}} \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{(p-1),(n-p)}.$$

Para um nível de significância $0 < \alpha < 1$, rejeita-se H_0 se $\mathrm{F} > \mathrm{F}_{(1-\alpha),(p-1),(n-p)}$, em que $\mathrm{F}_{(1-\alpha),(p-1),(n-p)}$ denota o quantil $(1-\alpha)$ da distribuição F com $(p-1)$ e $(n-p)$ graus de liberdade. É usual construir a tabela de análise de variância (ANOVA), conforme descrito na Tabela 1.

Tabela 1: Descrição da tabela de Análise de Variância (ANOVA).

F. Variação	S.Quadrados	G.L.	Q. Médio	F
Regressão	SQReg	$p-1$	QMReg	$\frac{\mathrm{QMReg}}{\mathrm{QMRes}}$
Resíduos	SQRes	$n-p$	QMRes	
Total	SQT	$n-1$		

Denote $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{C}$, em que $\mathbf{C} = (\mathbf{X}^\top\mathbf{X})^{-1}$. Então, pode-se expressar as variâncias e covariâncias dos estimadores $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ nas formas $\mathrm{Var}(\widehat{\beta}_j) = \sigma^2\mathrm{C}_{jj}$ e $\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_\ell) = \sigma^2\mathrm{C}_{j\ell}$, em que $\mathrm{C}_{j\ell}$ denota o elemento (j,ℓ) da matriz $\mathbf{C}$, para $j, \ell = 1, \ldots, p$. Supor então que o interesse é testar as hipóteses $\mathrm{H}_0 : \beta_j = 0$ contra $\mathrm{H}_1 : \beta_j \neq 0$, para algum $j = 1, \ldots, p$. A estatística t-Student fica expressa na forma

$$\mathrm{t} = \frac{\widehat{\beta}_j}{\widehat{\mathrm{EP}}(\widehat{\beta}_j)} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{(n-p)},$$

em que $\widehat{\mathrm{EP}}(\widehat{\beta}_j) = s\sqrt{\mathrm{C}_{jj}}$. Para um nível de significância $0 < \alpha < 1$, rejeita-se H_0 se $|\mathrm{t}| > \mathrm{t}_{(1-\alpha/2),(n-p)}$, em que $\mathrm{t}_{(1-\alpha/2),(n-p)}$ denota o quantil $(1-\alpha/2)$ de uma distribuição t-Student com $(n-p)$ graus de liberdade. Em particular, pode-se mostrar que t^2 segue sob H_0 distribuição $\mathrm{F}_{1,(n-p)}$.

Generalizando, supor que o interesse agora é testar $\mathrm{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{0}$ contra contra $\mathrm{H}_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{0}$, com pelo menos uma desigualdade estrita em H_1, em que $\mathbf{R}$ é uma matriz $r \times p$ com posto linha $r \leq p$. O acréscimo na soma de quadrados de resíduos devido à restrição $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$ (vide Montgomery et al., 2021, Cap. 3) é dado por

$$\mathrm{ASQ}(\mathbf{R}\boldsymbol{\beta} = \mathbf{0}) = (\mathbf{R}\widehat{\boldsymbol{\beta}})^\top \{\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\}^{-1}\mathbf{R}\widehat{\boldsymbol{\beta}}.$$

Portanto, tem-se que

$$\mathrm{F} = \frac{\mathrm{ASQ}(\mathbf{R}\boldsymbol{\beta} = \mathbf{0})/r}{\mathrm{SQRes}/(n-p)} \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{r,(n-p)}.$$

Logo, para um nível de significância $0 < \alpha < 1$, rejeita-se H_0 se $\mathrm{F} > \mathrm{F}_{(1-\alpha),r,(n-p)}$.

Um caso particular é considerar a regressão linear múltipla (2) com efeitos particionados

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{3}$$

em que $\mathbf{X}_1$ e $\mathbf{X}_2$ são matrizes de dimensões $n \times p_1$ e $n \times p_2$, respectivamente, enquanto $\boldsymbol{\beta}_1$ tem dimensão $p_1 \times 1$ e $\boldsymbol{\beta}_2$ tem dimensão $p_2 \times 1$. Logo, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ e $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. Supor que o interesse seja testar $\mathrm{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ contra $\mathrm{H}_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$, com pelo menos uma desigualdade estrita em H_1. A soma de quadrados de resíduos correspondente ao modelo (3) com p parâmetros será denotada por $\mathrm{SQRes}(\boldsymbol{\beta}) = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}$, enquanto que a soma de quadrados de resíduos sob o modelo em H_0 com p_1 parâmetros será denotada por $\mathrm{SQRes}(\boldsymbol{\beta}|\boldsymbol{\beta}_2 = \mathbf{0}) = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}$, em que $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top$. Note que $\mathrm{SQRes}(\boldsymbol{\beta}|\boldsymbol{\beta}_2 = \mathbf{0}) \geq \mathrm{SQRes}(\boldsymbol{\beta})$. Assim, o acréscimo na soma de quadrados de resíduos devido à restrição $\boldsymbol{\beta}_2 = \mathbf{0}$ pode ser expresso na forma

$$\mathrm{ASQ}(\boldsymbol{\beta}_2 = \mathbf{0}) = \mathrm{SQRes}(\boldsymbol{\beta}|\boldsymbol{\beta}_2 = \mathbf{0}) - \mathrm{SQRes}(\boldsymbol{\beta}) = \mathbf{y}^\top(\mathbf{H}_1 - \mathbf{H})\mathbf{y},$$

e consequentemente a estatística F para testar $\mathrm{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ contra $\mathrm{H}_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$ fica dada por

$$\mathrm{F} = \frac{\mathbf{y}^\top(\mathbf{H}_1 - \mathbf{H})\mathbf{y}/p_2}{\mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}/(n-p)} \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{p_2,(n-p)}.$$

Logo, para um nível de significância $0 < \alpha < 1$, rejeita-se H_0 se $\mathrm{F} > \mathrm{F}_{(1-\alpha),p_2,(n-p)}$.

4 Estimativa Intervalar

Um estimativa intervalar de coeficiente de confiança $(1-\alpha)$ para β_j fica dada por

$$[\widehat{\beta}_j \pm \mathrm{t}_{(1-\alpha/2),(n-p)}\widehat{\mathrm{EP}}(\widehat{\beta}_j)],$$

em que $j = 1, \ldots, p$. Como para n grande a t-Student se aproxima da normal, pode-se usar o quantil $(1-\alpha/2)$ da $\mathrm{N}(0,1)$ no lugar de $\mathrm{t}_{(1-\alpha/2),(n-p)}$.

É possível mostrar que

$$\frac{\mathrm{SQRes}}{\sigma^2} \overset{\text{modelo}}{\sim} \chi^2_{(n-p)}.$$

Logo, segue que $\mathrm{E}\left(\frac{\mathrm{SQRes}}{\sigma^2}\right) = (n-p)$ e portanto $s^2 = \frac{\mathrm{SQRes}}{(n-p)}$ é um estimador não tendencioso de σ^2. Após algumas manipulações com a distribuição $\chi^2_{(n-p)}$ tem-se que

$$P\left\{\frac{(n-p)s^2}{\chi^2_{(1-\alpha/2),(n-p)}} \leq \sigma^2 \leq \frac{(n-p)s^2}{\chi^2_{(\alpha/2),(n-p)}}\right\} = (1-\alpha),$$

em que $\chi^2_{(\alpha/2),(n-p)}$ e $\chi^2_{(1-\alpha/2),(n-p)}$ denotam, respectivamente, os quantis $\alpha/2$ e $(1-\alpha/2)$ da distribuição $\chi^2_{(n-p)}$. Assim, uma estimativa intervalar de coeficiente de confiança $(1-\alpha)$ para σ^2 fica dada por

$$\left[\frac{(n-p)s^2}{\chi^2_{(1-\alpha/2),(n-p)}}; \frac{(n-p)s^2}{\chi^2_{(\alpha/2),(n-p)}}\right].$$

5 Bandas de Confiança

Supor uma nova observação que não pertence à amostra com valores para as variáveis explicativas representados por $\mathbf{z} = (z_1, z_2, \ldots, z_p)^\top$. Portanto, tem-se que

$$y(\mathbf{z}) = \mathbf{z}^\top\boldsymbol{\beta} + \epsilon(\mathbf{z})$$

e valor esperado $\mathrm{E}\{Y(\mathbf{z})\} = \mu(\mathbf{z})$. Logo $\widehat{\mu}(\mathbf{z}) = \mathbf{z}^\top\widehat{\boldsymbol{\beta}}$ e

$$\mathrm{Var}\{\widehat{\mu}(\mathbf{z})\} = \mathrm{Var}(\mathbf{z}^\top\widehat{\boldsymbol{\beta}}) = \mathbf{z}^\top\mathrm{Var}(\widehat{\boldsymbol{\beta}})\mathbf{z} = \sigma^2\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}.$$

Desde que $\widehat{\mathrm{Var}}\{\widehat{\mu}(\mathbf{z})\} = s^2\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}$, uma estimativa intervalar de coeficiente de confiança $(1-\alpha)$ para $\mu(\mathbf{z})$ fica dada por

$$[\mathbf{z}^\top\widehat{\boldsymbol{\beta}} \pm \mathrm{t}_{(1-\alpha/2),(n-p)}s\{\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}\}^{\frac{1}{2}}],$$

em que $t_{(1-\alpha/2),(n-p)}$ denota o quantil $(1-\alpha/2)$ de uma distribuição t-Student com $(n-p)$ graus de liberdade. A banda de coeficiente de confiança $(1-\alpha)$ para $\mu(\mathbf{z})$ assume a forma

$$[\mathbf{z}^\top\widehat{\boldsymbol{\beta}} \pm \sqrt{c_\alpha}\sigma\{\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}\}^{\frac{1}{2}},\ \forall \mathbf{z} \in I\!R^p],$$

em que c_α é tal que $\mathrm{P}\{\chi^2_p \leq c_\alpha\} = 1-\alpha$ (vide, por exemplo, Rao, 1973).

Por outro lado, o valor predito de $Y(\mathbf{z})$ pode ser representado por $\widehat{y}(\mathbf{z}) = \mathbf{z}^\top\widehat{\boldsymbol{\beta}} + \epsilon(\mathbf{z})$ e portanto

$$\begin{aligned}
\mathrm{Var}\{\widehat{Y}(\mathbf{z})\} &= \mathrm{Var}\{\mathbf{z}^\top\widehat{\boldsymbol{\beta}} + \epsilon(\mathbf{z})\} = \mathrm{Var}\{\mathbf{z}^\top\widehat{\boldsymbol{\beta}}\} + \mathrm{Var}\{\epsilon(\mathbf{z})\} \\
&= \mathbf{z}^\top\mathrm{Var}(\widehat{\boldsymbol{\beta}})\mathbf{z} + \mathrm{Var}\{\epsilon(\mathbf{z})\} = \sigma^2\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z} + \sigma^2 \\
&= \sigma^2\{1 + \mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}\}.
\end{aligned}$$

Tem-se que $\widehat{\mathrm{Var}}\{\widehat{Y}(\mathbf{z})\} = s^2\{1 + \mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}\}$.

Assim, estimativa intervalar e banda de confiança de coeficiente de confiança $(1-\alpha)$ para $y(\mathbf{z})$ ficam, respectivamente, dadas por

$$[\mathbf{z}^\top\widehat{\boldsymbol{\beta}} \pm t_{(1-\alpha/2),(n-p)}s\{1 + \mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}\}^{\frac{1}{2}}]$$

e

$$[\mathbf{z}^\top\widehat{\boldsymbol{\beta}} \pm \sqrt{c_\alpha}\sigma\{1 + \mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z}\}^{\frac{1}{2}},\ \forall \mathbf{z} \in I\!R^p].$$

Na prática deve-se substituir σ^2 por s^2 e c_α é obtido tal que $\mathrm{P}\{F_{p,(n-p)} \leq c_\alpha\} = 1-\alpha$. Em particular, para regressão linear simples é possível mostrar que $\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z} = 1/n + (z-\bar{x})^2/S_{xx}$.

6 Métodos de Diagnóstico

Procedimentos de diagnóstico devem ser aplicados após o ajuste do modelo linear normal e têm como principais objetivos:

(i) avaliar se há afastamentos importantes das suposições feitas para o modelo, tais como independência, normalidade, homocedasticidade dos erros e linearidade da média com relação aos valores das variáveis explicativas;

(ii) avaliar se há presença de observações atípicas ou discrepantes. Essas observações podem ser classificadas como pontos de alavanca, pontos aberrantes ou pontos influentes.

Abaixo segue descrição dos três tipos de observações atípicas.

Pontos de alavanca: observações em que o vetor $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ está remoto no subespaço $\mathrm{C}(\mathbf{X})$ gerado pelas colunas da matriz $\mathbf{X}$. Essas observações têm influência desproporcional no próprio valor ajustado.

Pontos aberrantes: observações com resíduo alto, posicionadas fora da banda de confiança. Ou seja, observações mal ajustadas pelo modelo. Em geral essas observações têm influência desproporcional na predição das respostas.

Pontos influentes: observações com peso desproporcional nas estimativas dos coeficientes do componente sistemático do modelo. Em geral são pontos de alavanca mas a recíproca nem sempre é verdadeira.

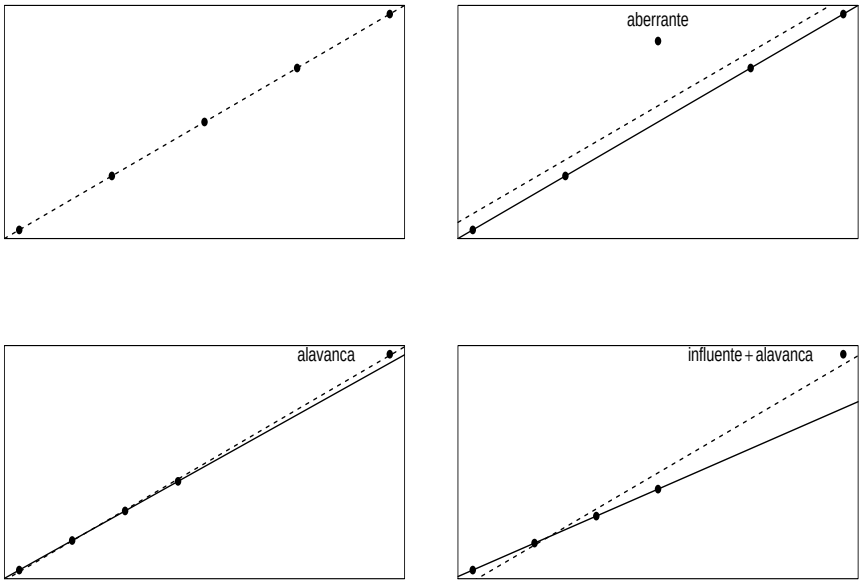

Figura 2: Representação gráfica para um conjunto de dados hipotéticos de pontos de alavanca, aberrantes e influentes. Reta ajustada com todos as observações $(\cdots)$ e sem a observação deslocada $(-)$.

Na Figura 2 há uma descrição gráfica de observações atípicas. No primeiro gráfico (acima à esquerda) tem-se uma regressão hipotética com a reta

ajustada passando pelas 5 observações, no segundo gráfico (acima à direita) a 3$^{\underline{a}}$ observação é deslocada verticalmente de forma a tornar- se aberrante, enquanto no terceiro e quarto gráficos (abaixo à esquerda e à direita) a 5$^{\underline{a}}$ observação é deslocada em direções diferentes de modo a tornar-se de alavanca e influente, respectivamente.

6.1 Pontos de Alavanca

Uma observação é definida como ponto de alavanca se tem uma alta influência no próprio valor ajustado. Essa influência é medida através da derivada $\partial\widehat{y}/\partial y$. Ou seja, mede o impacto que uma variação infinitesimal na resposta causa no valor ajustado. Da relação $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ obtém-se $\widehat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$, em que h_{ij} denota o elemento (i,j) da matriz $\mathbf{H}$ que é simétrica de dimensão $n \times n$. Daí segue que $\partial\widehat{y}_i/\partial y_i = h_{ii}$ e ainda pode-se mostrar que $h_{ii} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$.

Como a matriz $\mathbf{H}$ é idempotente ($\mathbf{H} = \mathbf{H}\mathbf{H}$) segue que

$$\sum_{j=1}^{n} h_{ij}^2 = h_{ii} \rightarrow \sum_{j\neq i} h_{ij}^2 = h_{ii} - h_{ii}^2 = h_{ii}(1 - h_{ii}),$$

então $h_{ii} \geq 0$ e $h_{ii}(1 - h_{ii}) \geq 0$ e portanto $0 \leq h_{ii} \leq 1$. Note que se $h_{ii} = 1$ então $h_{ij} = 0 \;\; \forall j \neq i$ e logo $\widehat{y}_i = y_i$. Hoaglin e Welsch (1978) propõem classificar pontos de alavanca segundo o critério $h_{ii} \geq 2\bar{h}$, em que $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n}$. Assim, desde que

$$\sum_{i=1}^{n} h_{ii} = \text{tr}(\mathbf{H}) = \text{tr}\{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\} = \text{tr}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\} = \text{tr}(\mathbf{I}_p) = p,$$

o critério fica dado por $h_{ii} \geq \frac{2p}{n}$. Para amostras grandes sugere-se $h_{ii} \geq \frac{3p}{n}$.

6.2 Limites para a Predição

Supor uma nova observação com valores para as variáveis explicativas representados por $\mathbf{z} = (z_1, z_2, \ldots, z_p)^\top$. Qual a condição para obter $\widehat{y}(\mathbf{z})$? Segundo Montgomery et al.(2021, p.110) pode-se fazer predição (interpolação) no modelo de regressão linear múltipla com segurança se a seguinte condição for satisfeita:

$$\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} \leq h_{\max} \;\forall \mathbf{x} \in I\!R^p,$$

em que $h_{\max} = \max\{h_{11}, \ldots, h_{nn}\}$. Logo, uma condição para predição de $y(\mathbf{z})$ é que $\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z} \leq h_{\max}$.

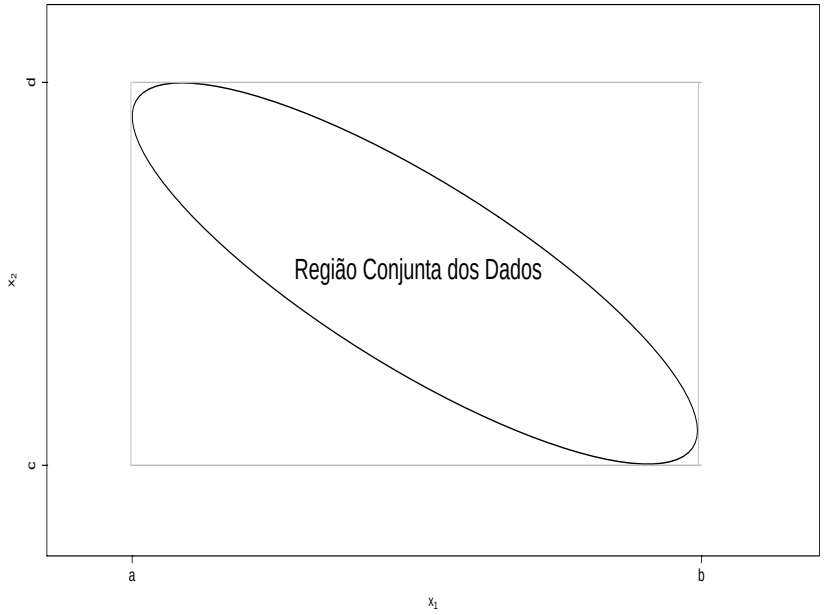

Figura 3: Representação geométrica para os limites de predição de um modelo de regressão (sem intercepto) com duas variáveis explicativas, com valores tais que $a \leq x_1 \leq b$ e $c \leq x_2 \leq d$.

Na Figura 3 tem-se a representação geométrica da "região conjunta dos dados"para a qual recomenda-se fazer as predições do modelo linear $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$, em que $a \leq x_1 \leq b$ e $c \leq x_2 \leq d$. Nota-se que há vários pares de valores (x_1, x_2) para os quais não é recomendado fazer interpolação.

6.3 Análise de Resíduos

Como visto anteriormente, o vetor de resíduos ordinários é definido por $\mathbf{r} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$, em que $\mathbf{r} = (r_1, \ldots, r_n)^\top$ com $r_i = y_i - \widehat{y}_i$, para $i = 1, \ldots, n$. Tem-se que

$$\begin{aligned} \mathrm{E}(\mathbf{r}) &= \mathrm{E}(\mathbf{Y}|\mathbf{X}) - \mathbf{H}\mathrm{E}(\mathbf{Y}|\mathbf{X}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}. \end{aligned}$$

A matriz de variância-covariância de $\mathbf{r}$ fica dada por

$$\begin{aligned}\text{Var}(\mathbf{r}) &= \text{Var}\{(\mathbf{I}_n - \mathbf{H})\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{I}_n - \mathbf{H})\text{Var}(\mathbf{Y}|\mathbf{X})(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H}).\end{aligned}$$

Portanto, segue que $\mathbf{r} \sim \text{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$, e consequentemente

(i) $r_i \sim \text{N}(0, \sigma^2(1 - h_{ii}))$;

(ii) $\text{Cov}(r_i, r_j) = -\sigma^2 h_{ij},\ i \neq j$ e

(iii) $\text{Corr}(r_i, r_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}},\ i \neq j,$

para $i, j = 1, \ldots, n$. Ou seja, os resíduos têm distribuição marginal normal de média zero, variâncias não constantes e são correlacionados.

Para que os resíduos sejam comparáveis é preciso padronizá-los. Uma padronização natural seria o resíduo normalizado

$$t_{r_i} = \frac{r_i}{\sigma\sqrt{1 - h_{ii}}} \sim \text{N}(0, 1),\ i = 1, \ldots, n.$$

Porém, é preciso estimar σ^2. Sabe-se que a estatística t-Student é construída da seguinte forma:

$$t = \frac{Z}{\sqrt{U/\nu}} \sim t_\nu,$$

em que $Z \sim \text{N}(0, 1)$, $U \sim \chi^2_\nu$ e Z e U são variáveis aleatórias independentes. Tem-se que $t_{r_i} \sim \text{N}(0, 1)$ e é possível mostrar que $(n - p)s^2/\sigma^2 \sim \chi^2_{(n-p)}$, porém t_{r_i} e s^2 não são independentes. Logo, o resíduo

$$t_i = \frac{r_i}{s\sqrt{1 - h_{ii}}} \nsim t_{(n-p)}.$$

Cook e Weisberg (1982) mostram que $\frac{t_i^2}{(n-p)} \sim \text{Beta}(\frac{1}{2}, \frac{(n-p-1)}{2})$. A sugestão é substituir s^2 por $s^2_{(i)}$, o erro quadrático médio do modelo sem a i-ésima observação. Agora, tem-se que $t_{r_i} \sim \text{N}(0, 1)$, $(n - p - 1)s^2_{(i)}/\sigma^2 \sim \chi^2_{(n-p-1)}$ e ainda t_{r_i} e $s^2_{(i)}$ são independentes. Logo, o resíduo

$$t_i^* = \frac{r_i}{s_{(i)}\sqrt{1 - h_{ii}}} \sim t_{(n-p-1)},$$

para $i = 1, \ldots, n$. É possível mostrar que

$$s^2_{(i)} = s^2 \left(\frac{n - p - t_i^2}{n - p - 1} \right).$$

Ou seja, $s^2_{(i)}$ pode ser obtido sem a necessidade de fazer o ajuste sem a i-ésima observação.

Gráficos sugeridos com o resíduo t_i^*:

(i) gráfico normal de probabilidades com banda de confiança empírica denominada envelope (Atkinson, 1981). Espera-se os pontos distribuídos de forma aleatória dentro da banda de confiança. Distorções no gráfico podem ser causadas por observações aberrantes e outras formas para o gráfico são indícios de afastamentos da normalidade dos erros;

(ii) gráfico de t_i^* contra valores ajustados $\widehat{y}_i$. Desde que $\mathrm{Cov}(\mathbf{r}, \widehat{\mathbf{y}}) = \mathbf{0}$, espera-se distribuição uniforme dos pontos conforma varia o valor ajustado. Afastamentos dessa tendência são indícios de que a variância dos erros não deve ser constante;

(iii) gráfico de t_i^* contra a ordem das observações para detectar (quando fizer sentido) correlação temporal dos dados;

(iv) gráfico de t_i^* contra valores de variáveis explicativas contínuas para avaliar se há algum termo que não foi incluído no componente sistemático do modelo.

A suposição de normalidade dos erros é crucial para fazer inferências quando o tamanho amostral n é pequeno ou moderado, contudo para n grande tem-se pelo Teorema Central do Limite (TCL) a normalidade assintótica de $\widehat{\boldsymbol{\beta}}$ desde que os erros tenham média zero e variância constante. Assim, quando há indícios de afastamentos importantes da suposição de normalidade dos erros pode-se tentar aplicar alguma transformaçao apropriada $g(Y)$ a fim de alcançar a normalidade mesmo que aproximadamente (vide exercícios 12 e 13). O incoveniente desse procedimento é que o novo modelo estará explicando $\mathrm{E}\{g(Y)\}$ ao invés de $\mathrm{E}(Y)$. Outra opção seria aplicar modelos lineares generalizados, em que procura-se uma distribuição apropriada para Y, porém tem-se em contrapartida a modelagem de $\mathrm{E}(Y)$. No caso da violação da suposição de variância constante para os erros, uma primeira opção seria aplicar regressão linear ponderada (Seção 9) que flexibiliza a variãncia dos erros sem comprometer os resultados da regressão linear. Alternativamente, pode-se aplicar a modelagem dupla em que $\mathrm{E}(Y)$

e Var(Y) são modelados conjuntamente. Para amostras pequenas e moderadas quando há violação da suposição de erros normais, pode-se aplicar procedimentos de reamostragem para estimação e inferência dos coeficientes da regressão (vide, por exemplo, Fox e Weisberg, 2019).

6.4 Outra Interpretação para t_i^*

Supor que o i-ésimo ponto é suspeito de ser aberrante. Essa hipótese pode ser testada através do modelo

$$y_j = \mathbf{x}_j^\top \boldsymbol{\beta} + \omega_j \gamma + \epsilon_j, \tag{4}$$

em que $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})^\top$ e $\epsilon_j \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ para $j = 1, \ldots, n$, com $\omega_j = 1$ para $j = i$ e $\omega_j = 0$ em caso contrário. Usando resultados da Seção 3 pode-se mostrar que sob a hipótese $\text{H}_0 : \gamma = 0$ o acréscimo na soma de quadrados de resíduos fica dado por

$$\text{ASQ}(\gamma = 0) = \widehat{\gamma}^2(1 - h_{ii}),$$

em que $\widehat{\gamma} = r_i(1 - h_{ii})^{-1}$ com $r_i = y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}$ e $h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$. Logo, a soma de quadrados de resíduos correspondente ao modelo (4) fica dada por $(n-p)s^2 - \widehat{\gamma}^2(1 - h_{ii}) = (n-p)s^2 - \frac{r_i^2}{1-h_{ii}}$ e a estatística F para testar $\text{H}_0 : \gamma = 0$ contra $\text{H}_1 : \gamma \neq 0$ assume a forma

$$\text{F} = \frac{\widehat{\gamma}^2(1 - h_{ii})}{\left\{(n-p)s^2 - \frac{r_i^2}{(1-h_{ii})}\right\}/(n-p-1)} \overset{\text{H}_0}{\sim} \text{F}_{1,(n-p-1)}.$$

Trabalhando um pouco a expressão acima chega-se ao seguinte resultado:

$$\text{F} = \frac{r_i^2(n-p-1)}{s^2(1-h_{ii})(n-p-t_i^2)} = t_i^{*2}.$$

Portanto, para um nível de significância α, rejeita-se H_0 se $|t_i^*| > \text{t}_{(1-\alpha/2),(n-p-1)}$.

6.5 Análise de Influência

O objetivo principal da análise de influência em regressão é avaliar o impacto de perturbações no modelo e/ou dados nos coeficientes da regressão, sendo esse impacto avaliado através de alguma medida de influência. A medida de influência mais conhecida, denominada distância de Cook (Cook, 1977), procura avaliar o impacto da retirada de cada observação nas estimativas

dos coeficientes. Uma vez detectadas as observações com maior variação para essa medida, deve-se proceder algum tipo de análise confirmatória a fim de avaliar a influência das observações destacadas e também o tipo de influência. Variações numéricas nas estimativas dos coeficientes são esperadas quando elimina-se observações, contudo quando essas variações são desproporcionais, muito acima $\frac{1}{n} \times 100\%$, as observações podem ser consideradas influentes. O mais grave é quando a eliminação individual de uma observação leva a mudanças inferenciais, ou seja, determinados coeficientes deixam ou passam a ser significativos. No primeiro caso a observação induz o efeito do coeficiente enquanto que no segundo caso há mascaramento do efeito pela observação.

Transformações dos valores das variáveis explicativas, inclusão de interação ou mesmo ponderação na regressão, dentre outros procedimentos, são comumente aplicados para reduzir a influência de observações na regressão. Contudo, quando esses procedimentos não levam a soluções satisfatórias recomenda-se a aplicação de procedimentos de estimação robusta. Montogomery et al. (2021, Cap.15) apresentam alguns procedimentos de estimação robusta para regressão linear múltipla.

Nesta seção será discutida a distância de Cook aplicada ao modelo de regressão linear múltipla (2). Essa medida pode ser motivada através da região de confiança de coeficiente $(1-\alpha)$ para $\boldsymbol{\beta}$, dada por

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{ps^2} \leq F_{(1-\alpha),p,(n-p)},$$

em que $F_{(1-\alpha),p,(n-p)}$, como definido anteriormente, denota o quantil $(1-\alpha)$ de uma distribuição F com p e $(n-p)$ graus de liberdade. Essa região de confiança é construída usando o resultado abaixo

$$P\left\{\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{ps^2} \leq F_{(1-\alpha),p,(n-p)}\right\} = 1-\alpha.$$

A distância de Cook é definida por

$$\mathrm{D}_i = \frac{(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)})^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)})}{ps^2},$$

em que $\widehat{\boldsymbol{\beta}}_{(i)}$ denota a estimativa de mínimos quadrados quando a i-ésima observação não é considerada no modelo. Após manipulações algébricas

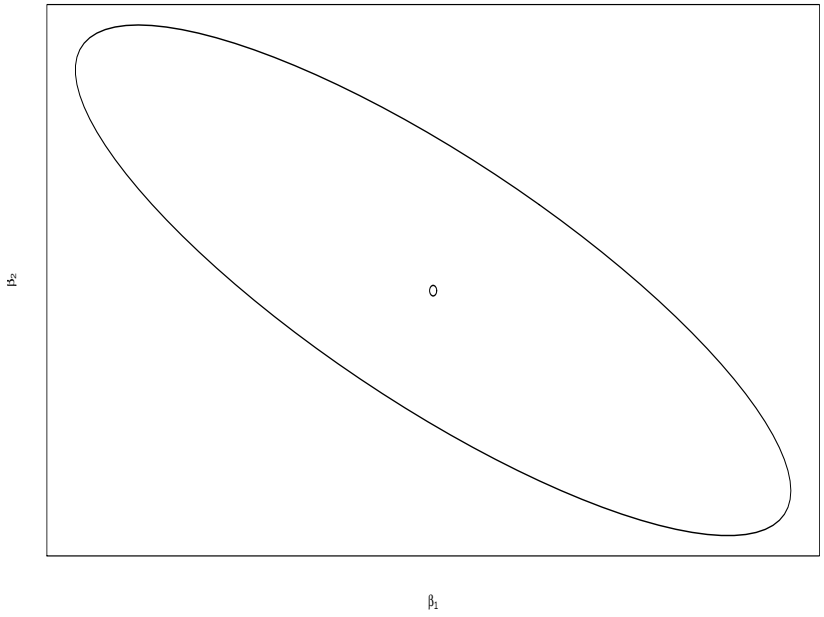

Figura 4: Representação geométrica para a região de confiança de 95% para os coeficientes de um modelo de regressão hipotético com $p = 2$.

obtém-se

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{(i)} &= \{\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\}^{-1}\mathbf{X}_{(i)}^\top \mathbf{y}_{(i)} \\
&= \{\mathbf{X}^\top\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^\top\}^{-1}\{\mathbf{X}^\top\mathbf{y} - \mathbf{x}_i y_i\} \\
&= \left\{(\mathbf{X}^\top\mathbf{X})^{-1} + \frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}}{1-h_{ii}}\right\}\{\mathbf{X}^\top\mathbf{y} - \mathbf{x}_i y_i\} \\
&= \widehat{\boldsymbol{\beta}} - \frac{r_i}{(1-h_{ii})}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i,
\end{aligned}$$

para $i = 1, \ldots, n$. Portanto, tem-se que

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{r_i}{(1-h_{ii})}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i.$$

Assim, a distância de Cook fica dada

$$\mathrm{D}_i = \frac{1}{p}t_i^2\frac{h_{ii}}{(1-h_{ii})}.$$

Como $h_{ii}/(1-h_{ii})$ é uma função crescente de h_{ii}, então D_i será grande se $|t_i|$ e/ou h_{ii} forem (for) grande(s). Uma proposta de pontos suspeitos de serem influentes, baseada na região de confiança para $\boldsymbol{\beta}$, é destacar as observações tais que $\mathrm{D}_i \geq F_{(1-\alpha),p,(n-p)}$. Outras sugestões se baseiam em obter limites superiores para a distância de Cook com base nas variações dos valores amostrais da distância e que levem em conta o tamanho amostral. Sugere-se destacar as observações tais que $\mathrm{D}_i \geq \bar{\mathrm{D}} + k\mathrm{DP}(\mathrm{D}_i)$, para $k = 2, 3, 4$. Deve-se aumentar o valor k à medida que aumenta o tamanho amostral.

Outra medida de influência proposta por Belsley et al. (1980), que é derivada da distância de Cook com s^2 substuído por $s^2_{(i)}$, é definida por

$$\begin{aligned} \mathrm{DFFITS}_i &= \frac{|r_i|}{s_{(i)}\sqrt{1-h_{ii}}}\left\{\frac{h_{ii}}{1-h_{ii}}\right\}^{\frac{1}{2}} \\ &= |t_i^*|\left\{\frac{h_{ii}}{1-h_{ii}}\right\}^{\frac{1}{2}}. \end{aligned}$$

Sugere-se destacar as observações tais que $\mathrm{DFFITS}_i \geq 2\{p/(n-p)\}^{\frac{1}{2}}$. Essa medida leva também em conta a influência das observações na estimativa de σ^2. Contudo, quando o interesse está apenas nos coeficientes da regressão sugere-se utilizar apenas a distância de Cook.

Finalmente, pode haver interesse em estudar a influência das observações em coeficientes específicos da regressão. Por exemplo, se há interesse em avaliar a influência da eliminação da i-ésima observação no j-ésimo coeficiente estimado da regressão, utiliza-se a seguinte medida de influência:

$$\begin{aligned} \mathrm{DFBETAS}_{ji} &= \frac{(\widehat{\beta}_j - \widehat{\beta}_{j(i)})}{s_{(i)}\sqrt{C_{jj}}} \\ &= \frac{\mathbf{C}_j^\top \mathbf{x}_i r_i}{s_{(i)}(1-h_{ii})\sqrt{C_{jj}}} \\ &= \frac{p_{ji}}{\sqrt{\mathbf{p}_j^\top \mathbf{p}_j}}\frac{t_i^*}{\sqrt{1-h_{ii}}}, \end{aligned}$$

em que $\mathbf{C} = (\mathbf{X}^\top\mathbf{X})^{-1}$, $\mathbf{C}_j$ denota a j-ésima coluna de $\mathbf{C}$, p_{ji} e $\mathbf{p}_j^\top$ denotam, respectivamente, o (j,i)-ésimo elemento e a j-ésima linha de $\mathbf{P} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, $i = 1, \ldots, n$ e $j = 1, \ldots, p$. Recomenda-se dar atenção àquelas observações tais que $\mathrm{DFBETAS}_{ji} > \frac{2}{\sqrt{n}}$ (vide Montgomery et al., 2021, Cap.6).

6.6 Gráfico da Variável Adicionada

Supor que uma variável explicativa é adicionada no modelo (2) obtendo-se o seguinte modelo de regressão linear:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}\gamma + \boldsymbol{\epsilon}$$

em que $\mathbf{X}$ denota a matriz modelo $n \times p$ do modelo reduzido, $\mathbf{w}$ denota vetor $n \times 1$ dos valores observados da variável adicionada, $\mathbf{y}$ é o vetor $n \times 1$ dos valores observados da variável resposta, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ e $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I_n})$. Definindo $\mathbf{Z} = (\mathbf{X}, \boldsymbol{\omega})$ como matriz do modelo ampliado, mostra-se facilmente que a estimativa de mínimos quadrados de $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma})^\top$ fica expressa na forma $\hat{\boldsymbol{\theta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$. Após algumas manipulações algébricas a estimativa de mínimos quadrados do coeficiente da variável adicionada fica dada por

$$\begin{aligned} \hat{\gamma} &= \frac{\boldsymbol{\omega}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}}{\boldsymbol{\omega}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\omega}} \\ &= \frac{\boldsymbol{\omega}^\top\mathbf{r}}{\boldsymbol{\omega}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\omega}}. \end{aligned}$$

Ou seja, $\hat{\gamma}$ pode ser expresso como sendo o coeficiente da regressão linear passando pela origem do vetor de resíduos $\mathbf{r} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ sobre o novo resíduo $\boldsymbol{\upsilon} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\omega}$, dado por

$$\begin{aligned} \hat{\gamma} &= (\boldsymbol{\upsilon}^\top\mathbf{v})^{-1}\boldsymbol{\upsilon}^\top\mathbf{r} \\ &= \{\boldsymbol{\omega}^\top(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H})\boldsymbol{\omega}\}^{-1}\boldsymbol{\omega}^\top(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H})\mathbf{y} \\ &= \frac{\boldsymbol{\omega}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}}{\boldsymbol{\omega}^T(\mathbf{I}_n - \mathbf{H})\boldsymbol{\omega}}. \end{aligned}$$

Portanto, o gráfico de $\mathbf{r}$ contra $\boldsymbol{\upsilon}$ pode fornecer informações sobre a evidência dessa regressão, indicando quais observações que estão contribuindo para a relação linear e quais observações que estão se desviando da mesma. Esse gráfico, conhecido como gráfico da variável adicionada (ver, por exemplo, Atkinson, 1985) pode revelar quais observações que estão influenciando (e de que maneira) a inclusão da nova variável explicativa no modelo.

A sugestão é que seja construído para cada variável explicativa contínua incluída de forma linear no modelo um gráfico da variável adicionada.

6.7 Aplicação

Para ilustrar um exemplo de regressão linear simples considere parte dos dados descritos em Neter et al. (1996, p.449) referentes à venda no ano

anterior de um tipo de telhado de madeira em $n = 26$ filiais de uma rede de lojas de construção civil. Apenas duas variáveis serão consideradas:

(i) Telhados: total de telhados vendidos (em mil metros quadrados) e

(ii) Nclientes: número de clientes cadastrados na loja (em milhares).

O interesse é explicar o número médio de telhados vendidos dado o número de clientes cadastrados. Na Tabela 2 são apresentadas algumas medidas resumo referentes às duas variáveis observadas.

Tabela 2: Medidas resumo referentes ao exemplo sobre venda de telhados.

Medida	Telhados	Nclientes
Média	170,20	51,85
D.Padrão	84,55	14,21
CV(em %)	49,68	27,41
Mínimo	30,90	26,00
1º Quartil	102,00	49,50
Mediana	159,80	51,50
3º Quartil	217,50	61,50
Máximo	339,40	75,00

Na Figura 5 tem-se o boxplot robusto (Hubert e Vandervierin, 2008) e a densidade estimada do total de telhados vendidos. Nota-se ausência de observações aberrantes e uma ligeira assimetria à direita. O diagrama de dispersão entre o total de telhados vendidos e o número de clientes cadastrados na loja (Figura 6) apresenta uma tendência aproximadamente linear e positiva. À medida que aumenta o número de clientes aumenta o total de telhados vendidos.

Portanto, sugere-se o seguinte modelo de regressão linear simples:

$$y_i = \beta_1 + \beta_2 \text{Nclientes}_i + \epsilon_i,$$

em que y_i denota o total de telhados vendidos na i-ésima filial e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, 26$. Nota-se pela Tabela 3 que o coeficiente estimado do número de clientes é altamente significativo e o intercepto é significativo ao nível de 10%. Assim, para um aumento de 1000 clientes em qualquer filial espera-se aumento de 4656 mil m^2 de telhados vendidos.

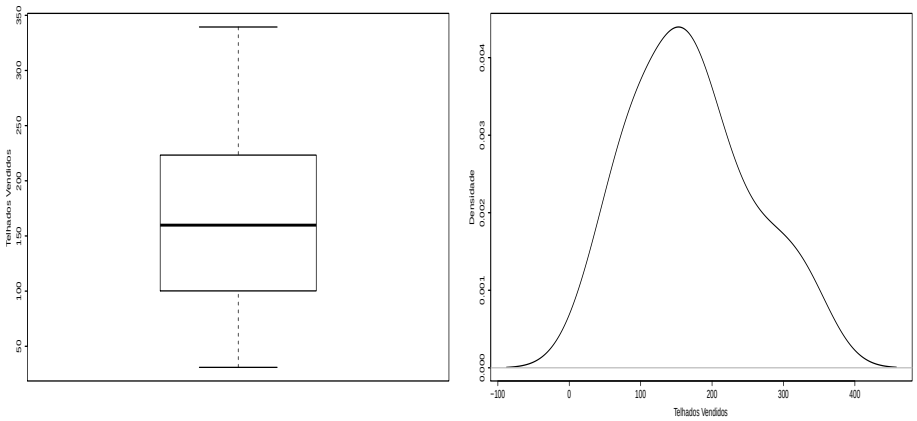

Figura 5: Boxplot robusto e densidade estimada do total de telhados vendidos.

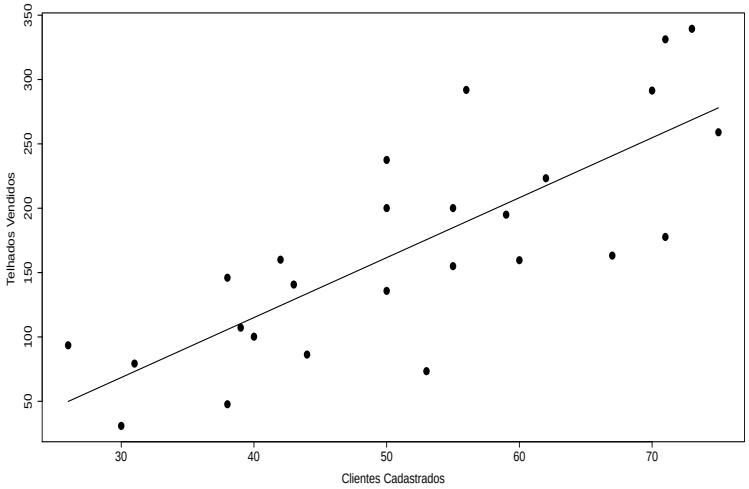

Figura 6: Diagrama de dispersão (com tendência) entre o total de telhados vendidos e o número de clientes cadastrados na loja.

Pela Figura 7, em que são apresentados o gráfico do resíduo t_i^* contra o valor ajustado $\widehat{y}_i$ e o gráfico normal de probabilidades para t_i^* com banda empírica de confiança (envelope) de 95%, não há indícios de variância não

Tabela 3: Estimativas dos parâmetros referentes ao modelo de regressão linear simples ajustado aos dados sobre venda de telhados.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	-71,208	40,558	-1,76	0,092
Nclientes	4,656	0,756	6,16	0,000
s	53,69			
R^2	0,61			
$\overline{R}^2$	0,60			

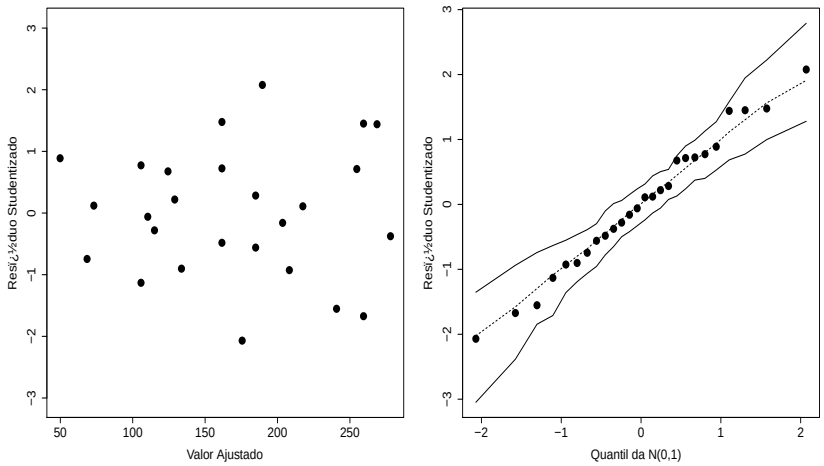

Figura 7: Gráficos de resíduos referentes ao modelo de regressão linear simples ajustado aos dados sobre venda de telhados.

constante nem de afastamentos da normalidade dos erros. Nota-se também ausência de observações aberrantes. O gráfico da distância de Cook com $k = 2$ (Figura 8) contra a ordem das observações destaca como possivelmente influentes as observações #6 e #10. O ajuste sem cada uma das observações traz variações nas estimativas dos coeficientes, como pode ser notado pela Figura 9, porém não há mudanças inferencias. Finalmente, tem-se na Figura 10 as bandas de confiança de 95% para o número esperado de telhados vendidos e para o número de telhados vendidos de uma filial qualquer, dado

o número de clientes cadastrados.

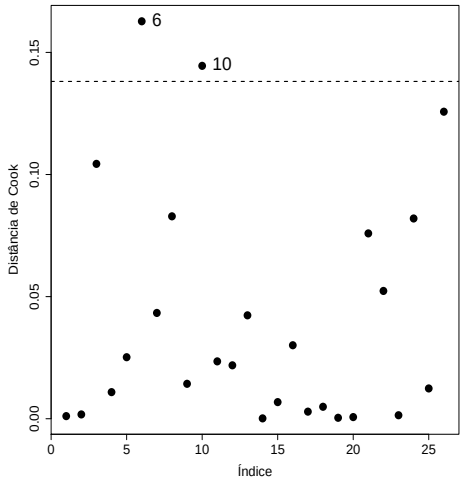

Figura 8: Distância de Cook contra a ordem das observações referente ao modelo de regressão linear simples ajustado aos dados sobre venda de telhados.

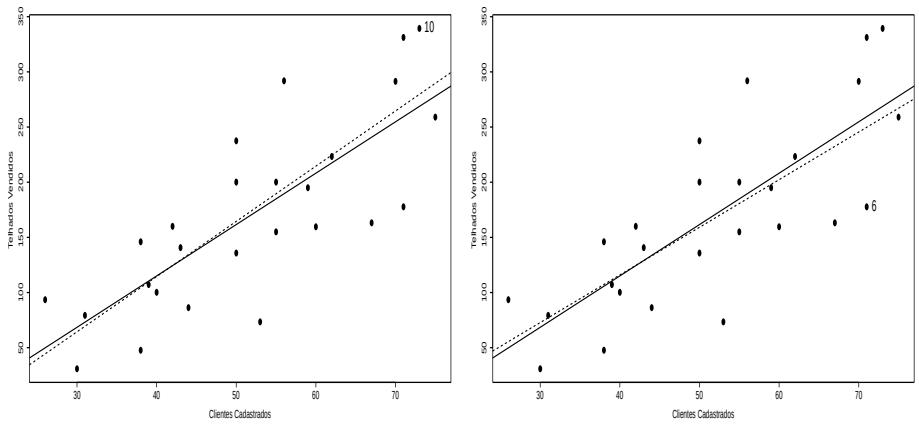

Figura 9: Retas ajustadas com todos os pontos (−) e sem as observações destacatadas pela distância de Cook (· · ·).

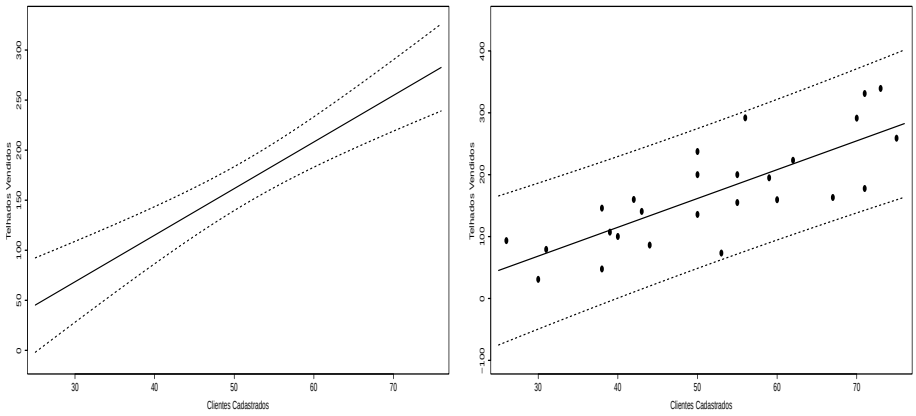

Figura 10: Bandas de confiança de 95% para o número esperado de telhados vendidos (esquerda) e para o número de telhados vendidos de uma filial qualquer (direita), dado o número de clientes cadastrados.

7 Variável Binária e Interação

Supor o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, x_{i2} representa os valores de uma variável aleatória binária tal que

$$x_{i2} = \begin{cases} 1 & \text{grupo A} \\ 0 & \text{grupo B}, \end{cases}$$

enquanto x_{i3} representa valores observados de uma variável contínua e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$.

Portanto, tem-se dois submodelos de regressão

- (Grupo A) $y_i = \beta_1 + \beta_2 + \beta_3 x_{i3} + \epsilon_i$
- (Grupo B) $y_i = \beta_1 + \beta_3 x_{i3} + \epsilon_i$

com valores esperados

- $\text{E}_\text{A}(Y_i | x_{i3}) = \beta_1 + \beta_2 + \beta_3 x_{i3}$

- $\mathrm{E_B}(Y_i|x_{i3}) = \beta_1 + \beta_3 x_{i3}$,

para $i = 1, \ldots, n$. Assim, $\mathrm{E_A}(Y_i|x_{i3}) - \mathrm{E_B}(Y_i|x_{i3}) = \beta_2$, que indica ausência de interação (paralelismo) entre as variáveis explicativas X_2 e X_3 (vide ilustração na Figura 11).

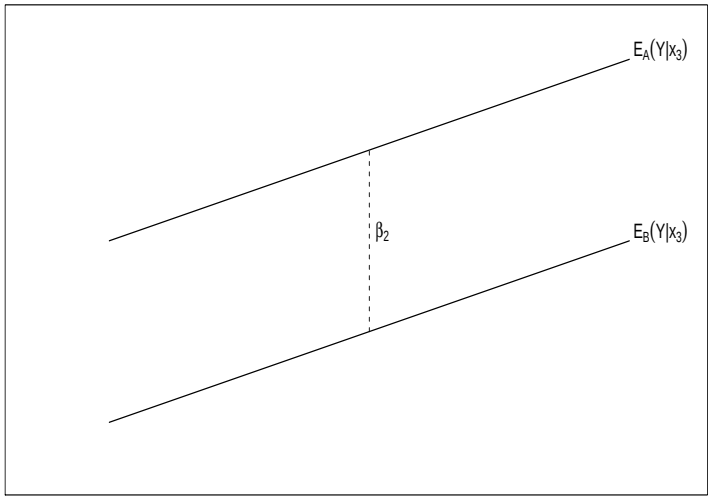

Figura 11: Descrição gráfica de ausência de interação (paralelismo) entre as variáveis explicativas X_2 e X_3.

Supor agora a inclusão de interação entre as variáveis explicativas X_2 e X_3, resultando no seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2} x_{i3} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. Tem-se os seguintes submodelos:

- (Grupo A) $y_i = \beta_1 + \beta_2 + \beta_3 x_{i3} + \beta_4 x_{i3} + \epsilon_i$
- (Grupo B) $y_i = \beta_1 + \beta_3 x_{i3} + \epsilon_i$

com valores esperados expressos por

- $\mathrm{E_A}(Y_i|x_{i3}) = \beta_1 + \beta_2 + \beta_3 x_{i3} + \beta_4 x_{i3}$

- $\mathrm{E_B}(Y_i|x_{i3}) = \beta_1 + \beta_3 x_{i3}$,

para $i = 1, \ldots, n$. Assim, a diferença entre os valores esperados, $\mathrm{E_A}(Y_i|x_{i3}) - \mathrm{E_B}(Y_i|x_{i3}) = \beta_2 + \beta_4 x_{i3}$, não é mais contante dependendo dos valores da variável explicativa X_3. Isso indica presença de interação (ausência de paralelismo) entre as variáveis explicativas X_2 e X_3 (vide Figura 12).

Supor agora variável explicativa categórica com três níveis

$$X = \begin{cases} 1 & \text{grupo A} \\ 2 & \text{grupo B} \\ 3 & \text{grupo C.} \end{cases}$$

Um maneira de representar essa variável explicativa num modelo de regressão é atribuindo a cada grupo uma variável binária da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ denotam os valores observados da variável resposta, x_{i1}, x_{i2} e x_{i3} são os valores observados das variáveis binárias representando os grupos e $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, para $j = 1, \ldots, n$.

Supondo que os grupos A, B e C têm n_1, n_2 e n_3 elementos, respectivamente, o modelo pode ser expresso na forma matricial $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, em que $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \mathbf{y}_3^\top)^\top$ com $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$, para $i = 1, 2, 3$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ e matriz $\mathbf{X}$ de dimensão $(n_1 + n_2 + n_3) \times 4$ dada por

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Note que a matriz $\mathbf{X}$ não tem posto coluna completo, a $1^{\underline{a}}$ coluna é a soma das outras três colunas. Uma solução é reduzir o número de colunas da matriz modelo impondo alguma restrição nos parâmetros.

Os seguintes procedimentos são mais utilizados:

- Restrição nos parâmetros: $\beta_1 + \beta_2 + \beta_3 = 0$, que implica em $\beta_1 = -\beta_2 - \beta_3$.

- Casela de referência: um dos coeficientes é fixado como sendo zero. Por exemplo, fazendo $\beta_1 = 0$ o grupo A será denominado casela de referência.

Nesses dois casos $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3)^\top$ e a matriz modelo terá dimensão $n \times 3$ com posto coluna completo.

Como exemplo, o modelo com casela de referência no grupo A pode ser expresso na forma

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ denotam os valores observados da variável resposta, x_{i2} e x_{i3} são valores de variáveis binárias representando os grupos B e C, respectivamente, enquanto x_{i4} representa os valores observados de uma variável contínua e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. Quando $x_{i2} = x_{i3} = 0$ tem-se o grupo A.

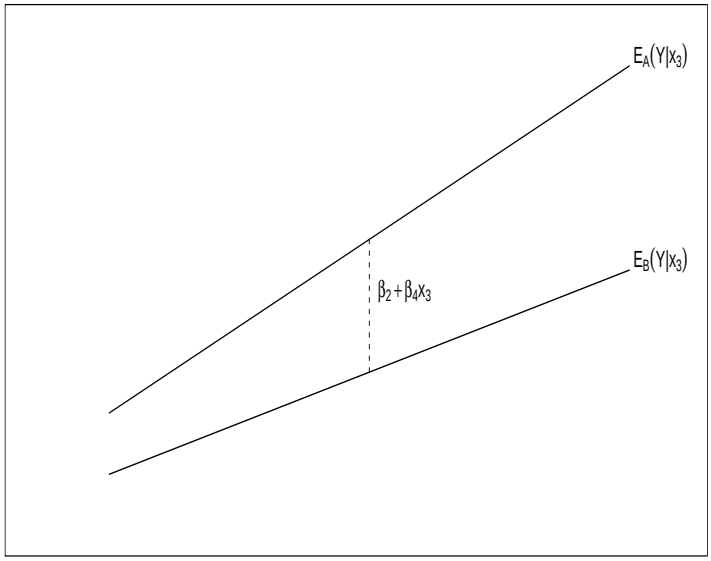

Figura 12: Descrição gráfica de presença de interação (ausência de paralelismo) entre as variáveis explicativas X_2 e X_3.

A matriz modelo nesse caso fica dada por

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}.$$

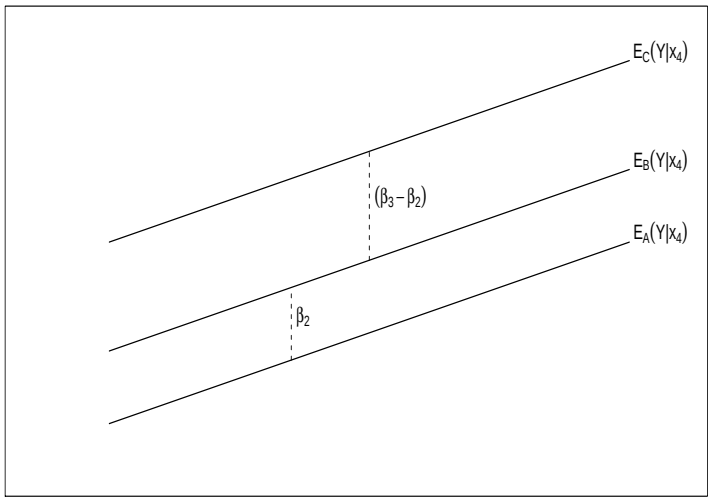

Figura 13: Descrição gráfica de ausência de interação (paralelismo) entre a variável categórica X e a variável contínua X_4.

Portanto, tem-se três submodelos

- (Grupo A) $y_i = \beta_0 + \beta_4 x_{i4} + \epsilon_i$
- (Grupo B) $y_i = \beta_0 + \beta_2 + \beta_4 x_{i4} + \epsilon_i$

- (Grupo C) $y_i = \beta_0 + \beta_3 + \beta_4 x_{i4} + \epsilon_i$

com diferenças de valores esperados

- $\mathrm{E_B}(Y_i|x_{i4}) - \mathrm{E_A}(Y_i|x_{i4}) = \beta_2$
- $\mathrm{E_C}(Y_i|x_{i4}) - \mathrm{E_A}(Y_i|x_{i4}) = \beta_3$,

para $i = 1, \ldots, n$. Assim, os efeitos β_2 e β_3 são incrementos nos valores esperados dos grupos B e C, respectivamente, com relação ao grupo A (vide ilustração na Figura 13).

Em forma matricial o modelo com ausência de interação fica dado por $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, em que $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \mathbf{y}_3^\top)^\top$ com $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$, para $i = 1, 2, 3$, $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, \beta_4)^\top$ e matriz modelo $\mathbf{X}$ terá adicionada a coluna $(x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2}, x_{n_1+n_2+1}, \ldots, x_n)^\top$.

O modelo com interação entre a variável categórica X e a variável contínua X_4 pode ser expresso na seguinte forma:

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i2} x_{i4} + \beta_6 x_{i3} x_{i4} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ denotam os valores observados da variável resposta, x_{i2} e x_{i3} são valores de variáveis binárias representando os grupos B e C, respectivamente, enquanto x_{i4} representa os valores observados de uma variável contínua e $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, para $i = 1, \ldots, n$.

Portanto, tem-se três submodelos

- (Grupo A) $y_i = \beta_0 + \beta_4 x_{i4} + \epsilon_i$
- (Grupo B) $y_i = \beta_0 + \beta_2 + \beta_4 x_{i4} + \beta_5 x_{i4} + \epsilon_i$
- (Grupo C) $y_i = \beta_0 + \beta_3 + \beta_4 x_{i4} + \beta_6 x_{i4} + \epsilon_i$,

com diferenças de valores esperados

- $\mathrm{E_B}(Y_i|x_{i4}) - \mathrm{E_A}(Y_i|x_{i4}) = \beta_2 + \beta_5 x_{i4}$
- $\mathrm{E_C}(Y_i|x_{i4}) - \mathrm{E_A}(Y_i|x_{i4}) = \beta_3 + \beta_6 x_{i4}$,

para $i = 1, \ldots, n$. Assim, nota-se que as diferenças entre os valores esperados dependem dos valores da variável explicativa X_4 (vide Figura 14). A matriz modelo $\mathbf{X}$ terá duas colunas adicionais com relação à matriz modelo sob ausência de interação.

O conceito de interação pode ser estendido para quaisquer tipos de variáveis explicativas e para mais do que duas variáveis explicativas. Contudo, devido a dificuldades na interpretação, em geral considera-se apenas interações de $1^{\underline{a}}$ ordem (entre duas variáveis eplicativas).

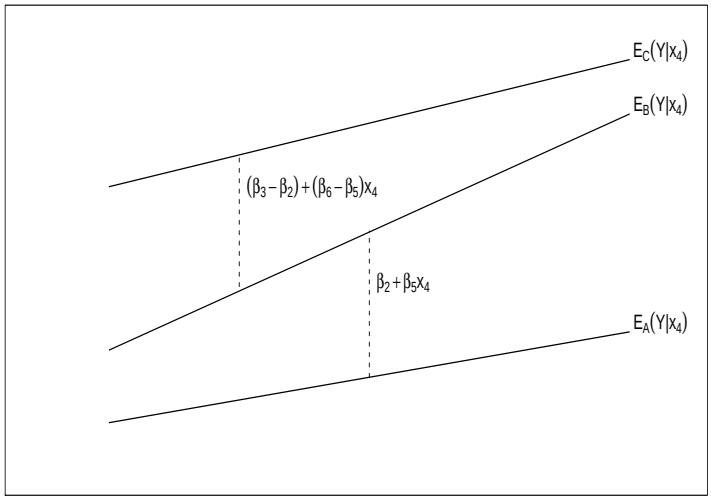

Figura 14: Descrição gráfica de interação entre a variável categórica X e a variável contínua X_4.

8 Comparação de Médias

Uma aplicação de modelos de regressão linear com variáveis binárias é na comparação das médias de k grupos. O modelo pode ser expresso na forma

$$y_{ij} = \alpha + \beta_i + \epsilon_{ij},$$

em que $\epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$, com a restrição $\beta_1 = 0$. O Grupo 1 é denominado casela de referência. Assim, tem-se os valores esperados

- $\text{E}(Y_{1j}) = \alpha$ para $j = 1, \ldots, n_1$
- $\text{E}(Y_{ij}) = \alpha + \beta_i$, para $i = 2, \ldots, k$ e $j = 1, \ldots, n_i$,

e daí segue que β_i é o incremento no valor médio do i-ésimo grupo com relação ao valor médio do grupo 1, para $i = 2, \ldots, k$. Testar a igualdade de médias equivale a testar $\text{H}_0 : \beta_2 = \cdots = \beta_k$ contra $\text{H}_1 : \beta_j \neq 0$ para pelo menos algum $j = 2, \ldots, k$.

Em forma matricial o modelo fica dado por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_k^\top)^\top$ com $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$, para $i = 1, \ldots, k$, $\boldsymbol{\beta} = (\alpha, \beta_2, \ldots, \beta_k)^\top$ e matriz $\mathbf{X}$ de dimensão $(\sum_{i=1}^k n_i) \times k$ dada abaixo.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 1 \end{bmatrix}.$$

A solução de mínimos quadrados leva às estimativas $\widehat{\alpha} = \bar{y}_1$ e $\widehat{\beta}_i = \bar{y}_i - \bar{y}_1$ para $i = 1, \ldots, k$, com variâncias e covariâncias

$$\mathrm{Var}(\widehat{\alpha}) = \frac{\sigma^2}{n_1}, \ \mathrm{Var}(\widehat{\beta}_j) = \sigma^2 \left\{ \frac{1}{n_j} + \frac{1}{n_1} \right\}, \ \mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}_j) = -\frac{\sigma^2}{n_1} \ \text{e} \ \mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_\ell) = \frac{\sigma^2}{n_1},$$

para $j \neq \ell = 2, \ldots, k$.

Tem-se a seguinte decomposição das somas de quadrados:

$$\begin{aligned} \mathrm{SQT} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2, \\ \mathrm{SQReg} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\widehat{y}_{ij} - \bar{y})^2 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \ \text{ e} \\ \mathrm{SQRes} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \widehat{y}_{ij})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \end{aligned}$$

Daí segue que a estatística F para testar a homogeneidade de médias H_0 : $\beta_2 = \cdots = \beta_k = 0$ contra H_1 : pelo menos duas médias diferentes fica expressa na forma

$$F = \frac{(n-k+1)}{(k-1)} \frac{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \overset{\mathrm{H}_0}{\sim} F_{(k-1),(n-k+1)}.$$

Rejeita-se H_0 se $F > F_{(1-\alpha),(k-1),(n-k+1)}$, em que $F_{(1-\alpha),(k-1),(n-k+1)}$ denota o quantil $(1-\alpha)$ da distribuição F com $(k-1)$ e $(n-k+1)$ graus de liberdade e $n = n_1 + \cdots + n_k$.

8.1 Comparações Múltiplas

Quando rejeita-se a hipótese nula deseja-se saber onde estão as diferenças entre as médias dos k grupos. As propostas mais conhecidas consistem em construir $m = \binom{k}{2}$ estimativas intervalares para as diferenças de médias, de modo que cada estimativa intervalar tenha coeficiente de confiança $(1-\alpha^*)$ sendo o coeficiente de confiança global $(1-\alpha)$.

Pelo método de Bonferroni (recomendado para m pequeno) cada estimativa intervalar deve ter coeficiente de confiança $(1-\alpha^*)$, sendo dadas por

$$(\bar{y}_i - \bar{y}_j) \pm t_{(1-\alpha^*/2),(n-k)} \sqrt{s^2 \left\{ \frac{1}{n_i} + \frac{1}{n_j} \right\}},$$

para $i \neq j$, em que $\alpha^* = \frac{\alpha}{m}$, de modo que o coeficiente global de confiança seja de pelo menos $(1-\alpha)$.

O método de Tukey é o mais utilizado na prática por ter um nível de significância global mais próximo de $(1-\alpha)$. As estimativas intervalares são expressas na forma

$$(\bar{y}_i - \bar{y}_j) \pm q(k, n-k) \sqrt{\frac{s^2}{2} \left\{ \frac{1}{n_i} + \frac{1}{n_j} \right\}},$$

para $i \neq j$, em que $q(k, n-k)$ é o quantil de uma distribuição denominada amplitude Studentizada.

8.2 Aplicação

Como ilustração serão considerados os dados referentes ao tempo de deslocamento (em minutos) antes de decolar de 184 aeronaves de 8 Cias Aéreas no aeroporto EWR (Newark) no período 1999-2001 (Venzani, 2004, Exemplo 11.7), descritas abaixo

- AA, American Airlines
- CO, Continental Airlines
- DL, Delta Airlines

- HP, American West Airlines
- NW, North West Airlines
- TW, Trans World Airlines
- UA, United Airlines
- US, US Airways.

Na Figura 15 tem-se os boxplots robustos dos tempos para a decolagem das Cias Aéreas. Nota-se tempos medianos distintos, porém em geral variabilidades similares. As Cias Aéreas NW e US apresentam os menores tempos medianos enquanto CO apresenta o maior tempo mediano. A fim de comparar os tempos médios supondo variabilidades homogêneas considere o modelo $y_{ij} = \alpha + \beta_i + \epsilon_{ij}$, em que $\epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, 8$ e $j = 1, \ldots, 23$, com a restrição $\beta_1 = 0$. AA como casela de referência.

É bastante razoável esperar pelo TCL que $\widehat{\alpha}$ e $\widehat{\beta}_i$ estejam bem aproximadas pela distribuição normal levando-se em conta o número de réplicas para cada Cia Aérea. Assim, como não há indícios pela Figura 15 de afastamentos importantes da suposição de variâncias contantes para os erros, pode-se esperar uma boa aprocimação da distribuiçao nula da estatística F para testar a homogeneidade de médias.

Tabela 4: Estimativas dos parâmetros referentes ao modelo de comparação dos tempos médios de deslocamento das Cias Aéreas.

Efeito	Estimativa	valor-t	valor-P
AA	27,056	37,56	0,000
CO	3,835	3,76	0,000
DL	-2,052	-2,01	0,045
HP	1,526	1,50	0,136
NW	-4,061	-3,99	0,000
TW	-1,652	-1,62	0,107
UA	-0,039	-0,04	0,969
US	-3,830	-3.76	0,000
s	3,455		
R^2	0,355		
$\overline{\text{R}}^2$	0,329		

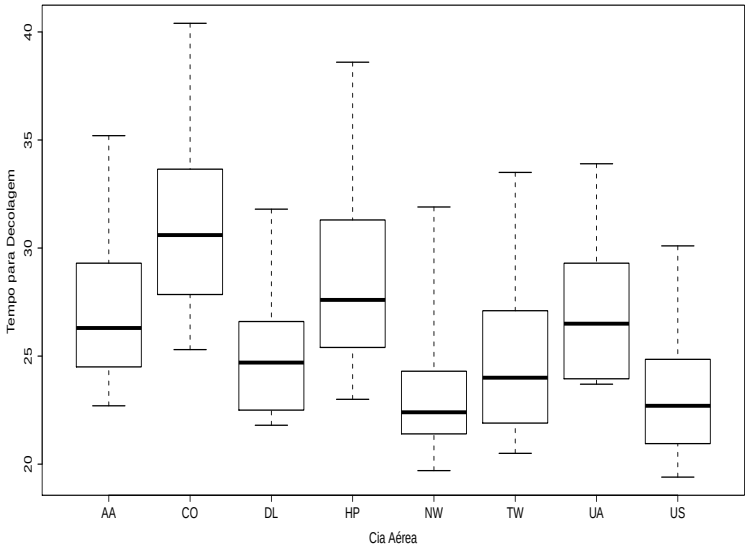

Figura 15: Boxplot do tempo de deslocamento segundo a Cia Aérea.

Pela Tabela 4 nota-se que o tempo de deslocamento médio de algumas Cias Aéreas é significativamente diferente do tempo médio da Cia AA. Por exemplo, o tempo médio de NW é significativamente menor enquanto o tempo médio de CO é significativamente maior. Porém, para algumas Cias Aéreas não foi possível detectar diferença significativa com AA. Isso é confirmado pelo teste F de homogeneidade de médias (vide Tabela 5), em que a hipótese nula é fortemente rejeitada. Logo, há tempos médios de deslocamento diferentes e resta saber entre quais Cias Aéreas.

Tabela 5: Tabela ANOVA referente à comparação dos tempos médios de deslocamento das Cias Aéreas.

F.Variação	S.Q.	G.L.	Q.M.	F	valor-P
Cia Aérea	1155,0	7	165,01	13,82	0,000
Resíduos	2100,9	176	11,94		
Total	3255,9	183			

Como há $m = \binom{8}{2} = 28$ pares de Cias Aéreas o método de Tukey é o

mais adequado para construir as estimativas intervalares para as diferenças das médias. Na Figura 16 tem-se um resumo das 28 estimativas intervalares com coeficiente global de confiança de 95%, construída através da biblioteca `UsingR` do R. Nota-se que 15 dessas estimativas intervalares cobrem o valor zero indicando que não foi possível detectar diferença significativa entre os deslocamentos médios das Cias Aéreas correspondentes. Por outro lado, há 13 estimativas intervalares que não cobrem o valor zero. Observando essas estimativas intervalares nota-se que as Cias Aéreas NW e US são aqueleas que mais diferem das demais no sentido de terem um tempo médio de deslocamento menor do que as demais. Isso vai ao encontro dos resultados da Tabela 4.

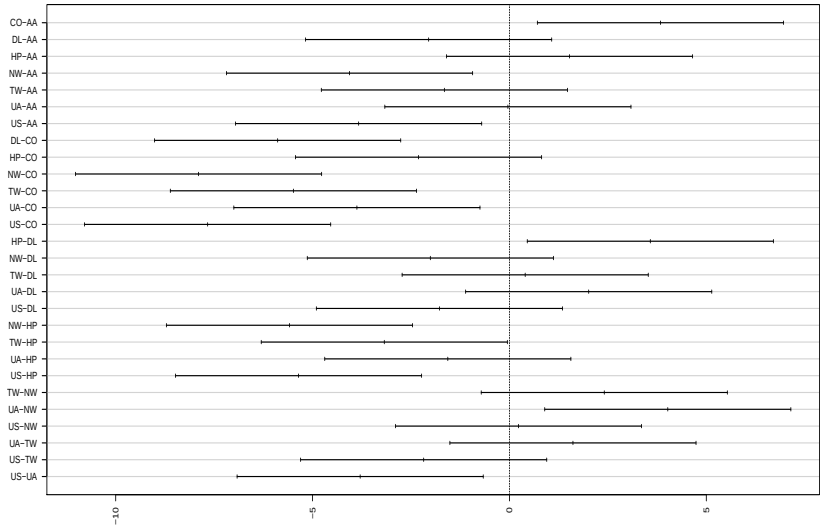

Figura 16: Estimativas intervalares para as diferenças entre os deslocamentos médios das Cias Aérea pelo método de Tukey com coeficiente global de confiança de 95%.

9 Regressão Linear Ponderada

Quando há indícios fortes de afastamentos da suposição de variâncias constantes dos erros (homocedasticidade), uma maneira de correção é através da

regressão linear ponderada em que a variância de cada erro é flexibilizada. A forma mais usual de regressão linear ponderada é a seguinte:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{5}$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, $x_{i1}, \ldots, x_{ip}$ são valores observados de variáveis explicativas e $\epsilon_i \overset{\text{ind}}{\sim} \text{N}(0, \sigma_i^2)$, com $\sigma_i^2 = \sigma^2 \omega_i$ e $\omega_i > 0$ (conhecido), para $i = 1, \ldots, n$. A soma dos quadrados dos erros (função objetivo) fica nesse caso expressa na forma

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \omega_i^{-1} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2,$$

em que em que $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. Matricialmente tem-se que

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{V} = \text{diag}\{\omega_1, \ldots, \omega_n\}$ e $\mathbf{X}$ é a matriz modelo.

Derivando a função objetivo $S(\boldsymbol{\beta})$ em relação a $\boldsymbol{\beta}$ obtém-se

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

que igualando a zero leva à seguinte solução de de mínimos quadrados ponderados:

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

Denotando $\widehat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y}$, em que $\mathbf{A} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1}$, tem-se a seguinte propriedade:

$$\begin{aligned} \text{E}(\widehat{\boldsymbol{\beta}}) &= \text{E}(\mathbf{A}\mathbf{Y}|\mathbf{X}) = \mathbf{A}\text{E}(\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

Logo, $\widehat{\boldsymbol{\beta}}$ é um estimador não tendencioso de $\boldsymbol{\beta}$. Desde que $\text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 \mathbf{V}$, segue a propriedade

$$\begin{aligned} \text{Var}(\widehat{\boldsymbol{\beta}}) &= \text{Var}(\mathbf{A}\mathbf{Y}|\mathbf{X}) = \mathbf{A}\text{Var}(\mathbf{Y}|\mathbf{X})\mathbf{A}^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{V}\mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}, \end{aligned}$$

e portanto $\widehat{\boldsymbol{\beta}} \sim \text{N}_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1})$.

As somas de quadrados ponderadas ficam expressas nas formas

$$\text{SQT} = \sum_{i=1}^{n} \omega_i^{-1}(y_i - \bar{y})^2, \ \text{SQReg} = \sum_{i=1}^{n} \omega_i^{-1}(\widehat{y}_i - \bar{y})^2 \ \text{e} \ \text{SQRes} = \sum_{i=1}^{n} \omega_i^{-1}(y_i - \widehat{y}_i)^2.$$

Similarmente ao caso homocedástico é possível mostrar que $s^2 = \frac{\text{SQRes}}{(n-p)}$ é um estimador não tendencioso de σ^2. Continuam valendo a decomposição das somas de quadrados e as interpretações do R^2 e $\overline{\text{R}}^2$.

É possível mostrar que o acréscimo na soma de quadrados de resíduos no modelo linear ponderado (5), devido às restrições lineares $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$, pode ser expresso na forma

$$\text{ASQ}(\mathbf{R}\boldsymbol{\beta} = \mathbf{0}) = (\mathbf{R}\widehat{\boldsymbol{\beta}})^\top \{\mathbf{R}(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{R}^\top\}^{-1}\mathbf{R}\widehat{\boldsymbol{\beta}},$$

em que $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y}$. Assim, se o interesse é testar $\text{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{0}$ contra $\text{H}_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{0}$, a estatística F fica dada por

$$\text{F} = \frac{\text{ASQ}(\mathbf{R}\boldsymbol{\beta} = \mathbf{0})/r}{\text{SQRes}/(n-p)} \overset{\text{H}_0}{\sim} \text{F}_{r,(n-p)}.$$

Rejeita-se H_0 se $\text{F} > \text{F}_{(1-\alpha),r,(n-p)}$, em que $\text{F}_{(1-\alpha),r,(n-p)}$ denota o quantil $(1-\alpha)$ da distribuição F com r e $(n-p)$ graus de liberdade.

9.1 Forma Equivalente

Os resultados da regressão linear ponderada (5) podem ser obtidos de forma equivalente através de uma regressão linear homocedástica aplicando as seguintes transformações:

- $z_i = y_i/\sqrt{\omega_i}$,
- $u_{ij} = x_{ij}/\sqrt{\omega_i}$,

para $i = 1, \ldots, n$ e $j = 1, \ldots, p$. Então, considere o modelo

$$z_i = \beta_1 u_{i1} + \beta_2 u_{i2} + \cdots + \beta_p u_{ip} + e_i,$$

com $e_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. Esse modelo em forma matricial fica dado por

$$\mathbf{z} = \mathbf{U}\boldsymbol{\beta} + \mathbf{e},$$

em que $\mathbf{z} = \mathbf{V}^{-\frac{1}{2}}\mathbf{y}$, $\mathbf{U} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}$ é a matriz modelo, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, enquanto $\mathbf{e} = \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\epsilon}$. Note que $\mathbf{e} \sim \text{N}_n(\mathbf{0}, \sigma^2\mathbf{I_n})$. Mostra-se facilmente que $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y}$. Assim, todos os resultados descritos nas seções anteriores podem ser estendidos facilmente para o modelo (5) através das transformações acima.

9.2 Aplicação

Como ilustração considere parte dos dados de um experimento desenvolvimento em 2006 nas Faculdades de Medicina e de Filosofia, Letras e Ciências Humanas da USP e analisado no Centro de Estatística Aplicada do IME-USP (CEA0P16) para avaliar o fluxo da fala de falantes do Português Brasileiro segundo o gênero, idade e escolaridade. Uma amostra de 595 indivíduos residentes na cidade de São Paulo com idade entre 2 e 99 anos foi avaliada segundo a fala auto-expressiva. O indivíduo era apresentado a uma figura e orientado a discorrer sobre a mesma durante um tempo mínimo de 3 minutos e máximo de 6 minutos. Para crianças de 2 e 3 anos, as amostras foram obtidas com a colaboração dos pais. As variáveis consideradas no estudo foram as seguintes: (i) idade (em anos), (ii) gênero (1:feminino, 2:masculino), (iii) interj (número de interjeições durante o discurso), (iv) fpm (fluxo de palavras por minuto) e (v) fsm (fluxo de sílabas por minuto).

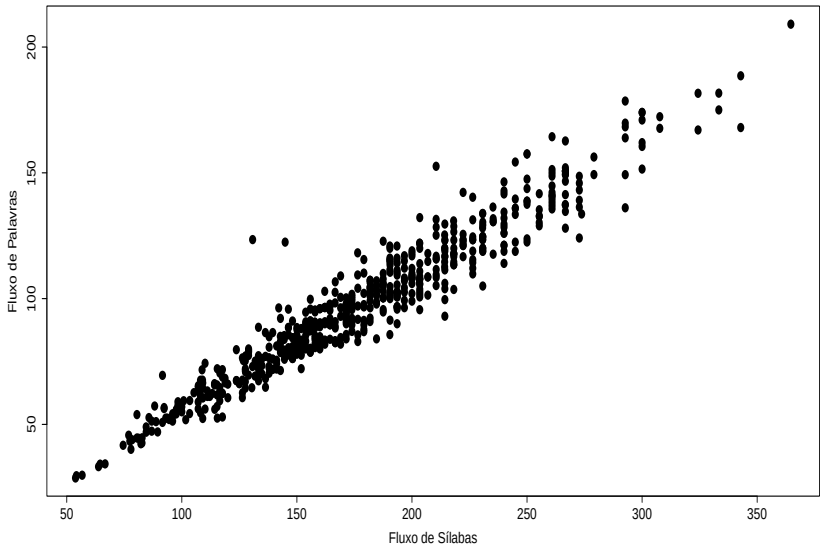

Figura 17: Diagrama de dispersão entre o fluxo de palavras por minuto e o fluxo de sílabas por minuto.

Como aplicação de regressão linear ponderada considere apenas duas variáveis, fpm e fsm. Na Figura 17 tem-se o diagrama de dispersão entre fpm e fsm e nota-se uma forte relação linear positiva e variabilidade não

constante da resposta fpm à medida que aumenta fsm. Isso sugere um modelo linear simples entre fpm e fsm. Nas Tabelas 6 e 7 tem-se as estimativas dos parâmetros do modelo

$$\text{fpm}_i = \beta_1 + \beta_2 \text{fsm}_i + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ ou $\epsilon_i \overset{\text{ind}}{\sim} \text{N}(0, \omega_i \sigma^2)$ com $\omega_i = \text{fsm}_i$, respectivamente, para $i = 1, \ldots, 594$. Nota-se uma redução na estimativa do intercepto e aumento do coeficiente de determinação sob o modelo linear ponderado. Há também um controle melhor da variabilidade sob esse modelo (Figura 18) e melhora na qualidade do ajuste (Figura 19). As três observações que aparecem destacadas como pontos aberrantes afetam muito pouco as estimativas quando são excluídas. Outros procedimentos para aprimoramento do controle da variabilidade poderiam ser aplicados, como por exemplo a modelagem dupla da média e variância.

Tabela 6: Estimativas dos parâmetros referentes ao modelo de regressão linear simples ajustado aos dados sobre fluxo da fala de falantes do Português Brasileiro.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	4,198	1,172	3,74	0,00
fsm	0,527	0,006	88,10	0,00
s	7,98			
R^2	0,93			
$\overline{\text{R}}^2$	0,93			

10 Ortogonalidade

Supor novamente o modelo de regressão linear múltipla

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, $x_{i1}, \ldots, x_{ip}$ são valores observados de variáveis explicativas e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$. Tem-se ortogonalidade entre as colunas da matriz modelo $\mathbf{X}$ se

$$\sum_{i=1}^{n} x_{ij} x_{i\ell} = 0, \ \forall j \neq \ell = 1, \ldots, p,$$

Tabela 7: Estimativas dos parâmetros referentes ao modelo de regressão linear simples ponderado ajustado aos dados sobre fluxo da fala de falantes do Português Brasileiro.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	3,663	0,974	3,76	0,00
fsm	0,530	0,006	92,57	0,00
s	0,59			
R^2	0,99			
$\overline{R}^2$	0,99			

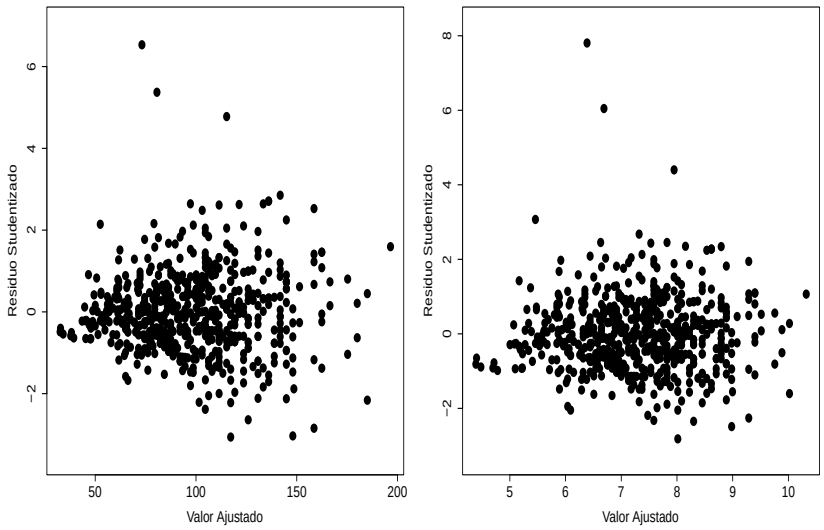

Figura 18: Gráficos entre o resíduo Studentizado e o valor ajustado referentes aos modelos homocedástico (esquerdo) e ponderado (direito) ajustados aos dados spbre fluxo da fala de falantes do Português Brasileiro.

ou seja, a matriz $\mathbf{X}^\top\mathbf{X}$ é bloco diagonal.

Quando a matriz modelo $\mathbf{X}$ tem posto coluna completo tem-se sob orto-

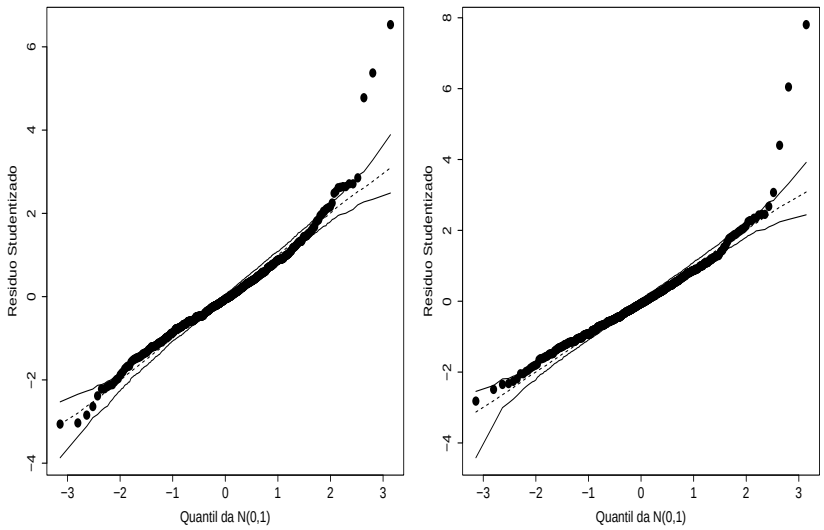

Figura 19: Gráficos normais de probabilidade com banda empírica de 95% referentes aos modelos homocedástico (esquerdo) e ponderado (direito) ajustados aos dados spbre fluxo da fala de falantes do Português Brasileiro.

gonalidade que

$$\mathbf{X}^\top\mathbf{X} = \text{diag}\{\sum_{i=1}^{n} x_{i1}^2, \ldots, \sum_{i=1}^{n} x_{ip}^2\} \quad \text{e} \quad \mathbf{X}^\top\mathbf{y} = (\sum_{i=1}^{n} x_{i1}y_i, \ldots, \sum_{i=1}^{n} x_{ip}y_i)^\top,$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$, e consequentemente

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_p \end{bmatrix} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \begin{bmatrix} \frac{\sum_{i=1}^{n} x_{i1}y_i}{\sum_{i=1}^{n} x_{i1}^2} \\ \vdots \\ \frac{\sum_{i=1}^{n} x_{ip}y_i}{\sum_{i=1}^{n} x_{ip}^2} \end{bmatrix}.$$

Logo, $\widehat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij}y_i}{\sum_{i=1}^{n} x_{ij}^2}$ depende apenas dos valores $y_1, \ldots, y_n$ e de $x_{1j}, \ldots, x_{nj}$, para $j = 1, \ldots, p$. Ou seja, dos valores da variável resposta e da variável explicativa X_j.

Além disso, a matriz de variância-covariância para $\widehat{\boldsymbol{\beta}}$ fica dada por

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} \frac{\sigma^2}{\sum_{i=1}^n x_{i1}^2} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{\sigma^2}{\sum_{i=1}^n x_{ip}^2} \end{bmatrix}.$$

Portanto, $\mathrm{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n x_{ij}^2}$ e $\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_\ell) = 0$, para $j \neq \ell$ e $j, \ell = 1, \ldots, p$. Tem-se independência mútua entre os estimadores dos coeficientes.

11 Multicolinearidade

Multicolinearidade é o oposto da ortogonalidade. Ocorre quando há uma alta correlação linear entre variáveis explicativas e consequentemente entre os estimadores dos coeficientes da regressão linear múltipla. Uma consequência prática é que $\det(\mathbf{X}^\top\mathbf{X}) \cong 0$. Agumas fontes de multicolinearidade são as seguintes:

- Método empregado na coleta de dados
  Os dados são coletados de um estrato da população onde há uma alta correlação linear entre duas variáveis explicativas. Por exemplo, num estudo de regressão em que tem-se como variáveis explicativas o consumo de um produto alimentício e o preço do produto alimentício. É razoável esperar nos estratos de renda mais baixa uma correlação mais alta entre as duas variáveis explicativas.

- Restrições no modelo ou na população
  Duas variáveis explicativas que têm uma correlação linear alta são incluídas no modelo. Por exemplo, consumo de energia elétrica e renda percapita. Notas referentes às avaliações sobre qualidade e clareza das aulas de um instrutor.

- Especificação do modelo
  No modelo são incluídos vários termos que estão em função de uma mesma variável explicativa. Por exemplo, numa regressão polinomial em que são incluídos termos $x + x^2 + x^3 + \cdots$.

- Modelo superdimensionado
  Estudos com amostras pequenas e uma grande quantidade de variáveis explicativas. Por exemplo, na área médica em geral tem-se amostras pequenas com uma grande quantidade de informações por paciente.

11.1 Efeitos da Multicoliearidade

Para ilustrar considere o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta com comprimento unitário, x_{i1} e x_{i2} são valores observados de variáveis explicativas com comprimento unitário, em que $\sum_{i=1}^n x_{ij} = 0$ e $\sum_{i=1}^n x_{ij}^2 = 1$ para $j = 1, 2$, e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$.

Para esse exemplo tem-se que

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} \\ \sum_{i=1}^n x_{i2} x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix},$$

em que r_{12} denota a correlação linear amostral entre X_1 e X_2. Além disso

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} y_i \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix},$$

em que r_{1y} e r_{2y} denotam, respectivamente, as correlações lineares amostrais entre X_1 e Y e X_2 e Y. Portanto, as estimativas de mínimos quadrados ficam dadas por

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \frac{r_{1y} - r_{12} r_{2y}}{(1 - r_{12}^2)} \\ \frac{r_{2y} - r_{12} r_{1y}}{(1 - r_{12}^2)} \end{bmatrix},$$

e dependem das correlações lineares r_{12}, r_{1y} e r_{2y}. Além disso, a matriz de variância-covariância para $\widehat{\boldsymbol{\beta}}$ assume a forma

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} \frac{\sigma^2}{(1 - r_{12}^2)} & -\frac{\sigma^2 r_{12}}{(1 - r_{12}^2)} \\ -\frac{\sigma^2 r_{12}}{(1 - r_{12}^2)} & \frac{\sigma^2}{(1 - r_{12}^2)} \end{bmatrix}.$$

Ou seja, $\text{Var}(\widehat{\beta}_1) = \text{Var}(\widehat{\beta}_2) = \frac{\sigma^2}{(1 - r_{12}^2)}$ e $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = -\frac{\sigma^2 r_{12}}{(1 - r_{12}^2)}$. E tem-se as seguintes consequências:

- Se $|r_{12}| \to 1$ então $\text{Var}(\widehat{\beta}_1)$ e $\text{Var}(\widehat{\beta}_2)$ ficam grandes.
- Se $r_{12} \to 1$ então $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \to -\infty$.
- Se $r_{12} \to -1$ então $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \to \infty$.

11.2 Procedimentos para Detectar Multicolinearidade

Fator de Inflação da Variância

É possível mostrar que

$$\mathrm{Var}(\widehat{\beta}_j) = \sigma^2 \mathrm{C}_{jj} = \sigma^2 (1 - \mathrm{R}_j^2)^{-1},$$

em que $\mathrm{C}_{j\ell}$ denota o (j,ℓ)-ésimo elemento da matriz $\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1}$ e R_j^2 denota o coeficiente de determinação da regressão linear da variável explicativa X_j contra as demais variáveis explicativas X_ℓ, em que $j \neq \ell$, para $j, \ell = 1, \ldots, p$. O fator de inflação de variância da j-ésima variável explicativa é definido por

$$\mathrm{VIF}_j = (1 - \mathrm{R}_j^2)^{-1}.$$

Assim, se $\mathrm{R}_j^2 \to 1$ então $\mathrm{VIF}_j \to \infty$, para $j = 1, \ldots, p$. Para ilustrar, supor três variáveis explicativas X_1, X_2 e X_3 cujos valores amostrais têm comprimento unitário. Os VIFs saem das seguintes regressões:

- VIF_1: da regressão $x_{i1} = \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$
- VIF_2: da regressão $x_{i2} = \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i$
- VIF_3: da regressão $x_{i3} = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, para $i = 1, \ldots, n$.

Critério: $\mathrm{VIF}_j \geq 10$ indica que $\widehat{\beta}_j$ está com variância inflacionada.

Número da Condição

Sejam $\lambda_1, \ldots, \lambda_p$ os autovalores da matrix $\mathbf{X}^\top \mathbf{X}$. Como é uma matriz simétrica positiva definida todos os seus autovalores são não negativos. Contudo, a existência de autovalores próximos de zero é indício de multicolinearidade. Uma medida resumo de multicolinearidade entre as colunas da matriz $\mathbf{X}$ é o número da condição definido por

$$k = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Portanto, se esta razão é muito grande há indícios de multicolinearidade com a matriz $\mathbf{X}^\top \mathbf{X}$.

Critério: (i) se $k \leq 100$ não há indícios de multicolinearidade, (ii) se $100 < k \leq 1000$ há indícios moderados de multicolinearidade e (iii) se $k > 1000$ há indícios fortes de multicolinearidade.

Índice da Condição

Quando há indícios de multicolinearidade através do número da condição, pode-se avaliar a contribuição de cada variável explicativa através do índice da condição definido por

$$k_j = \frac{\lambda_{\max}}{\lambda_j},$$

para $j = 1, \ldots, p$. Os mesmos critérios usados para o número da condição são usados para o índice da condição.

Determinante da Matrix $\mathbf{X}^\top\mathbf{X}$

Se as variáveis explicativas têm comprimento unitário, mostra-se que

$$0 \leq \det(\mathbf{X}^\top\mathbf{X}) \leq 1.$$

Logo, $\det(\mathbf{X}^\top\mathbf{X}) = 1$ indica ortogonalidade entre as colunas da matriz $\mathbf{X}$, enquanto $\det(\mathbf{X}^\top\mathbf{X}) = 0$ indica dependência linear entre as colunas da matrix $\mathbf{X}$. Valores próximos de zero são indícios de multicolinearidade.

11.3 Tratamentos da Multicolineridade

Alguns tratamentos para a multicolinearidade

- Coletar mais dados.
- Eliminação de variáveis explicativas.
- Transformação de variáveis explicativas.
- Regressão *ridge*.
- Regressão através de componentes principais.

Regressão *ridge*

O objetivo da regressão *ridge* é utilizar um estimador tendencioso que produza variâncias mais estáveis para os estimadores dos coeficientes da regressão. Assim, seja $\widehat{\boldsymbol{\beta}}^*$ um estimador tendencioso de $\boldsymbol{\beta}$. Mostra-se que o erro quadrático médio de $\widehat{\boldsymbol{\beta}}^*$ pode ser expresso na forma

$$\text{EQM}(\widehat{\boldsymbol{\beta}}^*) = \text{Var}(\widehat{\boldsymbol{\beta}}^*) + [\text{Viés}][\text{Viés}]^\top,$$

em que Viés $= \mathrm{E}(\widehat{\boldsymbol{\beta}}^*) - \boldsymbol{\beta}$. A fim de estabilizar as estimativas dos coeficientes da regressão linear múltipla bem com as respectivas variâncias é proposto o seguinte estimador:

$$\widehat{\boldsymbol{\beta}}_R = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^\top \mathbf{y},$$

em que $k > 0$ é uma constante desconhecida que é estimada separadamente. Em particular quando $k = 0$ recupera-se o estimador de mínimos quadrados. Estima-se k até estabilizar as estimativas dos coeficientes. Na Figura 20 tem-se um exemplo ilustrativo em que quatro coeficientes estão sendo ajustados e nota-se uma estabilidade das estimativas a partir de $k = 0,10$.

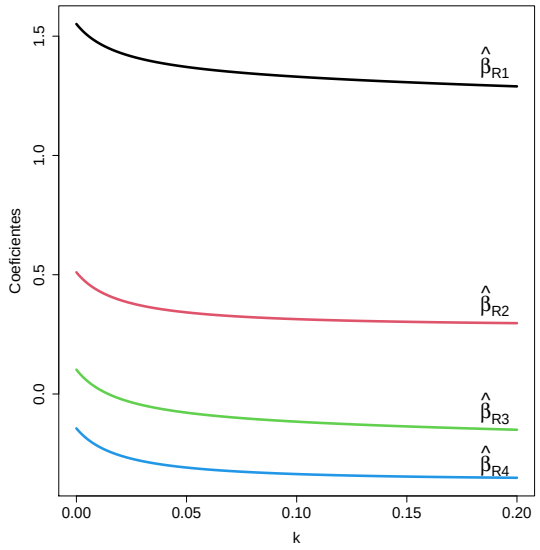

Figura 20: Ilustração dos coeficientes estimados através da regressão *ridge* variando-se o valor de k.

Denotando $\widehat{\boldsymbol{\beta}}_R = \mathbf{Z}_k\widehat{\boldsymbol{\beta}}$, em que $\mathbf{Z}_k = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I}_p)^{-1}(\mathbf{X}^\top \mathbf{X})$, tem-se as seguintes propriedades:

- $\mathrm{E}(\widehat{\boldsymbol{\beta}}_R) = \mathrm{E}(\mathbf{Z}_k\widehat{\boldsymbol{\beta}}) = \mathbf{Z}_k\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{Z}_k\boldsymbol{\beta}$.
- $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_R) = \mathrm{Var}(\mathbf{Z}_k\widehat{\boldsymbol{\beta}}) = \mathbf{Z}_k\mathrm{Var}(\widehat{\boldsymbol{\beta}})\mathbf{Z}_k^\top = \sigma^2\mathbf{Z}_k(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{Z}_k^\top$.

Em particular, se $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$ tem-se que $\mathbf{Z}_k = (1+k)^{-1}\mathbf{I}_p$. Logo, $\mathrm{E}(\widehat{\boldsymbol{\beta}}_R) = (1+k)^{-1}\boldsymbol{\beta}$ e $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_R) = \sigma^2(1+k)^{-2}\mathbf{I}_p$. Ou seja, à medida que k cresce o

estimador *ridge* fica mais tendencioso havendo um encolhimento com relação ao estimador de mínimos quadrados. A variância diminui com o aumento de k.

Tem-se ainda que $\widehat{\boldsymbol{\beta}}_R \sim \mathrm{N}_p(\mathrm{E}(\widehat{\boldsymbol{\beta}}_R), \mathrm{Var}(\widehat{\boldsymbol{\beta}}_R))$. Daí segue que $\widehat{\beta}_{R_j}$ são normais de média $\mathrm{E}(\widehat{\beta}_{R_j})$ e variância $\mathrm{Var}(\widehat{\beta}_{R_j})$, para $j = 1, \ldots, p$. É possível mostrar que

$$\begin{aligned} \mathrm{SQRes}(k) &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R)^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R) \\ &= \mathrm{SQRes} + (\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}}), \end{aligned}$$

em que SQRes denota a soma de quadrados de resíduos da regressão de mínimos quadrados. Portanto, na regressão *ridge* há um aumento na soma de quadrados de resíduos, logo uma redução no valor de R^2.

A constante k pode ser estimada através do processo iterativo

$$k^{(m+1)} = \frac{p\widehat{\sigma}^2}{\widehat{\boldsymbol{\beta}}_R^\top(k^{(m)})\widehat{\boldsymbol{\beta}}_R(k^{(m)})},$$

para $m = 0, 1, \ldots$, em que $\widehat{\sigma}^2$ é obtido através do estimador de mínimos quadrados $\widehat{\boldsymbol{\beta}}$. Para valor inicial utiliza-se o estimador de HKB (Montgomery et al., 2021, Cap.9) dado por $k^{(0)} = p\widehat{\sigma}^2/\widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\beta}}$.

Regressão dos Componentes Principais

A forma canônica da regressão linear múltipla $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ é definida por

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{Z} = \mathbf{X}\mathbf{T}$, $\boldsymbol{\alpha} = \mathbf{T}^\top\boldsymbol{\beta}$ e $\mathbf{Z}^\top\mathbf{Z} = \mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T} = \boldsymbol{\Lambda}$, com $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ sendo a matriz diagonal $p \times p$ com os autovalores da matriz $\mathbf{X}^\top\mathbf{X}$ e $\mathbf{T}$ a matriz $p \times p$ cujas colunas são os autovetores ortonormais (ortogonais com comprimento unitário) correspondentes aos autovalores $\lambda_1, \ldots, \lambda_p$. Sugere-se que $\mathbf{y}$ e a matriz $\mathbf{X}$ sejam centralizadas, assim não precisa de intercepto.

Portanto, a estimativa de mínimos quadrados de $\boldsymbol{\alpha}$ fica dada por

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} &= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y} \\ &= \boldsymbol{\Lambda}^{-1}\mathbf{Z}^\top\mathbf{y}, \end{aligned}$$

com matriz de variância-covariância expressa na forma

$$\begin{aligned} \mathrm{Var}(\widehat{\boldsymbol{\alpha}}) &= \sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1} \\ &= \sigma^2\boldsymbol{\Lambda}^{-1}. \end{aligned}$$

Daí segue que $\text{Var}(\widehat{\alpha}_j) = \sigma^2\lambda_j^{-1}$. Assim, λ_j próximo de zero inflaciona a variância de $\widehat{\alpha}_j$. Similarmente, segue que a matriz de variância-covariância de $\widehat{\boldsymbol{\beta}}$ pode ser expressa na forma

$$\begin{aligned}\text{Var}(\widehat{\boldsymbol{\beta}}) &= \text{Var}(\mathbf{T}\widehat{\boldsymbol{\alpha}}) \\ &= \mathbf{T}\text{Var}(\widehat{\boldsymbol{\alpha}})\mathbf{T}^\top \\ &= \sigma^2\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}^\top.\end{aligned}$$

E daí pode-se mostrar que $\text{Var}(\widehat{\beta}_j) = \sigma^2\sum_{\ell=1}^{p} t_{j\ell}^2/\lambda_\ell$, em que $t_{j\ell}$ denota o (j,ℓ)-ésimo elemento da matriz $\mathbf{T}$. Esse resultado confirma o efeito de autovalores próximos de zero na inflação da variância de $\widehat{\beta}_j$.

A partir da relação $\widehat{\boldsymbol{\beta}} = \mathbf{T}\widehat{\boldsymbol{\alpha}}$, a proposta da regressão dos componentes principais é considerar os coeficientes estimados

$$\widehat{\boldsymbol{\beta}}^{CP} = \mathbf{T}\widehat{\boldsymbol{\alpha}}^{CP},$$

em que $\widehat{\boldsymbol{\alpha}}^{CP}$ é um vetor $p \times 1$ que contém os coeficientes estimados correspondentes aos $p - s$ maiores autovalores da matriz $\mathbf{X}^\top\mathbf{X}$ e os demais s coeficientes como sendo iguais a zero. Assim, os novos coeficientes estimados $\widehat{\beta}_1^{CP}, \ldots, \widehat{\beta}_p^{CP}$ irão depender apenas das variáveis explicativas com menor potencial de estarem causando multicolinearidade. Esses coeficients estimados são interpretados de forma similar aos coeficientes estimados por mínimos quadrados.

Da relação $\mathbf{Z} = \mathbf{XT}$ segue que $\mathbf{Z}_j = \sum_{\ell=1}^{p}\mathbf{X}_\ell t_{\ell j}$, em que $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$ e $\mathbf{X}_1, \ldots, \mathbf{X}_p$ denotam, respectivamente, as colunas de $\mathbf{Z}$ e $\mathbf{X}$, enquanto $t_{1j}, \ldots, t_{pj}$ denotam os componentes do autovetor correspondente ao autovalor λ_j. Assim, se λ_j for próximo de zero os componentes de $\mathbf{Z}_j$ devem ser aproximadamente constantes. Deve-se portanto escolher os $p-s$ componentes principais $\mathbf{Z}_1, \ldots, \mathbf{Z}_{(p-s)}$ que correspondem aos $p-s$ maiores autovalores.

11.4 Aplicação

Como ilustração para o tópico de multicolinearidade será analisado um conjunto de dados proposto em Montgomery et al. (2021, Tabela B.21) em que o calor (em calorias por grama) de $n = 13$ amostras de cimento é relacionado com as seguintes variáveis explicativas referentes a ingredientes usados na mistura do cimento:

- X_1: aluminato tricálcico
- X_2: silicato tricálcico

- X_3: aluminato-ferrita tetracálcico
- X_4: silicato dicálcico.

Tabela 8: Matriz de correlações lineares amostrais de Pearson entre as variáveis do exemplo sobre o calor do cimento em amostras de cimento.

	Calor	X_1	X_2	X_3	X_4
Calor	1,00	0,73	0,82	-0,54	-0,82
X_1		1,00	0,23	-0,82	-0,25
X_2			1,00	-0,14	-0,97
X_3				1,00	0,03
X_4					1,00

Nota-se pela Tabela 8 correlações lineares altas entre a resposta calor do cimento e as variáveis explicativas X_2 e X_4, enquanto entre as variáveis explicativas nota-se correlação liner muito alta entre X_2 e X_4, indicando possível multicolineridade nos dados. Nota-se pelo boxplot robusto da Figura 21 que a distribuição da variável resposta é aproximadamente simétrica, enquanto os diagramas de dispersão da Figura 22 confirmam as correlações lineares descritas na Tabela 8.

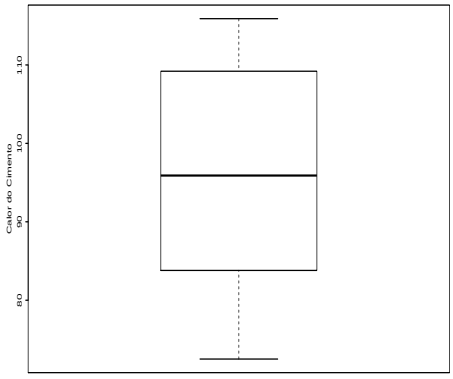

Figura 21: Boxplot robusto da variável resposta calor do cimento.

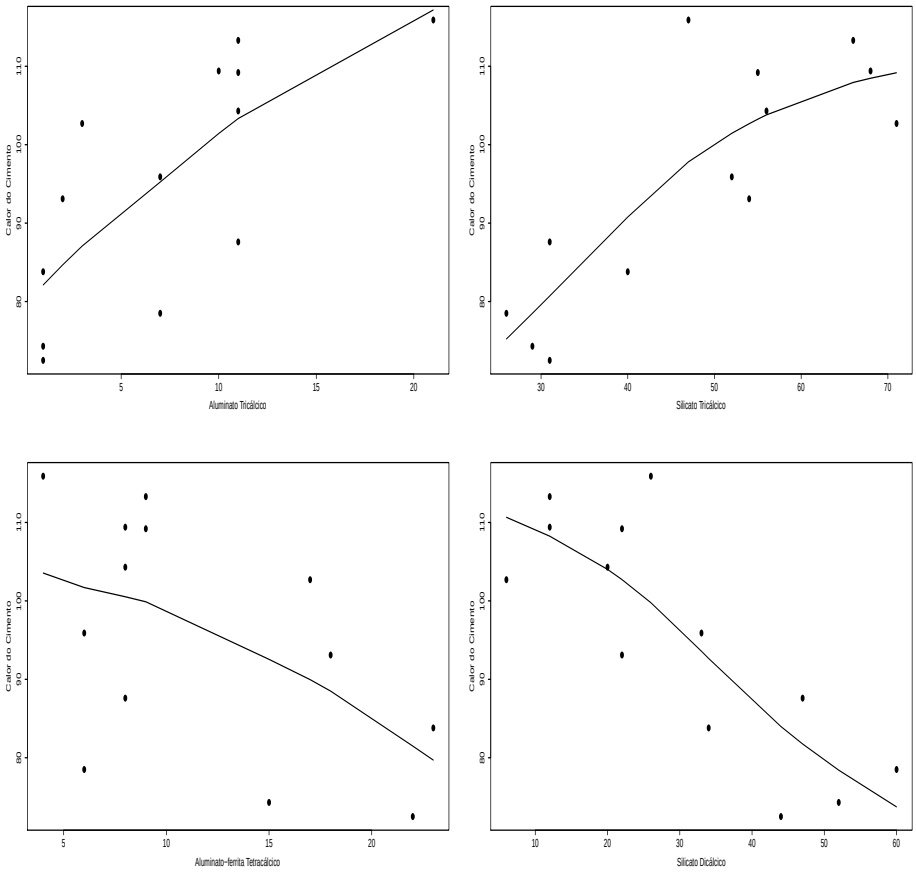

Figura 22: Diagramas de dispersão (com tendência) entre a variável resposta calor do cimento e as demais variáveis explicativas.

Com base nos diagramas de dispersão o seguinte modelo é proposto:

$$cy_i = \beta_1 \mathrm{cx}_{i1} + \beta_2 \mathrm{cx}_{i2} + \beta_3 \mathrm{cx}_{i3} + \beta_4 \mathrm{cx}_{i4} + \epsilon_i,$$

em que cy_i denota o calor da i-ésima amostra de cimento centralizada (subtraído da média amostral), bem como os valores das variáveis explicativas e $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, para $i = 1, \ldots, 13$. Dessa forma, não é necessário incluir o intercepto.

Pela Tabela 9 apenas a variável X_1 é marginalmente significativa. Os gráficos de resíduos são apresentados na Figura 23, não havendo indícios de

afastamentos da normalidade, de presença de observações aberrantes e de variância não constante dos erros. Como a amostra é pequena a suposição de normalidade dos erros é crucial para fazer inferência. A observação #8 aparece como possivelmente influente no gráfico da distância de Cook com $k = 2$ (Figura 24). Quando essa observação não é considerada na regressão o valor-P correspondente à estimativa do coeficiente da variável X_1 reduz para 0,02, porém os demais coeficientes continuam não significativos e todos com sinal positivo.

Tabela 9: Estimativas dos parâmetros referentes ao modelo de regressão linear ajustado aos dados sobre o calor do cimento em amostras de cimento.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
cx_1	1,551	0,702	2,21	0,06
cx_2	0,510	0,602	0,75	0,47
cx_3	0,102	0,716	0,14	0,89
cx_4	-0,144	0,669	-0,22	0,83
s	2,31			
R^2	0,98			
$\overline{R}^2$	0,97			

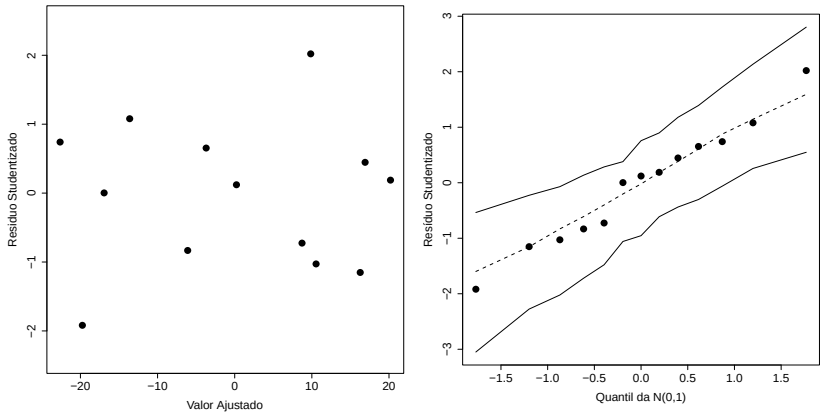

Figura 23: Gráficos de resíduos referentes ao ajuste do modelo de regressão linear aos dados sobre o calor do cimento em amostras de cimento.

Na Tabela 10 tem-se os VIFs corresponentes às 4 variáveis explicativos, confirmando os indícios de multicolinearidade. As estimativas da regressão *ridge* com $k = 0,076$ (vide comportamento dos coeficientes etimados na Figura 20) apresenta estimativas mais coerentes com a análise descritiva, porém apenas a variável explicativa X_1 é marginalmente significativa. Os autovalores da matriz $\mathbf{X}^\top\mathbf{X}$ são respectivamente dados por $\lambda_1 = 6213,56$, $\lambda_2 = 809,96$, $\lambda_3 = 148,86$ e $\lambda_4 = 2,84$ com autovalores ortonormais dados abaixo.

$\mathbf{T}_1$	$\mathbf{T}_2$	$\mathbf{T}_3$	$\mathbf{T}_4$
-0,067800	0,646018	-0,567315	0,506180
-0,678516	0,019993	0,543969	0.493268
0,029021	-0,755310	-0,403554	0,515567
0,730874	0,108480	0,468398	0,484416

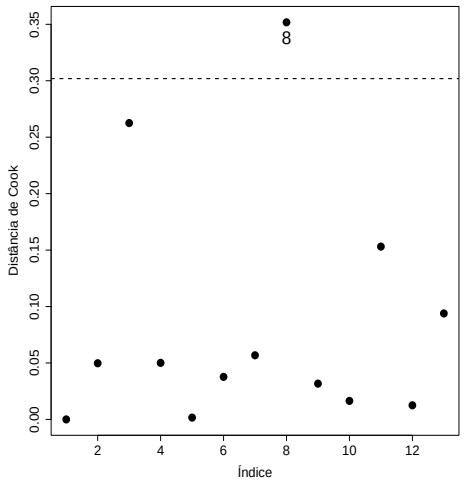

Figura 24: Gráfico da distância de Cook contra a ordem das observações referente ao ajuste do modelo de regressão linear aos dados sobre o calor do cimento em amostras de cimento.

Considerando apenas o primeiro componente principal, que explica 86,60%, tem-se a seguinte relação:

$$z_1 = -0,067800cx_1 - 0,678516cx_2 + 0,029021cx_3 + 0,730874cx_4.$$

Com base nos diagramas de dispersão da Figura 22, o componente z_1 aumenta à medida que os valores das variáveis explicativas diminuem. O modelo na forma canônica fica dado por

$$cy_i = z_{i1}\alpha + \epsilon_i,$$

em que cy_i denota o calor da i-ésima amostra de cimento centralizado e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, 13$. Desse ajuste obtém-se $\widehat{\alpha} = -0,5537(0,1043)$, que é altamente significativo. Assim, espera-se aumento do calor do cimento à medida que aumenta z_1.

Tabela 10: Fator de inflação da variância das variáveis explicativas do modelo de regressão linear ajustado aos dados sobre o calor do cimento em amostras de cimento.

Variável	VIF
cx_1	38,49
cx_2	254,42
cx_3	46,87
cx_4	282,51

Tabela 11: Estimativas dos parâmetros referentes ao modelo de regressão *ridge* ajustado aos dados sobre o calor do cimento em amostras de cimento.

Efeito	Estimativa	Erro padrão	valor-z
cx_1	1,3460	0,6844	1,967
cx_2	0,3236	0,6651	0,486
cx_3	-0,1018	0,6934	-0,147
cx_4	-0,3263	0,6514	-0,501

12 Seleção de Modelos

A seleção de modelos consiste em uma etapa importante e também complexa na análise de regressão, principalmente quando há um grande número de variáveis explicativas candidatas a entrarem no modelo. O fato das variáveis

explicativas em geral estarem correlacionadas dificulta a seleção de um subconjunto de coeficientes que além de serem significativos sejam de fácil interpretação. Sabe-se que a omissão de coeficientes significativos pode levar a estimativas tendenciosas para os demais coeficientes da regressão. Assim, a seleção de modelos pode ser considerado um procedimentos que envolve técnica e bom senso. Nesta seção serão apresentados alguns procedimentos tradicionais de seleção de modelos em regressão linear múltipla.

12.1 Todas Regressões Possíveis

Supor um total de $(p-1)$ variáveis explicativas a serem selecionadas num modelo de regressão e seja T o total de regressões possíveis. Tem-se que

$$T = 1 + \binom{p-1}{1} + \binom{p-1}{2} + \cdots + \binom{p-1}{p-1} = 2^{(p-1)}.$$

Por exemplo, se $p = 4$ (3 variáveis explicativas), haverá um total de $T = 1 + 3 + 3 + 1 = 8$ regressões possíveis.

Maior $\mathbf{R}_k^2$

Seja R_k^2 o coeficiente de determinação de um submodelo com k coeficientes ($(k-1)$ variáveis explicativas + intercepto), definido por

$$\begin{aligned} \mathrm{R}_k^2 &= \frac{\mathrm{SQReg}(k)}{\mathrm{SQT}} \\ &= 1 - \frac{\mathrm{SQRes}(k)}{\mathrm{SQT}}. \end{aligned}$$

Esse critério procura um submodelo com R_k^2 alto e k pequeno (vide Figura 25). Alternativamente, denote por $\overline{\mathrm{R}}_k^2$ o coeficiente de determinação ajustado do submodelo com k coeficientes. Tem-se que

$$\overline{\mathrm{R}}_k^2 = 1 - (1 - \mathrm{R}_k^2)\frac{(n-1)}{(n-k)}.$$

Pode-se adotar como critério a escolha de um submodelo com $\overline{\mathrm{R}}_k^2$ alto e k pequeno. Contudo, $\overline{\mathrm{R}}_k^2$ não necessariamente cresce com k.

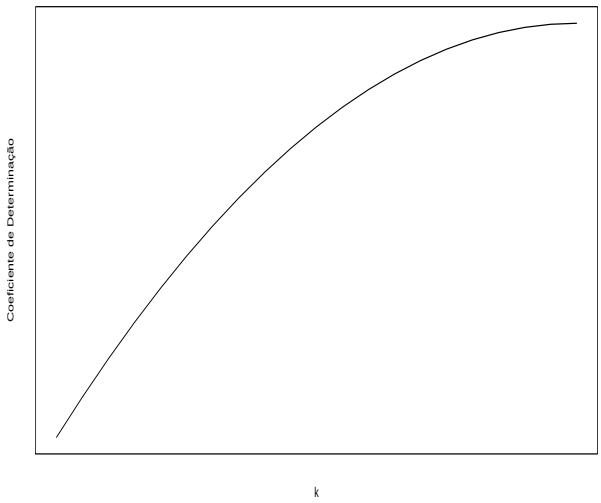

Figura 25: Comportamento do coeficiente de determinação R_k^2 com o número k de coeficientes.

Menor s_k^2

Seja s_k^2 o erro quadrático médio de um submodelo com k, sendo denotado por

$$s_k^2 = \frac{\mathrm{SQRes}(k)}{n-k}.$$

Esse critério procura um submodelo com s_k^2 pequeno e k pequeno. Conforme descrito pela Figura 26 nem sempre o erro quadrático médio decresce com o aumento do número de coeficientes.

Mostra-se que

$$\begin{aligned}
\overline{\mathrm{R}}_k^2 &= 1 - \frac{(n-1)}{(n-k)}(1-\mathrm{R}_k^2) \\
&= 1 - \frac{(n-1)}{(n-k)}\left\{1 - \frac{\mathrm{SQReg}(k)}{\mathrm{SQT}}\right\} \\
&= 1 - \frac{(n-1)}{(n-k)}\frac{\mathrm{SQRes}(k)}{\mathrm{SQT}} \\
&= 1 - \frac{(n-1)}{\mathrm{SQT}}s_k^2.
\end{aligned}$$

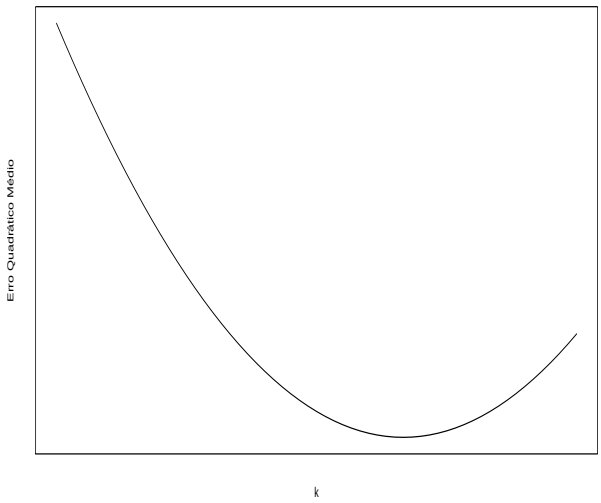

Figura 26: Comportamento do erro quadrático médio s_k^2 com o número k de coeficientes.

Assim, minimizar s_k^2 é equivalente a maximizar $\overline{\mathrm{R}}_k^2$.

Critério de Mallows

Um outro método, conhecido como critério de Mallows, está relacionado com o erro quadrático médio do i-ésimo valor ajustado $\widehat{Y}_i$ do submodelo com k coeficientes

$$\mathrm{E}\{\widehat{Y}_i - \mathrm{E}(Y_i)\}^2 = \mathrm{Var}(\widehat{Y}_i) + \{\mathrm{E}(\widehat{Y}_i) - \mathrm{E}(Y_i)\}^2.$$

A soma dos vieses ao quadrado do submodelo com k coeficientes fica dada por

$$\{\mathrm{Viés}(k)\}^2 = \sum_{i=1}^{n} \{\mathrm{E}(\widehat{Y}_i) - \mathrm{E}(Y_i)\}^2,$$

em que $\mathrm{E}(Y_i)$ denota o valor esperado do modelo correto. Uma forma padronizada para o erro quadrático médio do submodelo com k coeficientes é expressa na forma

$$\mathrm{EQM}(k) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} \{\mathrm{E}(\widehat{Y}_i) - \mathrm{E}(Y_i)\}^2 + \sum_{i=1}^{n} \mathrm{Var}(\widehat{Y}_i)\right].$$

Usando o resultado $\sum_{i=1}^{n} \mathrm{Var}(\widehat{Y}_i) = k\sigma^2$ obtém-se

$$\mathrm{EQM}(k) = \frac{\{\mathrm{Viés}(k)\}^2}{\sigma^2} + k.$$

Por outro lado

$$\mathrm{E}\{\mathrm{SQRes}(k)\} = \{\mathrm{Viés}(k)\}^2 + (n-k)\sigma^2.$$

Portanto, o erro quadrático médio padronizado assume a forma

$$\mathrm{EQM}(k) = \frac{\mathrm{E}\{\mathrm{SQRes}(k)\}}{\sigma^2} - n + 2k.$$

Deve-se escolher submodelos com $\mathrm{EQM}(k)$ pequeno.

A estatística C_k de Mallows é definida por

$$\mathrm{C}_k = \frac{\mathrm{SQRes}(k)}{\widehat{\sigma}^2} - n + 2k,$$

em que $\widehat{\sigma}^2$ deve ser obtido de um modelo bem ajustado. Sob viés zero tem-se que

$$\mathrm{E}(\mathrm{C}_k|\mathrm{Viés} = 0) = \frac{(n-k)\sigma^2}{\sigma^2} - n + 2k = k.$$

Portanto, deve-se escolher submodelos com C_k pequenos tais que $\mathrm{C}_k \cong k$. Para um mesmo k, submodelos com $\mathrm{C}_k < k$ têm uma SQRes menor, enquanto submodelos com $\mathrm{C}_k > k$ têm uma SQRes maior.

Na Figura 27 são ilustrados 3 submodelos hipóteticos, A, B e C. O submodelo A é o pior submodelo, tem C_k alto e viés alto. O submodelo B tem um C_k menor e viés pequeno. Já o submodelo C tem um viés um pouco maior do que o submodelo B, porém um C_k bem menor, assim poderia ser o submodelo escolhido.

Critério Press

Finalmente, tem-se o critério Press que consiste em escolher o submodelo com o menor valor para a estatística

$$\mathrm{Press}_k = \sum_{i=1}^{n} \{y_i - \widehat{y}_{(i)}\}^2,$$

em que $\widehat{y}_{(i)}$ denota o valor predito para y_i do ajuste do submodelo com k coeficientes sem a i-ésima observação. Desde que $\widehat{y}_{(i)} = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{(i)}$, usando a

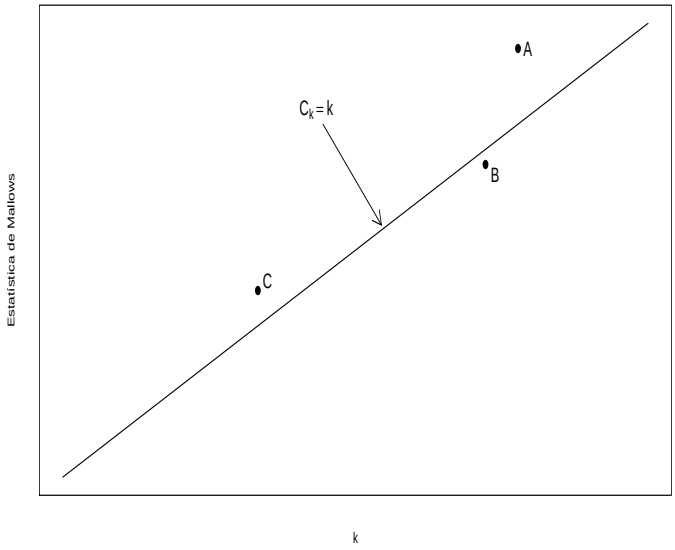

Figura 27: Descrição da reta $C_k = k$ e da estatística de Mallows para três submodelos hipotéticos A, B e C.

expressão para $\widehat{\boldsymbol{\beta}}_{(i)}$ descrita na Seção 6.5 obtém-se

$$\text{Press}_k = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}} \right)^2,$$

em que r_i e h_{ii} denotam, respectivamente, o i-ésimo resíduo ordinário e i-ésima medida de alavanca do submodelo com k coeficientes. Como a estatística Press_k cresce com o tamanho amostral n, uma proposta alternativa é considerar a estatística $\overline{\text{Press}}_k = \text{Press}_k/n$.

Assim, a fim de selecionar um submodelo usando os critérios: R_k^2 maior, s_K^2 menor, $C_k \cong k$ e pequeno e menor Press_k, deve-se ajustar todas as $T = 2^{(p-1)}$ regressões possíveis e selecionar um submodelo seguindo os 4 critérios descritos.

12.2 Métodos Sequenciais

Critérios de Akaike e de Schwartz

Seja $\mathrm{L}(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança de um modelo de regressão com p coeficientes a serem estimados. O método de Akaike consiste em escolher um submodelo que maximize $\mathrm{L}(\boldsymbol{\theta})$ minimizando o número de coeficientes. Isso é equivalente a minimizar a função penalizada abaixo

$$\mathrm{AIC}_k = -2\mathrm{L}(\widehat{\boldsymbol{\theta}}) + 2k,$$

em que $1 \leq k \leq p$ denota o número de coeficientes do submodelo. No caso de regressão linear múltipla mostra-se que $\mathrm{AIC}_k = n \log\left(\frac{\mathrm{SQRes}}{n}\right) + 2k$ (vide Exercício 10). Similarmente ao método de Akaike o método de Schwartz consiste em maximizar $\mathrm{L}(\boldsymbol{\theta})$ também minimizando o número de coeficientes da regressão, porém com uma penalização diferente. O método é equivalente a minimizar a função abaixo

$$\mathrm{BIC}_k = -2\mathrm{L}(\widehat{\boldsymbol{\theta}}) + k\log(n).$$

Para a regressão linear múltipla tem-se que $\mathrm{BIC}_k = n \log\left(\frac{\mathrm{SQRes}}{n}\right) + k\log(n)$.

Método LASSO

O método LASSO é utilizado para a seleção de variáveis explicativas (na forma padronizada) eliminando coeficientes da regressão cujas estimativas estejam próximas de zero. No contexto de mínimos quadrados o método é equivalente a minimizar a função abaixo

$$\mathrm{S}(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=2}^{p} |\beta_j|,$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\lambda \geq 0$ é o parâmetro de penalização. Quando $\lambda = 0$ tem-se o método de mínimos quadrados e quando $\lambda \to \infty$ todos os coeficientes tendem a zero.

Critério *Forward*

Passo 1

Ajustar todas as regressões possíveis com apenas 1 variável explicativa. Isto é, ajustar as regressões

$$y_i = \beta_1 + \beta_j x_{ij} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$ e $j = 2, \ldots, p$. Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$\text{F}_j = \frac{\text{SQReg}(x_j)}{s^2(x_j)} \overset{\text{H}_0}{\sim} F_{1,(n-2)}.$$

Denote P_j o valor-P do teste. Seja $P_{\min} = \min\{P_2, \ldots, P_p\}$. Se $P_{\min} \leq P_E$ então a variável explicativa correspondente entra no modelo. Supor que X_2 entra no modelo.

Passo 2

Ajustar todas as regressões possíveis com apenas X_2 mais uma variável explicativa. Isto é, ajustar as regressões

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_j x_{ij} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$ e $j = 3, \ldots, p$. Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$\text{F}_j = \frac{\text{SQReg}(x_j|x_2)}{s^2(x_2, x_j)} \overset{\text{H}_0}{\sim} F_{1,(n-3)}.$$

Denote P_j o valor-P do teste. Seja $P_{\min} = \min\{P_3, \ldots, P_p\}$. Se $P_{\min} \leq P_E$ então a variável explicativa correspondente entra no modelo. Supor que X_3 entra no modelo.

Passo 3

Ajustar todas as regressões possíveis com apenas X_2 e X_3 mais uma variável explicativa. Isto é, ajustar as regressões

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_j x_{ij} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$ e $j = 4, \ldots, p$. Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$\text{F}_j = \frac{\text{SQReg}(x_j|x_2, x_3)}{s^2(x_2, x_3, x_j)} \overset{\text{H}_0}{\sim} F_{1,(n-4)}.$$

Denote P_j o valor-P do teste. Seja $P_{\min} = \min\{P_4, \ldots, P_p\}$. Se $P_{\min} \leq P_E$ então a variável explicativa correspondente entra no modelo. Se $P_{\min} > P_E$ parar o processo, nenhuma variável entra no modelo.

Critério *Backward*

Passo 1

Ajustar a regressão com todas as variáveis explicativas. Isto é, ajustar o seguinte modelo:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$\text{F}_j = \frac{\text{SQReg}(x_j|\text{demais})}{s^2(x_2, \ldots, x_p)} \overset{\text{H}_0}{\sim} F_{1,(n-p)}.$$

Denote P_j o valor-P do teste, para $j = 2, \ldots, p$. Seja $P_{\max} = \max\{P_2, \ldots, P_p\}$. Se $P_{\max} \geq P_S$ então a variável explicativa correspondente sai do modelo. Supor que X_2 sai do modelo.

Passo 2

Ajustar a regressão sem a variável explicativa X_2. Isto é, ajustar o seguinte modelo:

$$y_i = \beta_1 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$\text{F}_j = \frac{\text{SQReg}(x_j|\text{demais})}{s^2(x_3, \ldots, x_p)} \overset{\text{H}_0}{\sim} F_{1,(n-p-1)}.$$

Denote P_j o valor-P do teste, para $j = 3, \ldots, p$. Seja $P_{\max} = \max\{P_3, \ldots, P_p\}$. Se $P_{\max} < P_S$ o processo é terminado, nenhuma variável sai do modelo.

Critério *Stepwise*

O critério *stepwise* é uma combinação dos critérios *forward* e *backward*.

Passo 1

Ajustar todas as regressões com apenas uma variável explicativa, além do intercepto. Verificar se alguma variável explicativa entra no modelo. Supor que X_2 entrou no modelo.

Passo 2

Ajustar todas as regressões com X_2 mais uma variável explicativa, além do intercepto. Verificar se alguma variável explicativa entra no modelo. Supor que X_3 entrou no modelo. Verificar se X_2 sai do modelo dado que X_3 está no modelo.

Passo 3

O processo *stepwise* deve continuar até que não seja possível incluir nenhuma variável no modelo, nem retirar nenhuma variável do modelo.

Critérios de Parada

Não há um consenso na área de regressão a respeito de critérios de parada para os processos sequenciais. Alguns critérios mais utilizados:

(i) usar $F_E = F_S = 4$ que equivale aproximadamente a usar $P_E = P_S = 0,05$;

(ii) ser mais flexível na entrada do que na saída $P_E = 0,25$ e $P_S = 0,10$, ou com os mesmos critérios na entrada e na saída $P_E = P_S = 0,15$.

12.3 Estratégias para a Seleção de Modelos

Portanto, não há uma receita pronta para a seleção de modelos a partir de um conjunto de variáveis explicativas. Em Montgomery et al. (2021, Seção 10.3) há uma longa discussão a respeito de possíveis estratégias para seleção de modelos através dos critérios propostos nesta seção.

Segundo os autores, quando o número de variáveis explicativas é relativamente pequeno pode ser factível ajustar todas as regressões possíves e selecionar algumas candidatas segundo os critérios R_k^2 maior, s_K^2 menor, $\mathrm{C}_k \cong k$ e pequeno e menor $\overline{\mathrm{Press}}_k$. Para as regressões selecionadas sugere-se fazer uma análise de diagnóstico e levar em conta aspectos como a importância, custo e facilidade de interpretação das variáveis explicativas, bem como da capacidade de predição do modelo.

Os métodos sequenciais *forward*, *backward* e *stepwise* são recomendados quando há um número médio ou alto de variáveis explicativas, contudo exigem os níveis de significância de entrada e saída das variáveis explicativas. Já os métodos de Akaike e de Schwartz são mais recomendados quando há um grande número de variáveis explicativas no sentido de se fazer uma pré-seleção de variáveis sem a necessidade de estabelecer níveis de significância.

Todos os métodos sequencias podem ser combinados com o ajuste de todas as regressões possíveis.

A seleção de modelos pode ficar mais complexa quando há interesse em selecionar variáveis explicativas que estejam relacionadas no sentido causa-efeito com a resposta, como ocorre por exemplo na área médica. Nesses casos, os algoritmos em geral são combinações de procedimentos sequenciais com procedimentos que procuram evitar a eliminação precoce de variáveis explicativas potenciais no sentido causa-efeito. Em Dunkler et al. (2014) há uma proposta de algoritmo híbrido que combina o procedimento de eliminação *backward* com procedimentos que levam em conta o efeito da eliminação de variáveis explicativas nos coeficientes das variáveuis mantidas no modelo.

13 Aplicações

13.1 Venda de Telhados

Considere novamente os dados descritos em Neter et al. (1996, p.449) referentes à venda no ano anterior de um tipo de telhado de madeira em $n = 26$ filiais de uma rede de lojas de construção civil, agora com as seguintes variáveis:

(i) Telhados: total de telhados vendidos (em mil metros quadrados),

(ii) Nclientes: número de clientes cadastrados na loja (em milhares),

(iii) Gastos: gastos pela loja com promoções do produto (em mil USD),

(iv) Marcas: número de marcas concorrentes do produto e

(v) Potencial: potencial da loja (quanto maior o valor maior o potencial).

O interesse é explicar o número médio de telhados vendidos dadas as demais variáveis. Na Tabela 12 tem-se as estimativas da correlação linear de Pearson entre as variáveis do exemplo vendas de telhados. Nota-se uma baixa correlação entre telhados e gastos, altas correlações entre telhados com número de clientes e marcas e uma correlação moderada com potencial da loja. Entre as variáveis explicativas nota-se correlações baixas, exceto uma correlação moderada entre número de clientes e potencial da loja. As correlações descritas na Tabela 12 estão coerentes com os diagramas de dispersão apresentados nas Figuras 28 e 29.

Tabela 12: Matriz de correlações lineares amostrais de Pearson entre as variáveis do exemplo vendas de telhados.

	Telhados	Gastos	Nclientes	Marcas	Potencial
Telhados	1,0	0,159	0,783	-0,833	0.407
Gastos		1,0	0,173	-0,038	-0,070
Nclientes			1,0	-0,324	0,468
Marcas				1,0	-0,202
Potencial					1,0

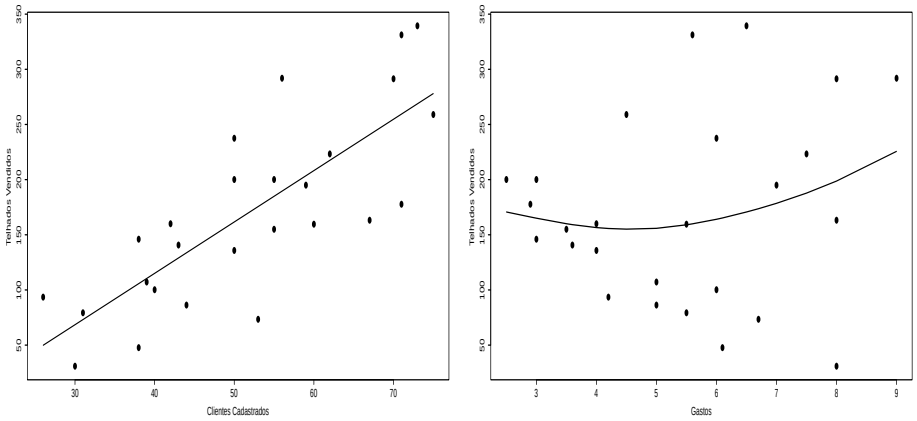

Figura 28: Diagramas de dispersão (com tendência) entre o total de telhados vendidos e o número de clientes cadastrados (esquerda) e gastos pela loja com promoções (direita).

O primeiro critério a ser aplicado para selecionar um submodelo linear normal é com todas as regressões possíveis, cujos resultados das medidas resumo são apresentados na Tabela 13. Dois submodelos se destacam segundo os 4 critérios utilizados: 1 + Nclientes + Marcas e 1 + Gastos + Ncientes + Marcas. Levando-se em conta o número de variáveis explicativas o submodelo 1 + Nclientes + Marcas poderia ser escolhido, contudo deve-se fazer antes uma análise de diagnóstico com cada submodelo.

Os dois submodelos selecionados 1 + Nclientes + Marcas e 1 + Gastos + Ncientes + Marcas apresentaram excelentes ajustes, conforme pode

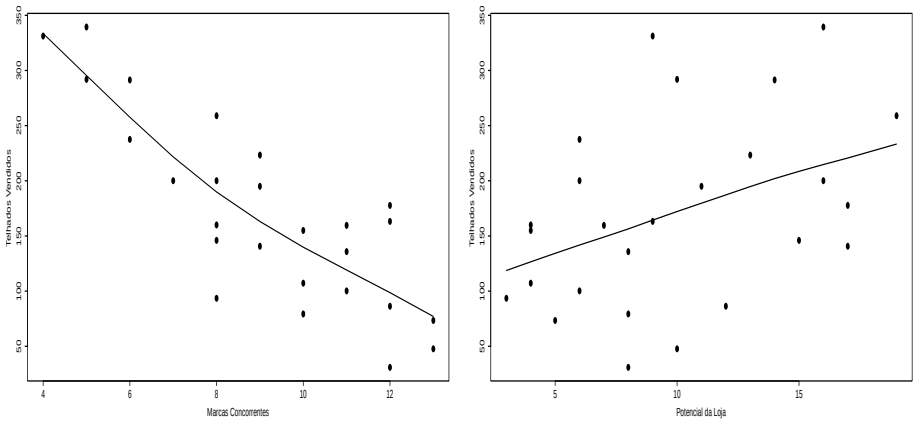

Figura 29: Diagramas de dispersão (com tendência) entre o total de telhados vendidos e o número de marcas concorrentes do produto (esquerda) e o potencial da loja (direita).

ser observado pelas Tabelas 14 e 15 e pelos gráficos de resíduos descritos nas Figuras 30 e 31. Porém, a variável explicativa gastos aparece marginalmente não significativa no 2º submodelo. Ambos os submodelos destacam os mesmos pontos potencialmente influentes pela distância de Cook com $k = 2$ (Figura 32). A eliminação da observação #21 deixa a variável explicativa gastos significativa ao nível de 5% no 2º submodelo. Portanto, essa observação está mascarando o efeito de gastos. Assim, deve-se escolher o submodelo 1 + Gastos + Nclientes + Marcas.

O segundo critério a ser aplicado é o método sequencial *stepwise* com $P_E = P_S = 0,15$. Na Tabela 16 tem-se um resumo dos 6 passos necessários para selecionar um submodelo. No 1º passo entra a variável marcas e no 2º passo entra a variável número de clientes. No 3º passo a variável marcas não sai do modelo. Já no 4º passo entra no modelo a variável gastos e no 5º passo nenhuma variável sai do modelo e finalmente no 6º passo a última variável potencial não entra no modelo. Assim, o submodelo selecionado pelo procedimento *stepwise* coincide com o submodelo selecionado pelo critério com todas as regrossões possíveis.

Finalmente, aplicando o critério de Akaike obtém-se como menor valor `AIC` $= 120,67$, que corresponde ao mesmo submodelo obtido com os dois procedimentos anteriores. Portanto, o submodelo selecionado contém as variáveis explicativas gastos, número de clientes e marcas, além da cons-

Tabela 13: Medidas resumo dos 16 submodelos para explicar o número médio de telhados vendidos, em que T:Telhados, G:Gastos, N:Nclientes, M:Marcas, P:Potencial e k denota o número de parâmetros.

Submodelo[1]	$k-1$	k	R_k^2	s_k	C_k	$\overline{Press}_k$
1	0	1	0,00	84,6	1960,2	7434,5
1 + G	1	2	0,025	85,2	1912,1	7829,8
1 + N	1	2	0,613	53,7	746,2	3115,0
1 + M	1	2	0,694	47,8	585,4	2428,8
1 + P	1	2	0,166	78.8	1633,1	6522,2
1 + G + N	2	3	0,613	54,8	747,0	3508,8
1 + G + M	2	3	0,710	47,5	555,4	2543,8
1 + G + P	2	3	0,201	78,8	1564,9	6770,1
1 + N + M	2	3	0,988	9,8	4,5	113,6
1 + N + P	2	3	0,615	54,7	744,0	3330,4
1 + M + P	2	3	0,753	43,8	469,3	2166,2
1 + G + N + M	3	4	0,989	9,5	4,0	115,4
1 + G + N + P	3	4	0,616	55,9	743,9	3726,5
1 + G + P + M	3	4	0,775	42,6	428,4	2222,4
1 + N + P + M	3	4	0,988	10,0	6,4	120,8
1 + G + N + P + M	4	5	0,989	9,6	5,5	119,5

Tabela 14: Estimativas referentes ao submodelo 1 + N + M.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	186,694	12,259	15,23	0,00
Nclientes	3,408	0,146	23,37	0,00
Marcas	-21,193	0,803	-26,40	0,00
s	9,803			
R^2	0,988			
$\overline{R}^2$	0,987			

tante, cujas estimativas são apresentadas na Tabela 15. Interpretando as estimativas tem-se que a cada aumento de USD 1000 nos gastos da loja com promoções e de 100 clientes cadastrados, espera-se aumento de 1677 mil m^2

Tabela 15: Estimativas referentes ao submodelo 1 + G + N + M.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	179,844	12,621	14,25	0,00
Gastos	1,677	1,052	1,59	0,12
Nclientes	3,369	0,143	23,52	0,00
Marcas	-21,217	0,773	-27,30	0,00
s	9,491			
R^2	0,989			
$\overline{R}^2$	0,987			

Tabela 16: Resumo dos passos do procedimento *stepwise* com $P_E = P_S = 0,15$ e valores-P em cada passo para selecionar as variáveis explicativas do exemplo venda de telhados.

Passo	Gastos	Nclientes	Marcas	Potencial
Passo 1	0,4382	0,0000	0,0000	0,0389
Passo 2	0,2693	0,0000	-	0,0274
Passo 3	-	-	0,0000	-
Passo 4	0,1252	-	-	0,6968
Passo 5	-	0,0000	0,0000	-
Passo 6	-	-	-	0,4854

e 337 mil m^2 de telhados vendidos, respectivamente. Por outro lado, um aumento de 10 marcas concorrentes leva a uma redução média de 212 mil m^2 de telhados vendidos.

13.2 Salário de Executivos

Considere os dados de uma pesquisa realizada por uma revista de negócios sobre o salário anual de executivos (em mil USD) descrita em Foster et al. (1998, pp. 180-188), em que uma amostra aleatória de 220 executivos (145 homens e 75 mulheres) foi coletada. Além do salário anual foram consideradas as seguintes variáveis explicativas:

(i) Gênero (1: masculino; 0: feminino),

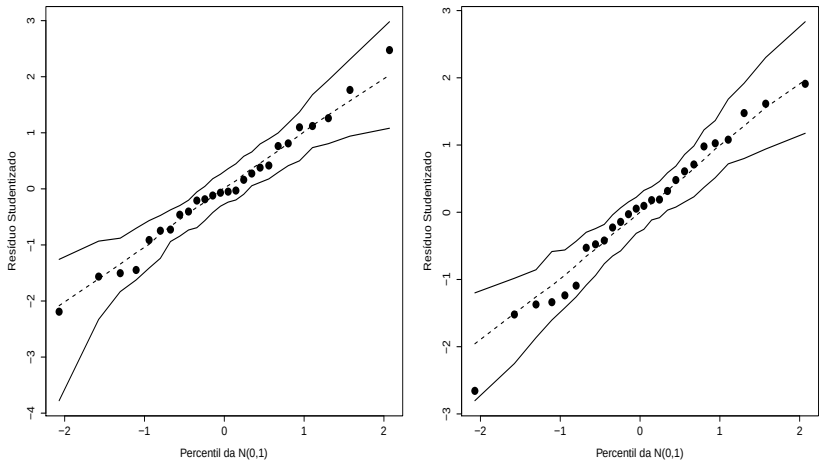

Figura 30: Gráficos normais de probabilidades referentes aos submodelos 1 + N + M (esquerda) e 1 + G + N + M (direita).

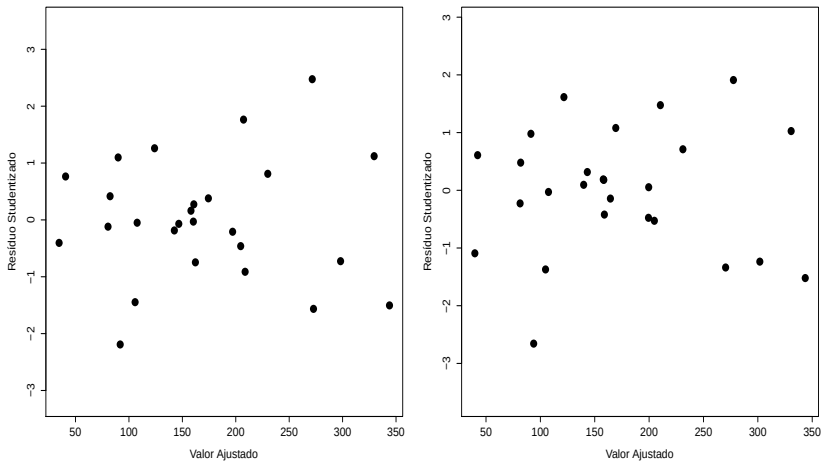

Figura 31: Gráficos do resíduo Studentizado contra o valor ajustado referentes aos submodelo 1 + N + M (esquerda) e 1 + G + N + M (direita).

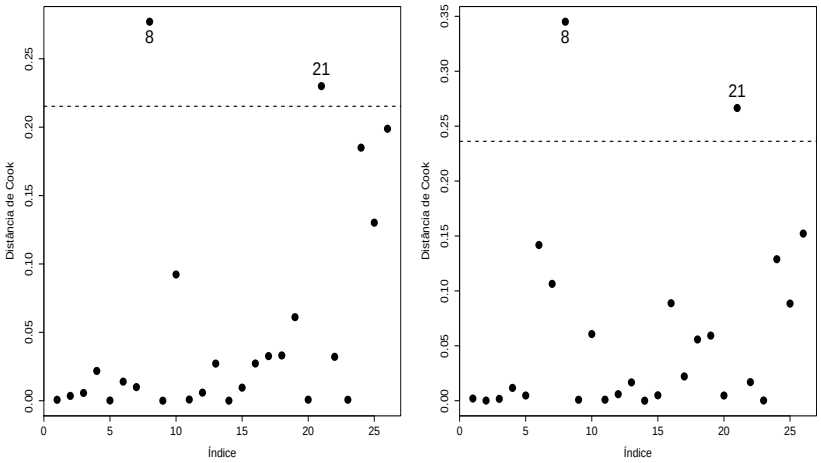

Figura 32: Gráficos da distância de Cook referentes aos submodelos 1 + N + M (esquerda) e 1 + G + N + M (direita).

(ii) Posição: posição na empresa (varia de 1 a 9), quanto maior o valor mais alta a posição e

(iii) Experiência: anos de experiência no cargo ou tempo no cargo.

Tabela 17: Descrição dos salários médios anuais com os respectivos erros padrão e do teste-t de igualdede de médias.

Gênero	Amostra	Média	E.Padrão
Masculino	145	144,11	1,03
Feminino	75	140,47	1,43
	Diferença	Teste-t	valor-P
Estimativa	3,64	2,06	0,04
E.Padrão	1,77		

O objetivo principal do estudo é explicar o salário médio anual segundo as três variáveis explicativas. As Figuras 33 e 34 descrevem, respectivamente, os bloxplots robustos do salário anual segundo o gênero e as respectivas

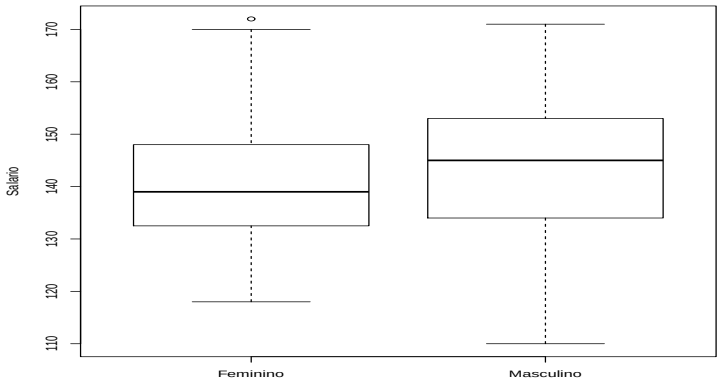

Figura 33: Boxplot robusto do salário anual segundo o gênero.

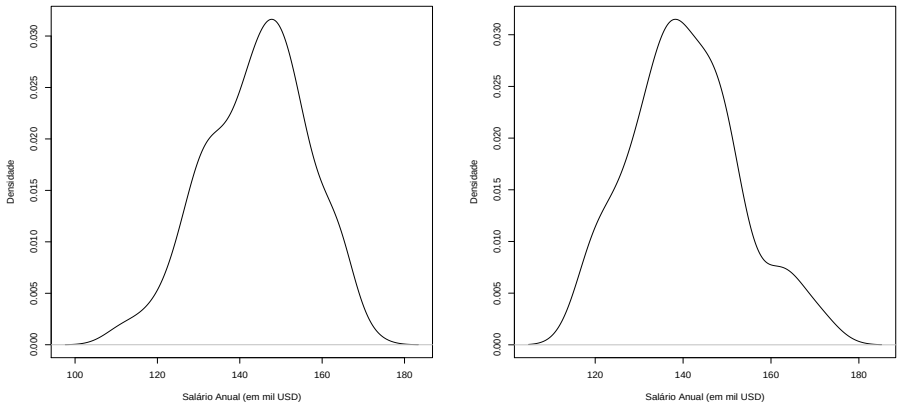

Figura 34: Densidade do salário anual dos executivos (esquerda) e das executivas (direita).

densidades empíricas. Nota-se uma ligeira superioridade dos salários anuais dos executivos. Isso é confirmado pela Tabela 17 onde são descritas as médias salariais com os respectivos erros padrão e o test-t para comparação de médias. A hipótese de igualdade de médias entre os dois grupos é rejeitada ao nível de significância de 5%. Há, portanto, indícios que os executivos em

média ganham mais do que as executivas.

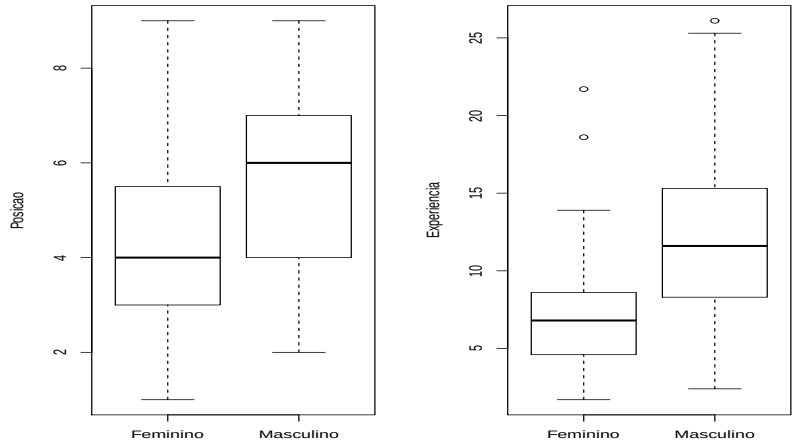

Figura 35: Boxplots robustos da posição e da experiência segundo o gênero.

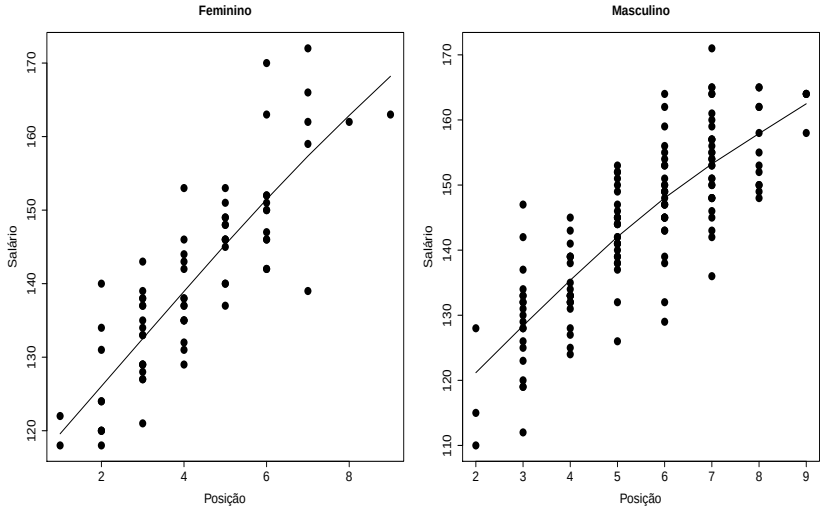

Figura 36: Diagrama de dispersão (com tendência) entre salário e posição segundo o gênero.

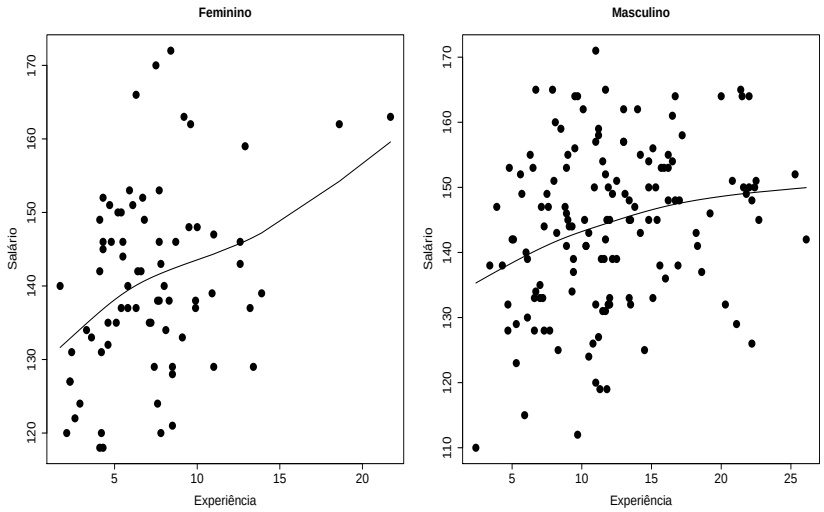

Figura 37: Diagrama de dispersão (com tendência) entre salário e experiência segundo o gênero.

Com relação à posição na empresa e experiência no cargo, nota-se pela Figura 35 que os executivos ecupam em geral posições mais altas e têm mais experiência do que as executivas. Os diagramas de dispersão entre o salário anual e a posição para ambos os gêneros (Figura 36) descrevem tendências crescentes, enquanto os diagramas de dispersão entre salário e experiência indicam também tendências crescentes (Figura 37), porém com menor intensidade.

Essas análises descritivas sugerem, em princípio, o seguinte modelo linear:

$$y_i = \beta_1 + \beta_2 \text{gênero}_i + \beta_3 \text{experiência}_i + \beta_4 \text{posição}_i + \epsilon_i, \qquad (6)$$

em que y_i denota o salário do i-ésimo executivo da amostra com $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, 220$.

As estimativas referentes ao modelo (6) estão descritas na Tabela 18 e pode-se notar que todos os efeitos são marginalmente significativos. Em particular, nota-se que à medida que aumenta a posição na empresa espera-se maior salário, fixados os demais efeitos. A experiência, segundo o modelo ajustado, à medida que aumenta tende a reduzir o salário médio e as executivas, quando comparadas com os executivos nos mesmos níveis de posição e experiência, têm um salário esperado maior. Esses resultados parecem con-

tradizer parte da análise descritiva, contudo são interpretações diferentes. A análise descritiva faz comparações marginais, enquanto a análise de regressão leva em conta todas as variáveis conjuntamente. Segundo as análises de resíduos (omitidas aqui) o modelo está bem ajustado, porém Foster et al.(1998) sugerem a inclusão de interações para agregar mais interpretações.

Tabela 18: Estimativas dos parâmetros referentes ao modelo de regressão linear múltipla (6) ajustado aos dados sobre salário de executivos.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	115,262	1,491	82,25	0,00
Experiência	-0,472	0,113	-4,17	0,00
GêneroM	-2,201	1,080	-2,04	0,04
Posição	6,710	0,313	21,46	0,00
s	6,77			
R^2	0,71			
$\overline{R}^2$	0,71			

Tabela 19: Teste F para a inclusão de interação no modelo (6).

Interação	valor-F	valor-P
gênero*experiência	1,615	0,20
gênero*posicão	0,001	0,97
experiência*posição	7,594	0,00

A Tabela 19 apresenta os valores da estatística F com os respectivos valores-P para a inclusão de cada interação no modelo (6). Nota-se que apenas a interação entre experiência e posição será incluída no modelo. Assim, o seguinte modelo será considerado:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 \text{gênero}_i + \beta_3 \text{experiência}_i + \beta_4 \text{posição}_i + \\ &\quad + \gamma \text{experiência}_i * \text{posição}_i + \epsilon_i, \end{aligned} \tag{7}$$

em que y_i denota o salário do i-ésimo executivo da amostra com $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, para $i = 1, \ldots, 220$. Na Tabela 20 são apresentadas as estimativas do ajuste do modelo (7) aos dados sobre salário de executivos. Nota-se

confirmação da inclusão da interação entre experiência e posição, contudo o efeito principal de experiência ficou não significativo. Não houve variações importantes nos coeficientes de determinação, indicando que a qualidade do ajuste permanece a mesma. Confirma-se pela estimativa do coeficiente de gênero que as executivas ganham em média mais do que os executivos, fixando-se os níveis de posição e experiência.

Tabela 20: Estimativas dos parâmetros referentes ao modelo de regressão linear múltipla (7) ajustado aos dados sobre salário de executivos.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	108,042	2,961	36,48	0,00
Experiência	0,336	0,314	1,07	0,28
GêneroM	-2,811	1,087	-2,59	0,01
Posição	8,096	0,590	13,73	0,00
Exper*Posição	-0,135	0,049	-2,76	0,00
s	6,67			
R^2	0,72			
R^2-ajustado	0,72			

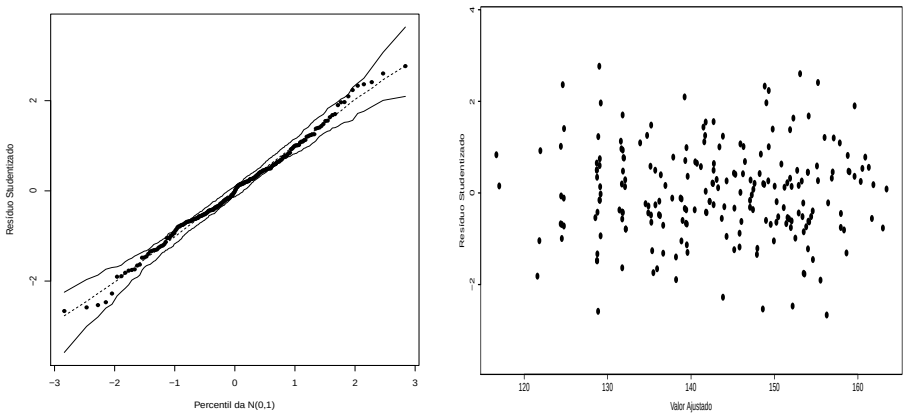

Figura 38: Análise de resíduos referente ao modelo (7) ajustado aos dados sobre salário de executivos.

Pela Figura 38 não há indícios de afastamentos da normalidade e da

constância de variância dos erros, bem como ausência de observações aberrantes. Contudo, pelo gráfico da distância de Cook com $k = 4$ (Figura 39) três observações são destacadas como possivelmente influentes. Apenas as observações #4 e #30 causam variações desproporcionais, respectivamente, de -14% e 11% na estimativa do coeficiente de gênero, embora não ocorram mudanças inferencias. A observação #4 é de uma executiva com salário anual de USD 139 mil (média USD 140,5 mil), posição 7 (média 4,3) e 13,9 anos de experiência (média 7,3 anos), enquanto a observação #30 é de um executivo com salário anual de USD 110 mil (média USD 144,1 mil), posição 2 (média 5,3) e 2,4 anos de experiência (média 12,2 anos).

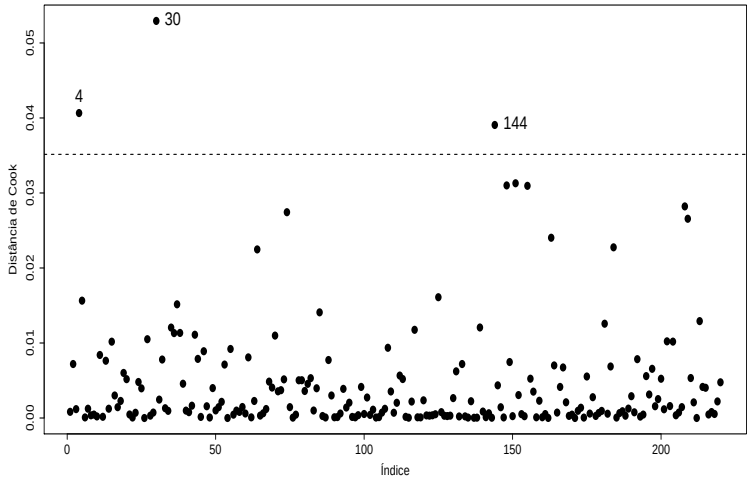

Figura 39: Distância de Cook contra a ordem das observações referente ao modelo (6) ajustado aos dados sobre salário de executivos.

O modelo ajustado fica então dado por

$$\begin{aligned}\widehat{y}(\mathbf{x}) &= 108,042 + 0,336\text{experiência} - 2,811\text{gênero} + \\ &\quad +8,096\text{posição} - 0,135\text{posição} * \text{experiência},\end{aligned}$$

em que $\mathbf{x} = (1, \text{experiência}, \text{gênero}, \text{posição})^\top$.

Finalmente, nas Figuras 40 e 41 tem-se os salários preditos para executivas e executivos, conforme variam a experiência e a posição. Nota-se que o salário predito para as executivas é sempre maior do que o salário predito para os executivos, fixados os níveis de experiência e posição. Para

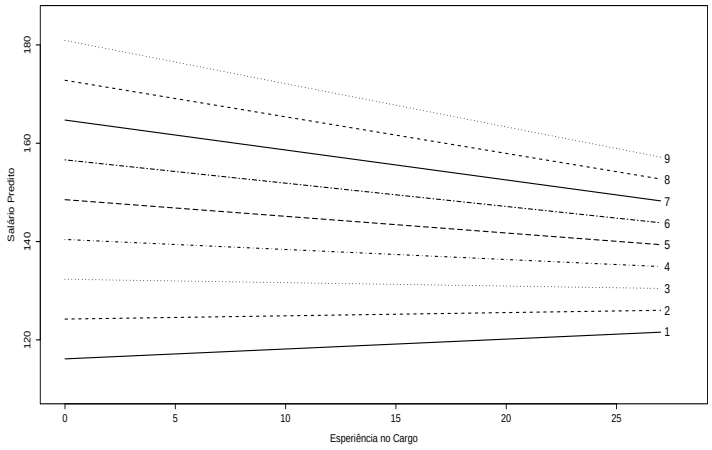

Figura 40: Salário médio estimado das executivas segundo a experiência e a posição.

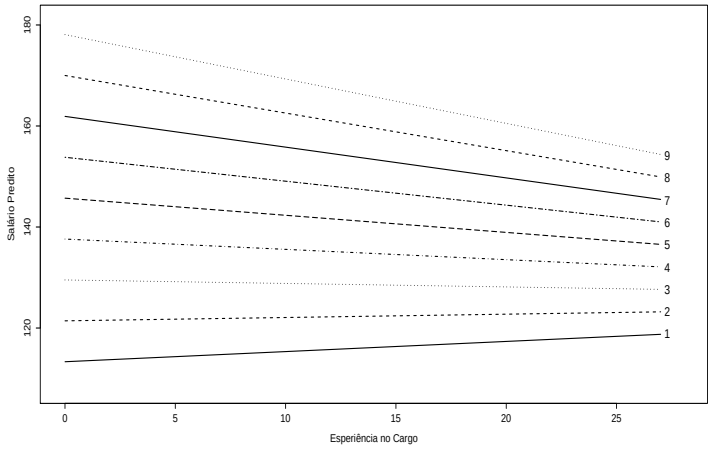

Figura 41: Salário médio estimado dos executivos segundo a experiência e a posição.

ambos os grupos o salário tende a crescer com o aumento do tempo no cargo nas posições iniciais 1 e 2. Contudo, nas demais posições o salário tende a

decrescer com o aumento do tempo no cargo. Fixando-se a experiência o salário aumenta à medida que aumenta a posição. Todavia, a diferença salarial entre duas posições quaisquer tende a diminuir à medida que aumenta a experiência. Portanto, uma conclusão que pode-se extrair da interação entre posição e experiência é que não vale a pena do ponto de vista salarial ficar muito tempo no mesmo cargo.

14 Estimação por Máxima Verossimilhança

Como visto anteriormente o modelo de regressão linear múltipla assume que $Y_i|\mathbf{x}_i \overset{\text{ind}}{\sim} \text{N}(\mu_i, \sigma^2)$ com $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, para $i = 1, \ldots, n$. Denotando $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$, em que $\phi = \sigma^2$, a função densidade de probabilidade de $Y_i|\mathbf{x}_i$ fica expressa na forma

$$f(y_i; \mathbf{x}_i, \boldsymbol{\theta}) = \left(\frac{1}{\sqrt{2\pi\phi}}\right) \exp\left\{-\frac{1}{2\phi}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\right\},$$

para $i = 1, \ldots, n$. Assim, o logaritmo da função de verossimilhança fica dado por

$$\begin{aligned}
\text{L}(\boldsymbol{\theta}) &= \log[\Pi_{i=1}^n \{f(y_i; \mathbf{x}_i, \boldsymbol{\theta})\}] \\
&= n \log\left(\frac{1}{\sqrt{2\pi\phi}}\right) - \frac{1}{2\phi}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \\
&= -\frac{n}{2}\log(2\pi\phi) - \frac{1}{2\phi}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2.
\end{aligned}$$

Para obter as estimativas de máxima verossimilhança de $\boldsymbol{\beta}$ e ϕ é preciso derivar a função escore

$$\mathbf{U}_\theta = \frac{\partial \text{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \mathbf{U}_\beta \\ \text{U}_\phi \end{pmatrix} = \begin{pmatrix} \frac{\partial \text{L}(\theta)}{\partial \beta} \\ \frac{\partial \text{L}(\theta)}{\partial \phi} \end{pmatrix}.$$

As estimativas de máxima verossimilhança são obtidas resolvendo-se as equações $\mathbf{U}_\beta = \mathbf{0}$ e $\text{U}_\phi = 0$.

A derivada parcial de $\text{L}(\boldsymbol{\theta})$ com relação a β_j fica dada por

$$\frac{\partial \text{L}(\boldsymbol{\theta})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}),$$

para $j = 1, \ldots, p$. Em forma matricial obtém-se

$$\mathbf{U}_\beta = \frac{\partial \mathrm{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$ e $\mathbf{X}$ é a matriz modelo. A estimativa de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$ é obtida tal que

$$\mathbf{U}_\beta = \mathbf{0} \Rightarrow \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Então, se $\mathbf{X}$ é uma matriz de posto coluna completo tem-se solução única

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

que coincide com a estimativa de mínimos quadrados. Por outro lado, a derivada parcial de $\mathrm{L}(\boldsymbol{\theta})$ com relação a ϕ fica dada por

$$\mathrm{U}_\phi = \frac{\partial \mathrm{L}(\boldsymbol{\theta})}{\partial \phi} = -\frac{n}{2\phi} + \frac{1}{2\phi^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2,$$

e fazendo $\mathrm{U}_\phi = 0$ obtém-se

$$\widehat{\phi} = \widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n},$$

em que $\widehat{y}_i = \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}$. Portanto, tem-se que $\widehat{\sigma}^2 = \frac{(n-p)}{n}s^2$ e $\mathrm{E}(\widehat{\sigma}^2) = \frac{(n-p)}{n}\sigma^2$. Logo, $\widehat{\sigma}^2$ é um estimador tendencioso de σ^2.

A matriz de informação de Fisher para $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$ é definada como sendo o valor esperado da curvatura de $\mathrm{L}(\boldsymbol{\theta})$

$$\mathbf{K}_{\theta\theta} = \mathrm{E}\left(-\frac{\partial^2 \mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right) = \begin{bmatrix} \mathbf{K}_{\beta\beta} & \mathbf{K}_{\beta\phi} \\ \mathbf{K}_{\phi\beta} & \mathrm{K}_{\phi\phi} \end{bmatrix}.$$

As submatrizes $\mathbf{K}_{\theta\theta}$ e $\mathbf{K}_{\phi\beta}$ ficam dadas por

$$\begin{aligned}
\mathbf{K}_{\beta\beta} &= \mathrm{E}\left(-\frac{\partial^2 \mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right) \\
&= \frac{1}{\phi}(\mathbf{X}^\top\mathbf{X}) \text{ e} \\
\mathbf{K}_{\beta\phi} &= \mathrm{E}\left(-\frac{\partial^2 \mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\phi}\right) \\
&= \frac{1}{\phi}\mathrm{E}\{\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})|\mathbf{X}\} \\
&= \mathbf{X}^\top\mathrm{E}\{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})|\mathbf{X}\} = \mathbf{0}.
\end{aligned}$$

Assim, os parâmetros $\boldsymbol{\beta}$ e ϕ são ortogonais. Ainda tem-se que

$$\begin{aligned} \mathrm{K}_{\phi\phi} &= \mathrm{E}\left(-\frac{\partial^2 \mathrm{L}(\boldsymbol{\theta})}{\partial \phi^2}\right) \\ &= -\frac{n}{2\phi^2} + \frac{1}{\phi^3}\sum_{i=1}^{n} \mathrm{E}\{(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\} \\ &= -\frac{n}{2\phi^2} + \frac{n}{\phi^2} = \frac{n}{2\phi^2}. \end{aligned}$$

Logo, a matriz de informação de Fisher para $\boldsymbol{\theta}$ assume a forma bloco diagonal

$$\mathbf{K}_{\theta\theta} = \begin{bmatrix} \mathbf{K}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \mathrm{K}_{\phi\phi} \end{bmatrix},$$

e pelas propriedades de estimação por máxima verosimilhança, tem-se para n grande que $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \mathbf{K}_{\beta\beta}^{-1})$ e $\widehat{\sigma}^2 \sim \mathrm{N}(\sigma^2, \mathrm{K}_{\phi\phi}^{-1})$. Além disso, $\widehat{\boldsymbol{\beta}}$ e $\widehat{\sigma}^2$ são independentes. No caso de $\widehat{\boldsymbol{\beta}}$ o resultado vale para todo n. Similarmente, segue que $(n-p)s^2/\sigma^2 \sim \chi^2_{(n-p)}$.

Exercícios

1. Seja T um estimador do parâmetro θ e supor a existência dos dois primeiros momentos de T. Mostre que

   $$\mathrm{E}\{(\mathrm{T} - \theta)^2\} = \mathrm{E}[\{T - \mathrm{E}(T)\}^2] + \{\mathrm{E}(\mathrm{T}) - \theta\}^2.$$

   Ou seja, EQM(T) = Var(T) + $\{\text{Viés(T)}\}^2$.

2. Com base numa amostra de $n = 3$ de uma variável aleatória X de média μ_X e variância σ_X^2 foram propostos para μ_X os seguintes estimadores:

   $$\mathrm{T}_1 = \frac{1}{5}(X_1 + 3X_2 + X_3), \quad \mathrm{T}_2 = \frac{1}{2}(X_1 + 2X_3),$$

   $$\mathrm{T}_3 = \frac{1}{4}(2X_1 + X_2 + X_3) \text{ e } \mathrm{T}_4 = \frac{1}{3}(X_1 + X_2 + X_3).$$

   Obtenha o erro quadrático médio, a variância e o viés de cada estimador. Entre os não tendenciosos qual escolher? Justifique.

3. Considere a seguinte regressão linear simples:

   $$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, $i = 1, \ldots$. Mostre que: (i) $\text{Cov}(\bar{Y}, \hat{\beta}_2) = 0$, (ii) $\sum_{i=1}^n r_i \hat{y}_i = 0$, (iii) $\sum_{i=1}^n r_i x_i = 0$, (iv) $\sum_{i=1}^n r_i = 0$ e (v) $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$, em que $r_i = y_i - \hat{y}_i$.

4. Supor que foi ajustado através de mínimos quadrados o modelo de regressão $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_2$, porém o modelo verdadeiro é dado por

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

em que $\epsilon \sim \text{N}(0, \sigma^2)$. Mostre que o estimador $\hat{\beta}_2$ obtido no primeiro ajuste é tendencioso. Expresse o viés de $\hat{\beta}_2$.

5. Supor uma amostra aleatória de tamanho n e o seguinte modelo de regressão:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots . n$. Supondo β_1 conhecido obtenha o estimador de mínimos quadrados de β_2 e o respectivo erro padrão. Compare esse estimador com o estimador de mínimos quadrados de β_2 quando β_1 é desconhecido. Comente.

6. Considere o modelo de regressão linear múltipla $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, para $i = 1, \ldots, n$. Mostre que a estatística F para testar $\text{H}_0 : \beta_2 = \cdots = \beta_p$ contra $\text{H}_1 : \beta_j \neq 0$, para pelo menos algum $j = 2, \ldots, p$, pode ser expressa na forma

$$\text{F} = \frac{R^2(n-p)}{(p-1)(1-R^2)}.$$

7. São apresentados na tabela abaixo o consumo (galão/milha)(Y) e a cilindrada (polegadas^3) (X) de uma amostra de $n = 32$ automóveis de marcas diferentes (Montgomery et al., 2021, Tabela B3).

y	x	y	x	y	x	y	x
18,90	350,0	17,00	350,0	20,00	250,0	18,25	351,0
20,07	225,0	11,20	440,0	22,12	231,0	21,47	262,0
34,70	89,7	30,40	96,9	16,50	350,0	36,50	85,3
21,50	171,0	19,70	258,0	20,30	140,0	17,80	302,0
14,39	500,0	14,89	440,0	17,80	350,0	16,41	318,0
23,54	231,0	21,47	360,0	16,59	400,0	31,90	96,9
29,40	140,0	13,27	460,0	23,90	133,6	19,73	318,0
13,90	351,0	13,27	351,0	13,77	360,0	16,50	350.0

Responda às seguintes questões: (i) construir o diagrama de dispersão entre o consumo e a cilindrada dos automóveis, comente; (ii) obter a correlação linear amostral de Pearson; (iii) ajustar o modelo de regressão linear simples de mínimos quadrados, obtendo as estimativas $\hat{\beta}_1$ e $\hat{\beta}_2$ e os respectivos erros padrão; (iv) traçar a reta de regressão no diagrama de dispersão; (v) interpretar a estimativa $\hat{\beta}_2$; (vi) obter as estimativas intervalares de 95% para β_1 e β_2 e (vii) obter a estimativa intervalar de 97% para o consumo de um automóvel com cilindrada de $x = 300$ polegadas[3]. Resultados úteis: $\bar{y} = 20,2231$, $\bar{x} = 284,7312$, $\sum y_i^2 = 14324,74$. $\sum x_i^2 = 3019001$ e $\sum x_i y_i = 164118,10$. Este exercício deve ser feito manualmente. O diagrama de dispersão pode ser feito no R.

8. No arquivo **capm.txt** estão os seguintes dados (Ruppert, 2004, Cap.7): Tbill (taxa de retorno livre de risco), retorno Microsoft, SP500 (retorno do mercado), retorno GE e retorno FORD de janeiro de 2002 a abril de 2003. Todos os retornos são diários e estão em porcentagem. Construir inicialmente os diagramas de dispersão (com tendência) entre o excesso de retorno $(y_{rt} - r_{ft})$ de cada uma das empresas Microsoft, GE e FORD e o excesso de retorno do mercado $(r_{mt} - r_{ft})$, em que y_{rt} denota o retorno da ação da empresa, r_{mt} é o retorno do mercado e r_{ft} indica a taxa livre de risco durante o t-ésimo período. Posteriormente, ajustar o seguinte modelo de regressão linear simples para cada ação:

$$y_t = \alpha + \beta x_t + \epsilon_t,$$

em que $y_t = y_{rt} - r_{ft}$, $x_t = r_{mt} - r_{ft}$ e $\epsilon_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$. No modelo acima, o parâmetro β é denominado risco sistemático com a seguinte interpretação: se $\beta = 1$ o excesso de retorno é equivalente ao mercado (volatilidade similar ao mercado), se $\beta > 1$ o excesso de retorno é maior do que o excesso de retorno do mercado (ação mais volátil do que o mercado), e se $\beta < 1$ o excesso de retorno é menor do que o excesso de retorno do mercado (ação menos volátil do que o mercado). O intercepto é incluído para controlar eventuais precificações incorretas, porém em geral $\alpha = 0$ não é rejeitado.

Para ler o arquivo no R use os comandos

```
capm = read.table("capm.txt", header=TRUE).
```

Para deixar o arquivo disponível use o comando

```
attach(capm).
```

Por exemplo, para ajustar o excesso de retorno da Microsoft use os comandos

```
ymsf = rmsf - tbill
xmerc = sp500 - tbill
ajuste.msf = lm(ymsf ~ xmerc)
summary(ajuste.msf).
```

Verifique se os modelos estão bem ajustados através de análise de resíduos. Para cada ação encontre uma estimativa intervalar de 95% para o risco sistemático e classifique o excesso de retorno em relação ao mercado. Finalmente, construa para cada ação a banda de confiança de 95% para prever o excesso de retorno num determinado dia, dado o excesso de retorno do mercado.

9. Suponha o modelo de comparação de médias

$$y_{ij} = \mu_i + \epsilon_{ij},$$

em que $\epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$. Mostre que $\hat{\mu}_i = \bar{y}_i$ e $\text{Var}(r_{ij}) = \sigma^2(1 - 1/n_i)$, em que $r_{ij} = y_{ij} - \bar{y}_i$.

10. Considere o modelo de regressão linear múltipla

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Mostre que o critério de Akaike é equivalente a minimizar a quantidade

$$\text{AIC} = n \log \left\{ \frac{\text{SQRes}}{n} \right\} + 2p,$$

com $\text{SQRes} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

11. Na tabela abaixo (Lawless, 1992, Seção 6.8) são apresentados os resultados de um experimento em que a resistência (em horas) de um determinado tipo de vidro foi avaliada segundo quatro níveis de voltagem (em kilovolts) e duas temperaturas (em graus Celsus). Esses dados estão também disponíveis no arquivo **vidros.txt**. Na primeira coluna do arquivo tem-se o tempo de resistência, na segunda coluna a voltagem (1: 200kV, 2: 250kV, 3: 300kV e 4: 350kV) e na terceira coluna a temperatura (1: 170^{o}C e 2: 180^{o}C). Seja Y_{ijk} o tempo de

resistência da k-ésima amostra de vidro submetida à i-ésima voltagem e à j-ésima temperatura.

Para ler o arquivo no R use os comandos

```
vidros = read.table("vidros.txt", header=TRUE)
voltagem = factor(voltagem)
temperatura = factor(temperatura).
```

	Voltagem(kV)			
Temperatura (^{o}C)	200	250	300	350
170	439	572	315	258
	904	690	315	258
	1092	904	439	347
	1105	1090	628	588
180	959	216	241	241
	1065	315	315	241
	1065	455	332	435
	1087	473	380	455

Faça inicialmente uma análise descritiva dos dados, por exemplo apresentando os perfis médios da resistência segundo a voltagem para os dois níveis de temperatura. Comente e verifique se há indícios de interação entre temperatura e voltagem.

Supor inicialmente o seguinte modelo:

$$y_{ijk} = \alpha + \beta_i + \gamma_j + \epsilon_{ijk},$$

em que β_i denota o efeito da i-ésima voltagem e γ_j o efeito da j-ésima temperatura em relação à casela de referência, sendo assumido $\beta_1 = 0$, $\gamma_1 = 0$ e $\epsilon_{ijk} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, 2, 3, 4$ e $j = 1, 2$. Testar a inclusão de interação entre temperatura e voltagem. Construir a tabela ANOVA. Para ajustar os modelos sem interação e com interação e gerar a tabela ANOVA use os comandos

```
fit1 = lm(resistencia ~ voltagem + temperatura)
fit2 = lm(resistencia ~ voltagem + temperatura +
temperatura*voltagem)
```

`anova(fit1,fit2).`

Fazer análises de resíduos e de sensibilidade. Construir o gráfico dos perfis ajustados. Comente.

12. Considere o arquivo **BigMac2003** da biblioteca `alr4` do R, em que são descritas as seguintes variáveis de 69 cidades de diversos países:

    - `BigMac`: minutos de trabalho para comprar um Big Mac
    - `Bread`: minutos de trabalho para comprar 1kg de pão
    - `Rice`: minutos de trabalho para comprar 1kg de arroz
    - `FoodIndex`: índice de preços de alimentos
    - `Bus`: valor da passagem de ônibus (em USD)
    - `Apt`: valor do aluguel (em USD) de um apartamento padrão de 3 dormitórios
    - `TeachGI`: salário bruto anual (em 1000 USD) de um professor de ensino fundamental
    - `TeachNI`: salário líquido anual (em 1000 USD) de um professor de ensino fundamental
    - `TaxRate`: imposto pago (em porcentagem) por um professor de ensino fundamental
    - `TeachHours`: carga horária semanal (em horas) de um professor de ensino fundamental.

    Para disponibilizar e visualizar um resumo dos dados use na sequência os seguintes comandos do R:

    ```
    require(alr4)

    require(MASS)

    attach(BigMac2003)

    summary(BigMac2003).
    ```

    O objetivo principal do estudo é relacionar a variável BigMac com as demais variáveis explicativas. A fim de obter uma melhor aproximação para a normalidade considere log(BigMac) como variável resposta. Apresente os diagramas de dispersão (com tendência) entre a variável resposta e cada uma das variáveis explicativas e comente. Padronize as variáveis explicativas. Por exemplo, para padronizar a variável explicativa Bread use o comando

`sBread = scale(Bread, center = TRUE, scale = TRUE).`

Através do procedimento `stepAIC` fazer uma selação das variáveis explicativas. Para o modelo selecionado aplicar análises de resíduos e de sensibilidade. Comente. Classifique as variáveis explicativas segundo o impacto na explicação da média da variável resposta.

13. No arquivo **motorins** da biblioteca `faraway` do R são descritas informações relacionadas a 1797 grupos de apólices de seguro de automóvel no ano de 1977 na Suécia. Em particular, há interesse em saber se há diferenças significativas entre o seguro médio pago por sinistro em 7 regiões do país. Para ler o arquivo no R utilize os comandos

    `require(faraway)`

    `summary(motorins)`

    `attach(motorins).`

    Considere as variáveis `Zone` (região do país) e perd valor pago por sinistro (em coroas suecas). A fim de obter uma melhor aproximação para a normalidade considere como resposta a variável log(perd). Construir boxplots de log(perd) segundo a região. Comente. Aplique em seguida um ajuste de comparação de médias através do comando

    `fit1.motor = lm(log(perd) ~ Zone).`

    Construa a tabela ANOVA através do comando

    `fit2.motor = aov(log(perd) ~ Zone).`

    Se for rejeitada a hipótese de homogeneidade de médias, aplique o método de Tukey para verificar quais contrastes são significativos através do comando

    `TukeyHSD(fit2.motor)`

    `plot(TukeyHSD(fit2.motor),las=2).`

    Comente.

14. No arquivo **fuel2001.txt** da biblioteca `alr4` do R, estão descritas as seguintes variáveis referentes aos 50 estados norte-americanos mais o Distrito de Columbia no ano de 2001:

    - `UF`: unidade da federação
    - `Drivers`: número de motoristas licenciados

- `FuelC`: total de gasolina vendida (em mil galões)
- `Income`, renda per capita em 2000 (em mil USD)
- `Miles`, total de milhas em estradas federais
- `MPC`, milhas per capita percorridas
- `Pop`, população ≥ 16 anos
- `Tax`, taxa da gasolina (em cents por galão).

A fim de possibilitar uma comparação entre as UFs duas novas variáveis são consideradas `Fuel = 1000*FuelC/Pop` e `Dlic = 1000*Drivers/Pop`, além da variável `Miles` ser substituída por `log(Miles)`. Para ler o arquivo no R use os comandos

```
require(alr4)

require(MASS)

attach(fuel2001)

summary(fuel2001).
```

Considere como resposta a variável Fuel e como variáveis explicativas Dlic, log(Miles), Income e Tax. Faça inicialmente uma análise descritiva dos dados. Por exemplo, boxplot robusto para a variável resposta e diagramas de dispersão (com tendência) entre cada variável explicativa e a variável resposta. Comente. Aplique o procedimento `stepAIC` para selecionar as variáveis explicativas. Verifique se é possível incluir alguma interação. Com o modelo selecionado faça uma análise de diagnóstico: análise de resíduos, pontos de alavanca, distância de Cook e DFFITS. Avalie o impacto dos pontos destacados. Interprete os coeficientes estimados.

15. No arquivo **wine.txt** (Montgomery et al., 2021, Tabela B.11) são descritas características de uma amostra aleatória de 38 vinhos da marca "Pinot Noir". O objetivo do estudo é relacionar a qualidade do vinho com as seguintes variáveis explicativas: (i) `claridade`, (ii) `aroma`, (iii) `corpo`, (iv) `sabor`, (v) `aromac`, aroma do tonel de carvalho e (vi) `regiao` (1: região 1, 2: região 2 e 3: região 3). Para ler o arquivo no R use os comandos

    ```
    wine = read.table("wine.txt", header=TRUE).
    ```

    A variável região é categórica com três níveis. Assim é possível através do comando `factor` do R transformá-la em duas variáveis binárias: `regiao2` $= 1$ para região 2 e 0 caso contrário e `regiao3` $= 1$ para

região 3 e 0 em caso contrário. A casela de referência será a região 1. Para acionar o procedimento use o comando

`regiao = factor(regiao).`

Faça inicialmente uma análise descritiva dos dados com boxplot robusto para a variável resposta e diagramas de dispersão (com tendência) entre a variável resposta e variáveis explicativas. Calcule também as correlações lineares de Peasorn entre as variáveis (exceto região). Selecione incialmente um submodelo através dos métodos de maior R_k^2, menor s_k, menor C_k e menor $\overline{\text{Press}}_k$. Em seguida selecione outro submodelo através do procedimento `stepwise` usando PE=PS=0,15. Compare os submodelos escolhidos e para o submodelo selecionado aplicar análise de resíduos e sensibilidade. Interpretar os coeficientes estimados.

16. Considere o modelo linear simples

$$y_i = \beta_1 + \beta_2(x_i - \bar{x}) + \epsilon_i,$$

para $i = 1, \ldots, n$ com $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$. Encontrar os estimadores *ridge* $\hat{\beta}_{R1}$ e $\hat{\beta}_{R2}$ como também suas variâncias e covariância assintóticas $\text{Var}(\hat{\beta}_{R1})$, $\text{Var}(\hat{\beta}_{R2})$ e $\text{Cov}(\hat{\beta}_{R1}, \hat{\beta}_{R2})$. Expresse os estimadores *ridge* em função dos estimadores de mínimos quadrados e mostre que são estimadores tendenciosos.

17. Para avaliar a relação entre a energia necessária diária e a produção de carne, uma amostra aleatória de 64 ovelhas em fase de crescimento foi considerada, sendo observado para cada animal o consumo médio diário de energia (mcal) e o peso (em kg). Esses dados estão descritos no arquivo **sheep.txt** (vide Lindsey, 1997, Seção 9.4). Para ler o arquivo no R use os comandos

`sheep = read.table("sheep.txt", header=TRUE).`

Fazer inicialmente uma análise descritivia dos dados, boxplot robusto da variável resposta (peso) e diagrama de dispersão entre o peso do animal e o consumo diário de energia (variável explicativa). Ajustar um modelo linear normal aos dados e verificar que há indícios de variância não constante dos erros. Ajustar um modelo normal ponderado com pesos apropriados. Fazer uma análise de diagnóstico e interpretar as estimativas.

18. Considere o modelo de regressão linear múltipla

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Mostre que SQRes(k) $\geq$ SQRes, em que SQRes(k) $= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R)^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R)$ e SQRes $= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ denotam, respectivamente, as somas de quadrados de resíduos da regressão *ridge* e da regressão de mínimos quadrados.

19. Supor o modelo linear ponderado $y_i = \alpha + \beta x_i + \epsilon_i$, em que $\epsilon_i \overset{\text{ind}}{\sim} \text{N}(0, \text{d}_i\sigma^2)$, $\text{d}_i > 0$, para $i = 1, \ldots, n$. Obter $\hat{\beta}$ e ASQ($\beta = 0$).

20. Considere os dados do arquivo **Rateprof** da biblioteca `alr4` do R, referentes a notas médias recebidas por 364 instrutores de uma universidade norte americana durante um período de 10 anos. O objetivo do estudo é relacionar o interesse do avaliador (`RaterInterest`) (escore de 1 a 5) com as seguintes avaliações feitas pelo avaliador:

    - `Quality`: qualidade das aulas do instrutor (escore de 1 a 5)
    - `Helpfulness`: prestatividade do instrutor (escore de 1 a 5)
    - `Clarity`: clareza das aulas do instrutor (escore de 1 a 5)
    - `Easiness`: facilidade que o instrutor tem com a matéria (escore de 1 a 5).

    Inicialmente centralize as 5 variáveis através do comando

    `cvariavel = variavel - mean(variavel).`

    Fazer uma análise descritiva com os dados apresentando a matriz de correlações lineares de Pearson e os diagramas de dispersão (com tendência). Comente. Ajustar agora um modelo de regressão linear da variável resposta centralizada contra as demais variáveis explicativas centralizadas e passando pela origem. Use o comando

    `fit1 = lm(cresposta ~ cv1 + cv2 + cv3 + cv4 -1).`

    Verifique se há indícios de multicolineridade através do VIF. Tente contornar o problema através de componentes principais, considerando apenas o $1^{\underline{o}}$ componente. Qual a explicação desse componente? Expresse esse componente em função das 4 variáveis explicativas centralizadas. Fazer um ajuste da regressão linear da variável resposta centralizada contra esse componente e passando pela origem. Interprete o coeficiente estimado e apresente análises de diagnóstico.

Referências

Atkinson AC (1981) Two graphical display for outlying and influential observations in regression. *Biometrika* 68:13-20.

Atkinson AC (1985) *Plots, Transformations and Regressions.* Oxford Statistical Science Series, Oxford.

Belsley DA, Kuh E, Welsch RE (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley, New York.

Cook RD (1977) Detection of influential observations in linear regressions. *Technometrics* 19:15-18.

Cook RD, Weisberg S (1982) *Residuals and Influence in Regression.* Chapman and Hall/CRC.

Dunkler D, Plischke M, Leffondré K, Heinze G (2014) Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *Plos One* 9(11):e113677.

Faraway JJ (2016) *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models, 2nd Edition.* Chapman and Hall/CRC.

Foster DP, Stine RA, Waterman RP (1998) *Business Analysis using Regression.* Springer.

Fox J, Weisberg S (2019) *An R Companion to Applied Regression, 3nd Edition.* Sage, Thousand Oaks, CA.

Hoaglin DC, Welsch RE (1978) The hat matrix in regression and ANOVA. *The American Statistician* 32:17-22.

Lawless JF (1982) *Statistical Models and Methods for Lifetime Data.* Wiley.

Lindsey JK (1997) Applying Generalized Linear Models. Springer, New York.

Hubert M, Vandervierin E (2008) An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis* 32:5186-5201.

Montgomery DC, Peck EA, Vining GG (2021) *Introduction to Linear Regression Analysis, 6th Edition.* Wiley.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) *Applied Linear Statistical Models, 4th Edition.* WCB McGraw-Hill.

Rao CR (1973) *Linear Statistical Inference and Its Applications, Second Edition.* Wiley, New York.

Ruppert D (2004) *Statistical and Finance.* Springer, New York.

Weisberg S (2014) *Applied Linear Regression, Fourth Edition.* Wiley.