Análise de dados longitudinais

Apêndices A, B e C Versão parcial preliminar

junho 2018

Julio M. Singer

Juvêncio S. Nobre

Francisco Marcelo M. Rocha

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

Conteúdo

Apêndice A

Matrizes e espaços vetoriais

A.1 Matrizes

Neste apêndice apresentamos a notação matricial utilizada no texto e listamos alguns resultados relacionados com álgebra e derivação de matrizes além de conceitos de espaços vetoriais. Para detalhes, o leitor deve consultar Searle (1982), Magnus & Neudecker (1988) e Harville (1997), por exemplo.

Uma **matriz A** de dimensão $m \times n$, é um arranjo retangular de elementos[1] com m linhas e n colunas, no qual o elemento a_{ij} situa-se no cruzamento da i-ésima linha com a j-ésima coluna:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = ((a_{ij}))_{1 \leq i \leq n, 1 \leq j \leq m}.$$

Um exemplo de uma matriz (2×4) é

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 4 & -8 \\ 3 & 5 & 2 & 0 \end{pmatrix}.$$

Um **vetor** de dimensão $(m \times 1)$ é uma matriz com m linhas e uma única coluna:

$$\mathbf{u} = \begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{m1} \end{pmatrix}.$$

[1] Neste texto consideramos apenas matrizes com elementos reais.

Matrizes serão representadas por letras maiúsculas em negrito (por exemplo, $\mathbf{A}, \mathbf{X}, \mathbf{G}$) e vetores, por letras minúsculas em negrito ($\mathbf{a}, \mathbf{x}, \mathbf{y}$, por exemplo). Quando necessário, a dimensão será especificada entre parênteses; por exemplo, $\mathbf{A}$ $(m \times n)$. Uma matriz $\mathbf{A}$ $(m \times n)$, pode ser expressa como $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_n)$, com $\mathbf{a}_j$ denotando sua j-ésima coluna, ou seja,

$$\mathbf{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}.$$

A.1.1 Operações básicas

Multiplicação por escalar: *Sejam k um número real e $\mathbf{A}$ uma matriz $(m \times n)$. O produto de $\mathbf{A}$ por k, denotado $\mathbf{B} = k\mathbf{A}$ é uma matriz $(m \times n)$ no qual o elemento $b_{ij} = ka_{ij}$, ou seja,*

$$\mathbf{B} = k\mathbf{A} = \begin{pmatrix} ka_{11} & ka_{12} & \dots & ka_{1n} \\ ka_{21} & ka_{22} & \dots & ka_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ ka_{m1} & ka_{m2} & \dots & ka_{mn} \end{pmatrix}.$$

Soma e subtração de matrizes: *Sejam $\mathbf{A}$ e $\mathbf{B}$ duas matrizes de mesma dimensão $(m \times n)$. Sua soma, representada por $\mathbf{A} + \mathbf{B}$, é a matriz $(m \times n)$ cujos elementos são dados por $c_{ij} = a_{ij} + b_{ij}$, ou seja,*

$$\mathbf{A} + \mathbf{B} \;=\; \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}.$$

Sejam $\mathbf{A}$, $\mathbf{B}$ matrizes de dimensão $(m \times n)$ e k um número real. Então valem as seguintes propriedades:

i) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$;

ii) $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$;

iii) $k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$.

A subtração de duas matrizes quaisquer $\mathbf{A}$ e $\mathbf{B}$, de mesma dimensão $(m \times n)$, denotada $\mathbf{A} - \mathbf{B}$ é uma matriz de dimensão $(m \times n)$ cujos elementos são dados por $d_{ij} = a_{ij} - b_{ij}$.

Produto de matrizes: *O produto de uma matriz* $\mathbf{A}$ *com dimensão* $(m \times n)$ *por uma matriz* $\mathbf{B}$ *com dimensão* $(n \times q)$ *é uma matriz* $\mathbf{C} = \mathbf{AB}$ *com dimensão* $(m \times q)$ *e elementos*

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj},$$

para $i = 1, \ldots, m$ e $j = 1, \ldots, q$.

Sejam $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ matrizes com produtos $\mathbf{AB}$, $\mathbf{AC}$ e $\mathbf{BC}$ bem definidos. Então:

i) $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$;

ii) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.

Em geral o produto de matrizes não é comutativo, ou seja, não necessariamente $\mathbf{AB} = \mathbf{BA}$. Por exemplo, dadas

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 1 \\ 0 & 0 & 5 \end{pmatrix} \text{ e } \mathbf{B} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

temos

$$\mathbf{AB} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 2 & 1 & 1 \end{pmatrix} \neq \mathbf{BA} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix}.$$

Por outro lado, dadas

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix} \text{ e } \mathbf{B} = \begin{pmatrix} -1 & 3 \\ 1 & 2 \\ -2 & -2 \end{pmatrix},$$

temos

$$\mathbf{AB} = \begin{pmatrix} -1 & 3 \\ 2 & 4 \\ -10 & -10 \end{pmatrix},$$

mas o produto $\mathbf{BA}$ não está definido.

Matriz transposta: *A matriz transposta (às vezes chamada apenas de transposta) de uma matriz* $\mathbf{A}$ $(m \times n)$*, denotada por* $\mathbf{A}^\top$*, é a matriz com dimensão* $(n \times m)$ *cujos elementos* a'_{ij} *são dados por* $a'_{ij} = a_{ji}$. Por exemplo, se

$$\mathbf{A} = \begin{pmatrix} -1 & 3 \\ 1 & 2 \\ -2 & -2 \end{pmatrix} \text{ então } \mathbf{A}^\top = \begin{pmatrix} -1 & 1 & -2 \\ 3 & 2 & -2 \end{pmatrix}.$$

Para quaisquer matrizes $\mathbf{A}$ e $\mathbf{B}$ (para as quais as operações matriciais abaixo estejam definidas) valem as seguintes propriedades:

i) $(\mathbf{A}^\top)^\top = \mathbf{A}$;

ii) $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$;

iii) $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.

A.1.2 Tipos especiais de matrizes

Matriz quadrada: *Uma matriz* $\mathbf{A}$ *com dimensão* $(n \times n)$ *é chamada de matriz quadrada de ordem* n*. Os elementos* $a_{11}, \ldots, a_{nn}$ *de uma matriz quadrada constituem sua* **diagonal principal**.

Matriz simétrica: *Uma matriz quadrada* $\mathbf{A}$ *é simétrica se* $\mathbf{A} = \mathbf{A}^\top$.

Matriz diagonal: *Uma matriz quadrada* $\mathbf{A}$ *é diagonal se todos os elementos não pertencentes à diagonal principal forem nulos, ou seja, se for da forma*

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & a_{nn} \end{pmatrix}.$$

Seja $\mathbf{A}$ uma matriz quadrada de ordem n e $\mathbf{a}$ um vetor $(n \times 1)$ formado pelos elementos de sua diagonal principal. Então, o operador diagonal é definido como

$$\text{diag}(\mathbf{A}) = \mathbf{a}$$

e

$$\text{diag}(\mathbf{a}) = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & a_{nn} \end{pmatrix}.$$

Matriz identidade: *Uma matriz diagonal de ordem n em que todos os elementos da diagonal principal são iguais a* 1 *é chamada matriz identidade de ordem n e é denotada por* $\mathbf{I}_n$ [2].

A matriz $\mathbf{I}$ é o elemento neutro na multiplicação de matrizes, isto é, para qualquer matriz quadrada $\mathbf{A}$ de ordem n

$$\mathbf{IA} = \mathbf{AI}.$$

Matriz triangular superior: *Uma matriz quadrada* $\mathbf{A}$ *é triangular superior se todos os elementos abaixo da diagonal principal forem iguais a zero, ou seja, se for da forma:*

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix},$$

Matriz triangular inferior: *Analogamente, se todos os elementos acima da diagonal principal de* $\mathbf{A}$ *forem nulos, a matriz* $\mathbf{A}$ *é triangular inferior.*

Matriz triangular: *Uma matriz é triangular se ela é triangular superior ou inferior.*

Se a matriz $\mathbf{A}$ é triangular inferior (superior), então $\mathbf{A}^\top$ é triangular superior (inferior); além disso, a matriz resultante da soma ou produto de matrizes triangulares superiores (inferiores) é triangular superior (inferior).

Matriz idempotente: *Uma matriz quadrada* $\mathbf{A}$ *de ordem n é idempotente se*

$$\mathbf{AA} = \mathbf{A}^2 = \mathbf{A}.$$

A.1.3 Submatrizes e matrizes particionadas

Submatriz: *Uma submatriz de uma matriz* $\mathbf{A}$ *é qualquer matriz obtida através da eliminação de linhas e/ou colunas.*

Por exemplo, se considerarmos

$$\mathbf{A} = \begin{pmatrix} 2 & 5 & 1 & 4 \\ -7 & 4 & 0 & 10 \\ 3 & 7 & 20 & 8 \end{pmatrix},$$

[2] Quando a ordem da matriz identidade for evidente, ela será denotada $\mathbf{I}$.

duas submatrizes de $\mathbf{A}$ são

$$\begin{pmatrix} 2 & 5 & 1 \\ -7 & 4 & 0 \\ 3 & 7 & 20 \end{pmatrix} \text{ e } \begin{pmatrix} 2 & 5 & 1 & 4 \\ -7 & 4 & 0 & 10 \end{pmatrix},$$

quando eliminamos, respectivamente, a quarta coluna ou terceira linha de $\mathbf{A}$.

Matriz particionada: *Uma matriz particionada de dimensão* $(m \times n)$ *é uma matriz expressa na forma*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \dots & \mathbf{A}_{1s} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \dots & \mathbf{A}_{2s} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \mathbf{A}_{p3} & \dots & \mathbf{A}_{ps} \end{pmatrix},$$

com $\mathbf{A}_{ij}$ *representando uma submatriz de dimensão* $(m_i \times n_j)$, $i = 1, \dots, p$; $j = 1, \dots, s$ *com* $m_1, \dots, m_p, n_1, \dots, n_s$ *representando números inteiros positivos, tais que* $\sum_{i=1}^{p} m_i = m$ *e* $\sum_{j=1}^{s} n_j = n$. Por exemplo, a matriz

$$\mathbf{A} = \begin{pmatrix} 2 & 5 & 1 & 4 \\ -7 & 4 & 0 & 10 \\ 3 & 7 & 20 & 8 \end{pmatrix},$$

pode ser representada como

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

em que

$$\mathbf{A}_{11} = \begin{pmatrix} 2 & 5 \\ -7 & 4 \end{pmatrix}, \quad \mathbf{A}_{12} = \begin{pmatrix} 1 & 4 \\ 0 & 10 \end{pmatrix}, \quad \mathbf{A}_{21}^{\top} = \begin{pmatrix} 3 \\ 7 \end{pmatrix} \text{ e } \mathbf{A}_{22}^{\top} = \begin{pmatrix} 20 \\ 8 \end{pmatrix},$$

com $m_1 = n_1 = n_2 = 2$ e $m_2 = 1$. Obviamente, uma matriz pode ser particionada de várias maneiras.

A.1.4 Independência linear e espaço coluna

Combinação linear: *Sejam* $\mathbf{x}_1, \dots, \mathbf{x}_p$ *vetores de dimensão* $(n \times 1)$. *O vetor* $\mathbf{u}$, *de dimensão* $(n \times 1)$, *é uma combinação linear dos vetores* $\mathbf{x}_1, \dots, \mathbf{x}_p$, *se existem* $c_1, \dots, c_p$, *números reais tais que* $\mathbf{u} = c_1\mathbf{x}_1 + \dots + c_p\mathbf{x}_p$.

Independência linear: *Os vetores* $\mathbf{x}_1, \dots, \mathbf{x}_p$ *de dimensão* $(n \times 1)$ *são linearmente independentes (L.I.) se, e somente se, a combinação linear* $c_1\mathbf{x}_1 + \dots + c_p\mathbf{x}_p = \mathbf{0}$ *implicar* $c_1 = \dots = c_p = 0$.

Espaço-coluna: *Seja* $\mathbf{X}$ *uma matriz de dimensão* $(n \times p)$*. O espaço-coluna da matriz* $\mathbf{X}$*, denotado* $\mathcal{C}(\mathbf{X})$*, é o conjunto de todas as combinações lineares dos vetores coluna da matriz* $\mathbf{X}$*, ou seja,* $\mathcal{C}(\mathbf{X}) = \{\mathbf{y} \in I\!R^n : \mathbf{y} = \mathbf{X}\mathbf{a},\ \mathbf{a} \in I\!R^p\}$*.*

Posto (*rank*) de uma matriz: *Seja* $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ *uma matriz de dimensão* $(m \times n)$*. O posto de* $\mathbf{A}$*, denotado* $r(\mathbf{A})$*, é o número máximo de colunas (linhas) linearmente independentes de* $\mathbf{A}$*, ou seja, é a dimensão do espaço-coluna (linha) de* $\mathbf{A}$*.*

Sejam $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ matrizes de dimensões $(m \times n)$, $(n \times q)$ e $(q \times t)$, respectivamente, então valem as seguintes propriedades:

i) $\mathrm{r}(\mathbf{A}) = \mathrm{r}(\mathbf{A}^\top)$;

ii) $\mathrm{r}(\mathbf{AB}) \leq \min[\mathrm{r}(\mathbf{A}), \mathrm{r}(\mathbf{B})]$;

iii) Se $\mathrm{r}(\mathbf{A}) = n$ e $\mathrm{r}(\mathbf{B}) = q < n$, então $\mathrm{r}(\mathbf{AB}) = q$;

iv) $\mathrm{r}(\mathbf{AB}) + \mathrm{r}(\mathbf{BC}) \leq \mathrm{r}(\mathbf{B}) + \mathrm{r}(\mathbf{ABC})$.

Matriz de posto completo: *Uma matriz* $\mathbf{A}$ *de dimensão* $(m \times n)$ *tem posto completo quando* $r(\mathbf{A}) = \min(m, n)$*.*

Matriz não singular: *Uma matriz matriz quadrada* $\mathbf{A}$ *de ordem* n *é não singular se* $r(\mathbf{A}) = n$*.*

A.1.5 Determinante de uma matriz

Determinante: *O determinante de uma matriz quadrada* $\mathbf{A}$ *de ordem* n*, denotado* $|\mathbf{A}|$*, é*

$$|\mathbf{A}| = \sum_{k=1}^{n} a_{ik}(-1)^{i+k}|\mathbf{A}_{ik}|,$$

em que $\mathbf{A}_{ik}$ *é obtida a partir da matriz* $\mathbf{A}$ *excluindo-se sua* i*-ésima linha e* k*-ésima coluna. O determinante* $|\mathbf{A}_{ik}|$ *é chamado* ***menor*** *de* $\mathbf{A}$*. Quando* $i = k$*, o determinante é chamado de* ***menor principal*** *de* $\mathbf{A}$*. O termo* $(-1)^{i+k}|\mathbf{A}_{ik}|$ *é denominado* ***cofator****.* Se

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

então,

$$\begin{aligned} |\mathbf{A}| &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} \\ &- a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} - a_{11}a_{23}a_{32}. \end{aligned}$$

O determinante satisfaz as seguintes propriedades:

i) $|\mathbf{A}| \neq 0$ se e somente se a matriz $\mathbf{A}$ tem posto completo. Quando $|\mathbf{A}| \neq 0$, a matriz $\mathbf{A}$ é **não singular**;

ii) $|\mathbf{A}| = |\mathbf{A}^\top|$;

iii) $|c\mathbf{A}| = c^n|\mathbf{A}|$, $c \in I\!R$;

iv) Se $\mathbf{A}$ é uma matriz triangular, então $|\mathbf{A}| = \prod_{i=1}^{n} a_{ii}$;

v) Sejam $\mathbf{A}$ e $\mathbf{B}$ duas matrizes quadradas de mesma ordem; então $|\mathbf{AB}| = |\mathbf{BA}| = |\mathbf{A}||\mathbf{B}|$;

vi) Sejam $\mathbf{A}$ uma matriz $(m \times n)$ e $\mathbf{B}$ uma matriz $(n \times m)$; então

$$|\mathbf{I}_m + \mathbf{AB}| = |\mathbf{I}_n + \mathbf{BA}|.$$

A.1.6 Inversão de matrizes

Matriz inversa: *A matriz inversa (quando existe) de uma matriz matriz quadrada* $\mathbf{A}$ *de ordem* n *é uma matriz* $\mathbf{A}^{-1}$ *quadrada de ordem* n *tal que* $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$.

O teorema abaixo, relaciona a existência da inversa de $\mathbf{A}$ com o fato de seu determinante ser diferente de zero.

***Teorema* A.1.1.** *Uma matriz quadrada* $\mathbf{A}$ *é inversível e sua inversa é única, se e somente se ela for não singular.*

Supondo que todas as matrizes inversas existam, valem as seguintes propriedades:

i) $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$;

ii) Se $|\mathbf{A}| \neq 0$, então $\mathbf{A}^\top$ e $\mathbf{A}^{-1}$ são matrizes não singulares e além disso $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$;

iii) $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$;

iv) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$;

v) Sejam $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ matrizes com dimensões $(k \times k)$, $(k \times n)$ e $(n \times k)$, respectivamente, com $\mathbf{A}$ não singular. Então

$$|\mathbf{A} + \mathbf{BC}| = |\mathbf{A}||\mathbf{I}_k + \mathbf{A}^{-1}\mathbf{BC}|;$$

vi) Sejam $\mathbf{A}, \mathbf{B}, \mathbf{C}$ e $\mathbf{D}$ matrizes com dimensões $(m \times m)$, $(m \times n)$, $(n \times n)$ e $(n \times m)$, respectivamente. Então

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}.$$

vii) Seja $\mathbf{J}_n$ uma matriz quadrada de ordem n com todos elementos iguais a 1 e a, b números reais positivos. Então

$$[a\mathbf{I}_n + b\mathbf{J}_n]^{-1} = \frac{1}{a}\left[\mathbf{I}_n - \frac{b}{nb + a}\mathbf{J}_n\right];$$

viii) Sejam $\mathbf{a}$ e $\mathbf{b}$ vetores de dimensão $(n \times 1)$ e $\mathbf{A}$ uma matriz quadrada não singular de ordem n. Se $1 \pm \mathbf{b}^\top\mathbf{A}^{-1}\mathbf{a} \neq 0$, então

$$(\mathbf{A} \pm \mathbf{ab}^\top)^{-1} = \mathbf{A}^{-1} \mp \frac{(\mathbf{A}^{-1}\mathbf{a})(\mathbf{b}^\top\mathbf{A}^{-1})}{1 \pm \mathbf{b}^\top\mathbf{A}^{-1}\mathbf{a}};$$

ix) Sejam $\mathbf{A}$ e $\mathbf{D}$ matrizes quadradas; então

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = \begin{cases} |\mathbf{A}||\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}|, \text{ se } \mathbf{A} \text{ for não singular} \\ |\mathbf{D}||\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}|, \text{ se } \mathbf{D} \text{ for não singular} \end{cases}$$

Se ambas $\mathbf{A}$ e $\mathbf{D}$ forem não singulares, então

$$|\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}| = \frac{|\mathbf{D}|}{|\mathbf{A}|}|\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}|.$$

x) Sejam $\mathbf{A}$ e $\mathbf{D}$ matrizes quadradas; então

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{F}^\top & \mathbf{G} \end{pmatrix},$$

com

$$\begin{aligned} \mathbf{E} &= \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{A}^{-1}, \\ \mathbf{F} &= -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1}, \\ \mathbf{G} &= (\mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})^{-1}. \end{aligned}$$

Inversa generalizada: *Uma matriz inversa generalizada da matriz* $\mathbf{A}$*,* $(m \times n)$*, é qualquer matriz* $\mathbf{G}$ *de dimensão* $(n \times m)$ *que satisfaz a relação*

$$\mathbf{AGA} = \mathbf{A}.$$

Para detalhes sobre essa classe de matrizes veja Harville (1997), por exemplo.

A.1.7 Traço de uma matriz

Traço de uma matriz: *O traço de uma matriz quadrada de ordem n é $tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.*

Por exemplo, se

$$\mathbf{A} = \begin{pmatrix} 6 & 7 & 4 \\ 5 & 9 & 1 \\ 3 & 8 & -2 \end{pmatrix},$$

então $\text{tr}(\mathbf{A}) = 6 + 9 - 2 = 13$.

Considere $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ matrizes quadradas de ordem n, $\mathbf{a}$ um vetor $(m \times 1)$ e a e b números reais. A função traço apresenta as seguintes propriedades:

i) $\text{tr}(\mathbf{I}_n) = n$;

ii) $\text{tr}(a\mathbf{A} \pm b\mathbf{B}) = a\text{tr}(\mathbf{A}) \pm b\text{tr}(\mathbf{B})$;

iii) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$;

iv) $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA})$;

v) Se $\mathbf{A}$ for idempotente, então $\text{tr}(\mathbf{A}) = \text{r}(\mathbf{A})$;

vii) $\text{tr}(\mathbf{A}^\top) = \text{tr}(\mathbf{A})$;

viii) $\text{tr}(\mathbf{AA}^\top) = \text{tr}(\mathbf{A}^\top\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$;

ix) $\text{tr}(\mathbf{aa}^\top) = \mathbf{a}^\top\mathbf{a} = \sum_{i=1}^{n} a_i^2$.

A.1.8 Soma direta e produto de Kronecker

Soma direta: *Sejam $\mathbf{A}$ uma matriz de dimensão $(m \times n)$ e $\mathbf{B}$ uma matriz de dimensão $(p \times q)$. A soma direta das matrizes $\mathbf{A}$ e $\mathbf{B}$ é a matriz diagonal em blocos de dimensão $[(m+p) \times (n+q)]$, definida por*

$$\mathbf{A} \oplus \mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}.$$

De uma forma mais geral, a soma direta de matrizes $\mathbf{A}_i$, com dimensão $(n_i \times m_i)$, $i = 1, \ldots, n$ é a matriz com dimensão $(\sum_{i=1}^{n} n_i \times \sum_{i=1}^{n} m_i)$ dada por

$$\bigoplus_{i=1}^{n} \mathbf{A}_i = \mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \ldots \oplus \mathbf{A}_n = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_n \end{pmatrix}.$$

Produto de Kronecker: *Sejam* $\mathbf{A}$ *e* $\mathbf{B}$ *matrizes de dimensões* $(m \times n)$ *e* $(p \times q)$, *respectivamente. O produto de Kronecker (**produto direto** ou **produto tensorial**) das matrizes* $\mathbf{A}$ *e* $\mathbf{B}$ *é a matriz de dimensão* $(mp \times nq)$, *definida por*

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix}.$$

Em geral $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$.

Sejam $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ matrizes de dimensões $(m \times n)$, $(p \times q)$, $(n \times u)$, $(q \times v)$, respectivamente, $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{d}$ vetores de dimensões $(m \times 1)$, $(n \times 1)$ e $(p \times 1)$, respectivamente, e x e y números reais. Então:

i) $x \otimes \mathbf{A} = \mathbf{A} \otimes x = x\mathbf{A}$;

ii) $\mathbf{a} \otimes \mathbf{b}^\top = \mathbf{b}^\top \otimes \mathbf{a} = \mathbf{a}\mathbf{b}^\top$;

iii) $\mathbf{0}_{p\times q} \otimes \mathbf{A} = \mathbf{A} \otimes \mathbf{0}_{p\times q} = \mathbf{0}_{mp\times nq}$;

iv) $\mathbf{I}_m \otimes \mathbf{I}_p = \mathbf{I}_{mp}$;

v) Se $\mathbf{F} = \text{diag}(f_{11}, \dots, f_{kk})$, então $\mathbf{F} \otimes \mathbf{A} = \bigoplus_{i=1}^{k} f_{ii}\mathbf{A}$;

vi) $\mathbf{I}_k \otimes \mathbf{A} = \bigoplus_{i=1}^{k} \mathbf{A}$;

vii) $x\mathbf{A} \otimes y\mathbf{B} = xy(\mathbf{A} \otimes \mathbf{B})$;

viii) $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$;

ix) $(\mathbf{A} \otimes \mathbf{B}) \otimes (\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$;

x) $\mathbf{A} \otimes \mathbf{B} = (\mathbf{A} \otimes \mathbf{I}_p)(\mathbf{I}_n \otimes \mathbf{B}) = (\mathbf{I}_m \otimes \mathbf{B})(\mathbf{A} \otimes \mathbf{I}_q)$;

xi) $(\mathbf{A} \otimes \mathbf{d}^\top)(\mathbf{b} \otimes \mathbf{B}) = (\mathbf{d}^\top \otimes \mathbf{A})(\mathbf{B} \otimes \mathbf{b}) = \mathbf{A}\mathbf{b}\mathbf{d}^\top\mathbf{B}$;

xii) $\mathbf{D} \otimes (\mathbf{A} + \mathbf{B}) = (\mathbf{D} \otimes \mathbf{A}) + (\mathbf{D} \otimes \mathbf{B})$;

xiii) $(\mathbf{A} \otimes \mathbf{B})^\top = (\mathbf{A}^\top \otimes \mathbf{B}^\top)$;

xiv) $\text{r}(\mathbf{A} \otimes \mathbf{B}) = \text{r}(\mathbf{A})\text{r}(\mathbf{B})$;

xv) $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}(\mathbf{A})\text{tr}(\mathbf{B})$;

Além disso,

xvi) Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes simétricas, então $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A} \otimes \mathbf{B}$;

xvii) Se $\mathbf{A}$ é uma matriz quadrada de ordem n e $\mathbf{a}$ é um vetor $(m \times 1)$, então $(\mathbf{I}_n \otimes \mathbf{a})\mathbf{A}(\mathbf{I}_n \otimes \mathbf{a}^\top) = \mathbf{A} \otimes \mathbf{a}\mathbf{a}^\top$;

xviii) Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes não singulares, temos $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$;

xix) Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes quadradas de ordem m e n, respectivamente, então $|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^n |\mathbf{B}|^m$;

xx) Seja $\mathbf{A} = [\mathbf{A}_1 \;\; \mathbf{A}_2]$; então $[\mathbf{A}_1 \;\; \mathbf{A}_2] \otimes \mathbf{B} = [\mathbf{A}_1 \otimes \mathbf{B} \;\; \mathbf{A}_2 \otimes \mathbf{B}]$, mas $\mathbf{W} \otimes [\mathbf{B}_1 \;\; \mathbf{B}_2] \neq [\mathbf{W} \otimes \mathbf{B}_1 \;\; \mathbf{W} \otimes \mathbf{B}_2]$.

A.1.9 Operadores vec e vech

Operador vec: *A operação de vetorização de uma matriz* $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_n)$ *consiste em "empilhar" seus elementos na forma*

$$vec(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}.$$

Sejam $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ matrizes reais de dimensões $(m \times n)$, $(n \times p)$, $(p \times q)$, respectivamente e vetores $\mathbf{a}$ e $\mathbf{b}$ de dimensões $(m \times 1)$ e $(n \times 1)$, respectivamente. Então:

i) $\text{vec}(\mathbf{a}^\top) = \text{vec}(\mathbf{a})$;

ii) $\text{vec}(\mathbf{a}\mathbf{b}^\top) = \mathbf{b} \otimes \mathbf{a}$;

iii) $\text{vec}(\mathbf{AB}) = (\mathbf{I}_p \otimes \mathbf{A})\text{vec}(\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I}_m)\text{vec}(\mathbf{A})$;

iv) $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B})$;

v) $\text{vec}(\mathbf{ABC}) = (\mathbf{I}_q \otimes \mathbf{AB})\text{vec}(\mathbf{C}) = (\mathbf{C}^\top \mathbf{B}^\top \otimes \mathbf{I}_n)\text{vec}(\mathbf{A})$.

Além disso,

vi) Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes de mesma dimensão, temos

$$\text{tr}(\mathbf{A}^\top \mathbf{B}) = \text{vec}(\mathbf{A})^\top \text{vec}(\mathbf{B})$$

e

$$\text{vec}(\mathbf{A}^\top)^\top \text{vec}(\mathbf{B}) = \text{vec}(\mathbf{B}^\top)^\top \text{vec}(\mathbf{A}) = \text{tr}(\mathbf{AB});$$

vii) Se $\mathbf{B}$ for uma matriz de dimensão $(n \times m)$, então

$$\text{tr}(\mathbf{AB}) = \text{vec}(\mathbf{A}^\top)^\top \text{vec}(\mathbf{B});$$

viii) Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes simétricas de ordem n, então

$$\text{vec}(\mathbf{A})^\top (\mathbf{B} \otimes \mathbf{B}) \text{vec}(\mathbf{A}) = [\text{tr}(\mathbf{BA})]^2;$$

ix) Se $\mathbf{C}$ é uma matriz de dimensão $(p \times m)$, temos

$$\begin{aligned}
\text{tr}(\mathbf{ABC}) &= \text{vec}(\mathbf{A}^\top)^\top (\mathbf{C}^\top \otimes \mathbf{I}_n) \text{vec}(\mathbf{B}) \\
&= \text{vec}(\mathbf{A}^\top)^\top (\mathbf{I}_m \otimes \mathbf{B}) \text{vec}(\mathbf{C}) \\
&= \text{vec}(\mathbf{B}^\top)^\top (\mathbf{A} \otimes \mathbf{I}_p) \text{vec}(\mathbf{C}) \\
&= \text{vec}(\mathbf{B}^\top)^\top (\mathbf{I}_n \otimes \mathbf{C}) \text{vec}(\mathbf{A}) \\
&= \text{vec}(\mathbf{C}^\top)^\top (\mathbf{B} \otimes \mathbf{I}_m) \text{vec}(\mathbf{A}) \\
&= \text{vec}(\mathbf{C}^\top)^\top (\mathbf{I}_p \otimes \mathbf{A}) \text{vec}(\mathbf{B})
\end{aligned}$$

O operador vech(.) aplicado a uma matriz simétrica $\mathbf{A}$ gera um vetor com os elementos distintos dessa matriz. Por exemplo, se

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad \text{então} \quad \text{vech}(\mathbf{A}) = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{22} \\ a_{32} \\ a_{33} \end{pmatrix}.$$

A.2 Tópicos de Álgebra Linear

Espaço vetorial: *Um espaço vetorial sobre $\mathbb{R}$ é um conjunto $\mathcal{V}$ não vazio de elementos chamados vetores no qual estão definidas:*

i) Uma operação de adição que associa a cada par de vetores $\mathbf{a}$ e $\mathbf{b}$ de $\mathcal{V}$ um vetor $\mathbf{a} + \mathbf{b} \in \mathcal{V}$ e para a qual são válidas as seguintes propriedades:

1. $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$ *(comutatividade);*
2. $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$ *(associatividade);*
3. *Existe um vetor nulo $\mathbf{0} \in \mathcal{V}$, tal que $\mathbf{a} + \mathbf{0} = \mathbf{a}$ para todo $\mathbf{a} \in \mathcal{V}$;*

4. *Para cada vetor* $\mathbf{a} \in \mathcal{V}$, *existe um vetor* $-\mathbf{a} \in \mathcal{V}$ *tal que* $\mathbf{a} + (-\mathbf{a}) = \mathbf{0}$;

ii) Uma operação de multiplicação por números reais que associa a cada $\alpha \in \mathbb{R}$ *e a cada vetor* $\mathbf{a} \in \mathcal{V}$ *um vetor* $\alpha\mathbf{a} \in \mathcal{V}$ *para a qual são válidas as seguintes propriedades:*

1. $1\mathbf{a} = \mathbf{a} \ \ \forall \mathbf{a} \in \mathcal{V}$;
2. $(\alpha\beta)\mathbf{a} = \alpha(\beta\mathbf{a}), \ \forall \alpha, \beta \in \mathbb{R}$
3. $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$;
4. $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$.

O espaço vetorial mais utilizado nas aplicações em Estatística é o espaço das matrizes reais de dimensão $(n \times p)$, denotado $\mathbb{R}^{n \times p}$.

Subespaço vetorial: *Sejam* $\mathcal{V}$ *um espaço vetorial sobre* $\mathbb{R}$ *e* $\mathcal{W}$ *um subconjunto não vazio de* $\mathcal{V}$. *Dizemos que que* $\mathcal{W}$ *é um subespaço vetorial de* $\mathcal{V}$ *se valem as seguintes propriedades:*

i) se $\mathbf{a}$ *e* $\mathbf{b} \in \mathcal{W}$, *então* $\mathbf{a} + \mathbf{b} \in \mathcal{W}$;

ii) se $\alpha \in \mathbb{R}$ *e* $\mathbf{a} \in \mathcal{W}$, *então* $\alpha\mathbf{a} \in \mathcal{W}$.

Base de um espaço vetorial: *Seja* $\mathcal{V}$ *um espaço vetorial. Se qualquer vetor* $\mathbf{x} \in \mathcal{V}$ *puder ser escrito como uma combinação linear de um conjunto de vetores linearmente independentes* $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subset \mathcal{V}$, *então dizemos que* $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ *é uma base do espaço vetorial* $\mathcal{V}$.

Dimensão de um espaço vetorial: *A dimensão do espaço vetorial* $\mathcal{V}$, *denotada* $dim\mathcal{V}$, *é igual ao número de vetores que formam uma base de* $\mathcal{V}$.

Transformação linear: *Sejam* $\mathcal{U}$ *e* $\mathcal{V}$ *espaços vetoriais. Uma transformação linear de* $\mathcal{U}$ *em* $\mathcal{V}$, *denotada por* $\mathbf{T} : \mathcal{U} \to \mathcal{V}$, *é uma função que associa a cada vetor* $\mathbf{v} \in \mathcal{U}$ *um vetor* $\mathbf{T}(\mathbf{v}) \in \mathcal{V}$, *de modo que, para quaisquer vetores* $\mathbf{a}, \mathbf{b} \in \mathcal{V}$ *e* $k \in \mathbb{R}$, *valem as seguintes propriedades:*

1. $\mathbf{T}(\mathbf{a} + \mathbf{b}) = \mathbf{T}(\mathbf{a}) + \mathbf{T}(\mathbf{b})$;
2. $\mathbf{T}(k\mathbf{a}) = k\mathbf{T}(\mathbf{a}), \ \forall k \in \mathbb{R}$.

As transformações lineares frequentemente utilizadas em Estatística são aquelas em que $\mathbf{T}$ é uma função vetorial do tipo $\mathbf{T} : \mathbb{R}^p \to \mathbb{R}^n$ definida por uma matriz $\mathbf{A}$

de dimensão $(p \times n)$, tal que $\forall \mathbf{x} \in I\!R^p$,

$$\mathbf{T}(\mathbf{x}) = \mathbf{A}\mathbf{x} = \begin{pmatrix} a_{11} & a_{11} & \dots & a_{1p} \\ a_{21} & a_{21} & \dots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \vdots & a_{np} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{p1} \end{pmatrix}.$$

***Teorema* A.2.1.** *Sejam $\mathbf{x}_1, \dots, \mathbf{x}_p$ vetores de dimensão $(n \times 1)$ pertencentes ao espaço vetorial $I\!R^n$ e $\mathcal{W}$ o conjunto definido por*

$$\mathcal{W} = \{\mathbf{b} \in I\!R^n | \ \mathbf{b} = \sum_{i=1}^{p} \alpha_i x_i = \mathbf{X}\boldsymbol{\alpha}, \ \mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in I\!R^{n \times p}, \ \boldsymbol{\alpha} \in I\!R^p\}$$

com $\mathbf{X}$ denotando a matriz da transformação linear de $I\!R^p$ em $I\!R^n$ $(p < n)$; então $\mathcal{W}$ é um subespaço de $I\!R^n$.

Espaço coluna de uma matriz: *Seja $\mathbf{X}$ uma matriz de dimensão $(n \times p)$; o espaço coluna de $\mathbf{X}$, denotado por $\mathcal{C}(\mathbf{X})$ é o espaço vetorial gerado por suas colunas.*

Espaço nulo de uma matriz: *O espaço nulo de uma matriz $\mathbf{X}$ com dimensão $(n \times p)$, denotado $\mathcal{N}(\mathbf{X})$, é o conjunto de vetores $\mathbf{a} \in I\!R^p$ tais que $\mathbf{X}\mathbf{a} = \mathbf{0}$, ou seja, $\mathcal{N}(\mathbf{X}) = \{\mathbf{a} \in I\!R^p; \ \mathbf{X}\mathbf{a} = \mathbf{0}\}$.*

***Teorema* A.2.2.** *Sejam $\mathbf{X}$ uma matriz de dimensão $(n \times p)$ com $r(\mathbf{X}) = r$ e $\mathcal{C}(\mathbf{X})$ seu espaço coluna. Então $r(\mathbf{X}) = dim[\mathcal{C}(\mathbf{X})] = r(\mathbf{X}^\top) = dim[\mathcal{C}(\mathbf{X}^\top)] = r$ e além disso, $dim[\mathcal{N}(\mathbf{X})] = p - r$ e $dim[\mathcal{N}(\mathbf{X}^\top)] = n - r$.*

***Teorema* A.2.3.** *Sejam $\mathbf{A}$ e $\mathbf{B}$ matrizes com dimensões $(n \times m)$ e $(m \times p)$, respectivamente. Então $\mathcal{C}(\mathbf{AB})$ é um subespaço de $\mathcal{C}(\mathbf{A})$.*

***Teorema* A.2.4.** *Seja $\mathbf{X}$ uma matriz de dimensão $(n \times p)$ com $r(\mathbf{X}) = r$. Então $\mathcal{C}(\mathbf{X}^\top) = \mathcal{C}(\mathbf{X}^\top\mathbf{X})$ e $r(\mathbf{X}^\top\mathbf{X}) = r$.*

Produto interno: *No espaço $I\!R^n$, o produto interno canônico dos vetores $\mathbf{x}$ e $\mathbf{y}$ é um número real dado por*

$$\mathbf{x} \bullet \mathbf{y} = \mathbf{x}^\top\mathbf{y} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Espaço euclidiano: *O espaço euclidiano $I\!R^n$ é o espaço vetorial $I\!R^n$ com soma e produto por escalar definidos da forma usual, munido do produto interno canônico.*

Norma de um vetor: *No espaço euclidiano $\mathbb{R}^n$, a norma (ou comprimento) do vetor* $\mathbf{x}$ *é o número* $\|\mathbf{x}\| = (\mathbf{x}^\top\mathbf{x})^{\frac{1}{2}}$. Quando a norma de um vetor $\mathbf{x}$ é igual a 1, diz-se que $\mathbf{x}$ é um vetor unitário.

Distância euclidiana: *A distância euclidiana entre os vetores* $\mathbf{x}$ *e* $\mathbf{y}$ *de* $\mathbb{R}^n$ *é o número* $\|\mathbf{x}-\mathbf{y}\|$.

Para $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c} \in \mathbb{R}^n$ e $k \in \mathbb{R}$, o produto interno canônico tem as seguintes propriedades:

i) $\mathbf{a}^\top\mathbf{b} = \mathbf{b}^\top\mathbf{a}$;

ii) $\mathbf{a}^\top(\mathbf{b}+\mathbf{c}) = \mathbf{a}^\top\mathbf{b} + \mathbf{a}^\top\mathbf{c}$;

iii) $k(\mathbf{a}^\top\mathbf{b}) = (k\mathbf{a})^\top\mathbf{b} = \mathbf{a}^\top(k\mathbf{b})$;

iv) $\mathbf{a}^\top\mathbf{a} = \|\mathbf{a}\|^2 > \mathbf{0}$ se $\mathbf{a} \neq \mathbf{0}$;

v) $\|\mathbf{a} \pm \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \pm 2\mathbf{a}^\top\mathbf{b}$;

vi) $|\mathbf{a}^\top\mathbf{b}| \leq \|\mathbf{a}\|\|\mathbf{b}\|$; **Desigualdade de Cauchy-Schwarz**

vii) $\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$; **Desigualdade triangular**

Ângulo entre vetores: *O ângulo* $\theta \in [0, \pi]$ *entre dois vetores* $\mathbf{a}$ *e* $\mathbf{b} \in \mathbb{R}^n$ *é*

$$\arccos(\theta) = \frac{\mathbf{a}^\top\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}.$$

Produto interno de matrizes: *No espaço $\mathbb{R}^{m\times n}$, o produto interno canônico das matrizes* $\mathbf{A}$ *e* $\mathbf{B}$ *é o número real* $tr(\mathbf{A}^\top\mathbf{B}) = tr(\mathbf{A}\mathbf{B}^\top)$.

Norma de uma matriz: *A norma da matriz* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *(comumente denominada* **norma de Frobenius***) é o número real*

$$\|\mathbf{A}\| = [tr(\mathbf{A}^\top\mathbf{A})]^{\frac{1}{2}} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}} = \|vec(\mathbf{A})\|.$$

Vetores ortogonais: *Se $\mathcal{V}$ é um espaço vetorial com produto interno canônico, dizemos que* $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ *são ortogonais (*$\mathbf{x} \perp \mathbf{y}$*), se, e somente se* $\mathbf{x}^\top\mathbf{y} = 0$.

Complemento ortogonal: *Se $\mathcal{V}$ é um espaço vetorial com produto interno e $\mathcal{W}$ é um subespaço de $\mathcal{V}$, o conjunto* $\mathcal{W}^\perp = \{\mathbf{a} \in \mathcal{V};\ \mathbf{a} \bullet \mathbf{v} = 0\ \forall \mathbf{v} \in \mathcal{V}\}$ *é um subespaço vetorial de $\mathcal{V}$, denominado complemento ortogonal de $\mathcal{W}$.*

***Teorema* A.2.5.** *Seja uma matriz* $\mathbf{X}$ *de dimensão* $(n \times p)$*. O espaço nulo de* $\mathbf{X}^\top$ *e o complemento ortogonal do espaço coluna de* $\mathbf{X}$ *são iguais, ou seja,* $\mathcal{N}(\mathbf{X}^\top) = \mathcal{C}(\mathbf{X})^\perp$ *e* $\mathcal{N}(\mathbf{X}) = \mathcal{C}(\mathbf{X}^\top)^\perp$*.*

Subespaço ortogonal: *Sejam* $\mathcal{V}$ *um espaço vetorial com produto interno,* $\mathcal{U}$ *e* $\mathcal{W}$*, subespaços de* $\mathcal{V}$*. O subespaço* $\mathcal{U}$ *é ortogonal ao subespaço* $\mathcal{W}$ $(\mathcal{U} \perp \mathcal{W})$*, se cada vetor de* $\mathcal{U}$ *for ortogonal a cada vetor de* $\mathcal{W}$*. Além disso, dizemos que* $\mathbf{v} \perp \mathcal{U}$ *se* $\mathbf{v} \bullet \mathbf{u} = 0 \ \ \forall \mathbf{u} \in \mathcal{U}$*.*

Sejam $\mathbf{y}$ um vetor de dimensão $(m \times 1)$ e $\mathbf{X}$ e $\mathbf{Z}$ matrizes com dimensões $(m \times n)$ e $(m \times p)$, respectivamente. Então $\mathbf{y}$ é ortogonal ao espaço coluna da matriz $\mathbf{X}$ (com relação ao produto interno canônico de $\mathbb{R}^m$), nomeadamente, $\mathcal{C}(\mathbf{X})$, se e somente se $\mathbf{X}^\top\mathbf{y} = \mathbf{0}$. De modo similar, o espaço coluna de $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$ é ortogonal ao espaço coluna de $\mathbf{Z}$, $\mathcal{C}(\mathbf{Z})$, se e somente se $\mathbf{X}^\top\mathbf{Z} = \mathbf{0}$.

Vetores ortonormais: *Seja* $\mathcal{V}$ *um espaço vetorial com produto interno e* $\mathbf{x}$ *e* $\mathbf{y} \in \mathcal{V}$*. Os vetores* $\mathbf{x}$ *e* $\mathbf{y}$ *são ortonormais se* $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ *e* $\mathbf{x} \perp \mathbf{y}$*.*

Base ortonormal: *Seja* $\mathcal{V}$ *um espaço vetorial de dimensão finita* n *com produto interno. Uma base* $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ *de* $\mathcal{V}$ *é dita ortonormal se seus elementos forem vetores de norma igual a 1 (*$\|\mathbf{x}_i\| = 1$*, para* $i = 1, \ldots, n$*) e forem ortogonais dois a dois.*

***Teorema* A.2.6.** *Sejam* $\mathbf{Y}$ *uma matriz no espaço vetorial* $\mathcal{V} \subset \mathbb{R}^{nm}$ *das matrizes* $(m \times n)$ *e* $\mathcal{U}$ *um subespaço de* $\mathcal{V}$*. Então existe uma única matriz* $\mathbf{Z} \in \mathcal{U}$*, tal que* $(\mathbf{Y} - \mathbf{Z}) \in \mathcal{U}^\perp$*. A matriz* $\mathbf{Z}$ *é a* **projeção ortogonal** *de* $\mathbf{Y}$ *em* $\mathcal{U}$*.*

Matriz base: *Uma matriz* $\mathbf{X}$ *de dimensão* $(m \times n)$ *é uma matriz base do subespaço* $\mathcal{U} \subset \mathbb{R}^m$*, se os vetores coluna de* $\mathbf{X}$ *formam uma base de* $\mathcal{U}$*; se os vetores coluna de* $\mathbf{X}$ *forem ortonormais, ela é uma* **base ortonormal** *de* $\mathcal{U}$*.*

***Lema* A.2.1.** *Se* $\mathbf{X}$ *de dimensão* $(n \times p)$ *é uma matriz base do subespaço* $\mathcal{U} \subset \mathbb{R}^n$*, então*

i) $\mathbf{X}$ *é uma matriz de posto* p *e* $\mathbf{X}^\top\mathbf{X}$ *é inversível;*

ii) $\mathbf{v} \in \mathcal{U}$ *se e somente se,* $\mathbf{v} = \mathbf{X}\mathbf{b}$ *para algum* $\mathbf{b} \in \mathbb{R}^p$*.*

Projeção ortogonal de um vetor: *Sejam* $\mathbf{y} \in \mathbb{R}^n$ *e* $\mathcal{U}$ *um subespaço do* $\mathbb{R}^n$*. A projeção ortogonal de* $\mathbf{y}$ *em* $\mathcal{U}$ *é um vetor* $\mathbf{x} \in \mathcal{U}$ *tal que* $\mathbf{y} - \mathbf{x} \in \mathcal{U}^\perp$*.*

***Teorema* A.2.7.** *Sejam* $\mathcal{U}$ *um subespaço do* $\mathbb{R}^n$ *e um vetor* $\mathbf{y} \in \mathbb{R}^n$*. Então,*

i) A projeção ortogonal de $\mathbf{y}$ *em* $\mathcal{U}$ *é única;*

ii) *Se* $\mathbf{X}$ *é a matriz base do subespaço* $\mathcal{U}$*, a projeção ortogonal de* $\mathbf{y}$ *em* $\mathcal{U}$ *é o vetor* $\mathbf{z} = \mathbf{X}\mathbf{b}^*$ *de dimensão* $(n \times 1)$ *em que* $\mathbf{b}^*$ *é a solução do sistema* $\mathbf{X}^\top\mathbf{X}\mathbf{b} = \mathbf{X}^\top\mathbf{y}$*, ou seja,* $\mathbf{z} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$*. A matriz* $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ *é denominada matriz de projeção.*

iii) $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$*;*

iv) $\mathbf{P}_\mathbf{X}$ *e* $\mathbf{I} - \mathbf{P}_\mathbf{X}$ *são matrizes simétricas e idempotentes;*

v) $\mathcal{C}(\mathbf{P}_\mathbf{X}) = \mathcal{C}(\mathbf{X})$*;*

vi) $r(\mathbf{P}_\mathbf{X}) = r(\mathbf{X})$ *e* $r(\mathbf{I} - \mathbf{P}_\mathbf{X}) = n - r(\mathbf{X})$

Autovalor: *Seja* $\mathbf{A}$ *uma matriz quadrada de ordem* n*. As raízes do polinômio característico* $|\mathbf{A} - \lambda\mathbf{I}|$*, denotadas* $\lambda_1, \ldots, \lambda_n$*, são denominadas autovalores (ou raízes características) da matriz* $\mathbf{A}$*. A equação* $|\mathbf{A} - \lambda\mathbf{I}| = 0$ *é denominada equação característica da matriz* $\mathbf{A}$*.*

Por exemplo, se

$$\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix},$$

seus autovalores correspondem às soluções da equação característica

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= \left| \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \\ &= (1-\lambda)^2 - 36 = 0, \end{aligned}$$

ou sejam, $\lambda_1 = -5$ e $\lambda_2 = 7$.

Autovetor (Vetor característico): *Seja* $\mathbf{A}$ *uma matriz quadrada de ordem* n *e* λ *um autovalor de* $\mathbf{A}$*. Se* $\mathbf{v}$ *é um vetor (não nulo) tal que* $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$*, então* $\mathbf{v}$ *é denominado autovetor (ou vetor característico) da matriz* $\mathbf{A}$*.*

Para o exemplo acima, o autovetor associado ao autovalor $\lambda_1 = -5$ é obtido do sistema

$$\begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = -5 \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix}$$

que tem infinitas soluções; um possível autovetor associado ao autovalor $\lambda_1 = -5$ é $\mathbf{v}_1 = (2 \ -3)^\top$. De modo similar, obtemos um autovetor associado ao autovalor $\lambda_2 = 7$, nomeadamente, $\mathbf{v}_2 = (2 \ 3)^\top$.

***Teorema* A.2.8.** *Seja* $\mathbf{A}$ *uma matriz quadrada de ordem* n *e* $\lambda_1, \ldots, \lambda_n$ *seus autovalores; então*

i) $|\mathbf{A}| = \prod_{i=1}^n \lambda_i$;

ii) $tr(\mathbf{A}) = \sum_{i=1}^n \lambda_i$.

A.3 Formas lineares, bilineares e quadráticas

Forma linear: *Uma forma linear é uma função* $f : \mathbb{R}^n \to \mathbb{R}$ *que associa a cada vetor* $\mathbf{x} \in \mathbb{R}^n$ *o número real*

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} = \sum_{i=1}^{n} a_i x_i$$

em que $\mathbf{a} \in \mathbb{R}^n$ *é denominado vetor de coeficientes.*

Forma bilinear: *Uma forma bilinear é uma função* $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ *que associa a cada par de vetores* $\mathbf{x} \in \mathbb{R}^m$ *e* $\mathbf{y} \in \mathbb{R}^n$ *o número real*

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{A} \mathbf{y} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_i y_j$$

em que $\mathbf{A}$ *é uma matriz de coeficientes de dimensão* $(m \times n)$.

Forma quadrática: *Uma forma quadrática é uma função* $f : \mathbb{R}^n \to \mathbb{R}$ *que associa ao vetor* $\mathbf{x} \in \mathbb{R}^n$ *o número real*

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

em que $\mathbf{A}$ *é uma matriz de coeficientes quadrada de ordem* n.

Matriz definida não negativa: *Uma matriz* $\mathbf{A}$*, quadrada de ordem* n *é denominada matriz definida não negativa se* $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ *para todo* $\mathbf{x} \in \mathbb{R}^n$.

Matriz definida positiva: *Uma matriz* $\mathbf{A}$*, quadrada de ordem* n *é denominada matriz definida positiva se* $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ *para todo vetor não nulo* $\mathbf{x} \in \mathbb{R}^n$ *e* $\mathbf{x}^\top \mathbf{A} \mathbf{x} = 0$ *somente quando* $\mathbf{x} = \mathbf{0}$.

Matriz semidefinida positiva: *Uma matriz* $\mathbf{A}$*, quadrada de ordem* n *é denominada matriz semidefinida positiva se* $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ *para* $\mathbf{x} \in \mathbb{R}^n$ *e* $\mathbf{x}^\top \mathbf{A} \mathbf{x} = 0$ *para algum vetor* $\mathbf{x}$ *não nulo.*

***Teorema* A.3.1.** *Seja* $\mathbf{A}$ *uma matriz de dimensão* $(n \times n)$ *e* $\mathbf{M}$ *uma matriz de dimensão* $(n \times m)$*. Então*

i) Se $\mathbf{A}$ *for definida não negativa, então* $\mathbf{M}^\top \mathbf{A} \mathbf{M}$ *é definida não negativa;*

ii) Se $\mathbf{A}$ *for definida não negativa e* $r(\mathbf{M}) < m$ *então* $\mathbf{M}^\top \mathbf{A} \mathbf{M}$ *é semidefinida positiva;*

iii) Se $\mathbf{A}$ *for definida positiva e* $r(\mathbf{M}) = m$ *então* $\mathbf{M}^\top \mathbf{A} \mathbf{M}$ *é definida positiva;*

Formas quadráticas envolvendo vetores com distribuição Normal são extremamente importantes para aplicações estatísticas. Nesse contexto, apresentaremos alguns resultados bastante úteis para inferência em modelos lineares em geral. O leitor poderá consultar Searle (1971) para detalhes e demonstrações.

***Teorema* A.3.2.** *Se* $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{V})$ *e* $\mathbf{A}$ *é uma matriz simétrica, então*

i) $\mathbb{E}(\mathbf{y}^\top \mathbf{A}\mathbf{y}) = tr(\mathbf{A}\mathbf{V}) + \mu^\top \mathbf{A}\boldsymbol{\mu}$*;*

ii) o cumulante de ordem r *de* $\mathbf{y}^\top \mathbf{A}\mathbf{y}$ *é*

$$K_r(\mathbf{y}^\top \mathbf{A}\mathbf{y}) = 2^{r-1}(r-1)![tr(\mathbf{A}\mathbf{V})^r + r\boldsymbol{\mu}^\top \mathbf{A}(\mathbf{V}\mathbf{A})^{r-1}\boldsymbol{\mu}];$$

iii) $\mathbb{C}\mathrm{ov}(\mathbf{y}, \mathbf{y}^\top \mathbf{A}\mathbf{y}) = 2\mathbf{V}\mathbf{A}\boldsymbol{\mu}$*;*

O item i) prescinde da suposição de normalidade. Tomando $r = 2$, uma aplicação direta desse resultado permite-nos calcular a variância de formas quadráticas envolvendo vetores com distribuição Normal, nomeadamente

$$\mathbb{V}(\mathbf{y}^\top \mathbf{A}\mathbf{y}) = 2tr(\mathbf{A}\mathbf{V})^2 + 4\boldsymbol{\mu}^\top \mathbf{A}(\mathbf{V}\mathbf{A})\boldsymbol{\mu};$$

se além disso, $\boldsymbol{\mu} = \mathbf{0}$ então $\mathbb{V}(\mathbf{y}^\top \mathbf{A}\mathbf{y}) = 2tr(\mathbf{A}\mathbf{V})^2$.

***Teorema* A.3.3.** *Se* $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{V})$ *e* $\mathbf{A}$ *é uma matriz simétrica com posto* r*, então* $\mathbf{y}^\top \mathbf{A}\mathbf{y} \sim \chi^2_r(\delta)$*, em que* $\delta = \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}$ *é o parâmetro de não centralidade, se e somente se* $\mathbf{A}\mathbf{V}$ *for idempotente.*

***Teorema* A.3.4.** *Se* $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{V})$*,* $\mathbf{A}$ *é uma matriz simétrica com posto* r *e* $\mathbf{B}$ *é uma matriz com dimensão* $(b \times p)$ *então* $\mathbf{y}^\top \mathbf{A}\mathbf{y}$ *e* $\mathbf{B}\mathbf{y}$ *têm distribuições independentes se e somente se* $\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{0}$*.*

Note que o teorema não envolve o produto $\mathbf{AVB}$, que pode não existir.

***Teorema* A.3.5.** *Se* $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{V})$*,* $\mathbf{A}$ *e* $\mathbf{B}$ *são matrizes simétricas então* $\mathbf{y}^\top \mathbf{A}\mathbf{y}$ *e* $\mathbf{y}^\top \mathbf{B}\mathbf{y}$ *têm distribuições independentes se e somente se* $\mathbf{A}\mathbf{V}\mathbf{B} = \mathbf{0}$ *ou equivalentemente se* $\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{0}$*.*

A.4 Decomposição de matrizes

***Teorema* A.4.1.** *Para toda matriz simétrica* $\mathbf{A}$ *de dimensão* $(n \times n)$ *existe uma matriz não singular* $\mathbf{Q}$ *tal que* $\mathbf{Q}^\top \mathbf{A}\mathbf{Q}$ *é uma matriz diagonal.*

***Teorema* A.4.2.** *Seja* $\mathbf{A}$ *uma matriz de dimensão* $(n \times n)$*. Então existem uma matriz não singular* $\mathbf{Q}$ *e uma matriz diagonal* $\mathbf{D}$ *tais que* $\mathbf{A} = \mathbf{Q}^\top \mathbf{D} \mathbf{Q}$*.*

***Teorema* A.4.3.** *Uma matriz* $\mathbf{A}$ *(não nula) de dimensão* $(n \times n)$ *é simétrica definida não negativa com* $r(\mathbf{A}) = r$ *se e somente se existe uma matriz* $\mathbf{Q}$ *de dimensão* $(r \times n)$ *com* $r(\mathbf{A}) = r$ *tal que* $\mathbf{A} = \mathbf{Q}^\top \mathbf{Q}$*.*

***Teorema* A.4.4.** *Uma matriz* $\mathbf{A}$ *(não nula) de dimensão* $(n \times n)$ *é simétrica definida positiva se e somente se existe uma matriz não singular* $\mathbf{Q}$ *tal que* $\mathbf{A} = \mathbf{Q}^\top \mathbf{Q}$*.*

***Teorema* A.4.5.** *Uma matriz simétrica definida não negativa* $\mathbf{A}$ *de dimensão* $(n \times n)$ *é definida positiva se e somente se ela for não singular (ou equivalentemente, ela é semidefinida positiva se e somente se ela for singular).*

***Teorema* A.4.6.** *Uma matriz definida positiva* $\mathbf{A}$ *de dimensão* $(n \times n)$ *tem uma única decomposição do tipo* $\mathbf{A} = \mathbf{L}^\top \mathbf{D} \mathbf{U}$ *em que* $\mathbf{L}$ *é uma matriz triangular inferior,* $\mathbf{U}$ *é uma matriz triangular superior e* $\mathbf{D}$ *é uma matriz diagonal com todos os elementos da diagonal principal positivos.*

***Teorema* A.4.7.** *Uma matriz simétrica definida positiva* $\mathbf{A}$ *de dimensão* $(n \times n)$ *tem uma única decomposição do tipo* $\mathbf{A} = \mathbf{U}^\top \mathbf{D} \mathbf{U}$ *em que* $\mathbf{U}$ *é uma matriz triangular superior e* $\mathbf{D}$ *é uma matriz diagonal com todos os elementos da diagonal principal positivos.*

***Teorema* A.4.8.** *Para qualquer matriz simétrica definida positiva* $\mathbf{A}$ *de dimensão* $(n \times n)$ *existe uma única matriz triangular superior* $\mathbf{A}^{1/2}$ *com todos os elementos da diagonal principal positivos tal que* $\mathbf{A} = [\mathbf{A}^{1/2}]^\top \mathbf{A}^{1/2}$*. Este resultado é conhecido como* **Decomposição de Cholesky**.

A.5 Derivadas de vetores e matrizes

Neste texto consideramos funções de várias variáveis expressas na forma de

i) escalares do tipo

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} = \sum_{i=1}^{n} a_i x_i;$$

ii) vetores do tipo

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ f_3(\mathbf{x}) \end{pmatrix}$$

em que, por exemplo, $f_1(\mathbf{x}) = x_1 + x_2$, $f_2(\mathbf{x}) = e^{x_1 x_2}$ e $f_3(\mathbf{x}) = x_1 x_2$.

iii) matrizes do tipo

$$\begin{aligned}\mathbf{F}(\mathbf{x}) &= [\ \mathbf{f}_1(\mathbf{x}) \quad \mathbf{f}_2(\mathbf{x}) \quad \ldots \quad \mathbf{f}_n(\mathbf{x})\ ] \\ &= \begin{pmatrix} f_{11}(\mathbf{x}) & f_{12}(\mathbf{x}) & \ldots & f_{1n}(\mathbf{x}) \\ f_{21}(\mathbf{x}) & f_{22}(\mathbf{x}) & \ldots & f_{2n}(\mathbf{x}) \\ \vdots & \vdots & \vdots & \vdots \\ f_{m1}(\mathbf{x}) & f_{m2}(\mathbf{x}) & \ldots & f_{mn}(\mathbf{x}) \end{pmatrix};\end{aligned}$$

por exemplo, se $\mathbf{f}_1(\mathbf{x}) = (x_1 + x_2, x_1x_2, x_1 - x_2)^\top$ e $\mathbf{f}_2(\mathbf{x}) = (x_1, x_1 + x_2, x_1x_2)^\top$, então

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} \mathbf{f}_1(\mathbf{x}) & \mathbf{f}_2(\mathbf{x}) \end{bmatrix} = \begin{pmatrix} f_{11}(\mathbf{x}) & f_{12}(\mathbf{x}) \\ f_{21}(\mathbf{x}) & f_{22}(\mathbf{x}) \\ f_{31}(\mathbf{x}) & f_{32}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} x_1 + x_2 & x_1 \\ x_1x_2 & x_1 + x_2 \\ x_1 - x_2 & x_1x_2 \end{pmatrix}. \tag{A.5.1}$$

Em muitas aplicações, é possível ainda encontrar funções do tipo $\mathbf{F}(\mathbf{X})$ com $\mathbf{X}$ denotando uma matriz de coeficientes com dimensão $(m \times n)$.

No restante desta subseção admitimos a existência de todas as derivadas mencionadas.

Vetor gradiente: *Seja $f(\mathbf{x})$ uma função do vetor $\mathbf{x}$ de dimensão $(p\times 1)$. A derivada de primeira ordem ou vetor gradiente de $f(\mathbf{x})$ é o vetor de dimensão $(p \times 1)$ dado por*

$$\nabla f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right)^\top.$$

Também podemos definir $\partial f(\mathbf{x})/\partial \mathbf{x}^\top = (\partial f(\mathbf{x})/\partial x_1, \ldots, \partial f(\mathbf{x})/\partial x_p) = (\partial f(\mathbf{x})/\partial \mathbf{x})^\top$.

Por exemplo, seja $\mathbf{x} = (x_1, x_2, x_3)^\top$ e $f(\mathbf{x}) = 2x_1^2 + 4x_2^2 + 5x_3^2$. O gradiente de f é dado por

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \frac{\partial f(\mathbf{x})}{\partial x_3} \right)^\top = (4x_1, 8x_2, 10x_3)^\top.$$

Matriz hessiana: *Seja $f(\mathbf{x})$ uma função do vetor $\mathbf{x}$ de dimensão $(p\times 1)$. A matriz de derivadas segundas ou matriz hessiana de $f(\mathbf{x})$ é a matriz quadrada de ordem p dada por*

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \begin{pmatrix} \partial^2 f(\mathbf{x})/\partial x_1^2 & \partial^2 f(\mathbf{x})/\partial x_1 \partial x_2 & \ldots & \partial^2 f(\mathbf{x})/\partial x_1 \partial x_p \\ \partial^2 f(\mathbf{x})/\partial x_2 \partial x_1 & \partial^2 f(\mathbf{x})/\partial x_2^2 & \ldots & \partial^2 f(\mathbf{x})/\partial x_2 \partial x_p \\ \vdots & \vdots & \vdots & \vdots \\ \partial^2 f(\mathbf{x})/\partial x_p \partial x_1 & \partial^2 f(\mathbf{x})/\partial x_p \partial x_2 & \ldots & \partial^2 f(\mathbf{x})/\partial x_p^2 \end{pmatrix}.$$

A matriz hessiana de $f(\mathbf{x})$ também é comumente denotada por $\nabla^2 f(\mathbf{x})$.

Por exemplo, se $\mathbf{x} = (x_1, x_2, x_3)^\top$ e $f(\mathbf{x}) = 2x_1^2 + 4x_2^2 + 5x_3^2$, a matriz hessiana de f é dada por

$$\begin{aligned}\frac{\partial f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} &= \begin{pmatrix} \partial^2 f(\mathbf{x})/\partial x_1^2 & \partial^2 f(\mathbf{x})/\partial x_1 \partial x_2 & \partial^2 f(\mathbf{x})/\partial x_1 \partial x_3 \\ \partial^2 f(\mathbf{x})/\partial x_2 \partial x_1 & \partial^2 f(\mathbf{x})/\partial x_2^2 & \partial^2 f(\mathbf{x})/\partial x_2 \partial x_3 \\ \partial^2 f(\mathbf{x})/\partial x_3 \partial x_1 & \partial^2 f(\mathbf{x})/\partial x_3 \partial x_2 & \partial^2 f(\mathbf{x})/\partial x_3^2 \end{pmatrix} \\ &= \begin{pmatrix} 4 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 10 \end{pmatrix}\end{aligned}$$

As definições acima podem ser estendidas para funções vetoriais ou matriciais.

Matriz jacobiana: *Seja $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})]^\top$ um vetor de dimensão $(m \times 1)$ de funções com argumento vetorial $\mathbf{x} = (x_1, \ldots, x_p)$. A matriz jacobiana de $\mathbf{f}(\mathbf{x})$ é a matriz de dimensão $(m \times p)$ dada por*

$$\nabla \mathbf{f}(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^\top} = \begin{pmatrix} \partial \mathbf{f}_1(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_1(\mathbf{x})/\partial x_2 & \ldots & \partial \mathbf{f}_1(\mathbf{x})/\partial x_p \\ \partial \mathbf{f}_2(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_2(\mathbf{x})/\partial x_2 & \ldots & \partial \mathbf{f}_2(\mathbf{x})/\partial x_p \\ \vdots & \vdots & \vdots & \vdots \\ \partial \mathbf{f}_m(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_m(\mathbf{x})/\partial x_2 & \ldots & \partial \mathbf{f}_m(\mathbf{x})/\partial x_p \end{pmatrix}.$$

Para o exemplo do início da seção, a matriz jacobiana de $\mathbf{f}(\mathbf{x})$ é

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^\top} = \begin{pmatrix} \partial \mathbf{f}_1(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_1(\mathbf{x})/\partial x_2 \\ \partial \mathbf{f}_2(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_2(\mathbf{x})/\partial x_2 \\ \partial \mathbf{f}_3(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_3(\mathbf{x})/\partial x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ x_2 e^{x_1 x_2} & x_1 e^{x_1 x_2} \\ x_2 & x_1 \end{pmatrix};$$

Similarmente, se $\mathbf{F}(\mathbf{x}) = [\mathbf{f}_1(\mathbf{x}) \ \mathbf{f}_2(\mathbf{x}) \ \ldots \ \mathbf{f}_n(\mathbf{x})]$ for uma matriz de dimensão $(m \times n)$ de funções $\mathbf{f}(\mathbf{x})$, sua derivada de primeira ordem é a matriz de dimensão $(m \times np)$ dada por

$$\nabla \mathbf{F}(\mathbf{x}) = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}^\top} = \begin{bmatrix} \partial \mathbf{f}_1(\mathbf{x})/\partial \mathbf{x}^\top & \partial \mathbf{f}_2(\mathbf{x})/\partial \mathbf{x}^\top & \ldots & \partial \mathbf{f}_n(\mathbf{x})/\partial \mathbf{x}^\top \end{bmatrix}.$$

Para o exemplo no início da seção, a derivada da função $\mathbf{F}(\mathbf{x})$ é dada por

$$\begin{aligned}\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}^\top} &= \begin{bmatrix} \partial \mathbf{f}_1(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_1(\mathbf{x})/\partial x_2 & \partial \mathbf{f}_2(\mathbf{x})/\partial x_1 & \partial \mathbf{f}_2(\mathbf{x})/\partial x_2 \end{bmatrix} \\ &= \begin{pmatrix} 1 & 1 & 1 & 0 \\ x_2 & x_1 & 1 & 1 \\ 1 & -1 & x_2 & x_1 \end{pmatrix}.\end{aligned}$$

Seja f uma função real da matriz $\mathbf{X}$ de dimensão $(m \times n)$ definida por

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix};$$

a derivada de f com relação à matriz $\mathbf{X}$ é uma matriz de dimensão $(m \times n)$ dada por

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = (\partial f/\partial x_{ij}) = \begin{pmatrix} \partial f/\partial x_{11} & \partial f/\partial x_{12} & \dots & \partial f/\partial x_{1n} \\ \partial f/\partial x_{21} & \partial f/\partial x_{22} & \dots & \partial f/\partial x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \partial f/\partial x_{m1} & \partial f/\partial x_{m2} & \dots & \partial f/\partial x_{mn} \end{pmatrix}.$$

As regras do produto e da cadeia podem ser empregadas para derivação vetorial e matricial. Se $\mathbf{f}_1(\mathbf{x})$ e $\mathbf{f}_2(\mathbf{x})$ são duas funções vetoriais diferenciáveis de dimensão $(m \times 1)$ com argumento $\mathbf{x}$, então

$$\partial[\mathbf{f}_1(\mathbf{x})^\top \mathbf{f}_2(\mathbf{x})]/\partial \mathbf{x}^\top = \mathbf{f}_1(\mathbf{x})^\top(\partial \mathbf{f}_2(\mathbf{x})/\partial \mathbf{x}^\top) + \mathbf{f}_2(\mathbf{x})^\top(\partial \mathbf{f}_1(\mathbf{x})/\partial \mathbf{x}^\top)$$

Se o vetor $\mathbf{g}(\mathbf{z})$ com dimensão $(p \times 1)$ é função de um vetor de variáveis $\mathbf{z}$ de dimensão $(q \times 1)$ e $\mathbf{f}(\mathbf{x})$ é uma função com argumento $\mathbf{g}(\mathbf{z})$, então

$$\partial \mathbf{f}[\mathbf{g}(\mathbf{z})]/\partial \mathbf{z}^\top = (\partial \mathbf{f}(\mathbf{x})/\partial \mathbf{x}^\top)|_{\mathbf{x}=\mathbf{g}(\mathbf{z})}(\partial \mathbf{g}(\mathbf{z})/\partial \mathbf{z}^\top).$$

Algumas das derivadas vetoriais e matriciais mais usadas em aplicações estatísticas são:

i) $\partial \mathbf{a}^\top \mathbf{x}/\partial \mathbf{x} = \mathbf{a}$;

ii) $\partial \mathbf{x}^\top \mathbf{x}/\partial \mathbf{x} = 2\mathbf{x}$;

iii) $\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}/\partial \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ $(= 2\mathbf{A}\mathbf{x}$, se $\mathbf{A}$ for simétrica);

iv) Para $\mathbf{A}$ e $\mathbf{B}$ simétricas,

$$\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x}/\mathbf{x}^\top \mathbf{B}\mathbf{x})/\partial \mathbf{x} = 2\mathbf{A}\mathbf{x}/\mathbf{x}^\top \mathbf{B}\mathbf{x} - 2[\mathbf{x}^\top \mathbf{A}\mathbf{x}/(\mathbf{x}^\top \mathbf{B}\mathbf{x})^2]\mathbf{B}\mathbf{x};$$

v) Para $\mathbf{A}$ simétrica,

$$\partial[\mathbf{y} - \mathbf{g}(\mathbf{x})]^\top \mathbf{A}[\mathbf{y} - \mathbf{g}(\mathbf{x})]/\partial \mathbf{x} = -2\mathbf{D}(\mathbf{x})^\top \mathbf{A}[\mathbf{y} - \mathbf{g}(\mathbf{x})]$$

em que $\mathbf{D}(\mathbf{x}) = \partial \mathbf{g}(\mathbf{x})/\partial \mathbf{x}^\top$;

vi) Para matrizes $\mathbf{A}$ com dimensão $(m \times n)$, e $\mathbf{B}$ com dimensão $(n \times q)$,

$$\partial \mathrm{tr}(\mathbf{A}(\mathbf{x})\mathbf{B}(\mathbf{x}))/\partial \mathbf{x} = \partial \mathrm{tr}(\mathbf{A}(\mathbf{x})\mathbf{B}(\mathbf{z}))/\partial \mathbf{x}|_{\mathbf{z}=\mathbf{x}} + \partial \mathrm{tr}(\mathbf{A}(\mathbf{z})\mathbf{B}(\mathbf{x}))/\partial \mathbf{x}|_{\mathbf{z}=\mathbf{x}}.$$

vii) $\partial |\mathbf{X}|/\partial x_{ij} = |\mathbf{X}_{ij}|$, em que $|\mathbf{X}_{ij}|$ é o cofator de x_{ij};

viii) $\partial \ln |\mathbf{X}|/\partial x_{ij} = \mathrm{tr}(\mathbf{X}^{-1}(\partial \mathbf{X}/\partial x_{ij}))$;

ix) $\partial \mathbf{X}^{-1}/\partial x_{ij} = -\mathbf{X}^{-1}(\partial \mathbf{X}/\partial x_{ij})\mathbf{X}^{-1}$;

x) $\partial [\mathrm{tr}\mathbf{X}]/\partial x_{ij} = \mathrm{tr}[\partial \mathbf{X}/\partial x_{ij}]$;

xi) $\partial \mathrm{tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})/\partial \mathbf{X} = -(\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1})^\top$;

xii) $\partial |\mathbf{X}|/\partial \mathbf{X} = |\mathbf{X}|(\mathbf{X}^{-1})^\top$;

xiii) $\partial \ln(|\mathbf{X}|)/\partial \mathbf{X} = \{1/|\mathbf{X}|\}\{\partial |\mathbf{X}|/\partial \mathbf{X}\} = (\mathbf{X}^{-1})^\top$;

xiv) $\partial \mathrm{tr}(\mathbf{A}\mathbf{X})/\partial \mathbf{X} = \mathbf{A}^\top$;

xv) $\partial |\mathbf{A}\mathbf{X}|/\partial \mathbf{X} = |\mathbf{A}\mathbf{X}|((\mathbf{A}\mathbf{X})^{-1}\mathbf{A})^\top$.

Se $\mathbf{X}$ é uma matriz com dimensão $(p \times q)$ e $\mathbf{U}(\mathbf{X})$ é uma matriz quadrada de ordem p,

xvi) $\partial \mathrm{tr}(\mathbf{U}(\mathbf{X})^{-1}\mathbf{A})/\partial \mathbf{X} = -(\partial/\partial \mathbf{X})\mathrm{tr}(\mathbf{U}(\mathbf{Z})^{-1}\mathbf{A}\mathbf{U}(\mathbf{Z})^{-1}\mathbf{U}(\mathbf{X}))|_{\mathbf{Z}=\mathbf{X}}$;

xvii) $\partial |\mathbf{U}(\mathbf{X})|/\partial \mathbf{X} = |\mathbf{U}(\mathbf{X})|(\partial/\partial \mathbf{X})\mathrm{tr}(\mathbf{U}(\mathbf{Z})^{-1}\mathbf{U}(\mathbf{X}))|_{\mathbf{Z}=\mathbf{X}}$.

Algumas derivadas envolvendo o operador vec são:

xviii) $\partial \mathrm{vec}(\mathbf{A}\mathbf{X}\mathbf{B})/\partial \mathrm{vec}(\mathbf{X})^\top = \mathbf{B}^\top \otimes \mathbf{A}$;

xix) Sejam $\mathbf{U}(\mathbf{X})$ uma matriz com dimensão $(m \times n)$ e $\mathbf{V}(\mathbf{X})$ uma matriz com dimensão $(n \times r)$, então

$$\frac{\partial \mathrm{vec}[\mathbf{U}(\mathbf{X})\mathbf{V}(\mathbf{X})]}{\partial \mathrm{vec}(\mathbf{X})^\top} = (\mathbf{V} \otimes \mathbf{I}_m)^\top \frac{\partial \mathrm{vec}[\mathbf{U}(\mathbf{X})]}{\partial \mathrm{vec}(\mathbf{X})^\top} + (\mathbf{I}_r \otimes \mathbf{U})^\top \frac{\partial \mathrm{vec}[\mathbf{V}(\mathbf{X})]}{\partial \mathrm{vec}(\mathbf{X})^\top};$$

xx) Seja $\mathbf{X}$ uma matriz quadrada; então

$$\frac{\partial \mathrm{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})}{\partial \mathrm{vec}(\mathbf{X})^\top} = -(\mathbf{X}\mathbf{B})^\top \otimes (\mathbf{A}\mathbf{X}^{-1});$$

xxi) Para uma matriz não singular $\mathbf{U}(\mathbf{X})$,

$$\frac{\partial \text{vec}[\mathbf{U}(\mathbf{X})^{-1}]}{\partial \text{vec}(\mathbf{X})^\top} = -[(\mathbf{U}(\mathbf{X})^{-1})^\top \otimes \mathbf{U}(\mathbf{X})^{-1}] \frac{\partial \text{vec}[\mathbf{U}(\mathbf{X})]}{\partial \text{vec}(\mathbf{X})^\top};$$

xxii) $$\frac{\partial \text{vec}\{\mathbf{F}[\mathbf{G}(\mathbf{X})]\}}{\partial \text{vec}(\mathbf{Z})^\top} = \frac{\partial \text{vec}[\mathbf{F}(\mathbf{X})]}{\partial \text{vec}(\mathbf{X})^\top}\bigg|_{\mathbf{X}=\mathbf{G}(\mathbf{Z})} \frac{\partial \text{vec}[\mathbf{G}(\mathbf{Z})]}{\partial \text{vec}(\mathbf{Z})^\top}.$$

Para ilustrar a aplicação de derivadas de vetores e matrizes, inicialmente, consideremos a função

$$\begin{aligned} g(\boldsymbol{\beta}, \boldsymbol{\theta}) &= [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] + \ln |\mathbf{V}(\boldsymbol{\theta})| \\ &= \text{tr}\{[\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})][\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top\} + \ln |\mathbf{V}(\boldsymbol{\theta})| \end{aligned}$$

com argumentos $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$, com $\boldsymbol{\theta}$ funcionalmente independente de $\boldsymbol{\beta}$, em que $\mathbf{V}(\boldsymbol{\theta})$ é uma matriz simétrica definida positiva. Além disso, considere que $\mathbf{f}(\boldsymbol{\beta})$ e $\mathbf{V}(\boldsymbol{\theta})$ sejam funções diferenciáveis. O gradiente de $g(\boldsymbol{\beta}, \boldsymbol{\theta})$ é dado por:

$$\nabla g(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{pmatrix} (\partial/\partial \boldsymbol{\beta}) g(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ (\partial/\partial \boldsymbol{\theta}) g(\boldsymbol{\beta}, \boldsymbol{\theta}) \end{pmatrix}$$

em que

i) a derivada em relação a β é

$$\begin{aligned} \frac{\partial g(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{\partial [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} \\ &= -2 \left(\frac{\partial [\mathbf{f}(\boldsymbol{\beta})]^\top}{\partial \boldsymbol{\beta}} \right) [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]. \end{aligned}$$

ii) a derivada em relação a θ_j para $j = 1, \ldots, k$ é

$$\begin{aligned}
\frac{\partial g(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_j} &= \frac{\partial [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]}{\partial \theta_j} + \frac{\partial \ln |\mathbf{V}(\boldsymbol{\theta})|}{\partial \theta_j} \\
&= [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top \frac{\partial [\mathbf{V}(\boldsymbol{\theta})]^{-1}}{\partial \theta_j} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] + \frac{\partial \ln |\mathbf{V}(\boldsymbol{\theta})|}{\partial \theta_j} \\
&= -[\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top [\mathbf{V}(\boldsymbol{\theta})]^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \theta_j} [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] \\
&\quad + \mathrm{tr} \left\{ [\mathbf{V}(\boldsymbol{\theta})]^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \theta_j} \right\} \\
&= -\mathrm{tr} \left\{ [\mathbf{V}(\boldsymbol{\theta})]^{-1} \frac{\partial [\mathbf{V}(\boldsymbol{\theta})]}{\partial \theta_j} [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] [(\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top \right\} \\
&\quad + \mathrm{tr} \left\{ [\mathbf{V}(\boldsymbol{\theta})]^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \theta_j} \right\} \\
= \quad & \mathrm{tr} \left\{ [\mathbf{V}(\boldsymbol{\theta})]^{-1} \left(-[\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})][\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top + \mathbf{V}(\boldsymbol{\theta}) \right) [\mathbf{V}(\boldsymbol{\theta})]^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \theta_j} \right\}.
\end{aligned}$$

A matriz de segundas derivadas de $g(\boldsymbol{\beta}, \boldsymbol{\theta})$ em relação a $\boldsymbol{\beta}$ é:

$$\begin{aligned}
\frac{\partial^2 g(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \frac{\partial \left\{ -2 \left(\partial [\mathbf{f}(\boldsymbol{\beta})]^\top / \partial \boldsymbol{\beta} \right) [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] \right)}{\partial \boldsymbol{\beta}^\top} \\
&= -2 \frac{\partial^2 [\mathbf{f}(\boldsymbol{\beta})]^\top}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})] \\
&\quad + 2 \left(\frac{\partial [\mathbf{f}(\boldsymbol{\beta})]^\top}{\partial \boldsymbol{\beta}} \right) [\mathbf{V}(\boldsymbol{\theta})]^{-1} \left(\frac{\partial \mathbf{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \right).
\end{aligned}$$

Para obter $\partial^2 g(\boldsymbol{\beta}, \boldsymbol{\theta}) / \partial \beta_i \partial \theta_j$, notemos que para $i = 1, \ldots, p$

$$\begin{aligned}
\frac{\partial g(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_i} &= \frac{\partial [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]^\top [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})]}{\partial \beta_i} \\
&= -2 \left(\frac{\partial [\mathbf{f}(\boldsymbol{\beta})]^\top}{\partial \beta_i} \right) [\mathbf{V}(\boldsymbol{\theta})]^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\beta})].
\end{aligned}$$

e que $\partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\beta_i\partial\theta_j = \partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\theta_j\partial\beta_i$ para $i = 1,\ldots,p$, $j = 1,\ldots,k$; então

$$\begin{aligned}
\frac{\partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta_i\partial\theta_j} &= \frac{\partial\left\{-2\left(\partial[\mathbf{f}(\boldsymbol{\beta})^\top/\partial\beta_i\right)[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]\right\}}{\partial\theta_j} \\
&= -2\left(\frac{\partial[\mathbf{f}(\boldsymbol{\beta})]^\top}{\partial\beta_i}\right)\frac{\partial[\mathbf{V}(\boldsymbol{\theta})]^{-1}}{\partial\theta_j}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})] \\
&= -2\left(\frac{\partial[\mathbf{f}(\boldsymbol{\beta})]^\top}{\partial\beta_i}\right)[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})].
\end{aligned}$$

Fazendo uso da propriedade (x), as derivadas $\partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\theta_s\partial\theta_j$ para $j,s = 1,\ldots,k$ são

$$\begin{aligned}
\frac{\partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j} &= \frac{\partial}{\partial\theta_s}\left\{\mathrm{tr}\left([\mathbf{V}(\boldsymbol{\theta})]^{-1}\left\{-[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})][\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top+\mathbf{V}(\boldsymbol{\theta})\right\}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\right)\right\} \\
&= \frac{\partial\left\{-[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top\,[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_j][\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]\right\}}{\partial\theta_s} \\
&\quad+\frac{\partial\left\{\mathrm{tr}\left([\mathbf{V}(\boldsymbol{\theta})]^{-1}[\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_j]\right)\right\}}{\partial\theta_s} \\
&= -[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top\frac{\partial\left([\mathbf{V}(\boldsymbol{\theta})]^{-1}[\partial[\mathbf{V}(\boldsymbol{\theta})]/\partial\theta_j][\mathbf{V}(\boldsymbol{\theta})]^{-1}\right)}{\partial\theta_s}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})] \\
&\quad+\mathrm{tr}\left\{\frac{\partial\left([\mathbf{V}(\boldsymbol{\theta})]^{-1}[\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_j]\right)}{\partial\theta_s}\right\}.
\end{aligned}$$

Utilizando a regra do produto para derivadas de matrizes, temos que

$$\begin{aligned}
\frac{\partial\left\{[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_j]\right\}}{\partial\theta_s} &= \frac{\partial[\mathbf{V}(\boldsymbol{\theta})]^{-1}}{\partial\theta_s}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}+[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial}{\partial\theta_s}\left[\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\right] \\
&= -[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}+[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial^2\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j}
\end{aligned}$$

e

$$\begin{aligned}\frac{\partial}{\partial\theta_s}\left\{[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\right\} &= \frac{\partial\left\{[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_j]\right\}}{\partial\theta_s} \\ &\quad + [\mathbf{V}(\boldsymbol{\theta})]^{-1} + \mathbf{V}(\boldsymbol{\theta})\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial[\mathbf{V}(\boldsymbol{\theta})]V}{\partial\theta_s} \\ &= -[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1} \\ &\quad + [\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial^2\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1} \\ &\quad - [\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}.\end{aligned}$$

Então

$$\begin{aligned}\frac{\partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j} &= [\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})] \\ &- [\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial^2\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})] \\ &+ [\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})] \\ &+ \operatorname{tr}\left\{-[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j} + [\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial^2\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j}\right\}.\end{aligned}$$

Além disso, como $[\mathbf{V}(\boldsymbol{\theta})]^{-1}$, $\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_j$ e $\partial\mathbf{V}(\boldsymbol{\theta})/\partial\theta_s$ são matrizes simétricas e

$$[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]$$

é um escalar, obtemos a igualdade

$$\begin{aligned}&[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})] = \\ &[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})].\end{aligned}$$

Portanto, utilizando as propriedades do traço de matrizes e os resultados acima, as derivadas de $g(\boldsymbol{\beta},\boldsymbol{\theta})$ em relação a θ_j e θ_s para $j,s=1,\ldots,k$ são dadas por

$$\begin{aligned}\frac{\partial^2 g(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j} &= \operatorname{tr}\left\{\mathbf{A}\left(2[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})][\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top - \mathbf{V}(\boldsymbol{\theta})\right)\right\} \\ &+ \operatorname{tr}\left\{\mathbf{B}\left(-[\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})][\mathbf{y}-\mathbf{f}(\boldsymbol{\beta})]^\top + \mathbf{V}(\boldsymbol{\theta})\right)\right\},\end{aligned}$$

em que

$$\begin{aligned}\mathbf{A} &= [\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\theta_s}[\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1},\\ \mathbf{B} &= [\mathbf{V}(\boldsymbol{\theta})]^{-1}\frac{\partial^2 \mathbf{V}(\boldsymbol{\theta})}{\partial\theta_s\partial\theta_j}[\mathbf{V}(\boldsymbol{\theta})]^{-1}.\end{aligned}$$

A.6 Exercícios

A.6.1. Sejam $\mathbf{x}$ e $\mathbf{y}$ vetores com dimensão $(n\times 1)$. Prove a desigualdade de Cauchy-Schwarz:

$$(\mathbf{x}^\top\mathbf{y})^2 \leq (\mathbf{x}^\top\mathbf{x})(\mathbf{y}^\top\mathbf{y}).$$

Em que condições a igualdade é válida?

Sugestão: Use a desigualdade $\|\mathbf{x}+\lambda\mathbf{y}\|^2 \geq 0,\ \forall \mathbf{x},\mathbf{y}$ e $\lambda \in I\!R$.

A.6.2. Seja $\mathbf{x} = (x_1,\ldots,x_n)^\top$ um vetor com dimensão $(n\times 1)$ e defina $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top\mathbf{x}} = (\sum_{i=1}^n x_i^2)^{1/2}$. Mostre que

i) $\mathbf{x}^\top\mathbf{x} \geq 0,\ \forall \mathbf{x}$;

ii) $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = (0,\ldots,0)^\top = \mathbf{0}$;

iii) $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|,\ \forall c \in I\!R$;

iv) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|,\ \forall \mathbf{x},\mathbf{y} \in I\!R^n$.

Sugestão: Use os resultados do Exercício A.6.1.

A.6.3. Sejam $\mathbf{A}$ e $\mathbf{B}$ duas matrizes quadradas de ordem n, $\mathbf{C}$ e $\mathbf{D}$ duas matrizes com dimensões $(m\times n)$ e $(n\times m)$, respectivamente e $\mathbf{x} = (x_1,\ldots,x_n)^\top$, um vetor com dimensão $(n\times 1)$. Prove que

i) $\mathrm{tr}(\mathbf{A}^\top) = \mathrm{tr}(\mathbf{A})$;

ii) $\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$;

iii) $\mathrm{tr}(\mathbf{CD}) = \mathrm{tr}(\mathbf{DC})$;

iv) $\mathrm{tr}(\mathbf{A}\mathbf{A}^\top) = \mathrm{tr}(\mathbf{A}^\top\mathbf{A}) = \sum_{i=1}^n\sum_{j=1}^n a_{ij}^2$;

v) $\|\mathbf{x}\|^2 = \sum_{i=1}^n x_i^2 = \mathbf{x}^\top\mathbf{x} = \mathrm{tr}(\mathbf{x}\mathbf{x}^\top)$.

vi) Construa exemplos numéricos para ilustrar as propriedades acima.

A.6.4. Seja $\mathbf{A}$ uma matriz quadrada *idempotente*, *i.e.*, tal que $\mathbf{A}^2 = \mathbf{AA} = \mathbf{A}$. Além disso, seja $\mathbf{0}_n$ a matriz nula de dimensão $(n \times n)$. Prove que

i) $|\mathbf{A}| = 0$ ou $|\mathbf{A}| = 1$;

ii) Os autovalores de $\mathbf{A}$ são iguais a 0 ou 1;

iii) $\mathbf{I}_n - \mathbf{A}$ é idempotente e $\mathbf{A}(\mathbf{I}_n - \mathbf{A}) = (\mathbf{I}_n - \mathbf{A})\mathbf{A} = \mathbf{0}_n$;

iv) Suponha adicionalmente que $\mathbf{A}$ é simétrica e prove que

   a) $\mathrm{r}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$;
   b) $\mathrm{r}(\mathbf{A}) = n \Rightarrow \mathbf{A} = \mathbf{I}_n$.

   Sugestão: Utilize a decomposição espectral de $\mathbf{A}$.

   Observação: Os resultados são válidos mesmo quando a matriz $\mathbf{A}$ não é simétrica.

v) Construa exemplos numéricos (com matrizes diferentes da matriz identidade) para ilustrar as propriedades acima.

A.6.5. Mostre que se $\mathbf{X}$ é uma matriz tal que $\mathbf{X}^\top\mathbf{X} = \mathbf{X}$ então ela é simétrica e idempotente. Construa um exemplo numérico (com $\mathbf{X}$ diferente da matriz identidade) para ilustrar a propriedade acima.

A.6.6. Seja $\mathbf{A}$ uma matriz quadrada cujos autovalores são representados por $\lambda_1, \ldots, \lambda_n$. Prove que

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i.$$

Construa um exemplo numérico (com $\mathbf{X}$ diferente da matriz identidade) para ilustrar a propriedade acima.

A.6.7. Uma matriz $\mathbf{A}$ quadrada de ordem n é positiva definida se $\mathbf{x}^\top\mathbf{Ax} > 0 \;\; \forall \mathbf{x} \neq \mathbf{0}$. Seja $\mathbf{X}$ uma matriz $(n \times p)$, $n > p$, de posto completo. Mostre que

i) $\mathbf{X}^\top\mathbf{X}$ é simétrica;

ii) $\mathbf{X}^\top\mathbf{X}$ é p.d.;

iii) como uma matriz é positiva definida se e somente se todos os seus autovalores são positivos, prove que $\mathbf{X}^\top\mathbf{X}$ é inversível;

iv) $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ é uma matriz simétrica e idempotente;

v) $\mathbf{I}_n - \mathbf{H}$ é uma matriz simétrica e idempotente;

vi) $\mathrm{r}(\mathbf{I}_n - \mathbf{H}) = n - p$.

vii) Construa exemplos numéricos (com matrizes diferentes da matriz identidade) para ilustrar as propriedades acima.

A.6.8. Seja a matriz

$$\mathbf{A} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Para que valores de ρ a matriz $\mathbf{A}$ é positiva definida?

A.6.9. Seja $f(\mathbf{X})$ uma função real de uma matriz $\mathbf{X}$ com dimensão $(m \times n)$ e elementos x_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$. A derivada de f com respeito a $\mathbf{X}$ é definida como sendo a matriz $(m \times n)$ de derivadas $\partial f(\mathbf{X})/\partial x_{ij}$:

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} := \begin{pmatrix} \dfrac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \dfrac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \dfrac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{pmatrix}.$$

Sejam $\mathbf{x} = (x_1, \ldots, x_n)^\top$ e $\mathbf{a} = (a_1, \ldots, a_n)^\top$, vetores com dimensão $(n \times 1)$, $\mathbf{A}$, uma matriz quadrada de ordem n e $\mathbf{B}$, uma matriz com dimensão $m \times n$. Mostre que

i) $\partial \mathbf{a}^\top\mathbf{x}/\partial \mathbf{x} = \partial \mathbf{x}^\top\mathbf{a}/\partial \mathbf{x} = \mathbf{a}$;

ii) $\partial \mathbf{B}\mathbf{x}/\partial \mathbf{x} = \mathbf{B}^\top$;

iii) $\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}/\partial \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$;

iiv) $\partial \mathbf{x}^\top\mathbf{A}\mathbf{x}/\partial \mathbf{x} = 2\mathbf{A}\mathbf{x}$ se $\mathbf{A}$ for simétrica.

A.6.10. Seja $\mathbf{x} = (x_1, \ldots, x_n)^\top$ um vetor de n observações de uma certa variável X. Mostre que a média e a variância amostral das observações podem ser escritas na seguinte forma matricial:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{1}_n^\top\mathbf{x} \\ s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}(\mathbf{x} - \mathbf{1}_n\overline{x})^\top(\mathbf{x} - \mathbf{1}_n\overline{x}) \\ &= \frac{1}{n-1}\mathbf{x}^\top(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{x}, \end{aligned}$$

em que $\mathbf{1}_n$ representa um vetor de dimensão $(n \times 1)$ com todos elementos iguais a 1 e $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n^\top$.

A.6.11. Mostre que a matriz $\mathbf{I}_n - n^{-1}\mathbf{J}_n$ é simétrica, idempotente e não negativa definida (n.n.d.), *i.e.*, $\mathbf{x}^\top(\mathbf{I}_n - \mathbf{J}_n)\mathbf{x} \geq 0$, $\forall \mathbf{x} \neq \mathbf{0}$

A.6.12. Considere as seguintes funções das variáveis aleatórias Y_1, Y_2, Y_3 e Y_4:

$$\begin{aligned} W_1 &= Y_1 - Y_2 \\ W_2 &= Y_1 + Y_3 \\ W_3 &= Y_1 - Y_4 \end{aligned}$$

i) Expresse as relações acima em notação matricial.

ii) Obtenha a esperança do vetor $\mathbf{W} = (W_1, W_2, W_3)^\top$ em termos das esperanças de Y_1, Y_2, Y_3 e Y_4.

iii) Obtenha a matriz de covariâncias do vetor $\mathbf{W}$ em termos das variâncias e covariâncias de Y_1, Y_2, Y_3 e Y_4.

Apêndice B

O método Delta

Em muitas situações, o Teorema Limite Central pode ser empregado para obtenção de distribuições assintóticas de funções de variáveis aleatórias. No entanto, há casos, como aquele que é objeto da Seção C.7, em que as condições de regularidade não permitem sua aplicação. Uma alternativa conveniente é o chamado método Delta. Embora casos multivariados sejam os mais comuns, apresentaremos inicialmente o caso univariado por razões didáticas. Detalhes podem ser obtidos em Sen, Singer & Pedroso-de Lima (2009).

B.1 O caso univariado

Consideremos uma função contínua $f(x) : \mathbb{R} \longrightarrow \mathbb{R}$, com derivadas contínuas até a ordem $k > 0$ em torno de um ponto $x_0 \in \mathbb{R}$. Sua expansão de Taylor de ordem k em torno do ponto x_0 é

$$f(x) = f(x_0) + \sum_{j=1}^{k} \frac{(x - x_0)^j}{j!} f^{(j)}(x_0) + R_k(x, x_0),$$

em que $f^{(j)}(x)$ denota a derivada de ordem k de f calculada no ponto x e

$$R_k(x, x_0) = \frac{(x - x_0)^k}{k!}\{f^{(k)}[hx_0 + (1 - h)x] - f^{(k)}(x_0)\}$$

para algum $0 < h < 1$ é o resto. O caso $k = +\infty$ é conhecido como série de Taylor.

Tomando $k = 1$, quando $x \to x_0$, podemos desprezar o resto e escrever

$$f(x) \approx f(x_0) + f^{(1)}(x - x_0)$$

em que $f'(x) = f^{(1)}(x)$.

***Teorema* B.1.1.** *Sejam* $\sqrt{n}(T_n - \theta)/\sigma \xrightarrow{\mathcal{D}} N(0,1)$ *e* g *uma função contínua tal que* $g'(\theta)$ *existe e* $g'(\theta) \neq 0$. *Então*

$$\sqrt{n}[g(T_n) - g(\theta)]/\sigma g'(\theta) \xrightarrow{\mathcal{D}} N(0,1).$$

Notemos que por (B.1),

$$\mathbb{E}[g(T_n)] \approx \mathbb{E}[g(\theta)] + \mathbb{E}[g'(\theta)(T_n - \theta)] = g(\theta)$$

e que

$$\mathbb{V}[g(T_n)] \approx \mathbb{V}[g(\theta)] + \mathbb{V}[g'(\theta)(T_n - \theta)] = [g'(\theta)]^2 \mathbb{V}(T_n - \theta) = [g'(\theta)]^2 \sigma^2. \quad \text{(B.1.1)}$$

B.2 O caso multivariado

Consideremos agora uma função contínua $f(\mathbf{x}) : I\!R^p \longrightarrow I\!R$ com derivadas contínuas até a ordem $k \geq 1$. Então para cada $\mathbf{x} \in I\!R$, $\mathbf{x}_0 \in I\!R$, temos a seguinte expansão de Taylor multivariada:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \sum_{j=1}^{k} \frac{1}{j!} \sum_{i_1=1}^{p} \cdots \sum_{i_j=1}^{p} \frac{\partial^j}{\partial x_{i_1} \cdots \partial x_{i_j}} f(\mathbf{x}) \Big|_{\mathbf{x}_0} \prod_{l=1}^{j} (x_{i_l} - x_{0_{i_l}}) + R_k(\mathbf{x}, \mathbf{x}_0),$$

em que $R_k(\mathbf{x}, \mathbf{x}_0)$ é

$$\frac{1}{(k+1)!} \sum_{i_1=1}^{p} \cdots \sum_{i_{k+1}=1}^{p} \frac{\partial^{k+1}}{\partial x_{i_1} \cdots \partial x_{i_{k+1}}} f[h\mathbf{x}_0 + (1-h)\mathbf{x}] \prod_{l=1}^{k+1} (x_{i_l} - x_{0_{i_l}})$$

para algum $0 < h < 1$. No caso $k = 1$, a expansão de Taylor de primeira ordem pode ser escrita em notação matricial como

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^\top \mathbf{u}(\mathbf{x}_0) + o(||\mathbf{x} - \mathbf{x}_0||).$$

em que $\mathbf{u}(\mathbf{x}_0) = \partial f(\mathbf{x})/\partial \mathbf{x}|_{\mathbf{x}=\mathbf{x}_0}$ e $o(||\mathbf{x} - \mathbf{x}_0||)$ denota um termo que converge para 0 quando $\mathbf{x} \longrightarrow \mathbf{x}_0$. Desprezando esse termo, na vizinhança de $\mathbf{x}_0$, a função $f(\mathbf{x})$ pode ser aproximada por

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^\top \mathbf{u}(\mathbf{x}_0). \quad \text{(B.2.1)}$$

Esta é a base para o seguinte teorema.

***Teorema* B.2.1.** *Seja* $\{\mathbf{T}_n\}$ *uma sequência de variáveis aleatórias com dimensão* p *tal que* $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma})$ *e consideremos uma função* $g(\mathbf{T}_n) : I\!R^p \longrightarrow I\!R$ *tal que* $\mathbf{u}(\boldsymbol{\theta}) = \partial g(\mathbf{x})/\partial \mathbf{x}$ *é não nula e contínua numa vizinhança de* $\boldsymbol{\theta}$*. Então*

$$\sqrt{n}[g(\mathbf{T}_n) - g(\boldsymbol{\theta})] \xrightarrow{\mathcal{D}} N(0, \gamma^2) \quad \textit{com} \quad \gamma^2 = \mathbf{u}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma} \mathbf{u}(\boldsymbol{\theta}).$$

Notemos que por (B.2.1),

$$\mathbb{E}[\mathbf{g}(\mathbf{T}_n)] \approx \mathbb{E}[\mathbf{g}(\boldsymbol{\theta})] + \mathbb{E}[(\mathbf{T}_n - \boldsymbol{\theta})^\top \mathbf{u}(\boldsymbol{\theta})] = \mathbf{g}(\boldsymbol{\theta})$$

e que

$$\mathbb{V}[\mathbf{g}(\mathbf{T}_n)] \approx \mathbb{V}[\mathbf{g}(\boldsymbol{\theta})] + \mathbb{V}[(\mathbf{T}_n - \boldsymbol{\theta})^\top \mathbf{u}(\boldsymbol{\theta})] = \mathbf{u}(\boldsymbol{\theta})^\top [\mathbb{V}(\mathbf{T}_n - \boldsymbol{\theta})] \mathbf{u}(\boldsymbol{\theta}) = \mathbf{u}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma} \mathbf{u}(\boldsymbol{\theta}). \tag{B.2.2}$$

Apêndice C

Análise de Regressão

C.1 Introdução

O objetivo da análise de regressão é estudar a associação entre uma variável, denominada resposta (ou dependente ou explicada ou ainda endógena, como é geralmente conhecida em textos de econometria), e uma ou mais variáveis explicativas (independentes, preditoras, de controle ou exógenas). Especificamente, esse tipo de análise visa estimar o valor médio (ou predizer o valor) da variável resposta condicionalmente aos valores das variáveis explicativas.

Sejam y_i e $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ respectivamente os valores de uma variável resposta e de p variáveis explicativas observadas na i-ésima unidade de um conjunto de n unidades amostrais. Esses valores podem ser dispostos na forma de um vetor $\mathbf{y} = (y_1, \ldots, y_n)^\top$ e de uma matriz $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$.[1]

Suponhamos que a resposta média da i-ésima unidade amostral condicionalmente às variáveis explicativas $\mathbf{x}_i$ pode ser expressa na forma $\mathbb{E}(y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta})$ em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denota um vetor de parâmetros desconhecidos e fixos, denominados **coeficientes de regressão**. A função $\mu(\mathbf{X}, \boldsymbol{\beta})$ é denominada função de regressão ou simplesmente, regressão. A forma de $\mu(\mathbf{X}, \boldsymbol{\beta})$ pode ser proveniente de alguma teoria (como no caso do espaço y_i percorrido por um corpo em queda livre após um tempo x_i, em que essa função seria quadrática) ou sugerida por meio de uma análise descritiva. A função de regressão pode ser:

i) linear nos parâmetros e nas variáveis explicativas, *e.g.*,

$$\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip};$$

[1] Embora seja possível desenvolver toda a metodologia em termos de matrizes $\mathbf{X}$ com posto incompleto, isso não é necessário para efeitos práticos, e neste texto assumiremos que $\mathbf{X}$ tem posto completo, ou seja, que $r(\mathbf{X}) = p$.

ii) linear somente nos parâmetros, *e.g.*,

$$\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 \log(x_{i2}) + \ldots + \beta_p \sqrt{x_{ip}};$$

iii) linearizável, *e.g.*,

$$\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})$$

ou

$$\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 x_{i1}^{\beta_1} \ldots x_{ip}^{\beta_p};$$

iv) não linear, *e.g.*,

$$\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 + \exp(\beta_1 x_{i1}^2) + \beta_2 \log(x_{i2}) + \ldots + \beta_p \sqrt{x_{ip}}.$$

Nós trataremos somente do caso de funções de regressão lineares (ou linearizáveis) nos parâmetros. Detalhes sobre os modelos de regressão não lineares podem ser encontrados em Seber & Wild (1989) ou Souza (1998), por exemplo.

Condicionalmente aos valores das variáveis explicativas, $\mathbf{x}_i$, a resposta da i-ésima unidade amostral se situa em torno de $\mu(\mathbf{x}_i, \boldsymbol{\beta})$; expressando o desvio de y_i em relação a $\mu(\mathbf{x}_i, \boldsymbol{\beta})$ como $e_i = y_i - \mu(\mathbf{x}_i, \boldsymbol{\beta})$, o modelo pode ser expresso na forma

$$y_i = \mu(\mathbf{x}_i, \boldsymbol{\beta}) + e_i,$$

$i = 1, \ldots, n$, ou na forma matricial, como

$$\mathbf{y} = \boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}) + \mathbf{e}$$

com $\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}) = [\mu_1(\mathbf{X}, \boldsymbol{\beta}), \ldots, \mu_n(\mathbf{X}, \boldsymbol{\beta})]^\top$ e $\mathbf{e} = [e_1, \ldots, e_n]^\top$. Quando $\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, temos

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{C.1.1}$$

que é o **modelo de regressão linear múltipla**. Explicitamente esse modelo corresponde a

$$y_i = \mu(\mathbf{x}_i, \boldsymbol{\beta}) + e_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i,$$

$i = 1, \ldots, n$. Quando $p = 1$, ele é denominado **modelo de regressão linear simples**.

Modelos da forma

$$y_i = \mu(\mathbf{x}_i, \boldsymbol{\beta}) + e_i = \beta_0 + \beta_1 x_i + \beta_2 x_1^2 + \ldots + \beta_p x_i^p + e_i,$$

$i = 1, \ldots, n$, $p \geq 2$ são conhecidos como **modelos de regressão polinomial**.

Usualmente assumimos que

$$\mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \mathbb{V}(\mathbf{e}) = \sigma^2 \mathbf{I}_n \tag{C.1.2}$$

ou seja, que os erros têm média nula, são homocedásticos e não correlacionados. Nesse contexto, convém lembrar que as variáveis explicativas são consideradas fixas e observadas sem erro. Quando isso não acontece, a análise de regressão ainda pode ser realizada por meio de modelos condicionais. Por exemplo, num estudo em que o objetivo é avaliar a relação entre peso (Y) e altura (X) de adolescentes, podemos obter os valores (possivelmente com erro) de ambas as variáveis numa amostra de n indivíduos, nomeadamente $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ e analisar os dados por meio do modelo condicional $y_i|x_i = \alpha + \beta x_i + e_i,\ i = 1, \ldots, n$ em que desconsideramos possíveis erros de medida da altura X. A ideia aqui é avaliar a distribuição do peso Y para dados valores da altura X. Alternativamente, é possível considerar os chamados **modelos com erros nas variáveis** ou **modelos com erros de medida** que também acomodam variáveis explicativas observadas com erro. A análise sob esse tipo de modelo foge aos objetivos deste texto e o leitor interessado pode consultar Fuller (1987) ou Chesher (1991) para detalhes.

Exemplo C.1.1: Os dados da Tabela C.1.1 correspondem à idade e a uma medida de pressão arterial sistólica para um conjunto de 20 pacientes de uma determinada clínica.

Tabela C.1.1: Idade (anos) e pressão arterial sistólica (mmHg)

Paciente	Pressão sistólica	Idade	Paciente	Pressão sistólica	Idade
1	114	17	11	156	47
2	134	18	12	159	47
3	116	20	13	142	50
4	139	23	14	156	52
5	110	34	15	164	57
6	150	38	16	185	60
7	152	41	17	162	64
8	138	42	18	176	66
9	142	46	19	175	69
10	145	47	20	180	70

Para estudar a associação entre idade (X) e pressão arterial sistólica (Y), consideramos o modelo de regressão linear simples

$$y_i = \alpha + \beta x_i + e_i,$$

$i = 1, \ldots, 20$ que pode ser expresso na forma matricial (C.1.1) com

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{20} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{20} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{20} \end{bmatrix}.$$

O parâmetro α corresponde ao valor esperado da pressão arterial sistólica de recém-nascidos e o β representa a variação na pressão arterial sistólica esperada por ano para pacientes com as mesmas características daqueles investigados. Como não há dados para pacientes com idades menores do que 17 anos, a extrapolação do modelo para aquém dessa idade não pode ser baseada em argumentos estatísticos (o modelo poderia, e deve ser, não linear quando consideradas todas as idades). Consequentemente, para efeito de interpretação, convém reescrever o modelo na forma

$$y_i = \alpha^* + \beta(x_i - x_0) + e_i,$$

$i = 1, \ldots, 20$ em que x_0 é uma constante conhecida, por exemplo $x_0 = \overline{x}$ com $\overline{x}$ denotando a idade média dos pacientes. Nesse caso, o parâmetro α^* corresponde ao valor esperado da pressão arterial sistólica para pacientes cuja idade é $\overline{x}$. Se escolhermos $x_0 = 17$, o parâmetro seria interpretado como valor esperado da pressão arterial sistólica para pacientes cuja idade é 17 anos. Nesses casos, a matriz de especificação do modelo seria dada por

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 - x_0 \\ 1 & x_2 - x_0 \\ \vdots & \vdots \\ 1 & x_{20} - x_0 \end{bmatrix}.$$

Tendo em conta que esse modelo é equivalente a

$$y_i = \alpha + \beta x_i + e_i,$$

$i = 1, \ldots, 20$ com $\alpha = \alpha^* - \beta x_0$, fica evidente que a alteração corresponde apenas a uma translação do eixo das ordenadas e que nem o valor do parâmetro β nem sua interpretação são afetados.

Exemplo C.1.2: A Tabela C.1.2 contém dados de capacidade instalada (ton), potência instalada (1000 kW) e área construída (100 m^2) de 10 empresas de uma certa indústria.

Tabela C.1.2: Capacidade instalada (ton), potência instalada (1000 kW) e área construída (100 m^2) de empresas de uma certa indústria

Produção	4.5	5.9	4.2	5.2	8.1	9.7	10.7	11.9	12.0	12.3
Potência	0.9	2.5	1.3	1.8	3.1	4.6	6.1	6.0	5.9	6.1
Área	7.1	10.4	7.2	8.2	8.5	11.9	12.1	12.5	12.0	11.3

Para representar a distribuição da capacidade instalada (Y) a partir das informações sobre potência instalada (X_1) e área construída (X_2), consideramos o seguinte modelo:

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

$i = 1, \ldots, 10$, que genericamente, *i.e.,* para n empresas, pode ser expresso na forma matricial (C.1.1) com

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \\ 1 & x_{1n} & x_{2n} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

Aqui, o parâmetro β_1 representa a variação na capacidade instalada esperada por unidade de potência instalada para empresas com a mesma área construída. O parâmetro β_2 tem interpretação semelhante com a substituição de potência instalada por área construída e vice-versa. Como não temos dados para empresas com potência instalada menor que $0.9 \times 1000kW$ e área construída menor que $7.1 \times 100m^2$, o parâmetro β_0 deve ser interpretado como um fator de ajuste do plano que aproxima a verdadeira função que relaciona o valor esperado da variável resposta com as variáveis explicativas na região em que há dados disponíveis. Se essa função for linear em todo o campo de variação das variáveis explicativas, um modelo linear mais adequado não deveria conter o termo β_0, pois potência instalada ou área construída iguais a zero implicariam capacidade de produção nula.

Exemplo C.1.3: A Tabela C.1.3 contém dados de um estudo cujo objetivo é avaliar o efeito de diferentes níveis de um determinado fertilizante no número de frutos de boa qualidade por macieira de uma certa variedade.

Tabela C.1.3: Concentração do fertilizante e número de frutos de boa qualidade

Concentração do fertilizante	Número de frutos de boa qualidade	Concentração do fertilizante	Número de frutos de boa qualidade
10	16	40	24
10	18	40	25
10	19	40	28
10	20	50	26
20	18	50	28
20	20	50	30
20	21	50	32
20	22	50	35
20	23	50	25
30	22		
30	25		
30	26		

Se encararmos a variável Concentração de fertilizante como um fator (no espírito dos modelos usuais de ANOVA), o estudo pode ser modelado (de forma genérica) por intermédio de

$$y_{ij} = \mu_i + e_{ij}, \tag{C.1.3}$$

$i = 1, \ldots, I,\ j = 1, \ldots, n_i$, em que y_{ij} representa o número de frutos de boa qualidade na j-ésima macieira tratada com a i-ésima concentração do fertilizante e μ_i o valor esperado correspondente.[2] Esse modelo pode ser escrito na forma matricial (C.1.1) com

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{In_I} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{bmatrix}$$

[2] Aqui expressamos o modelo sob a parametrização de médias de celas. Outras parametrizações podem ser consideradas; o leitor encontrará mais detalhes na Seção C.6.

e $\mathbf{e} = [e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}, \ldots, e_{k1}, \ldots, e_{kn_k}]^\top$. Embora esse seja um modelo linear (nos parâmetros), ele não impõe uma relação linear entre o valor esperado do número de frutos de boa qualidade (μ_i) e a concentração do fertilizante (x_{ij}), ou seja, permite acomodar um efeito não linear desse fator (no número esperado de frutos de boa qualidade). Se considerarmos a natureza quantitativa dos níveis do fator Concentração de fertilizante e quisermos adotar um modelo linear (expresso na forma genérica) que inclua um efeito linear desse fator, podemos utilizar

$$y_{ij} = \beta_0 + \beta_1 x_i + e_{ij}, \tag{C.1.4}$$

$i = 1, \ldots, I,\ j = 1, \ldots, n_i$ que pode ser expresso na forma matricial (C.1.1) com

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{In_I} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_I \\ \vdots & \vdots \\ 1 & x_I \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ e } \mathbf{e} = \begin{bmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ \vdots \\ e_{2n_2} \\ \vdots \\ e_{I1} \\ \vdots \\ e_{In_I} \end{bmatrix}.$$

Exemplo C.1.4: As ondas epidêmicas estacionais de gripe constituem uma carga considerável para os serviços de saúde e caracterizam-se por sua grande variabilidade de ano a ano. O uso dos dados históricos pode servir para estabelecer um modelo preditor do número de médicos indispensáveis para atender a totalidade das consultas de forma eficiente. Os dados da Tabela C.1.4 correspondem aos números de consultas (acumuladas semanalmente) da temporada invernal realizadas nos anos de 2000 a 2004 numa certa localidade.

Tabela C.1.4: Consultas acumuladas na temporada invernal de 2000 a 2004

Semana	2000	2001	2002	2003	2004
1	23470	46041	18284	20868	18726
2	26101	86018	23800	24574	21674
3	29178	105393	30764	28399	25771
4	32460	124280	40738	31496	29270
5	34949	138060	59248	34459	32563
6	37698	151779	93457	37703	40797
7	41216	162918	132469	42823	56698
8	44661	169642	162121	46930	72239
9	48030	174107	179232	50527	87008
10	55254	175520	185623	64897	84654

Supondo que para cada ano, a relação entre o tempo (em semanas) e a distribuição do número de consultas possa ser representado por intermédio de regressões lineares simples, poderíamos considerar um modelo do tipo

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + e_{ij}, \tag{C.1.5}$$

$i = 1, \ldots, I,\ j = 1, \ldots, n_i$, em que β_{0i} representa o número esperado de consultas no início do inverno do ano i ($i = 1$ correspondendo ao ano 2000) e β_{1i} denota a variação no número esperado de consultas por semana para o ano i. Neste caso particular, $I = 5$ e $n_i = 10$. Na forma matricial (C.1.1), esse modelo pode ser expresso como

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{In_I} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 0 & \ldots & 0 & x_{11} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \ldots & 0 & x_{1n_1} & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & x_{21} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \ldots & 0 & 0 & x_{2n_2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{I1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{In_I} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \vdots \\ \beta_{0I} \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1I} \end{bmatrix}$$

e

$$\mathbf{e} = [e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}, \ldots, e_{I1}, \ldots, e_{In_I}]^\top.$$

Se considerarmos que existe um número fixo de consultas de pacientes gripados no início do inverno (semana 1), e que a variação desse número de gripados depende

das condições climáticas de cada ano, o modelo poderia ser escrito como

$$y_{ij} = \beta_0 + \beta_{1i} x_{ij} + e_{ij}, \tag{C.1.6}$$

$i = 1, \ldots, I,\ j = 1, \ldots, n_i$ ou na forma matricial (C.1.1), com

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{In_I} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n_1} & 0 & \ldots & 0 \\ 1 & 0 & x_{21} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & x_{2n_2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & x_{I1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & x_{In_I} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1I} \end{bmatrix}$$

e

$$\mathbf{e} = [e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}, \ldots, e_{I1}, \ldots, e_{In_I}]^\top.$$

Em ambas as situações descritas acima, podemos dizer que existe **interação** entre o tempo decorrido desde o início do inverno e o ano (possivelmente por causa de diferenças climáticas), pois as taxas de variação no número esperado de consultas (β_{1i}) dependem do ano em que foram colhidos os dados (as retas que representam a variação no número esperado de consultas ao longo das 10 semanas de observação não são paralelas). Num modelo sem interação, essas taxas são iguais para todos os anos, ou seja, $\beta_{1i} = \beta_1$. Nesse caso, admitindo efeito de ano, o modelo poderia ser escrito como

$$y_{ij} = \beta_{0i} + \beta_1 x_{ij} + e_{ij},$$

$i = 1, \ldots, I,\ j = 1, \ldots, n_i$ e na forma matricial (C.1.1), com

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{In_I} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 0 & \ldots & 0 & x_{11} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 & x_{1n_1} \\ 0 & 1 & \ldots & 0 & x_{21} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & x_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & x_{I1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & x_{In_I} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \vdots \\ \beta_{0I} \\ \beta_1 \end{bmatrix}$$

e

$$\mathbf{e} = [e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}, \ldots, e_{I1}, \ldots, e_{In_I}]^\top .$$

Exemplo C.1.5: Em muitas situações, o modelo de **regressão segmentada** proporciona maior flexibilidade para a caracterização do comportamento da resposta. Esse modelo se caracteriza pela existência de um ponto x_0 (conhecido) em que a taxa de variação da resposta média se altera. Para justificar o modelo, suponhamos que

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + e_i, & x_i \leq x_0, \quad i = 1, \ldots, j \\ \beta_3 + \beta_2 x_i + e_i, & x_i \geq x_0, \quad i = j+1, \ldots, n. \end{cases}$$

Admitindo que as duas retas têm em comum o ponto (x_0, y_0), temos $\beta_0 + \beta_1 x_0 = \beta_3 + \beta_2 x_0$, ou seja $\beta_3 = \beta_0 + \beta_1 x_0 - \beta_2 x_0$. Substituindo essa expressão de β_3 no modelo original, obtém-se o seguinte modelo com três parâmetros $(\beta_0, \beta_1, \beta_2)$:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + e_i, & x_i \leq x_0, \quad i = 1, \ldots, j \\ \beta_0 + \beta_1 x_0 + \beta_2 (x_i - x_0) + e_i, & x_i \geq x_0, \quad i = j+1, \ldots, n. \end{cases}$$

que pode ser expresso na forma matricial (C.1.1) com

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ y_{j+1} \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_j & 0 \\ 1 & x_0 & (x_{j+1} - x_0) \\ \vdots & \vdots & \vdots \\ 1 & x_0 & (x_n - x_0) \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \text{ e } \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ e_{j+1} \\ \vdots \\ e_n \end{bmatrix}.$$

Aqui, o coeficiente β_0 corresponde ao valor esperado da resposta Y quando o valor da variável explicativa X é zero, o coeficiente β_1 pode ser interpretado como a variação esperada da resposta Y por unidade de variação da variável explicativa X para valores $X \leq x_0$ e o coeficiente β_2 corresponde à variação esperada da resposta Y por unidade de variação da variável explicativa X para valores $X \geq x_0$.

Sob a formulação matricial (C.1.1), hipóteses lineares de interesse podem ser expressas na forma

$$H : \mathbf{C}\boldsymbol{\beta} = \mathbf{m} \tag{C.1.7}$$

em que $\mathbf{C}$ é uma matriz $(c \times p)$ de constantes conhecidas com $r(\mathbf{C}) = c$ e $\mathbf{m}$ é um vetor $(c \times 1)$ de constantes conhecidas. Em geral, $\mathbf{m} = \mathbf{0}$.

No Exemplo C.1.1, a hipótese de que o intercepto é nulo é expressa como (C.1.7) com $\mathbf{C} = [1\ 0]$ e $\mathbf{m} = 0$ e a hipótese de que o coeficiente angular é nulo, com $\mathbf{C} = [0\ 1]$ e $\mathbf{m} = 0$. Além disso, supondo que, sob a parametrização alternativa lá

mencionada, $x_0 = 17$, a hipótese de que o valor esperado para a pressão sistólica de pacientes com idade igual a 30 anos é de 120 mmHg seria expressa na forma (C.1.7) com $\mathbf{C} = [1\ 13]$ e $\mathbf{m} = 120$. Em todos esses casos, a matriz $\mathbf{C}$ tem apenas uma linha, e consequentemente, $c = 1$.

No Exemplo C.1.2, uma hipótese de interesse é a de que nenhuma das duas variáveis é significativamente linearmente associada ao valor esperado da variável resposta. Esta hipótese pode ser expressa na forma (C.1.7) com

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

No Exemplo C.1.3, sob o modelo (C.1.3), a hipótese de que os valores esperados de frutos de boa qualidade por macieira sejam iguais, independentemente da concentração de fertilizante utilizada pode ser expressa na forma (C.1.7), em que $\mathbf{m} = \mathbf{0}$ é um vetor $[(I-1) \times 1]$ e $\mathbf{C}$ é uma matriz $[(I-1) \times I]$ construída de forma a gerar $I-1$ contrastes linearmente independentes dos elementos de $\boldsymbol{\beta} = [\mu_1\ \mu_2 \ldots \mu_I]^\top$. Como exemplo desse tipo de matriz, temos

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

Sob o modelo (C.1.4), uma hipótese equivalente seria expressa com $\mathbf{C} = [0\ 1]$ e $\mathbf{m} = 0$.

No Exemplo C.1.4, a hipótese a ser testada para avaliar se o modelo (C.1.5) pode ser reduzido ao modelo (C.1.6) também pode ser expressa na forma (C.1.7) em que $\mathbf{C}$ é uma matriz $[(I-1) \times 2I]$ com $r(\mathbf{C}) = I-1$ dada por

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & \ldots & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & -1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 1 & -1 & 0 & \ldots & 0 & 0 \end{bmatrix},$$

por exemplo, e $\mathbf{m} = \mathbf{0}$, um vetor de dimensão $[(I-1) \times 1]$.

No Exemplo C.1.5, a hipótese de que os coeficientes angulares correspondentes a valores da variável explicativa menores ou maiores do que x_0 são iguais pode ser expressa na forma (C.1.7) com $\mathbf{C} = [0\ \ 1\ \ -1]$.

C.2 Método de mínimos quadrados

O método de mínimos quadrados (ordinários) pode ser utilizado para estimar o vetor de parâmetros $\boldsymbol{\beta}$ no modelo (C.1.1)-(C.1.2). A base do método é a minimização da forma quadrática

$$Q(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2. \tag{C.2.1}$$

Com essa finalidade, consideremos

$$Q(\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y}$$

e utilizemos as expressões para derivadas matriciais apresentadas no Apêndice A para obter

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\top \mathbf{y}, \\ \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= 2\mathbf{X}^\top \mathbf{X}. \end{aligned}$$

Igualando a primeira derivada parcial a zero, obtemos o **sistema de equações de estimação** (também conhecido por **sistema de equações normais**) é [3]

$$\mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}.$$

Como assumimos que $\mathbf{X}$ tem posto completo, $\mathbf{X}^\top \mathbf{X}$ é uma matriz não singular e a solução do sistema de equações normais

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{C.2.2}$$

Como a matriz $\mathbf{X}^\top \mathbf{X}$ é definida positiva, o estimador (C.2.2) corresponde ao ponto de mínimo de (C.2.1).

Sob a suposição de que $\mathbb{E}(\mathbf{e}) = \mathbf{0}$ e $\mathbb{V}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$, temos

i) $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,

ii) $\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

[3] Aqui, o termo "normal" não se refere à distribuição normal e sim ao conceito de ortogonalidade. A razão para isso é que a teoria de mínimos quadrados pode ser desenvolvida por meio de projeções ortogonais.

Quando a matriz $\mathbf{X}$ não tem posto completo, o sistema de equações de estimação tem infinitas soluções dadas por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{y} + [\mathbf{I} - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}]\mathbf{g},$$

em que $(\mathbf{X}^\top\mathbf{X})^-$ é uma inversa generalizada de $\mathbf{X}^\top\mathbf{X}$ e $\mathbf{g}$ é um vetor arbitrário. Pode-se mostrar que $\mathbf{X}\widehat{\boldsymbol{\beta}}$ é invariante com relação ao vetor $\mathbf{g}$ e $Q(\boldsymbol{\beta})$ obtém seu ponto mínimo para qualquer solução de $\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}$. Nesse caso, $\widehat{\boldsymbol{\beta}}$ deve ser encarado como uma solução das equações normais e não como um estimador de β, que é dito **não estimável**. [4]

Há casos em que a matriz $\mathbf{X}$ tem posto completo, mas suas colunas estão muito próximas de serem linearmente dependentes; isso ocorre, por exemplo, em alguns modelos de regressão polinomial. Como consequência, a matriz $\mathbf{X}^\top\mathbf{X}$ é "quase" não singular, com autovalores próximos de zero de forma a gerar estimadores $\widehat{\boldsymbol{\beta}}$ muito imprecisos. A presença de um certo nível de dependência entre as colunas de $\mathbf{X}$ é conhecida na literatura como **multicolinearidade**. Procedimentos para lidar com esse problema serão discutidos na Seção C.5.

O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é o estimador linear não enviesado de variância mínima, (*best linear unbiased estimator* - BLUE) como se pode deduzir por meio do seguinte teorema:

Teorema* C.2.1. *Teorema de Gauss-Markov: *Considere o modelo linear* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$*, com as hipóteses* $\mathbb{E}(\mathbf{e}) = \mathbf{0}$ *e* $\mathbb{V}(\mathbf{e}) = \sigma^2\mathbf{I}_n$*. Seja* $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ *o estimador de mínimos quadrados de* $\boldsymbol{\beta}$*. Se* $\widetilde{\boldsymbol{\beta}}$ *é um outro estimador de* $\boldsymbol{\beta}$ *tal que* $\widetilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$*, em que* $\mathbf{C}$ *é uma matriz* $(p \times n)$*, e* $\mathbb{E}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$*, então*

$$\mathbf{r}^\top\mathbb{V}(\widetilde{\boldsymbol{\beta}})\mathbf{r} \geq \mathbf{r}^\top\mathbb{V}(\widehat{\boldsymbol{\beta}})\mathbf{r},\ \forall\mathbf{r} \in I\!R^p.$$

Sob a suposição adicional de que os erros e_i têm distribuição Normal, a linearidade do estimador (C.2.2) permite-nos concluir que

$$\widehat{\boldsymbol{\beta}} \sim N_p[\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}]. \tag{C.2.3}$$

Em geral, a variância σ^2 é desconhecida e inferências sobre $\boldsymbol{\beta}$ dependem de um estimador desse parâmetro; um candidato é o estimador não enviesado

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

[4] Uma função $\mathbf{a}^\top\boldsymbol{\beta}$ é estimável se tiver um estimador linear não enviesado, digamos $\mathbf{b}^\top\mathbf{y}$. Em muitos casos, a matriz $\mathbf{X}$ define modelos não identificáveis cujos parâmetros são não estimáveis embora haja funções desses parâmetros que são estimáveis. Para detalhes, veja a Seção C.6 e Freedman (2005).

Como a matriz $\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ é idempotente pode-se recorrer ao Teorema A.3.3 para mostrar que $(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}$.

Dados uma matriz $\mathbf{C}$ de constantes conhecidas com dimensão $(c \times p)$ e $r(\mathbf{C}) = c$ e um vetor $\mathbf{m}$ de constantes conhecidas com dimensão $(c \times 1)$ consideremos a forma quadrática

$$\begin{aligned} Q_{\mathbf{C}} &= (\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m})^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m}) \\ &= [\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{m}]^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{m}]. \end{aligned}$$

Tendo (C.2.3) em conta, podemos novamente recorrer ao Teorema A.3.3 para mostrar que $\sigma^{-2}Q_{\mathbf{C}} \sim \chi^2_c(\delta)$ com parâmetro de não centralidade

$$\delta = \frac{1}{2}(\mathbf{C}\boldsymbol{\beta} - \mathbf{m})^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{m}).$$

Como $\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{m} = \mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top[\mathbf{y} - \mathbf{X}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{m}]$, é possível expressar $Q_{\mathbf{C}}$ como uma forma quadrática na variável $\mathbf{y} - \mathbf{X}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{m}$, ou seja,

$$Q_{\mathbf{C}} = [\mathbf{y} - \mathbf{X}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{m}]^\top\mathbf{A}[\mathbf{y} - \mathbf{X}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{m}]$$

com

$$\mathbf{A} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

Também é possível escrever

$$\begin{aligned} (n-p)s^2 &= \mathbf{y}^\top[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{y} \\ &= [\mathbf{y} - \mathbf{X}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{m}]^\top\mathbf{B}[\mathbf{y} - \mathbf{X}\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{m}] \end{aligned}$$

com $\mathbf{B} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Observando que $\mathbf{AB} = \mathbf{0}$, podemos utilizar o Teorema A.3.5 para mostrar que

$$F = c^{-1}s^{-2}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m})^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m}) \sim F_{(c,n-p)}(\delta), \qquad \text{(C.2.4)}$$

com parâmetro de não centralidade

$$\delta = \frac{1}{2}(\mathbf{C}\boldsymbol{\beta} - \mathbf{m})^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{m}).$$

Esse resultado nos dá todos os ingredientes necessários tanto para construir intervalos (ou regiões) de confiança para combinações lineares dos elementos de $\boldsymbol{\beta}$ quanto para testar hipóteses lineares de interesse. Por exemplo, para testar hipóteses na forma (C.1.7), basta notar que sob a hipótese nula $\delta = 0$ e a estatística (C.2.4) tem distribuição $F_{(c,n-p)}$.

Sob o modelo de regressão linear simples, $y_i = \alpha + \beta x_i + e_i$, $i = 1, \ldots, n$ com as suposições mencionadas acima (incluindo a de normalidade de e_i), um teste da hipótese de que o coeficiente angular é igual a uma constante β_0 pode ser realizado por meio de (C.2.4) com $\mathbf{C} = [0\ 1]$ e $\mathbf{m} = \beta_0$. Nesse caso, a estatística (C.2.4) pode ser expressa como

$$(\widehat{\beta} - \beta_0)^2 / \{s^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{22}\}$$

em que $[(\mathbf{X}^\top \mathbf{X})^{-1}]_{22}$ denota o elemento $(2,2)$ da matriz $(\mathbf{X}^\top \mathbf{X})^{-1}$ e segue uma distribuição $F_{(1,n-p)}$ quando a hipótese $H : \beta = \beta_0$ é verdadeira. A raiz quadrada dessa estatística munida de sinal$(\widehat{\beta} - \beta_0)$[5], nomeadamente

$$[\text{sinal}(\widehat{\beta} - \beta_0)](\widehat{\beta} - \beta_0) / \{s \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{22}}\}$$

segue uma distribuição t_{n-p} quando a hipótese $H : \beta = \beta_0$ é verdadeira. O resultado para o caso particular em que $\beta_0 = 0$ é obtido de forma automática quando se utilizam os pacotes computacionais mais comuns.

Nos casos de dúvidas quanto à validade da suposição de normalidade, podemos recorrer ao Teorema Limite Central para efeito de inferência sobre os parâmetros do modelo. Para salientar que o resultado depende do tamanho da amostra, adicionamos o índice n aos componentes do modelo. Consideremos agora, uma sequência de elementos $\{\mathbf{y}_n, \mathbf{X}_n, \mathbf{e}_n\}$ com $\mathbf{X}_n = [\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nn}]^\top$ satisfazendo o modelo $\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta} + \mathbf{e}_n$ sob as suposições do Teorema de Gauss-Markov e admitamos que

a) $\max_{1 \le i \le n} \mathbf{x}_{ni}^\top (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{x}_{ni} \to 0$ com $n \to \infty$, [6]

b) $\lim_{n \to \infty} \mathbf{X}_n^\top \mathbf{X}_n = \mathbf{V}$, com $\mathbf{V}$ finita e definida positiva.

Então, utilizando o Teorema Limite Central de Hájek-Šidak (ver Sen et al. (2009), por exemplo), podemos mostrar que

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \sigma^2 \mathbf{V}^{-1})$$

ou equivalentemente, que

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \sigma^2 \mathbf{I}_p).$$

[5] sinal$(a) = 1$ se $a > 0$, sinal$(a) = -1$ se $a < 0$ e sinal$(a) = 0$ se $a = 0$.

[6] Esta suposição, conhecida como **condição de Noether** é válida na maioria dos casos de interesse prático. Em particular, no caso de problemas envolvendo a comparação de I médias, ela corresponde à exigência de que os tamanhos n_i das amostras colhidas de cada subpopulação investigada sejam tais que $n_i / \sum_{i=1}^I n_i$ convirja para uma constante λ_i com $n = \sum_{i=1}^I n_i \to \infty$.

Em termos práticos, esse resultado significa que para amostras com tamanhos suficientemente "grandes",

$$\widehat{\boldsymbol{\beta}} \approx N_p[\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}]$$

ou seja, que a distribuição do estimador $\widehat{\boldsymbol{\beta}}$ pode ser aproximada por uma distribuição normal com média $\boldsymbol{\beta}$ e matriz de covariâncias $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

Neste caso, para testar hipóteses da forma (C.1.7) podemos empregar a **estatística de Wald**

$$Q_W = s^{-2}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m})^\top[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m}), \tag{C.2.5}$$

cuja distribuição sob a hipótese nula pode ser aproximada por uma distribuição χ^2_c.

Existem situações em que a suposição de **homocedasticidade** *i.e.,* variâncias constantes, não é válida. Como ilustração, tomemos o Exemplo C.1.1, em que um aumento da variância da pressão arterial sistólica com a idade não seria um fato inesperado. Por exemplo, poderíamos supor que $\mathbb{V}(e_i) = x_i\sigma^2$, ou seja, que $\mathbb{V}(\mathbf{e}) = \text{diag}\{x_1, \ldots, x_{20}\}\sigma^2$. Nesse caso, o modelo é dito **heterocedástico**. De uma forma mais geral, podemos ter modelos em que a matriz de covariâncias de $\mathbf{e}$ é uma matriz simétrica definida positiva, $\mathbf{V}$. Nesses casos, o ajuste do modelo pode ser concretizado por meio do método dos **mínimos quadrados generalizados** (MQG) que consiste em minimizar

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

Quando a matriz $\mathbf{V}$ é uma matriz diagonal com elementos $w_i\sigma^2$ ao longo da diagonal principal, o método é chamado de **mínimos quadrados ponderados** e a função a ser minimizada pode ser expressa como

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^{n} w_i^{-1}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2.$$

Admitindo que $\mathbf{V}$ é conhecida e considerando um procedimento similar àquele usado no caso de mínimos quadrados ordinários, podemos mostrar que o estimador de mínimos quadrados generalizados de $\boldsymbol{\beta}$ é

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

Se $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V})$ com $\mathbf{V}$ conhecida, então podemos demonstrar que

$$\widetilde{\boldsymbol{\beta}} \sim N_p[\boldsymbol{\beta}, (\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}].$$

Quando há dúvida com relação à suposição de normalidade, mas $\mathbb{V}(\mathbf{e}) = \mathbf{V}$, com $\mathbf{V}$ finita e conhecida, o Teorema Limite Central permite-nos concluir que para n suficientemente grande,

$$\tilde{\boldsymbol{\beta}} \approx N_p[\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}].$$

Como, em geral, $\mathbf{V}$ não é conhecida, podemos substituí-la por um estimador consistente $\widehat{\mathbf{V}}$ e considerar o estimador

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \widehat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{V}}^{-1} \mathbf{y}. \tag{C.2.6}$$

O estimador consistente de $\mathbf{V}$ pode ser obtido de fontes externas, *e.g.*, outros estudos com características similares; em muitos casos, é possível utilizar a matriz de covariâncias dos resíduos obtidos do ajuste de um modelo homocedástico com os mesmos parâmetros $\boldsymbol{\beta}$. Apelando para o Teorema Limite Central e para o Teorema de Sverdrup (ver Sen et al. (2009)), por exemplo, podemos mostrar que para para n suficientemente grande

$$\widehat{\boldsymbol{\beta}} \approx N_p[\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}]. \tag{C.2.7}$$

Intervalos de confiança aproximados para combinações lineares da forma $\mathbf{C}\boldsymbol{\beta}$ ou testes para hipóteses do tipo (C.1.7) podem ser concretizados utilizando (C.2.7) ou a estatística de Wald

$$Q_W = (\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m})^\top [\mathbf{C}(\mathbf{X}^\top \widehat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{C}^\top]^{-1} (\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{m}), \tag{C.2.8}$$

cuja distribuição sob a hipótese nula pode ser aproximada por uma distribuição χ^2_c.

C.3 Método de máxima verossimilhança

Sob a suposição de normalidade, podemos considerar o método de máxima verossimilhança para a estimação dos parâmetros do modelo (C.1.1)-(C.1.2). Neste caso, o método consiste em maximizar a função de verossimilhança

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]. \tag{C.3.1}$$

com relação aos parâmetros $\boldsymbol{\beta}$ e σ^2. Lembrando que a determinação dos pontos de máximo de (C.3.1) é equivalente à obtenção dos pontos de máximo de seu logaritmo e utilizando as regras de derivação matricial apresentadas no Apêndice A, podemos mostrar que os estimadores de máxima verossimilhança desejados são

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

e

$$\widehat{\sigma}^2 = n^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = n^{-1}\mathbf{y}^\top[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{y}.$$

Dado que a maximização de (C.3.1) relativamente a $\boldsymbol{\beta}$ corresponde à minimização da forma quadrática de seu expoente, não é de se estranhar que os estimadores de máxima verossimilhança e de mínimos quadrados sejam idênticos. Por outro lado, o estimador de máxima verossimilhança de σ^2 corresponde ao estimador proposto anteriormente (s^2) multiplicado por um fator $(n-p)/n$ e é, consequentemente, enviesado. A razão para isso é que o estimador $\widehat{\sigma}^2$ depende de $\boldsymbol{\beta}$ e uma alternativa para se conseguir um estimador não enviesado é utilizar o método de máxima verossimilhança restrita que consiste na minimização de uma transformação linear dos dados que não dependa desse parâmetro. Uma transformação com essas características é dada por $\mathbf{y}^* = [\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}]\mathbf{y}$. A maximização da verossimilhança correspondente gera o estimador não enviesado s^2. Mais detalhes sobre esse método podem ser encontrados em Diggle, Heagerty, Liang & Zeger (2002), por exemplo.

A metodologia de máxima verossimilhança também pode ser aplicada a modelos do tipo (C.1.1) com $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V})$ em que os elementos não redundantes de $\mathbf{V}$ são expressos na forma de um vetor de parâmetros (de covariâncias), $\boldsymbol{\theta}$, *i.e.*, $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$. A função de verossimilhança correspondente é

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = (2\pi)^{-n/2}|\mathbf{V}(\boldsymbol{\theta})|^{-1/2}\exp[-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]. \quad \text{(C.3.2)}$$

Nesse caso, o estimador de $\boldsymbol{\beta}$ tem a mesma forma que o estimador de mínimos quadrados generalizados (C.2.6) mas com $\widehat{\mathbf{V}} = \mathbf{V}(\widehat{\boldsymbol{\theta}})$ representando o estimador de máxima verossimilhança de $\mathbf{V}$. É possível impor estruturas particulares à matriz de covariâncias, ou seja, é possível considerar situações em que $\mathbf{V}$ é função de um vetor de parâmetros de covariâncias com dimensão reduzida, *e.g.*, 2 ou 3 em vez de $n(n-1)/2$, que é a dimensão de $\boldsymbol{\theta}$ quando não se impõem restrições. Modelos com essa natureza são considerados para a análise de medidas repetidas ou de dados longitudinais.

C.4 Partição da soma de quadrados

A utilidade do modelo de regressão $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ para explicação da variação da resposta esperada como função das variáveis explicativas pode ser avaliada a partir da identidade

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

em que $\widehat{y}_i = \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}$ com $\mathbf{x}_i$ denotando a i-ésima linha da matriz $\mathbf{X}$ e $\overline{y} = n^{-1}\sum_{i=1}^n y_i$. Pode-se demonstrar que

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^n (y_i - \widehat{y}_i)^2.$$

Os termos

$$SQT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \mathbf{y}^\top[\mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^\top]\mathbf{y},$$

$$SQR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \mathbf{y}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - n^{-1}\mathbf{1}\mathbf{1}^\top)]\mathbf{y}$$

e

$$SQE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \mathbf{y}^\top[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{y}$$

correspondem, respectivamente, à **soma de quadrados total**, **soma de quadrados devida à regressão** e **soma de quadrados residual** e representam a variabilidade (expressa em termos de quadrados de diferenças) da resposta em torno de sua média, a parcela dessa variabilidade explicada pelo modelo $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ e a variabilidade da resposta em torno deste modelo, ou seja a parcela da variabilidade total não explicada.

Quando o ajuste do modelo $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ é perfeito, *i.e.*, quando $\mathbf{y} = \widehat{\mathbf{y}}$ temos $SQE = 0$ e quando o modelo $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ não melhora a explicação da variabilidade relativamente à resposta média, $SQR = 0$. Uma medida da parcela da variabilidade explicada pelo modelo $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ é o **coeficiente de determinação**

$$R^2 = \frac{SQR}{SQT} = \frac{SQT - SQE}{SQT} = 1 - \frac{SQE}{SQT}.$$

Como $0 \leq SQE \leq SQT$, temos $0 \leq R^2 \leq 1$ e quanto maior o valor de R^2, maior é a redução da variabilidade da resposta explicada pela introdução das variáveis explicativas.

Um dos problemas associados à utilização do coeficiente de determinação como medida da qualidade do modelo é que ele não leva em consideração o número de parâmetros envolvidos. Quanto mais variáveis explicativas forem introduzidas no modelo, mais o coeficiente R^2 se aproximará de 1. Para contornar esse problema, podemos considerar o **coeficiente de determinação ajustado**:

$$R^2_a = 1 - \left(\frac{n-1}{n-p}\right)\frac{SQE}{SQT}.$$

O coeficiente R^2_a pode diminuir quando adicionamos variáveis explicativas ao modelo pois o decréscimo que isso acarretará em SQE pode ser compensado pela perda de graus de liberdade no denominador $n - p$.

C.5 Diagnóstico

Modelos estatísticos são utilizados como aproximações de processos complexos e são construídos sobre um conjunto de suposições. Para efeitos práticos, é importante avaliar se tais aproximações são aceitáveis. Isto pode ser concretizado por meio de **técnicas de diagnóstico**, que englobam a **avaliação do ajuste** e a **análise de sensibilidade**. No primeiro caso, o objetivo é avaliar se as suposições adotadas são compatíveis com os dados; no segundo, o objetivo é estudar a variação dos resultados da análise quando a formulação inicial do modelo é ligeiramente modificada. Se esta variação for "substancial" no sentido de mudar as conclusões, diz-se que o modelo não é **robusto**. Nesse caso, ou as conclusões devem ser tomadas (se tomadas) de forma cautelosa, ou então deve-se optar por outro modelo.

Para ilustrar a importância do uso das técnicas de diagnóstico, apresentamos na Tabela C.5.1, quatro conjuntos de dados (A, B, C e D) extraídos de Anscombe (1973). Cada um desses conjuntos contém os valores de uma variável explicativa (x) e de uma variável resposta (y).

Tabela C.5.1: Dados obtidos de Anscombe (1973)

A		**B**		**C**		**D**	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

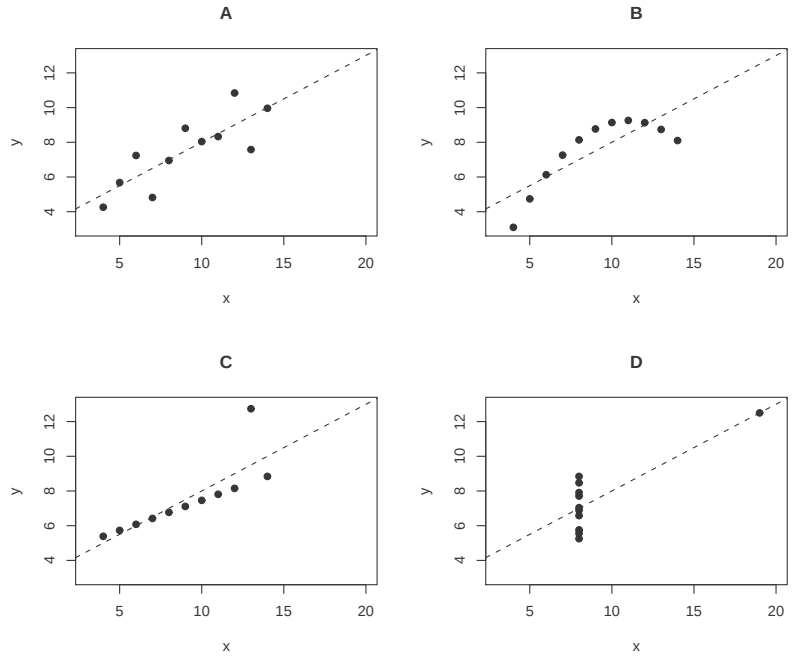
Consideremos o ajuste do seguinte modelo de regressão linear simples

$$y_i = \alpha + \beta x_i + e_i,$$

$i = 1, ..., 11$ com $\mathbb{E}(e_i) = 0$, $\mathbb{V}(e_i) = \sigma^2$ e $\mathbb{C}\text{ov}(e_i, e_j) = 0$, $i \neq j$ a cada um dos quatro subconjuntos de dados. Na Figura C.5.1 apresentamos diagramas de dispersão correspondentes a cada subconjunto; as linhas tracejadas correspondem às retas

ajustadas pelo método de mínimos quadrados. No conjunto A, o modelo parece adequado; no conjunto B, um modelo quadrático seria mais conveniente e no conjunto C, o modelo parece ajustar-se bem aos dados, com exceção do ponto (13.0, 12.74) que se caracteriza como uma observação **discrepante** (*outlier*)[7]. No conjunto D, o coeficiente de regressão β é significativo apenas em função do ponto $(19.0, 12.50)$. No entanto, para os quatro conjuntos de dados, temos $\widehat{\alpha} = 3.00$, $\widehat{\beta} = 0.50$, $\widehat{\sigma}^2 = 1.53$ e $R^2 = 0.667$ o que evidencia que o coeficiente de determinação (R^2) nem sempre é uma boa medida para a avaliação da qualidade do ajuste. Essa avaliação precisa ser complementada por outros métodos, alguns dos quais descrevemos a seguir.

Figura C.5.1: Diagramas de dispersão (dados da Tabela C.5.1)



C.5.1 Análise de resíduos

Resíduos são utilizados para avaliar a validade de determinadas suposições de modelos estatísticos. No caso de modelos de regressão linear clássicos, podemos utilizá-los para verificar homocedasticidade, existência de pontos discrepantes, normalidade e independência dos erros. Nós adotamos a proposta de Cox & Snell (1968), que apre-

[7] Grosso modo, uma observação é chamada discrepante, se apresenta um comportamento distinto das demais.

sentam uma forma geral para definir resíduos para modelos que contêm uma única fonte de variação.

Consideremos o modelo de regressão linear

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{C.5.1}$$

com $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Seja $\widehat{\boldsymbol{\beta}}$ o estimador de mínimos quadrados (ordinários) de $\boldsymbol{\beta}$ e $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y}$, o vetor dos valores ajustados. A matriz

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

é simétrica e idempotente e representa a matriz de projeção de $\mathbf{y}$ no subespaço $\mathcal{C}(\mathbf{X})$ (para maiores detalhes, veja o Teorema A.2.7). O vetor de **resíduos ordinários** é definido como

$$\widehat{\mathbf{e}} = [\widehat{e}_1, ..., \widehat{e}_n]^\top = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{C.5.2}$$

$$= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = (\mathbf{I} - \mathbf{H})\mathbf{e}, \tag{C.5.3}$$

de onde podemos concluir que $\widehat{\mathbf{e}} \sim N_n(\mathbf{0}, \sigma^2[\mathbf{I} - \mathbf{H}])$. A relação entre o vetor de erros e o vetor de resíduos ordinários depende somente da **matriz chapéu** (*hat matrix*) $\mathbf{H}$. [8]

Apresentamos a seguir algumas transformações dos resíduos ordinários apropriadas para fins de diagnóstico. O leitor interessado pode consultar Cook & Weisberg (1982), por exemplo, para detalhes.

A distribuição do vetor de resíduos ordinários depende tanto de σ^2 quanto da matriz $\mathbf{H}$; consequentemente, os resíduos podem ter variâncias distintas. Para efeito de comparação, convém construir **resíduos padronizados** que não dependam dessas quantidades. Se σ^2 for conhecido, uma padronização natural consistiria na divisão de $\widehat{e}_i$ pelo seu desvio padrão, a saber, $\sigma\sqrt{1 - h_{ii}}$, em que h_{ii} denota o i-ésimo elemento da diagonal principal de $\mathbf{H}$. Pode-se mostrar que $h_{ii} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$ em que $\mathbf{x}_i^\top$ é a i-ésima linha de $\mathbf{X}$. Com essa padronização, a distribuição conjunta dos resíduos transformados não depende da variância. Como, em geral, σ^2 é desconhecida, uma alternativa é considerar os **resíduos studentizados**, definidos por

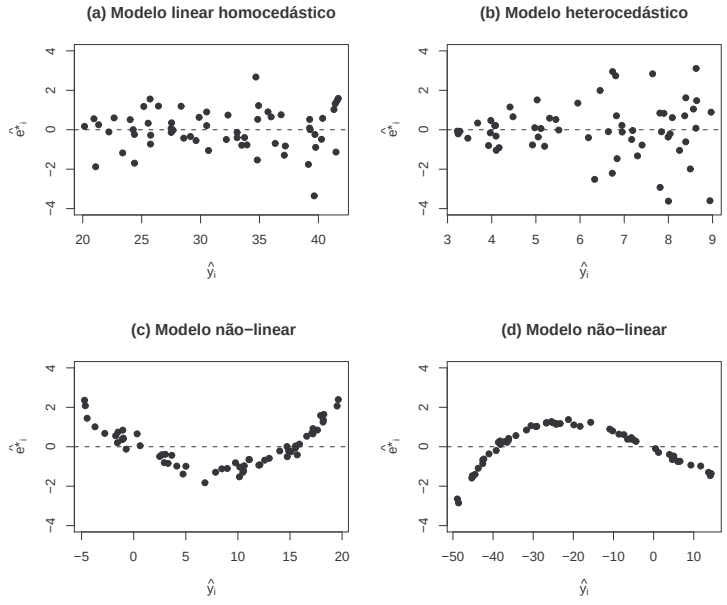
$$\widehat{e}_i^* = \frac{\widehat{e}_i}{s\sqrt{1 - h_{ii}}},\ i = 1, ..., n.$$

Um gráfico de $\widehat{e}_i^*$ *versus* $\widehat{y}_i$ (o valor predito da i-ésima observação) é útil para verificar a plausibilidade da suposição de homocedasticidade; quando esta suposição é verdadeira, espera-se que o comportamento de $\widehat{e}_i^*$ em torno do valor zero seja aleatório

[8] Essa denominação, dada por J.W. Tukey, se deve ao fato de $\mathbf{H}$ ser base da transformação linear do vetor dos valores observados no vetor de valores ajustados, $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, ou seja, $\mathbf{H}$ "coloca" o chapéu ($\widehat{\ }$) no vetor $\mathbf{y}$. Ela também é chamada de **matriz de predição**.

e que sua variabilidade tenha magnitude independente do valor de $\widehat{y}_i$. Esse tipo de gráfico também pode ser utilizado para avaliar a suposição de linearidade assumida pelo modelo, além de revelar para quais observações essa suposição parece inadequada. Se a suposição for verdadeira, esperamos que nenhuma tendência na disposição dos resíduos seja detectada. Como ilustração, considere a Figura C.5.2 que corresponde aos resíduos associados a quatro ajustes fictícios. No gráfico C.5.2(a) não detectamos nenhum tipo de violação das suposições de homocedasticidade e de linearidade; no gráfico C.5.2(b), podemos notar que a variabilidade dos resíduos aumenta com a magnitude do valor ajustado, indicando que o modelo em questão deve ser heterocedástico e nos gráficos C.5.2(c) e C.5.2(d) podemos notar uma tendência não linear dos resíduos, o que sugere a má especificação da forma funcional. Nesses casos, uma função não linear talvez fosse mais adequada para representar a relação entre o valor esperado da variável resposta e os valores da variável explicativa.

Figura C.5.2: Resíduos studentizados *versus* valores ajustados



Os resíduos studentizados que apresentam um valor absoluto "muito grande" (maior que 2 ou 3, por exemplo) identificam observações discrepantes. Este critério tem cunho puramente descritivo, dado que na realidade $(\widehat{e}_i^*)^2/(n-p) \sim Beta[1/2, (n-p-1)/2]$[9], indicando que $|\widehat{e}_i^*| \leq \sqrt{n-p}$. Além disso, pode-se mostrar, que $\mathbb{E}[\widehat{e}_i^*] = 0$, $\mathbb{V}[\widehat{e}_i^*] = 1$ e que $\mathbb{C}\mathrm{ov}[\widehat{e}_i^*, \widehat{e}_j^*] = -h_{ij}/[(1-h_{ii})(1-h_{jj})]^{1/2}$. Para detalhes e demonstração destas propriedades, veja Cook & Weisberg (1982), por exemplo.

[9] $Beta(a, b)$ denota a distribuição beta com parâmetros $a, b > 0$.

Na definição dos resíduos studentizados, utilizamos o quadrado médio do resíduo (s^2) como estimador de σ^2. Dada a dependência entre $\widehat{e}_i$ e s^2, a distribuição de $\widehat{e}_i^*$ não possui uma densidade com tratamento matemático fácil. Isto motiva outra forma de padronização dos resíduos ordinários em que um estimador de σ^2 independente de e_i é utilizado. Com essa finalidade, podemos considerar o quadrado médio do resíduo obtido com a omissão da i-ésima observação, que pode ser calculado como

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

Pode-se provar que $\widehat{e}_i^*$ e $s^2_{(i)}$ são independentes, de forma que definimos os **resíduos studentizados externamente** como

$$t_i = \frac{\widehat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}},\ i = 1, ..., n.$$

Cook & Weisberg (1982) mostram que $t_i \sim t(n-p-1)$ e que

$$t_i = \widehat{e}_i^* \left(\frac{n-p-1}{n-p-(\widehat{e}_i^*)^2} \right)^{1/2},$$

indicando que t_i é uma transformação monótona crescente de $\widehat{e}_i^*$ e proporcionando uma forma de cálculo que prescinde do ajuste do modelo sem a i-ésima observação.

Além disso, é possível mostrar que valores "grandes" de $|t_i|$, e consequentemente de $|\widehat{e}_i^*|$, podem ser utilizados como identificadores de observações significativamente discrepantes; detalhes que incluem sugestões para pontos de corte podem ser obtidos em Cook & Weisberg (1982).

Outros tipos de resíduos, tais como **resíduos preditos** (*predicted residuals*) e **resíduos recursivos** (*recursive residuals*), também são discutidos em Cook & Weisberg (1982).

C.5.2 Análise da suposição de normalidade

Na teoria clássica de regressão, tanto intervalos de confiança quanto testes de hipóteses sobre os parâmetros de modelos lineares são baseados na suposição de normalidade dos erros. A verificação da plausibilidade dessa suposição é fundamental para a validade dos procedimentos inferenciais (exatos). Como os resíduos são essencialmente preditores dos erros do modelo, nada mais natural do que utilizá-los com essa finalidade. Nesse sentido, gráficos do tipo **QQ** (**quantis-quantis**), em que dispomos os resíduos studentizados ordenados (quantis observados) no eixo das ordenadas e os quantis obtidos da distribuição normal padrão (quantis teóricos) no

eixo das abscissas, são ferramentas utilíssimas. Quando a distribuição dos erros é gaussiana, espera-se que esses resíduos estejam dispostos numa vizinhança da reta com inclinação de 45 graus. Esse tipo de gráfico também pode ser útil para detectar a presença de observações discrepantes, para avaliar se a distribuição dos erros possui caudas mais pesadas que a distribuição normal, para avaliar se os erros são heterocedásticos etc.

Esses gráficos podem ser obtidos por meio do seguinte procedimento:

i) Ajustar o modelo (C.5.1), obter o vetor de resíduos studentizados $\widehat{\mathbf{e}}^* = (\widehat{e}_1^*, \ldots, \widehat{e}_n^*)^\top$ e o vetor de resíduos studentizados ordenados, $(\widehat{e}_{(1)}^* \leq \widehat{e}_{(2)}^* \leq \ldots \leq \widehat{e}_{(n)}^*)^\top$.

ii) Calcular $p_i = (i - 0.375)/n$, $i = 1, \ldots, n$ e definir os quantis amostrais de ordem p_i, $i = 1, \ldots, n$ como $Q_{p_i} = \widehat{e}_{(i)}^*$.

iii) Obter os quantis normais $Z_{p_i} = P[Z \leq Z(p_i)] = p_i$, $i = 1, \ldots, n$ em que Z é uma variável com distribuição $N(0, 1)$.

iv) Construir o gráfico de dispersão $Q_{p_i} \times Z_{p_i}$.

Como em gráficos do tipo QQ, é difícil avaliar afastamentos da normalidade visualmente, Atkinson (1985), sugere a construção de bandas de confiança, denominadas **envelopes simulados** (*simulated envelopes*). Essas bandas de confiança são obtidas por meio de simulação de dados com distribuição normal com vetor de médias iguais a zero e matriz de covariâncias $(\mathbf{I} - \mathbf{H})$.

Um algoritmo para sua construção é o seguinte

i) Ajustar o modelo (C.5.1), obtendo $\widehat{\boldsymbol{\beta}}$, s^2 e o vetor de resíduos studentizados $\widehat{\mathbf{e}}^*$.

ii) Construir o gráfico QQ correspondente.

iii) Gerar um vetor $\mathbf{z}$ com n elementos correspondentes a valores independentes de uma distribuição $N(0, 1)$.

iv) Obter um vetor de observações simuladas $\mathbf{y}_s = \mathbf{X}\widehat{\boldsymbol{\beta}} + s\mathbf{z}$.

v) Ajustar o modelo $\mathbf{y}_s = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_s$, obter o vetor de resíduos studentizados (simulados) $\widehat{\mathbf{e}}_s^*$ e ordenar seus componentes.

v) Repetir os itens (iii)-(v) m vezes, gerando para cada $p_i = (i - 0.375)/n$, um conjunto de m resíduos simulados.

vi) Selecionar para cada p_i, o menor e o maior resíduo simulado e incluí-los no gráfico QQ construído no item (ii).

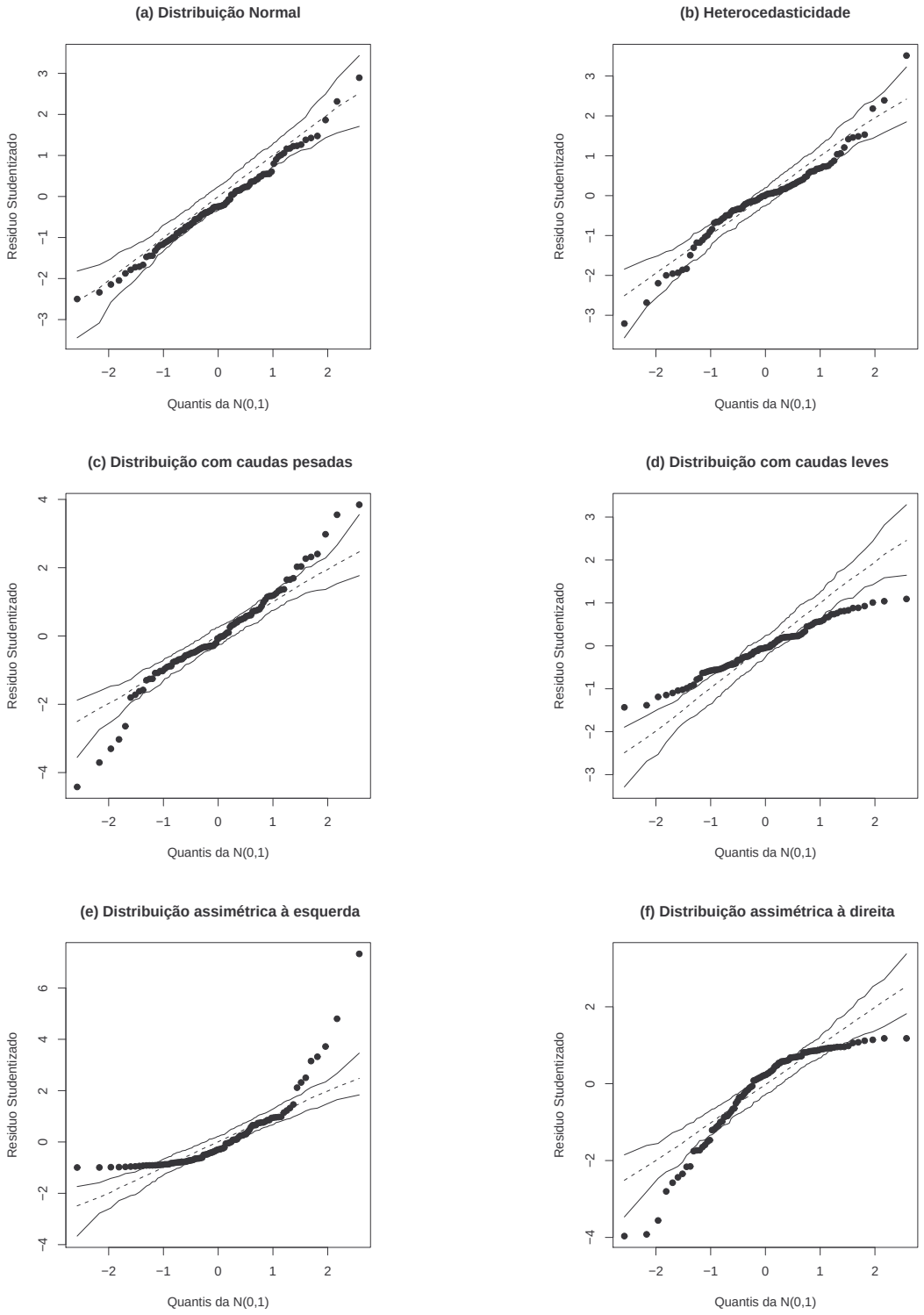
vii) Construir o limite inferior (superior) do envelope simulado ligando os pontos do conjunto de menores (maiores) resíduos simulados.

Se o modelo estiver correto, espera-se que todos os pontos observados no gráfico QQ fiquem contidos no envelope simulado em $[m/(m+1)] \times 100\%$ das vezes. Quando $m = 19$, isso corresponde a 95% das vezes.

Para detalhes a respeito de sua construção, veja Atkinson (1985), Atkinson & Riani (2000) ou Paula (2004), por exemplo. O mesmo tipo de procedimento pode ser empregado em situações em que outras distribuições para os erros são adotadas no modelo. Paula (2004) apresenta vários algoritmos (em linguagem `S-Plus` e `R`) para construção de gráficos envelope simulados, baseados em várias distribuições (Gama, Binomial, Poisson etc.).

Na Figura C.5.3, apresentamos exemplos de gráficos QQ (e envelopes simulados) com diferentes padrões de afastamento das suposições de normalidade, homocedasticidade ou simetria. Por exemplo, no gráfico C.5.3(a), não há indícios contrários às hipóteses de normalidade e homocedasticidade; no gráfico C.5.3(b), os resíduos estão próximos dos limites das bandas de confiança, evidenciando que a distribuição dos erros deve ser heterocedástica. O gráfico C.5.3(c) [C.5.3(d)] apresenta características típicas de casos em que a distribuição dos erros padronizados tem caudas mais pesadas (mais leves) que a distribuição normal padrão. Quando a distribuição dos erros padronizados possui caudas mais pesadas (leves) que a distribuição normal padrão, o gráfico QQ assume a forma de **S**, com os quantis extremos do resíduo studentizado maiores (menores) que os quantis teóricos da distribuição normal padrão. Por fim, o gráfico C.5.3(e) [C.5.3(f)] mostra uma situação em que a distribuição dos erros é assimétrica à direita (à esquerda). Quando se tem uma distribuição assimétrica à direita, o gráfico apresenta a forma de um **J**; se a distribuição for assimétrica à esquerda, o gráfico apresenta a forma de um **J** invertido.

Figura C.5.3: Envelopes simulados com coeficiente de confiança 95%



C.5.3 Análise de sensibilidade

A análise de sensibilidade visa avaliar o comportamento do ajuste de um modelo sujeito a algum tipo de perturbação, ou seja, sob alguma mudança nas hipóteses ou nos dados. Como cada observação não tem a mesma influência em todas as características do ajuste do modelo, é natural que se defina aquela na qual se quer focar a análise. Se o objetivo for fazer previsões, então é razoável medir a influência das observações nos valores preditos e não nos parâmetros de localização, como mencionam Chatterjee & Hadi (1986) e Chatterjee & Hadi (1988).

Existem medidas de influência baseadas nos resíduos, na curva de influência[10], na verossimilhança, no volume dos elipsoides de confiança, em um subconjunto do vetor de parâmetros de localização (influência parcial) e nos pontos remotos do espaço vetorial gerado pelas colunas da matriz de especificação $\mathbf{X}$.

Dentre as abordagens mais utilizadas na prática para medir influência em modelos lineares, destacam-se aquelas baseadas em **influência local** considerada em Cook (1986) e aquelas obtidas por intermédio da eliminação de observações (**influência global**).

O **poder de alavanca** (*leverage*) da i-ésima observação (y_i) é a derivada parcial $\partial\widehat{y}_i/\partial y_i$ e indica a taxa de variação do i-ésimo valor predito quando a i-ésima observação é acrescida de um infinitésimo. No modelo linear clássico, o vetor de valores preditos (ajustados) é dado por $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ e consequentemente,

$$\partial\widehat{\mathbf{y}}/\partial\mathbf{y} = \mathbf{H},$$

indicando que o poder de alavanca da i-ésima observação é dado por h_{ii}, ou seja, pelo i-ésimo elemento da diagonal principal da matriz de projeção $\mathbf{H}$. Como essa matriz é simétrica e idempotente, é possível mostrar que[11]

$$\begin{aligned} 0 &\leq h_{ii} \leq 1, \;\; i = 1, ..., n, \\ h_{ij} &\leq h_{ii}(1 - h_{ii}), \;\; 1 \leq i < j \leq n. \end{aligned}$$

Desta forma, se $h_{ii} = 1$, então $\widehat{y}_i = y_i$, implicando que a i-ésima observação tem influência total no seu valor predito. Além disso, como $\text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p$, então o valor médio do poder de alavanca é p/n. Quando todos os n elementos da diagonal principal de $\mathbf{H}$ são próximos de p/n, nenhuma observação influencia o seu valor predito de forma desproporcional; então podemos dizer que a i-ésima

[10] A curva de influência de uma estatística $T(x_1, \ldots, x_n)$ corresponde a um gráfico com os valores de uma nova observação amostral x_{n+1} no eixo das abscissas e com os valores da estatística $T(x_1, \ldots, x_n, x_{n+1})$ no eixo das ordenadas; ela dá uma ideia de como a estatística é influenciada pelo acréscimo de uma única observação amostral.

[11] Também é possível mostrar que $n^{-1} \leq h_{ii} \leq 1$ se o modelo incluir intercepto.

observação tem alto poder de alavanca (*high leverage*) se $h_{ii} \geq 2p/n$ ou $h_{ii} \geq 3p/n$, por exemplo. Esse critério é arbitrário e deve ser encarado com espírito puramente descritivo. Em função de (C.5.2), podemos concluir que, em geral, observações com alto poder de alavanca apresentam resíduos pequenos relativamente aos demais.

Para exemplificar, consideremos o modelo de regressão $y_i = \alpha + \beta x_i + e_i,\ \ i = 1, ..., n$, com $\mathbb{E}[e_i] = 0$, $\mathbb{V}[e_i] = \sigma^2$ e $\mathbb{Cov}(e_i, e_j) = 0$, $i \neq j$. Nesse caso, o poder de alavanca da i-ésima observação é dado por

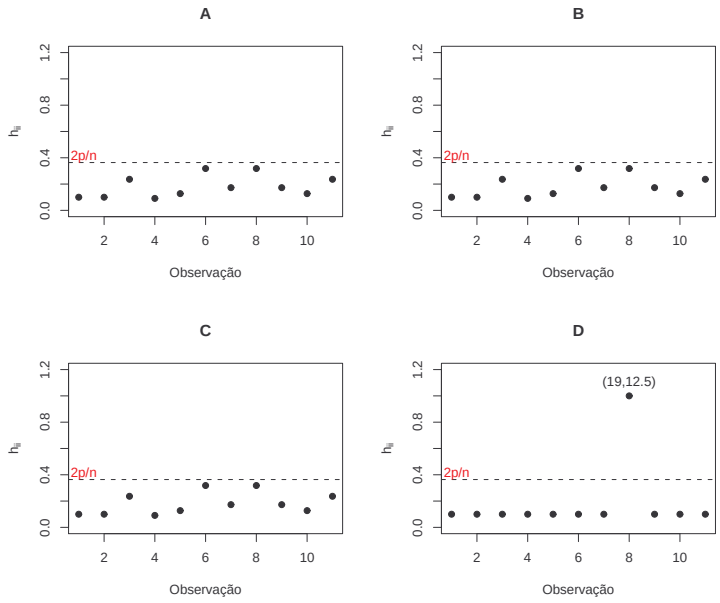
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

e consequentemente, a i-ésima observação tem alto poder de alavanca se x_i for um valor discrepante no conjunto $\{x_1, ..., x_n\}$.

De uma forma geral, é possível mostrar que observações com alto poder de alavanca são aquelas associadas a covariáveis discrepantes em $\mathcal{C}(\mathbf{X})$ como observado por Chatterjee & Hadi (1988) e Wei, Hu & Fung (1998). Mais detalhes, podem ser obtidos em Cook & Weisberg (1982), Chatterjee & Hadi (1988), Wei et al. (1998), Atkinson & Riani (2000) e Paula (2004), por exemplo.

A título de ilustração, na Figura C.5.4, mostramos os gráficos de h_{ii} *versus* índice das observações para os quatro subconjuntos de dados apresentados na Tabela C.5.1. Nos subconjuntos A, B e C não existe indício de observações com alto poder de alavanca. Por outro lado, no subconjunto D, a observação 8 apresenta $h_{88} = 1$ indicando que o coeficiente angular da regressão é significativo apenas pela sua presença. Como os valores da variável explicativa são idênticos nos conjuntos A, B e C, as matrizes $\mathbf{H}$ correspondentes também são as mesmas e consequentemente, os poderes de alavanca associados são iguais.

Figura C.5.4: Poder de alavanca *versus* índice das observações



Consideremos agora um modelo de regressão linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ com $\mathbb{E}(\mathbf{e}) = \mathbf{0}$ e $\mathbb{V}(\mathbf{e}) = \sigma^2\mathbf{I}$ e suponhamos que há interesse em incluir uma nova variável explicativa, digamos w, de forma que o modelo passa a ser

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{w} + \mathbf{e}^* = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{e}^* \tag{C.5.4}$$

com γ representando o coeficiente da nova variável, $\mathbf{w} = (w_1, ..., w_n)^\top$, $\mathbf{X}^* = (\mathbf{X}, \mathbf{w})$, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^\top, \gamma)^\top$ e $\mathbf{e}^*$ representando um conjunto de erros aleatórios homocedásticos e não correlacionados. Sob normalidade dos erros, podemos verificar a plausibilidade da inclusão da nova covariável, por exemplo, por meio de um teste t para $\gamma = 0$. Uma forma alternativa para avaliar a importância da inclusão da covariável é descrita a seguir.

Inicialmente, observemos que as equações de estimação (equações normais) para a obtenção do estimador de mínimos quadrados ordinários de $\boldsymbol{\beta}^*$ sob o modelo (C.5.4) são

$$\begin{aligned} \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{X}^\top\mathbf{w}\widehat{\gamma} &= \mathbf{X}^\top\mathbf{y} \\ \mathbf{w}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{w}^\top\mathbf{w}\widehat{\gamma} &= \mathbf{w}^\top\mathbf{y}. \end{aligned}$$

Resolvendo esse sistema, obtemos o estimador de mínimos quadrados

$$\widehat{\gamma} = \frac{\mathbf{w}^\top(\mathbf{I} - \mathbf{H})\mathbf{y}}{\mathbf{w}^\top(\mathbf{I} - \mathbf{H})\mathbf{w}}.$$

Lembrando que $(\mathbf{I}-\mathbf{H})$ é uma matriz idempotente, podemos reexpressar esse estimador como

$$\widehat{\gamma} = \frac{\mathbf{w}^\top(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{y}}{\mathbf{w}^\top(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{w}} = \frac{\widehat{\mathbf{e}}_w^\top\widehat{\mathbf{e}}}{\widehat{\mathbf{e}}_w^\top\widehat{\mathbf{e}}_w},$$

com $\widehat{\mathbf{e}} = (\mathbf{I}-\mathbf{H})\mathbf{y}$ e $\widehat{\mathbf{e}}_w = (\mathbf{I}-\mathbf{H})\mathbf{w}$ de forma que ele pode ser interpretado como o estimador de mínimos quadrados (ordinários) do coeficiente angular de uma regressão (sem intercepto) tendo como variável resposta os resíduos ordinários ($\widehat{\mathbf{e}}$) obtidos do modelo inicial (sem a covariável w) e como variável explicativa os resíduos ($\widehat{\mathbf{e}}_w$) obtidos do modelo de regressão $\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_w$.

O gráfico de dispersão entre os resíduos $\widehat{\mathbf{e}}$ e $\widehat{\mathbf{e}}_w$, conhecido como **gráfico da variável adicionada** ou **gráfico de regressão parcial** (*partial regression plot*) fornece informação sobre os ganhos com a inclusão da covariável w no modelo. Ele também pode ser útil para identificar pontos que se desviam da relação linear entre os resíduos, e que podem ser encarados como observações influentes na estimação de γ. Mais detalhes e extensões, que fogem ao escopo deste texto, podem ser encontrados em Cook & Weisberg (1989) e Cook (1996), por exemplo.

Em modelos de regressão linear múltipla, uma estratégia para identificar observações influentes na estimação dos coeficientes do vetor de parâmetros, consiste em construir um gráfico desse tipo para cada covariável do modelo.

Cook (1977), por outro lado, sugere que a influência de uma particular observação, ou de um conjunto de observações, seja avaliada por intermédio dos efeitos provocados por sua eliminação do conjunto de dados.

Consideremos o modelo de regressão linear (C.5.1) e denotemos por $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{(I)}$, respectivamente, os estimadores de mínimos quadrados de $\boldsymbol{\beta}$ obtidos com todos os dados da amostra e com a eliminação do conjunto de observações I. Nesse contexto, uma das medidas mais utilizadas é a **distância de Cook** definida por

$$D_I = \frac{\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}\right]^\top (\mathbf{X}^\top\mathbf{X}) \left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}\right]}{ps^2} = \frac{\left[\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(I)}\right]^\top \left[\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(I)}\right]}{ps^2}, \qquad \text{(C.5.5)}$$

com $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ e $\widehat{\mathbf{y}}_{(I)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(I)}$ representando respectivamente, o vetor de valores preditos pelo ajuste do modelo com todos os dados da amostra e aquele obtido com a eliminação do conjunto de observações I. A estatística D_I mede a influência das observações do conjunto I na estimativa de $\boldsymbol{\beta}$ segundo a métrica definida por $(ps^2)^{-1}\mathbf{X}^\top\mathbf{X}$ ou equivalentemente, a influência dessas observações no vetor de valores preditos. Valores grandes de D_I indicam que as observações do conjunto I são influentes na estimação de $\boldsymbol{\beta}$ ou nos valores de $\widehat{\mathbf{y}}$. Sob essa abordagem, é essencial obter expressões que relacionem o estimador do parâmetro de interesse calculado com base em toda amostra com o respectivo estimador calculado após a eliminação de um

conjunto de observações sem a necessidade de reajustar o modelo. Em particular, quando apenas a i-ésima observação é eliminada, é possível mostrar que

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

de tal forma que a distância de Cook correspondente é

$$D_i = \frac{\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right]^\top (\mathbf{X}^\top\mathbf{X}) \left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right]}{ps^2} = \frac{\widehat{e}_i^{*2}}{p}\frac{h_{ii}}{(1-h_{ii})^2}. \tag{C.5.6}$$

Dada a expressão (C.5.6), a influência da i-ésima observação depende tanto do respectivo resíduo studentizado quanto do grau de alavanca da observação em questão; se $h_{ii} \approx 0$ (indicando um baixo poder de alavanca) então D_i assume um valor "pequeno", mesmo quando a i-ésima observação for altamente discrepante, indicando que a distância (C.5.6) pode não ser adequada nessas situações.

Outras propostas são sugeridas na literatura para contornar este fato. Por exemplo, Belsley, Kuh & Welsch (1980) sugerem a utilização de

$$\text{DFFITS}_i = \frac{|\widehat{e}_i^*|}{s_{(i)}(1-h_{ii})^{1/2}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} = |t_i|\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}, \tag{C.5.7}$$

enquanto Atkinson (1981) sugere uma versão modificada da distância de Cook, obtida com a substituição de s^2 por $s^2_{(i)}$ e ajustada pelo tamanho da amostra, a saber,

$$C_i = \left(\frac{n-p}{p}\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}|t_i|. \tag{C.5.8}$$

Quando todos os h_{ii} são iguais, $C_i = |t_i|$. Chatterjee & Hadi (1988) sugerem outros usos para esta versão da distância de Cook. Em particular, esses autores comentam que $C_i^* = \text{sinal}(y_i - \widehat{y}_i)C_i$ $(i = 1, ...n)$ podem ser utilizados como resíduos.

Em tese, essas três medidas de influência competem entre si. Todavia, como destacam Cook, Peña & Weisberg (1988) e Paula (2004) elas servem para avaliar diferentes aspectos da influência das observações nos resultados do ajuste. Por exemplo, (C.5.6) é mais adequada para medir a influência das observações nos parâmetros de localização ($\boldsymbol{\beta}$), enquanto que (C.5.7) tem o objetivo de medir a influência das observações nos parâmetros de localização e escala simultaneamente embora possa falhar em algumas situações, como indicam Cook et al. (1988). Segundo Paula (2004), observações discrepantes com baixo poder de alavanca dificilmente influem na estimação de $\boldsymbol{\beta}$ e não comprometem o potencial uso de (C.5.5).

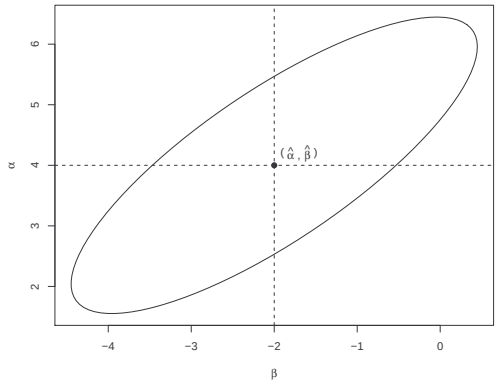
Gráficos das medidas de influência *versus* índices das observações são ferramentes úteis para a identificação observações influentes, *i.e.*, aquelas com valores da medida de influência "grandes" em relação aos demais.

Medidas de influência como a distância de Cook são baseadas na mudança do centro (definido em termos dos parâmetros de regressão) dos **elipsóides de confiança** para o vetor de parâmetros $\boldsymbol{\beta}$, nomeadamente

$$\left\{\boldsymbol{\beta} \in I\!R^p;\ (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq p s^2 F_{p,(n-p)}(\alpha)\right\} \tag{C.5.9}$$

em que $F_{p,(n-p)}(\alpha)$ denota o quantil de ordem $1-\alpha$ da distribuição F com p graus de liberdade no numerador e $n-p$ no denominador. A avaliação da influência de um conjunto de observações também pode ser concretizada a partir do volume desses elipsóides. Na Figura C.5.5, ilustramos um elipsóide de confiança com coeficiente de confiança de 95% para os parâmetros de uma regressão linear simples. O volume do elipsóide está diretamente relacionado com a estimativa da matriz de covariâncias de $\widehat{\boldsymbol{\beta}}$ e por conseguinte, a avaliação da influência das observações no respectivo volume reveste-se de particular importância, uma vez que a variância dos estimadores dos coeficientes podem ser extremamente afetadas por poucas observações. Aqui apresentamos duas propostas para tal fim; para detalhes, sugerimos uma consulta a Andrews & Pregibon (1978), Belsley et al. (1980), Cook & Weisberg (1982) ou Chatterjee & Hadi (1988), por exemplo.

Figura C.5.5: Elipsóide de confiança com coeficiente de confiança 95%



O volume do elipsóide (C.5.9) é inversamente proporcional à raiz quadrada de $\left|\mathbf{X}^\top \mathbf{X}\right|$. Por isso, Andrews & Pregibon (1978) sugeriram avaliar a influência da i-ésima observação por meio de

$$\mathrm{AP}_i = \frac{s^2_{(i)} \left|\mathbf{X}^\top_{(i)} \mathbf{X}_{(i)}\right|}{s^2 \left|\mathbf{X}^\top \mathbf{X}\right|} = h_{ii} + (1 - h_{ii}) \frac{\widehat{e}^{*2}_i}{n-p}. \tag{C.5.10}$$

Valores de AP_i muito distantes de 1 indicam que a observação em questão é potencialmente influente com relação à matriz de covariâncias de $\widehat{\boldsymbol{\beta}}$. Propriedades desta

medida podem ser encontrados em Cook & Weisberg (1982) ou Chatterjee & Hadi (1988), por exemplo.

Belsley et al. (1980) sugerem que se avalie a influência da i-ésima observação na matriz de covariâncias de $\widehat{\boldsymbol{\beta}}$ por meio de

$$\begin{aligned} \text{COVRATIO}_i &= \frac{\left|\widehat{\mathbb{V}}[\widehat{\boldsymbol{\beta}}_{(i)}]\right|}{\left|\widehat{\mathbb{V}}[\widehat{\boldsymbol{\beta}}]\right|} = \frac{\left|s^2_{(i)}(\mathbf{X}^\top_{(i)}\mathbf{X}_{(i)})^{-1}\right|}{|s^2(\mathbf{X}^\top\mathbf{X})^{-1}|} \\ &= \left(\frac{n-p-\widehat{e}^*_i}{n-p-1}\right)^p \frac{1}{1-h_{ii}}. \end{aligned} \tag{C.5.11}$$

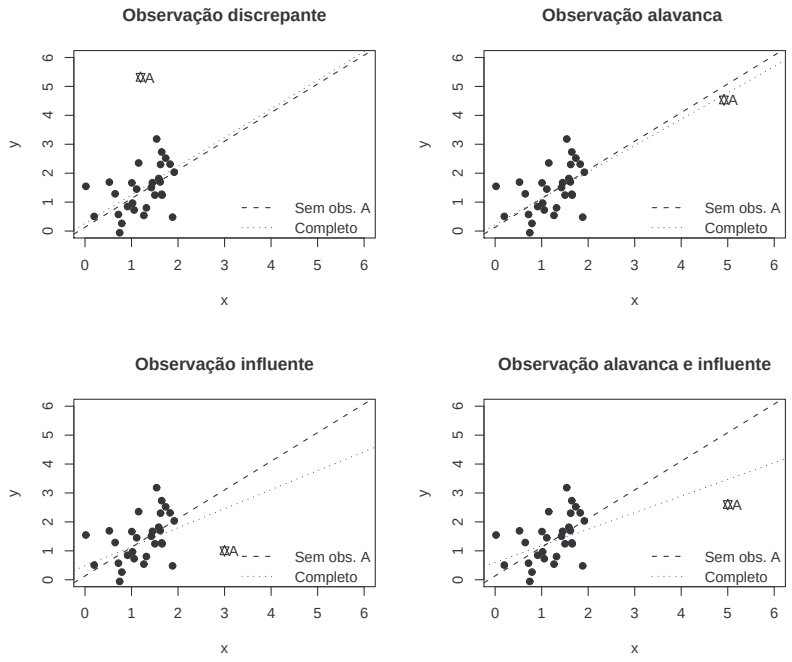
Se todas as observações tiverem o mesmo grau de influência, espera-se que $\text{COVRATIO}_i \approx 1$; afastamentos da unidade indicam que a a observação correspondente é potencialmente influente com relação à matriz de covariâncias de $\widehat{\boldsymbol{\beta}}$. Belsley et al. (1980) sugerem utilizar o seguinte limiar para identificar observações influentes

$$|\text{COVRATIO}_i - 1| \geq 3p/n.$$

Esse limiar, assim como aqueles mencionados anteriormente, tem cunho totalmente descritivo.

Na Figura C.5.6 mostramos gráficos que ilustram a diferença entre observações discrepantes, alavanca e influentes.

Figura C.5.6: Diferença entre observações alavanca, discrepantes e influentes



Todos os procedimentos aqui discutidos são baseados na eliminação de conjuntos de observações e são conhecidos na literatura como métodos de **influência global**. Todavia existem técnicas de diagnóstico baseados na discrepância da função de verossimilhança quando perturbamos as observações de alguma forma. Tais técnicas, conhecidas sob a denominação geral de **influência local**, foram propostas por Cook (1986). Para detalhes, veja, por exemplo, Cook (1986) ou Paula (2004).

C.5.4 Análise da suposição de correlação nula

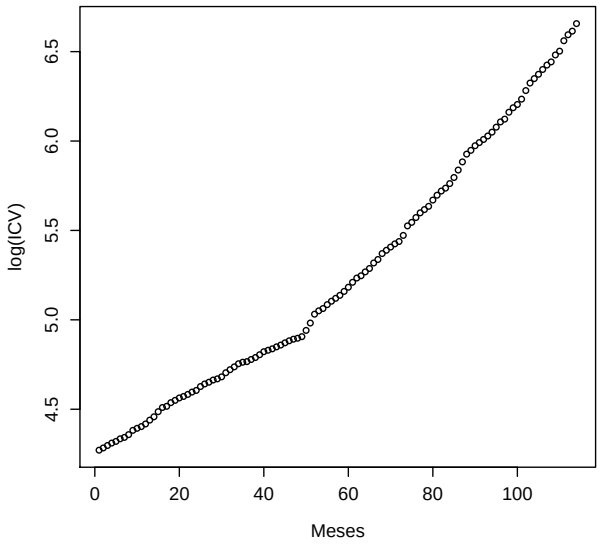
Em geral, a suposição de que os erros do modelo linear são não correlacionados deve ser questionada com base no procedimento de coleta de dados. Como ilustração, consideramos dois exemplos nos quais essa característica justifica a dúvida. O primeiro exemplo é um caso simples dos problemas abordados pelas técnicas de análise de séries cronológicas; o segundo exemplo é o caso típico daqueles que constituem o objeto do núcleo deste texto. Ambos são apresentados aqui com a finalidade de mostrar como as técnicas de análise de regressão podem ser empregadas para analisar modelos mais gerais do que aqueles governados pelo paradigma de Gauss-Markov.

Exemplo C.5.1: Na Tabela C.5.2 apresentamos valores do índice de custo de vida (ICV) na cidade de São Paulo colhidos pela Fundação Getúlio Vargas entre janeiro de 1970 e junho de 1979 com o objetivo de avaliar seu crescimento nesse período. O gráfico de dispersão correspondente está disposto na Figura C.5.7 Com base em argumentos de teoria econômica, é razoável supor que o ICV num determinado mês seja correlacionado com aqueles obtidos em meses anteriores.

Tabela C.5.2: Índice de custo de vida para São Paulo (jan/70 a jul/79)

Obs	ICV	Obs	ICV	Obs	ICV	Obs	ICV	Obs	ICV
1	71.6	24	100	47	133	70	223	93	415
2	72.5	25	102	48	134	71	227	94	424
3	73.5	26	104	49	135	72	230	95	436
4	74.5	27	105	50	140	73	238	96	449
5	75.2	28	106	51	146	74	251	97	456
6	76.3	29	107	52	153	75	256	98	474
7	76.9	30	108	53	156	76	263	99	486
8	78.1	31	110	54	158	77	270	100	495
9	80	32	112	55	162	78	275	101	510
10	80.9	33	114	56	165	79	280	102	535
11	81.7	34	116	57	167	80	290	103	558
12	82.9	35	117	58	170	81	298	104	572
13	84.7	36	118	59	174	82	305	105	586
14	86.3	37	119	60	178	83	310	106	602
15	88.8	38	120	61	183	84	318	107	617
16	90.9	39	122	62	188	85	329	108	628
17	91.5	40	124	63	190	86	343	109	653
18	93.4	41	125	64	194	87	359	110	667
19	94.6	42	126	65	198	88	375	111	707
20	95.9	43	128	66	204	89	383	112	731
21	96.7	44	129	67	208	90	393	113	746
22	97.8	45	131	68	215	91	400	114	778
23	99.1	46	132	69	219	92	407		

Figura C.5.7: Gráfico de dispersão para os dados do Exemplo C.5.1



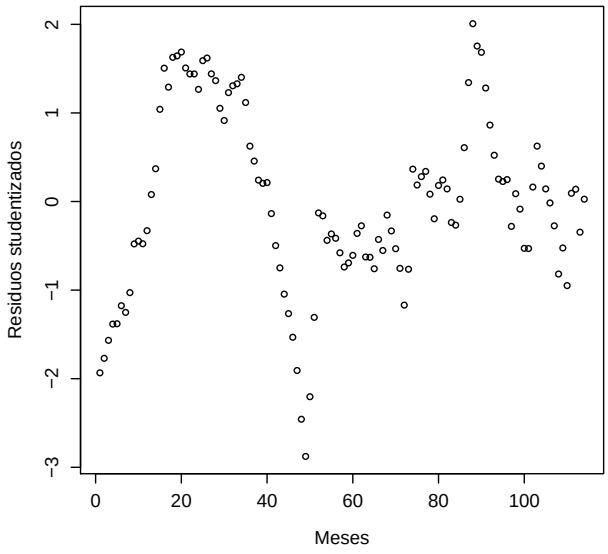
Tendo em vista o gráfico de dispersão apresentado na Figura C.5.7, uma primeira abordagem para a análise dos dados do Exemplo C.5.1 poderia envolver um modelo da forma

$$\log(y_t) = \alpha + \beta t + \gamma t^2 + e_t, \tag{C.5.12}$$

$t = 1, \ldots, n$ em que y_t representa o ICV no instante t, α denota o valor esperado do ICV no tempo $t = 0$, β e γ representam os componentes linear e quadrático da curva que rege a variação temporal do logaritmo do ICV e e_t denota um erro aleatório. Utilizamos t como índice para salientar que as observações são colhidas sequencialmente.

O coeficiente de determinação $R^2 = 0.9986$ indica que o ajuste (por mínimos quadrados) do modelo com $\widehat{\alpha} = 4.310$ (EP $= 0.007$), $\widehat{\beta} = 0.008$ (EP< 0.001) e $\widehat{\gamma} = 0.0001$ (EP < 0.00001) é excelente (sob essa ótica, obviamente). Por outro lado, o gráfico de resíduos apresentado na Figura C.5.8 mostra sequências de resíduos positivos seguidas de sequências de resíduos negativos, sugerindo uma possível correlação positiva entre eles (autocorrelação).

Figura C.5.8: Resíduos studentizados obtidos do ajuste do modelo (C.5.12)



Uma maneira de contornar esse problema, é modificar os componentes aleatórios do modelo para incorporar essa possível autocorrelação nos erros. Nesse contexto, podemos considerar o modelo (C.5.12) com

$$e_t = \rho e_{t-1} + u_t, \quad t = 1, \ldots, n \tag{C.5.13}$$

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

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

$$D = \sum_{t=2}^{n} (\widehat{e}_t - \widehat{e}_{t-1})^2 / \sum_{t=1}^{n} \widehat{e}_t^2, \tag{C.5.14}$$

em que $\widehat{e}_t$, $t = 1, \ldots, n$ são os resíduos obtidos do ajuste do modelo (C.5.12) por mínimos quadrados. Expandindo (C.5.14) obtemos

$$\begin{aligned} D &= \frac{\sum_{t=2}^{n} \widehat{e}_t^2}{\sum_{t=1}^{n} \widehat{e}_t^2} + \frac{\sum_{t=2}^{n} \widehat{e}_{t-1}^2}{\sum_{t=1}^{n} \widehat{e}_t^2} - 2\frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=1}^{n} \widehat{e}_t^2} \\ &\approx 2 - 2\frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=1}^{n} \widehat{e}_t^2}, \end{aligned} \tag{C.5.15}$$

Se os resíduos não forem correlacionados, então $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx 0$ e consequentemente, $D \approx 2$; se, por outro lado, os resíduos forem altamente correlacionados, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx \sum_{t=2}^{n} \widehat{e}_t^2$ e então $D \approx 0$; finalmente, se os resíduos tiverem uma grande correlação negativa, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx -\sum_{t=2}^{n} \widehat{e}_t^2$ e nesse caso, $D \approx 4$. Durbin & Watson (1950), Durbin & Watson (1951) e Durbin & Watson (1971) produziram tabelas da distribuição da estatística D que podem ser utilizados para avaliar a suposição de que os erros são não correlacionados.

Se a análise indicar que os erros são correlacionados, o modelo (C.5.12) - (C.5.13) poderá ser ajustado pelo método de mínimos quadrados generalizados. Com esse intuito, notemos primeiramente que (C.5.13) sugere o seguinte modelo de regressão linear simples sem intercepto

$$\widehat{e}_t = \rho \widehat{e}_{t-1} + u_t$$

de onde podemos obter estimadores de σ^2 e ρ; mais especificamente,

$$\widehat{\rho} = \sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} / \sum_{t=1}^{n} \widehat{e}_t^2$$

e

$$\widehat{\sigma}^2 = (n-1)^{-1} \sum_{t=1}^{n} (\widehat{e}_t - \widehat{\rho}\widehat{e}_{t-1})^2.$$

Expressando o modelo (C.5.12) - (C.5.13) na forma matricial, o vetor de parâmetros

é $\boldsymbol{\beta} = (\alpha, \beta, \gamma)^\top$ e a matriz de covariâncias é

$$\mathbf{V} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}.$$

Substituindo os elementos de $\mathbf{V}$ por seus estimadores, podemos utilizar (C.2.6) para estimar $\boldsymbol{\beta}$.

O valor da estatística de Durbin-Watson para os dados do Exemplo C.5.1 sob o modelo (C.5.12) é $D = 0.1259$ ($p < 0.0001$), sugerindo um alto grau de autocorrelação dos resíduos.

Exemplo C.5.2: Na Tabela C.5.3 apresentamos dados provenientes de um estudo em que o objetivo é avaliar a variação do peso de bezerros submetidos a diferentes dietas (tipos de pasto) entre 12 e 26 semanas após o nascimento. Como animais mais pesados (ou mais leves) no início do estudo tendem a permanecer mais pesados (mais leves) ao longo do tempo (pelo menos, ao longo das primeiras observações) é razoável supor que o peso de cada animal numa determinada semana seja correlacionado com seu peso na semana anterior.

Tabela C.5.3: Peso de bezerros (kg)

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

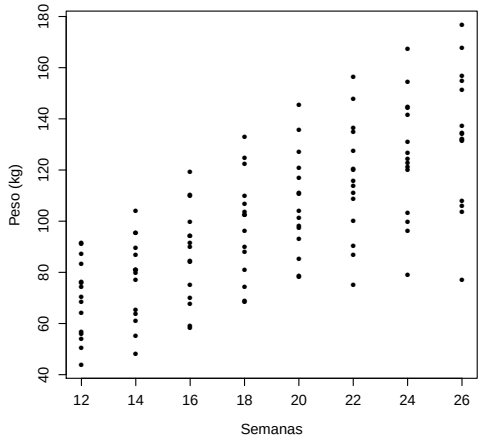
O **gráfico do desenhista** (*draftman's plot*) para os dados do Exemplo C.5.2 disposto na Figura C.5.9 sugere uma variação linear do peso médio dos animais ao longo do período estudado, para a qual podem-se usar modelos do tipo

$$y_{ij} = \alpha + \beta x_{ij} + e_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m_i, \tag{C.5.16}$$

em que y_{ij} denota a j-ésima medida do peso do i-ésimo animal, x_{ij} indica o número de semanas pós-nascimento em que foi realizada essa medida, α representa o peso esperado dos animais ao nascer (admitindo que o modelo linear possa ser extrapolado para o período entre o nascimento e a décima segunda semana), β representa a variação esperada do peso dos animais por semana e e_{ij} corresponde a um erro aleatório com média nula e variância constante σ_e^2. Esse modelo, particularizado para o Exemplo C.5.2, em todos os animais foram observados nos mesmos instantes, é tal que $n = 15$, $m_i = m = 8$ e $x_{ij} = x_j$.

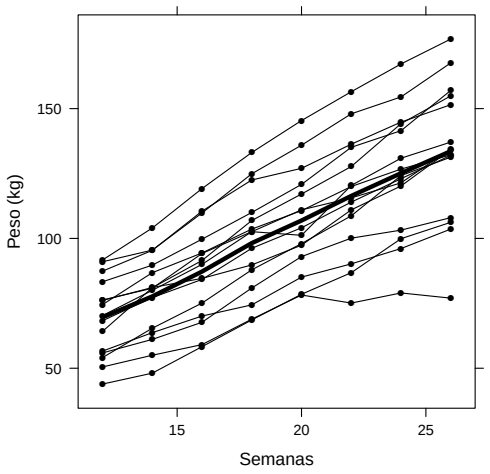
Sob a suposição de que os erros e_{ij} são não correlacionados, estimativas (e erros padrões) dos parâmetros de localização do modelo (C.5.16) obtidas por mínimos quadrados são $\widehat{\alpha} = 13.5$ (EP= 7.8), $\widehat{\beta} = 4.7$ (EP= 0.4); a estimativa do desvio padrão é $\widehat{\sigma}_e = 20.1$ e o coeficiente de determinação ajustado é $R_a^2 = 0.53$. Uma análise de resíduos não indica violações das suposições de heterocedasticidade e normalidade dos erros.

Figura C.5.9: Gráfico de dispersão para os dados do Exemplo C.5.2



Tendo em vista que várias observações são realizadas em cada animal, podemos construir um gráfico de perfis como aquele apresentado na Figura C.5.10. Esse tipo de gráfico serve como ferramenta adicional para avaliação das suposições adotadas.

Figura C.5.10: Gráfico de perfis para os dados do Exemplo C.5.2



Em primeiro lugar, observa-se que não há razões para duvidar da hipótese de homocedasticidade, pois a variabilidade das observações é similar nos oito instantes de observação. Também não há evidências nem de unidades amostrais (animais) com perfis discrepantes nem de observações discrepantes. Além disso, a proposta de uma reta para explicar a variação do peso médio ao longo do período de observação não é contrariada, dada a forma do perfil médio. No entanto, observa-se um certo "paralelismo" dos perfis individuais, sugerindo uma correlação positiva entre as observações intraunidades amostrais; os animais mais pesados (leves) tendem a manter essa característica ao longo das quatorze semanas em que se desenvolveu o estudo.

Para incorporar essa informação no modelo a ser adotado, uma alternativa é substituir o termo aleatório do modelo (C.5.16), fazendo

$$e_{ij} = a_i + d_{ij} \tag{C.5.17}$$

com a_i e d_{ij} denotando variáveis aleatórias não correlacionadas com médias nulas e variâncias σ_a^2 e σ^2, respectivamente. As suposições adotadas implicam

a) $\mathbb{V}(y_{ij}) = \sigma_a^2 + \sigma^2$

b) $\mathbb{C}\text{ov}(y_{ij}, y_{kl}) = \sigma_a^2$ se $i = k$ e $j \neq l$

c) $\mathbb{C}\text{ov}(y_{ij}, y_{kl}) = 0$ se $i \neq k$

de forma que observações realizadas na mesma unidade amostral são correlacionadas e aquelas realizadas em unidades amostrais diferentes, não. Embora a suposição de

homocedasticidade possa ser questionável as covariâncias (correlações) podem ser consideradas constantes, como sugerem tanto com a matriz de covariâncias (correlações) amostral disposta na Tabela C.5.4 quanto com a matriz de gráficos de dispersão (*draftman's plot*) apresentada na Figura C.5.11. Isso é uma indicação de que o modelo proposto é uma alternativa razoável para representar os dados.

Figura C.5.11: Gráfico do desenhista para o Exemplo C.5.2

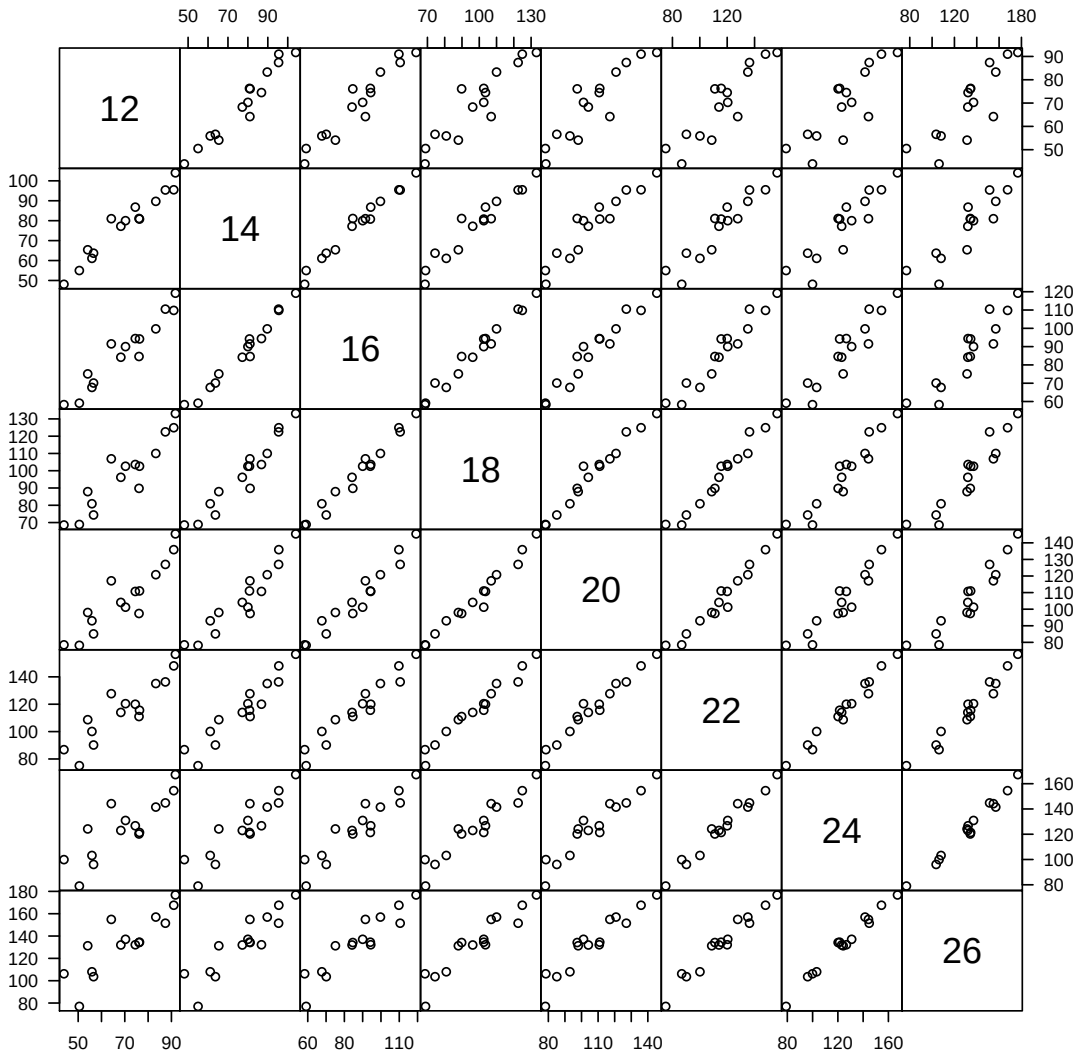


Tabela C.5.4: Matriz de covariâncias (correlações) para os dados do Exemplo C.5.2

Semanas	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**
12	229.7	0.97	0.96	0.93	0.91	0.90	0.84	0.85
14	236.3	257.2	0.99	0.96	0.94	0.95	0.91	0.90
16	270.1	293.7	344.3	0.99	0.97	0.97	0.93	0.93
18	280.2	309.1	366.4	398.9	0.98	0.98	0.96	0.94
20	274.2	301.8	359.0	392.2	397.5	0.98	0.95	0.94
22	309.1	342.8	405.1	442.9	441.0	509.8	0.99	0.98
24	300.6	343.8	409.7	452.2	449.0	526.3	558.5	0.99
26	341.1	384.2	456.0	498.0	497.3	587.1	622.1	705.5

Essencialmente, o modelo (C.5.16)-(C.5.17) é um modelo misto com um efeito aleatório (a_i) e sugere que o peso esperado do i-ésimo animal varia linearmente com o tempo segundo o modelo condicional

$$y_{ij}|a_i = (\alpha + a_i) + \beta x_{ij} + d_{ij} = \alpha_i + \beta x_{ij} + d_{ij} \tag{C.5.18}$$

i.e., com a mesma taxa de crescimento β, porém com pesos esperados ao nascer, $\alpha_i = \alpha + a_i$, diferentes.

Embora o ajuste de modelos mistos do tipo (C.5.16)-(C.5.17) possa ser facilmente realizado utilizando a metodologia descrita no Capítulo 2, no caso particular em que as unidades amostrais são observadas nos mesmos instantes, *i.e.*, em que $x_{ij} = x_j$ e $m_i = m$, é possível utilizar um enfoque ingênuo baseado na técnica de mínimos quadrados generalizados para concretizá-lo. Nesse contexto, o núcleo do problema é a estimação dos componentes de variância σ_a^2 e σ^2. Com essa finalidade, convém escrever o modelo (C.5.18) na forma matricial

$$\mathbf{y} = \mathbf{W}\boldsymbol{\gamma} + \mathbf{e},$$

em que $\mathbf{W} = [\mathbf{I}_n \otimes \mathbf{1}_m,\ \mathbf{1}_n \otimes \mathbf{x}]$, com dimensão $nm \times (n+1)$, é a matriz de variáveis explicativas do modelo condicional, $\mathbf{x} = [x_1, \ldots, x_m]^\top$, $\mathbf{1}_a$ é um vetor de ordem a com todos os elementos iguais a 1, $\mathbf{y} = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top]^\top$ com $\mathbf{y}_i = [y_{i1}, \ldots, y_{im}]^\top$ e $\boldsymbol{\gamma} = [\alpha_1, \ldots, \alpha_n, \beta]^\top$.

Um estimador consistente de σ^2 é

$$\widehat{\sigma}^2 = [nm - (n+1)]^{-1}\mathbf{y}^\top[\mathbf{I}_{nm} - \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top]\mathbf{y}. \tag{C.5.19}$$

Consideremos agora o modelo

$$\overline{y}_i = \alpha + a_i + \beta\overline{x} + \overline{d}_i \tag{C.5.20}$$

em que, $\overline{y}_i$ e $\overline{x}$ representam respectivamente a média dos m valores da variável resposta e da covariável para a i-ésima unidade amostral e $\overline{d}_i = m^{-1}\sum_{j=1}^{m} d_{ij}$, $i = 1, \ldots, n$. Esse modelo pode ser expresso como

$$\overline{y}_i = \mu + d_i^*$$

em que $\mu = \alpha + \beta\overline{x}$ e $d_i^* = a_i + \overline{d}_i$, $i = 1, \ldots, n$ são erros aleatórios independentes com média nula e variância $\tau^2 = \sigma_a^2 + \sigma^2/m$. Sob esse modelo, um estimador consistente de τ^2 é

$$\widehat{\tau}^2 = (n-1)^{-1}\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2 \tag{C.5.21}$$

com $\overline{y} = n^{-1}\sum_{i=1}^{n}\overline{y}_i$. Consequentemente, um estimador de σ_a^2 pode ser obtido de (C.5.19) e (C.5.21) por meio de

$$\widehat{\sigma_a}^2 = \widehat{\tau}^2 - \widehat{\sigma}^2/m.$$

Na forma matricial, o modelo (C.5.16)-(C.5.17) pode ser escrito como

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

em que $\mathbf{X} = [\mathbf{1}_n \otimes \mathbf{1}_m,\ \mathbf{1}_n \otimes \mathbf{x}]$, $\boldsymbol{\beta} = (\alpha, \beta)^\top$ e $\mathbf{e}$ é um vetor aleatório com média nula e matriz de covariâncias $\mathbb{V}(\mathbf{e}) = \mathbf{I}_n \otimes \mathbf{R}$ com $\mathbf{R} = \sigma^2\mathbf{I}_m + \sigma_a^2\mathbf{1}_m\mathbf{1}_m^\top$. Um estimador consistente, $\widehat{\mathbf{R}}$, da matriz de covariâncias intraunidades amostrais $\mathbf{R}$ pode ser obtido com a substituição dos parâmetros σ_a^2 e σ^2 pelos estimadores $\widehat{\sigma_a}^2$ e $\widehat{\sigma}^2$. O vetor de parâmetros de localização $\boldsymbol{\beta}$ pode ser facilmente estimado por meio de (C.2.6) tendo em vista que, segundo a propriedade *vii*) da Seção A.1.6,

$$\mathbf{R}^{-1} = \frac{1}{\sigma^2}[\mathbf{I}_m - \frac{\sigma_a^2}{m\sigma_a^2 + \sigma^2}\mathbf{1}_m\mathbf{1}_m^\top].$$

Mais especificamente, pode-se mostrar que, neste caso,

$$\widehat{\boldsymbol{\beta}} = [\sum_{i=1}^{n}\mathbf{X}_i\widehat{\mathbf{R}}^{-1}\mathbf{X}_i^\top]^{-1}\sum_{i=1}^{n}\mathbf{X}_i\widehat{\mathbf{R}}^{-1}\mathbf{y}_i$$

em que $\mathbf{X}_i = [\mathbf{1}_m \ \ \mathbf{x}]$. Sob as condições de regularidade detalhadas na Seção C.2, concluímos que a distribuição aproximada de $\widehat{\boldsymbol{\beta}}$ é dada por (C.2.7).

Para os dados do Exemplo C.5.2, obtemos $\widehat{\sigma}^2 = 34.6$, $\widehat{\tau}^2 = 393.5$ e $\widehat{\sigma_a}^2 = 389.2$, $\widehat{\alpha} = 13.5$ (EP= 5.6) e $\widehat{\beta} = 4.7$ (EP= 0.1), resultados que coincidem com aqueles gerados pela metodologia de máxima verossimilhança (sob normalidade) descrita no Capítulo 2. Observamos que, em conformidade com os perfis apresentados na Figura C.5.10, a variância interunidades amostrais é cerca de 10 vezes a variância intraunidades amostrais. Além disso, notamos que, embora as estimativas dos coeficientes de regressão α e β sejam iguais àquelas obtidas por mínimos quadrados ordinários, os erros padrões correspondentes são menores, evidenciando o ganho de precisão proporcionado pelo ajuste de um modelo mais adequado.

C.5.5 Multicolinearidade

A maneira mais fácil para avaliar a existência de colinearidade entre duas variáveis explicativas é a construção de gráficos de dispersão e o exame dos coeficientes de correlação linear entre elas. Multicolinearidade (entre várias variáveis explicativas) é mais difícil de ser detectada, mas pode ser avaliada por meio da observação dos seguintes fenômenos:

i) a inclusão de alguma variável altera consideravelmente as estimativas dos coeficientes de regressão;

ii) os erros padrões das estimativas dos coeficientes de regressão são grandes;

iii) as estimativas dos coeficientes de regressão não são significativas mesmo quando se espera uma associação linear entre a variável resposta e as variáveis preditoras;

iv) as estimativas dos coeficientes de regressão têm sinais diferentes daqueles esperados.

Consideremos o modelo de regressão linear múltipla

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i,$$

$i = 1, \ldots, n$ com erros e_i não correlacionados e $\mathbb{E}(e_i) = 0$ e $\mathbb{V}(e_i) = \sigma^2$. Se as variáveis X_i forem não correlacionadas, teremos

$$\mathbb{V}(\widehat{\beta}_i) = \sigma^2/nS^2_{X_i}$$

em que $S^2_{X_i}$ é a variância da variável X_i (lembremos que nesse caso $\mathbf{X}^\top\mathbf{X}$ é uma matriz diagonal). No caso geral, *i.e.*, em que as variáveis explicativas podem ser correlacionadas, teremos

$$\mathbb{V}(\widehat{\beta}_i) = \sigma^2[\mathbf{X}^\top\mathbf{X}]^{-1}_{i+1,i+1}.$$

O quociente entre as duas expressões da variância, nomeadamente,

$$FIV_i = nS^2_{X_i}[\mathbf{X}^\top\mathbf{X}]^{-1}_{i+1,i+1}$$

é o conhecido como **fator de inflação da variância** (*variance inflation factor*) para o i-ésimo coeficiente. Valores grandes de FIV_i sugerem alto grau de multicolinearidade. Pode-se mostrar que

$$FIV_i = 1/(1 - R^2_i) \geq 1$$

em que R_i^2 é o coeficiente de determinação correspondente à regressão que tem X_i como variável resposta e as demais variáveis como explicativas. FIV_i aumenta com o aumento da correlação entre X_i e alguma combinação linear das demais variáveis explicativas. Como medida resumo, é comum calcular o fator de inflação da variância médio, $\overline{FIV} = \sum_{i=1}^{p} FIV_i/p$; valores de $\overline{FIV} >> 1$ sugerem problemas sérios de multicolinearidade.

Para contornar esses problemas, algumas sugestões envolvem

i) utilizar variáveis centradas em modelos de regressão polinomial;

ii) eliminar algumas variáveis explicativas que sejam correlacionadas com as demais;

iii) quando possível, adicionar observações que permitam romper a estrutura de correlação entre as variáveis explicativas.

iv) substituir as variáveis originais por **componentes principais**.

Detalhes podem ser obtidos em Kutner, Nachtsheim, Neter & Li (2005).

C.6 Parametrização de modelos lineares

Consideremos um estudo em que se deseja comparar o "efeito" de I tratamentos (drogas, por exemplo) na distribuição de uma variável resposta Y (pressão diastólica, por exemplo). Há muitas situações práticas em que esse "efeito" consiste na modificação do valor esperado da resposta sem alteração da forma de sua distribuição ou de sua variância. Nesse contexto, supondo que m unidades amostrais escolhidas aleatoriamente sejam submetidas a cada um dos I tratamentos, um modelo bastante comum é

$$y_{ij} = \mu_i + e_{ij}, \tag{C.6.1}$$

$i = 1, \ldots, I,\ j = 1, \ldots, m$, em que $\mathbb{E}(e_{ij}) = 0$, $\mathbb{V}(e_{ij}) = \sigma^2$ e $\mathbb{E}(e_{ij}e_{kl}) = 0,\ i \neq k,\ j \neq l$ ou seja, os e_{ij} são não correlacionados. Esta parametrização do modelo é conhecida como **parametrização de médias de celas** pois o parâmetro de localização μ_i pode ser interpretado como o valor esperado (médio) da resposta de unidades amostrais submetidas ao tratamento i. O "efeito" do i-ésimo tratamento é definido como uma função desses valores esperados; por exemplo, podemos definir o efeito do tratamento i em relação ao tratamento I como a diferença $\mu_i - \mu_I$.

Para permitir que a definição de "efeito" seja incorporada diretamente nos parâmetros do modelo, é comum utilizarem-se outras parametrizações. Muitos autores

sugerem que se escreva $\mu_i = \mu + \alpha_i$, o que implica o modelo

$$y_{ij} = \mu + \alpha_i + e_{ij}, \tag{C.6.2}$$

$i = 1, \ldots, I,\ j = 1, \ldots, m$, e interpretam o parâmetro μ como "média geral" e α_i como o "efeito" do tratamento i. O problema é que esse modelo é **inidentificável**[12] e tanto μ quanto α_i nem podem ser interpretados dessa forma nem são **estimáveis**[13] Nesses casos, é possível trabalhar com funções estimáveis (e em geral, aquelas nas quais se tem interesse, o são), mas a ideia de acomodar parâmetros interpretáveis no modelo não é concretizada. Uma possível solução para esse problema consiste na adoção de **restrições de identificabilidade** que não só implicam a identificabilidade do modelo como a estimabilidade de seus parâmetros. Dentre elas, as mais utilizadas na prática são

$$\sum_{i=1}^{I} \alpha_i = 0, \tag{C.6.3}$$

que induz a chamada **parametrização de desvios de médias** e

$$\alpha_1 = 0, \tag{C.6.4}$$

que induz a chamada **parametrização de cela de referência**. Definindo $\overline{y} = (Im)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{m} y_{ij}$ e $\overline{y}_i = m^{-1} \sum_{j=1}^{m} y_{ij}$, e utilizando (C.6.3), obtemos

$$\mathbb{E}(\overline{y}) = (Im)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{m} E(y_{ij}) = \mu + I^{-1} \sum_{i=1}^{I} \alpha_i = \mu,$$

e consequentemente o termo μ pode ser interpretado como média geral (que essencialmente é uma média dos valores esperados da resposta associados aos I tratamentos). Além disso,

$$\mathbb{E}(\overline{y}_i) = m^{-1} \sum_{j=1}^{m} E(y_{ij}) = \mu + \alpha_i$$

[12] Um modelo $F(\theta)$, dependendo do parâmetro $\theta \subset \Theta$, é identificável se para todo $\theta_1, \theta_2 \subset \Theta$, $\theta_1 \neq \theta_2$ temos $F(\theta_1) \neq F(\theta_2)$. Em caso contrário, o modelo é dito inidentificável. Por exemplo, consideremos o modelo $y_i \sim N(\mu + \alpha_i, \sigma^2)$, $i = 1, 2$ em que y_1 e y_2 são independentes. Tomando $\boldsymbol{\theta} = (\mu, \alpha_1, \alpha_2)^\top$ como parâmetro, o modelo é inidentificável, pois tanto para $\boldsymbol{\theta}_1 = (5, 1, 0)^\top$ quanto para $\boldsymbol{\theta}_2 = (4, 2, 1)^\top \neq \boldsymbol{\theta}_1$, a distribuição conjunta de (y_1, y_2) é $N_2[(6, 6)^\top, \sigma^2 \mathbf{I}_2]$. O leitor poderá consultar Bickel & Doksum (2001), entre outros, para detalhes.

[13] Uma função linear de um vetor de parâmetros $\boldsymbol{\theta}$, nomeadamente, $\mathbf{c}^\top \boldsymbol{\theta}$, é estimável se ela for identicamente igual a uma combinação linear do valor esperado do vetor de observações, $\mathbf{y}$, *i.e.*, se existir um vetor $\mathbf{t}$ tal que $\mathbf{c}^\top \boldsymbol{\theta} = \mathbf{t}^\top \mathbb{E}(\mathbf{y})$. No modelo (C.6.2), nem μ nem α_i, $i = 1, \ldots, I$ são funções estimáveis, embora tanto $\mu + \alpha_i$, $i = 1, \ldots, I$ quanto $\alpha_i - \alpha_k$, $i \neq k$ o sejam. O leitor poderá consultar Searle (1971), entre outros, para detalhes.

o que implica que $\alpha_i = \mathbb{E}(\overline{y}_i) - \mu$, *i.e.*, a diferença entre o valor esperado das observações submetidas ao tratamento i e a média geral μ. Essa diferença pode ser interpretada como o efeito do tratamento i.

Se utilizarmos a restrição de identificabilidade (C.6.4), obtemos

$$\mathbb{E}(\overline{y}_1) = m^{-1} \sum_{j=1}^{m} E(y_{1j}) = \mu + \alpha_1 = \mu$$

e consequentemente, o termo μ pode ser interpretado como o valor esperado das observações submetidas ao tratamento 1. Além disso, para $k \neq 1$,

$$\mathbb{E}(\overline{y}_k) = m^{-1} \sum_{j=1}^{m} E(y_{kj}) = \mu + \alpha_k$$

o que sugere que α_k, $k \neq 1$, neste caso correspondendo à diferença entre o valor esperado das observações submetidas ao tratamento k e o valor esperado das observações submetidas ao tratamento 1, tomado como referência, pode ser interpretado como o efeito do tratamento k. Obviamente, qualquer dos tratamentos pode servir de referência, bastando para isso, modificar a restrição de identificabilidade convenientemente. Em geral, quando existe um tratamento controle, é ele que serve de referência.

Voltemos agora nossa atenção para estudos em que se deseja avaliar o efeito de dois fatores A (droga, por exemplo) e B (faixa etária, por exemplo), o primeiro com a e o segundo com b níveis, na distribuição de uma resposta (pressão diastólica, por exemplo). Admitamos que m unidades amostrais tenham sido observadas para cada tratamento, *i.e.*, para cada combinação dos níveis dos fatores A e B. Com base nos mesmos argumentos utilizados no caso anterior, suponhamos que o "efeito" de cada um dos fatores e sua interação possa ser definido apenas em termos dos valores esperados das distribuições da resposta sob os diferentes tratamentos. Um modelo comumente considerado para análise de dados com essa estrutura é

$$y_{ijk} = \mu_{ij} + e_{ijk}, \tag{C.6.5}$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, m$, em que $\mathbb{E}(e_{ijk}) = 0$, $\mathbb{V}(e_{ijk}) = \sigma^2$ e $\mathbb{E}(e_{ijk}e_{i'j'k'}) = 0$, $i \neq i'$ ou $j \neq j'$ ou $k \neq k'$, ou seja, os e_{ijk} são não correlacionados. Esta é a parametrização de médias de celas pois o parâmetro de localização μ_{ij} pode ser interpretado como o valor esperado (médio) da resposta de unidades amostrais submetidas ao tratamento correspondente ao nível i do fator A e nível j do fator B.

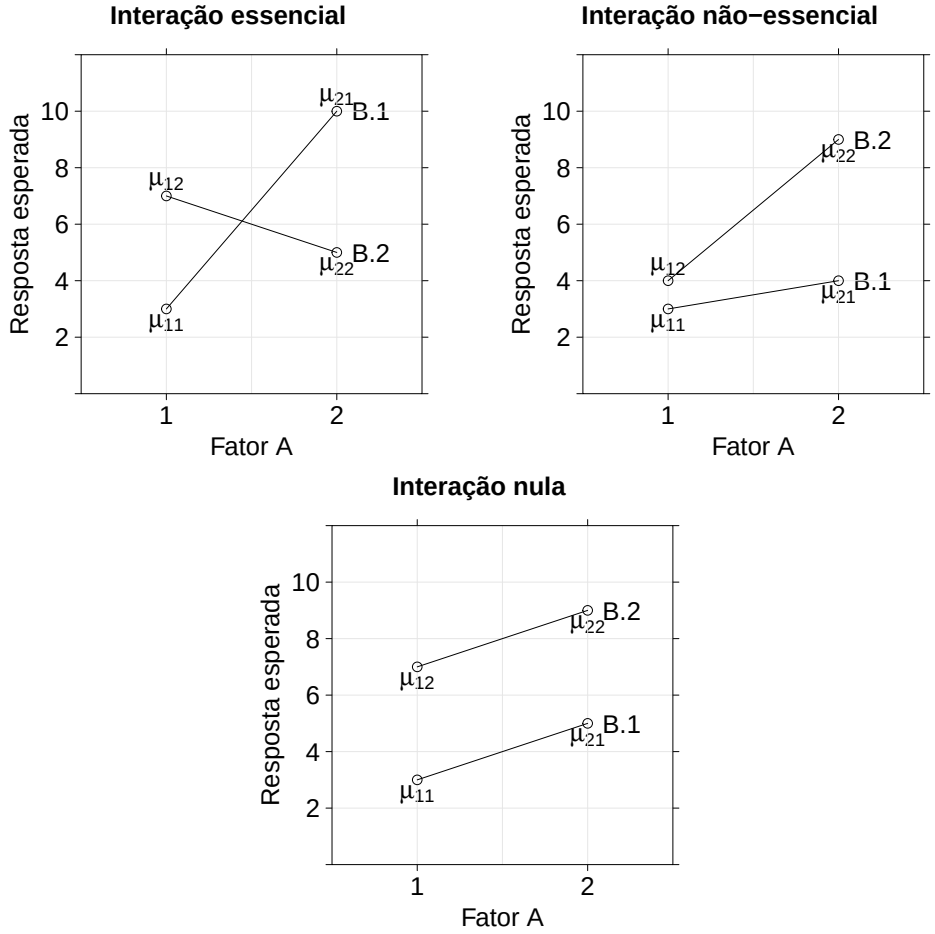
Tomando $a = b = 2$ para facilidade de exposição, o "efeito" do fator A para unidades amostrais no nível j do fator B pode ser definido como a diferença $\mu_{1j} - \mu_{2j}$, que, por exemplo, corresponde à diferença entre o valor esperado da pressão

diastólica de unidades amostrais na faixa etária j submetidas à droga 1 e o valor esperado da pressão diastólica de unidades amostrais na faixa etária j submetidas à droga 2. Analogamente, o "efeito" do fator B para unidades amostrais no nível i do fator A pode ser definido como a diferença $\mu_{i1} - \mu_{i2}$.

A interação entre os fatores A e B pode ser definida como a diferença entre o efeito do fator A para unidades amostrais no nível 1 do fator B e o efeito do fator A para unidades amostrais no nível 2 do fator B, nomeadamente, $(\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$. Outras definições equivalentes, como $(\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$ podem ser utilizadas. A escolha entre as alternativas deve ser feita em função dos detalhes do problema; por exemplo, se a droga 1 for uma droga padrão e a faixa etária 1 corresponder a indivíduos mais jovens, esta última proposta pode ser mais conveniente.

Quando a interação é nula, o efeito do fator A é o mesmo para unidades amostrais submetidas a qualquer dos níveis do fator B e pode-se definir o **efeito principal** do fator A como $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$, que corresponde à diferença entre o valor esperado da resposta para unidades amostrais submetidas ao nível 1 do fator A e o valor esperado da resposta para unidades amostrais submetidas ao nível 2 do fator A (independentemente do nível do fator B). Similarmente, o efeito principal do fator B pode ser definido como $(\mu_{11}+\mu_{21})/2-(\mu_{12}+\mu_{22})/2$. Em muitos casos, essas definições de efeitos principais podem ser consideradas mesmo na presença de interação, desde que ela seja não essencial. A interação entre os fatores A e B é não essencial quando as diferenças $\mu_{11} - \mu_{21}$ e $\mu_{12} - \mu_{22}$ têm o mesmo sinal, mas magnitudes diferentes. Por exemplo, se $\mu_{11} - \mu_{21} = K_1 > 0$ e $\mu_{12} - \mu_{22} = K_2 > 0$ com $K_1 \neq K_2$, a resposta esperada sob o nível 1 do fator A é maior que a resposta esperada sob o nível 2 do fator A tanto no nível 1 quanto no nível 2 do fator B, embora as magnitudes das diferenças não sejam iguais. Se essas magnitudes tiverem sinais diferentes, a interação é essencial. Por outro lado, se $K_1 = K_2$, não há interação. O leitor pode consultar Lencina, Singer & Stanek III (2005) e Lencina, Singer & Stanek III (2008) para uma discussão sobre a consideração de efeitos principais em situações com interação não essencial. Na Figura C.6.1 apresentamos gráficos de perfis médios com interações essencial e não essencial.

Figura C.6.1: Gráfico de perfis médios com diferentes tipos de interação



Com a finalidade de explicitar efeitos principais e interação no modelo, é comum considerar-se a reparametrização $\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$, o que implica o modelo

$$y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}, \tag{C.6.6}$$

$i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, m$. Muitos autores, como Nelder (1998), interpretam erroneamente os parâmetros μ, α_i, β_j, $\alpha\beta_{ij}$ como "média geral", "efeito principal do nível i do fator A", "efeito principal do nível j do fator B" e "interação entre os níveis i do fator A e j do fator B". Como no caso discutido acima, esse modelo também é inidentificável e seus parâmetros são não estimáveis e as restrições de identificabilidade mais frequentemente utilizadas e correspondentes às parametrizações de desvios de médias e cela de referência são, respectivamente,

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \alpha\beta_{ij} = \sum_{j=1}^{b} \alpha\beta_{ij} = 0 \tag{C.6.7}$$

e

$$\alpha_1 = \beta_1 = \alpha\beta_{11} = \ldots = \alpha\beta_{1b} = \alpha\beta_{21} = \ldots = \alpha\beta_{a1} = 0 \qquad \text{(C.6.8)}$$

Sob as restrições (C.6.7), pode-se mostrar que

$$\mu = (ab)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}, \;\; \alpha_i = b^{-1}\sum_{j=1}^{b}\mu_{ij} - \mu, \;\; \beta_j = a^{-1}\sum_{i=1}^{a}\mu_{ij} - \mu$$

e que

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

de forma que esses parâmetros podem ser interpretados como se desejava inicialmente. Sob as restrições (C.6.8), temos

$$\mu = \mu_{11}, \;\; \alpha_i = \mu_{ij} - \mu_{1j}, \;\; i = 2, \ldots, a, \;\; \beta_j = \mu_{ij} - \mu_{i1}, \;\; j = 2, \ldots, b,$$

e que

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

de forma que os parâmetros α_i, $i = 2, \ldots, a$ podem ser interpretados como efeitos diferenciais entre as respostas esperadas das unidades amostrais submetidas ao nível i do fator A relativamente àquelas obtidas por unidades amostrais submetidas ao tratamento associado ao nível 1 do fator A, mantido fixo o nível correspondente ao fator B. Analogamente, os parâmetros β_j, $j = 2, \ldots, b$ podem ser interpretados como efeitos diferenciais entre as respostas esperadas das unidades amostrais submetidas ao nível j do fator B relativamente àquelas obtidas por unidades amostrais submetidas ao tratamento associado ao nível 1 do fator B, mantido fixo o nível correspondente do fator A. Os parâmetros $\alpha\beta_{ij}$, $i = 2, \ldots, a$, $j = 2, \ldots b$ podem ser interpretados como diferenças entre as respostas esperadas das unidades amostrais submetidas ao tratamento correspondente à cela (i, j) e aquela esperada sob um modelo sem interação.

C.7 Regressão logística

Os dados da Tabela C.7.1 são extraídos de um estudo realizado no Hospital Universitário da Universidade de São Paulo com o objetivo de avaliar se algumas medidas obtidas ultrassonograficamente poderiam ser utilizadas como substitutas de medidas obtidas por métodos de ressonância magnética, considerada como padrão áureo, para avaliação do deslocamento do disco da articulação temporomandibular (doravante referido simplesmente como disco). Distâncias cápsula-côndilo (em mm) com boca

aberta ou fechada (referidas, respectivamente, como distância aberta ou fechada no-restante do texto) foram obtidas ultrassonograficamente de 104 articulações e o disco correspondente foi classificado como deslocado (1) ou não (0) segundo a avaliação por ressonância magnética. A variável resposta é o *status* do disco (1 = deslocado ou 0 = não). Mais detalhes podem ser obtidos em Elias, Birman, Matsuda, Oliveira & Jorge (2006).

A diferença fundamental entre este problema e os demais abordados neste apêndice é a natureza da variável resposta, que é discreta em vez de contínua. Denotando a resposta da i-ésima unidade amostral por y_i, podemos definir $y_i = 1$ se houve deslocamento do disco e $y_i = 0$, em caso contrário. Admitindo que para a i-ésima unidade amostral, o disco pode estar deslocado com probabilidade θ_i, temos $P(y_i = 1) = \theta_i$ e $P(y_i = 0) = 1 - \theta_i$, de maneira que $\mathbb{E}(y_i) = \theta_i$ e $\mathbb{V}(y_i) = \theta_i(1 - \theta_i)$.

O objetivo da análise é modelar $\mathbb{E}(y_i)$ como função das variáveis explicativas. O modelo (linear) correspondente é

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{C.7.1}$$

$i = 1, \ldots, n$ em que $\mathbf{x}_i^\top$ é a i-ésima linha da matriz $\mathbf{X}$, que contém os valores p variáveis explicativas para as n unidades amostrais. Em notação matricial, o modelo pode ser escrito como

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \;\text{ com }\; \mathbb{V}(\mathbf{y}) = n^{-1}\text{diag}[\theta_1(1-\theta_1), \ldots, \theta_n(1-\theta_n)]$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X}$ e $\boldsymbol{\beta}$ têm as interpretações usuais. Tanto o problema é claramente heterocedástico quanto a distribuição da variável resposta é claramente não gaussiana. No entanto, ele tem a mesma estrutura daqueles que podem ser analisados por intermédio de métodos de mínimos quadrados ponderados, desde que estimativas consistentes dos parâmetros θ_i estejam disponíveis, o que em geral, não é verdade. Uma situação na qual essa opção pode ser aplicada é aquela em que todas as variáveis explicativas são discretas. Detalhes sobre análises específicas para esse caso podem ser encontrados em Paulino & Singer (2006).

O método de máxima verossimilhança é uma alternativa para o ajuste desses modelos no caso mais geral (em que as variáveis explicativas podem ter natureza discreta ou contínua). A função de verossimilhança a ser maximizada é

$$L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} (\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i} (1 - \mathbf{x}_i^\top \boldsymbol{\beta})^{1-y_i}.$$

Com essa finalidade, podemos considerar seu logaritmo,

$$l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} [y_i \log(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log(1 - \mathbf{x}_i^\top \boldsymbol{\beta})],$$

Tabela C.7.1: Dados de um estudo odontológico

Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco
2.2	1.4	0	0.9	0.8	0	1.0	0.6	0
2.4	1.2	0	1.1	0.9	0	1.6	1.3	0
2.6	2.0	0	1.4	1.1	0	4.3	2.3	1
3.5	1.8	1	1.6	0.8	0	2.1	1.0	0
1.3	1.0	0	2.1	1.3	0	1.6	0.9	0
2.8	1.1	1	1.8	0.9	0	2.3	1.2	0
1.5	1.2	0	2.4	0.9	0	2.4	1.3	0
2.6	1.1	0	2.0	2.3	0	2.0	1.1	0
1.2	0.6	0	2.0	2.3	0	1.8	1.2	0
1.7	1.5	0	2.4	2.9	0	1.4	1.9	0
1.3	1.2	0	2.7	2.4	1	1.5	1.3	0
1.2	1.0	0	1.9	2.7	1	2.2	1.2	0
4.0	2.5	1	2.4	1.3	1	1.6	2.0	0
1.2	1.0	0	2.1	0.8	1	1.5	1.1	0
3.1	1.7	1	0.8	1.3	0	1.2	0.7	0
2.6	0.6	1	0.8	2.0	1	1.5	0.8	0
1.8	0.8	0	0.5	0.6	0	1.8	1.1	0
1.2	1.0	0	1.5	0.7	0	2.3	1.6	1
1.9	1.0	0	2.9	1.6	1	1.2	0.4	0
1.2	0.9	0	1.4	1.2	0	1.0	1.1	0
1.7	0.9	1	3.2	0.5	1	2.9	2.4	1
1.2	0.8	0	1.2	1.2	0	2.5	3.3	1
3.9	3.2	1	2.1	1.6	1	1.4	1.1	0
1.7	1.1	0	1.4	1.5	1	1.5	1.3	0
1.4	1.0	0	1.5	1.4	0	0.8	2.0	0
1.6	1.3	0	1.6	1.5	0	2.0	2.1	0
1.3	0.5	0	4.9	1.2	1	3.1	2.2	1
1.7	0.7	0	1.1	1.1	0	3.1	2.1	1
2.6	1.8	1	2.0	1.3	1	1.7	1.2	0
1.5	1.5	0	1.5	2.2	0	1.6	0.5	0
1.8	1.4	0	1.7	1.0	0	1.4	1.1	0
1.2	0.9	0	1.9	1.4	0	1.6	1.0	0
1.9	1.0	0	2.5	3.1	1	2.3	1.6	1
2.3	1.0	0	1.4	1.5	0	2.2	1.8	1
1.6	1.0	0	2.5	1.8	1			

Dist aberta: distância cápsula-côndilo com boca aberta (mm)
Dist fechada: distância cápsula-côndilo com boca fechada (mm)
Desloc disco: deslocamento do disco da articulação temporomandibular (1=sim, 0=não)

cujas primeira e segunda derivadas são, respectivamente,

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\mathbf{x}_i^\top \boldsymbol{\beta}(1 - \mathbf{x}_i^\top \boldsymbol{\beta})} \mathbf{x}_i,$$

e

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^\top} l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \mathbf{H}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \frac{(\sqrt{y_i} - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{[\mathbf{x}_i^\top \boldsymbol{\beta}(1 - \mathbf{x}_i^\top \boldsymbol{\beta})]^2} \mathbf{x}_i \mathbf{x}_i^\top.$$

As equações de estimação correspondentes são

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\mathbf{x}_i^\top \boldsymbol{\beta}(1 - \mathbf{x}_i^\top \boldsymbol{\beta})} \mathbf{x}_i = \mathbf{0}$$

e o estimador de máxima verossimilhança de $\boldsymbol{\beta}$ pode ser obtido por meio do algoritmo de Newton-Raphson, ou seja, iterando

$$\boldsymbol{\beta}^{(l)} = \boldsymbol{\beta}^{(l-1)} - [\mathbf{H}(\boldsymbol{\beta}^{(l-1)})]^{-1} \mathbf{U}(\boldsymbol{\beta}^{(l-1)}),\ l = 1, 2, \ldots \tag{C.7.2}$$

até que $||\boldsymbol{\beta}^{(l)} - \boldsymbol{\beta}^{(l-1)}|| < \varepsilon$ com $\varepsilon > 0$ e $\boldsymbol{\beta}^{(0)}$ é um valor inicial, arbitrário. O valor $\widehat{\boldsymbol{\beta}}$ obtido na convergência é o estimador de máxima verossimilhança de $\boldsymbol{\beta}$.

O problema com o modelo (C.7.1) é que ele não garante que as estimativas $\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$ das probabilidades θ_i fiquem restritas ao intervalo $(0, 1)$. Uma maneira de evitar esse problema é apelar para um modelo não linear do tipo $\theta_i = \theta(\mathbf{x}_i; \boldsymbol{\beta}) = F^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$ em que F é uma função distribuição, *i.e.*, com imagem no intervalo desejado. Em particular, a função distribuição logística,

$$F(x) = [1 + \exp(-x)]^{-1},\ x \in I\!R$$

é uma candidata com excelentes propriedades. Nesse contexto, o modelo conhecido como **regressão logística** é

$$\theta(\mathbf{x}_i; \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \tag{C.7.3}$$

ou, equivalentemente,

$$\log[\theta(\mathbf{x}_i; \boldsymbol{\beta})]/[1 - \theta(\mathbf{x}_i; \boldsymbol{\beta})] = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{C.7.4}$$

Os termos $\log[\theta/(1 - \theta)]$ são conhecidos como **logitos** (*logits*).

Para efeito de interpretação, consideremos o seguinte modelo de regressão logística com apenas uma variável explicativa,

$$\log[\theta(x; \alpha, \beta)]/[1 - \theta(x; \alpha, \beta)] = \alpha + \beta x.$$

Então, $\alpha = \log[\theta(0;\alpha,\beta)]/[1-\theta(0;\alpha,\beta)]$ e $\exp(\alpha)$ pode ser interpretado como a **chance** (*odds*) de resposta $Y=1$ *versus* $Y=0$ para unidades amostrais com valor da variável explicativa $x=0$. Por outro lado, $\beta = \log[\theta(x+1;\alpha,\beta)]/[1-\theta(x+1;\alpha,\beta)] - \log[\theta(x;\alpha,\beta)]/[1-\theta(x;\alpha,\beta)]$ e então $\exp(\beta)$ pode ser interpretado como a **razão de chances** (*odds ratio*) correspondente a unidades amostrais com valor da variável explicativa $x+1$ relativamente a unidades amostrais com valor da variável explicativa x.

A função de verossimilhança correspondente é

$$L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \prod_{i=1}^{n}\left[\frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}\right]^{y_i}\left[\frac{1}{1+\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}\right]^{1-y_i}$$

e seu logaritmo é

$$l(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n}[y_i\mathbf{x}_i^\top\boldsymbol{\beta} - \log(1+\mathbf{x}_i^\top\boldsymbol{\beta})],$$

com primeira e segunda derivadas dadas, respectivamente, por

$$\frac{\partial}{\partial\boldsymbol{\beta}}l(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i - \frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}]\mathbf{x}_i = \sum_{i=1}^{n}[y_i - \theta(\mathbf{x}_i;\boldsymbol{\beta})]\mathbf{x}_i$$

e

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^\top}l(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \mathbf{H}(\boldsymbol{\beta}) &= -\sum_{i=1}^{n}\frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}{[1+\exp(\mathbf{x}_i^\top\boldsymbol{\beta})]^2}\mathbf{x}_i\mathbf{x}_i^\top \\ &= -\sum_{i=1}^{n}\theta(\mathbf{x}_i;\boldsymbol{\beta})[1-\theta(\mathbf{x}_i;\boldsymbol{\beta})]\mathbf{x}_i\mathbf{x}_i^\top.\end{aligned}$$

O estimador de máxima verossimilhança de $\boldsymbol{\beta}$ é a solução $\widehat{\boldsymbol{\beta}}$ das equações de verossimilhança

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^\top[\mathbf{y} - \boldsymbol{\theta}(\boldsymbol{\beta})] = \mathbf{0} \tag{C.7.5}$$

em que $\boldsymbol{\theta}(\boldsymbol{\beta}) = [\theta(\mathbf{x}_1;\boldsymbol{\beta}),\ldots,\theta(\mathbf{x}_n;\boldsymbol{\beta})]^\top$. Observando que $\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^\top\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})]\mathbf{X}$ em que

$$\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})] = \text{diag}\{\theta(\mathbf{x}_1;\boldsymbol{\beta})[1-\theta(\mathbf{x}_1;\boldsymbol{\beta})],\ldots,\theta(\mathbf{x}_n;\boldsymbol{\beta})[1-\theta(\mathbf{x}_n;\boldsymbol{\beta})]\},$$

pode-se demonstrar que

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p\{\mathbf{0}, [\mathbf{I}(\boldsymbol{\beta})]^{-1}\}$$

com $\mathbf{I}(\boldsymbol{\beta}) = -\mathbb{E}[\mathbf{H}(\boldsymbol{\beta})] = -\mathbb{E}[\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^\top}l(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})] = \mathbf{X}^\top\mathbf{W}(\boldsymbol{\theta})\mathbf{X}$ denotando a matriz de informação de Fisher. Em termos práticos, isto significa que para n suficientemente grande

$$\widehat{\boldsymbol{\beta}} \approx N_p\{\boldsymbol{\beta}, (\mathbf{X}^\top\mathbf{W}[\boldsymbol{\theta}(\widehat{\boldsymbol{\beta}})]\mathbf{X})^{-1}\}.$$

Em notação matricial, as equações de estimação, obtidas por meio de uma aproximação de Taylor de primeira ordem de (C.7.5), podem ser escritas como

$$\widehat{\boldsymbol{\beta}} \approx \boldsymbol{\beta} - \{\mathbf{X}^\top \mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})]\mathbf{X}\}^{-1}\mathbf{U}[\boldsymbol{\theta}(\boldsymbol{\beta})]$$

sugerindo o seguinte algoritmo de Newton-Raphson, que neste caso coincide com o algoritmo "Scoring"/ de Fisher,

$$\boldsymbol{\beta}^{(l)} = \boldsymbol{\beta}^{(l-1)} - \{\mathbf{X}^\top \mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})\mathbf{X}\}^{-1}\mathbf{X}^\top \mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})][\mathbf{y} - \boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})],\ l = 1, 2, \ldots$$

Fazendo $\mathbf{z} = \mathbf{X}^\top\boldsymbol{\beta}^{(l-1)} + \{\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})]\}^{-1}[\mathbf{y} - \boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})]$, o algoritmo pode ser escrito como

$$\boldsymbol{\beta}^{(l)} = \{\mathbf{X}^\top \mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})\mathbf{X}\}^{-1}\mathbf{X}^\top \mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})]\mathbf{z},\ l = 1, 2, \ldots.$$

Em cada passo, esse algoritmo tem uma estrutura equivalente à solução de mínimos quadrados ponderados com pesos explicitados em $\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta}^{(l-1)})]$ e pseudo-variáveis $\mathbf{z}$, ambos recalculados em cada iteração. Por isso é conhecido como algoritmo de **mínimos quadrados iterativamente ponderados** (*iteratively reweighted least squares*). Mais detalhes podem ser obtidos em Sen et al. (2009), lembrando que o modelo de regressão logística é um caso particular dos chamados **modelos lineares generalizados**.

Uma das características do modelo de regressão logística é que ele permite a estimação das probabilidades de sucesso θ_i, bastando para isto, substituir $\boldsymbol{\beta}$ por $\widehat{\boldsymbol{\beta}}$ em (C.7.3), ou seja

$$\widehat{\theta}_i = \frac{\exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})}. \tag{C.7.6}$$

A distribuição aproximada de $\widehat{\theta}_i$ pode ser obtido por meio do **Método Delta** (ver Apêndice B). Para isto, notemos que

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}\theta(\mathbf{x}_i;\boldsymbol{\beta}) &= \frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})[1 + \exp(\mathbf{x}_i^\top\boldsymbol{\beta})] - [\exp(\mathbf{x}_i^\top\boldsymbol{\beta})]^2}{[1 + \exp(\mathbf{x}_i^\top\boldsymbol{\beta})]^2}\mathbf{x}_i \\ &= \frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^\top\boldsymbol{\beta})]^2}\mathbf{x}_i = \theta(\mathbf{x}_i;\boldsymbol{\beta})[1 - \theta(\mathbf{x}_i;\boldsymbol{\beta})]\mathbf{x}_i.\end{aligned}$$

Então, utilizando (B.2.2), obtemos

$$\begin{aligned}\mathbb{V}(\widehat{\theta}_i) &= \{\theta(\mathbf{x}_i;\boldsymbol{\beta})[1 - \theta(\mathbf{x}_i;\boldsymbol{\beta})]\}^2\mathbf{x}_i^\top\mathbb{V}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i \\ &= \{\theta(\mathbf{x}_i;\boldsymbol{\beta})[1 - \theta(\mathbf{x}_i;\boldsymbol{\beta})]\}^2\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})]\mathbf{X})^{-1}\mathbf{x}_i.\end{aligned}$$

Consequentemente, a matriz de covariâncias de $\widehat{\boldsymbol{\theta}}$ pode ser escrita compactamente como

$$\mathbb{V}(\widehat{\boldsymbol{\theta}}) = \mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})]\mathbf{X}(\mathbf{X}^\top\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})]\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}[\boldsymbol{\theta}(\boldsymbol{\beta})] \tag{C.7.7}$$

e pelo Método Delta, obtemos,

$$\widehat{\boldsymbol{\theta}} \approx N_n\{\boldsymbol{\theta}, \mathbb{V}(\widehat{\boldsymbol{\theta}})\},$$

em que $\mathbb{V}(\widehat{\boldsymbol{\theta}})$ é dada por (C.7.7). Para efeito de aplicações, uma estimativa de (C.7.7) pode ser obtida por meio da substituição de $\boldsymbol{\beta}$ por $\widehat{\boldsymbol{\beta}}$. Para detalhes, o leitor poderá consultar Hosmer & Lemeshow (2000) entre outros.

Com intuito didático, voltemos aos dados da Tabela C.7.1 e consideremos um modelo logístico para a chance de deslocamento do disco, tendo apenas a distância aberta como variável explicativa. Nesse contexto, o modelo (C.7.4) corresponde a

$$\log[\theta(x_i;\alpha,\beta)]/[1-\theta(x_i;\alpha,\beta)] = \alpha + x_i\beta \tag{C.7.8}$$

$i = 1,\ldots,104$ em que $\theta(x_i;\alpha,\beta)$ representa a probabilidade de deslocamento do disco quando o valor da distância aberta é x_i, α denota o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0$ e β é interpretado como a variação no logaritmo da chance de deslocamento do disco por unidade de variação da distância aberta. Consequentemente, a razão de chances do deslocamento do disco correspondente a uma diferença de d unidades da distância aberta será $\exp(d \times \beta)$. Como não temos dados correspondentes a distâncias abertas menores que 0.50, convém substituir os valores x_i por valores "centrados", ou seja por $x_i^* = x_i - x_0$. Uma possível escolha para x_0 é o mínimo de x_i, que é 0.50. Essa transformação na variável explicativa altera somente a interpretação do parâmetro α que passa a ser o logaritmo o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0.50$.

Estimativas (com erros padrões entre parênteses) dos parâmetros desse modelo ajustado por máxima verossimilhança aos dados da Tabela C.7.1, são, $\widehat{\alpha} = -5.86$ (1.10) e $\widehat{\beta} = 3.16$ (0.66) e então, segundo o modelo, uma estimativa da chance de deslocamento do disco para articulações com distância aberta $x = 0.50$ (que corresponde à distância aberta transformada $x^* = 0.00$) é $\exp(-5.86) = 0.003$; uma estiva da para a razão entre a chance de deslocamento do disco para articulações com distância aberta $x+1$ e um intervalo de confiança (95%) para essa chance pode ser obtido exponenciando os limites (LI e LS) do intervalo para o parâmetro α, nomeadamente

$$\begin{aligned} LI &= \exp[\widehat{\alpha} - 1.96EP(\widehat{\alpha})] = \exp(-5.16 - 1.96 \times 1.10) = 0.000 \\ LS &= \exp[\widehat{\alpha} + 1.96EP(\widehat{\alpha})] = \exp(-5.16 + 1.96 \times 1.10) = 0.024. \end{aligned}$$

Os limites de um intervalo de confiança para a razão de chances correspondentes a um variação de uma unidade no valor da distância aberta podem ser obtidos de maneira similar e são 6.55 e 85.56.

Com base em (C.7.6) podemos estimar a probabilidade de sucesso (deslocamento do disco, no exemplo sob investigação); por exemplo, para uma articulação cuja distância aberta seja 2.10 (correspondente à distância aberta transformada igual a 1.60), a estimativa dessa probabilidade é

$$\widehat{\theta} = \exp(-5.86 + 3.16 \times 1.60)/[1 + \exp(-5.86 + 3.16 \times 1.60)] = 0.31.$$

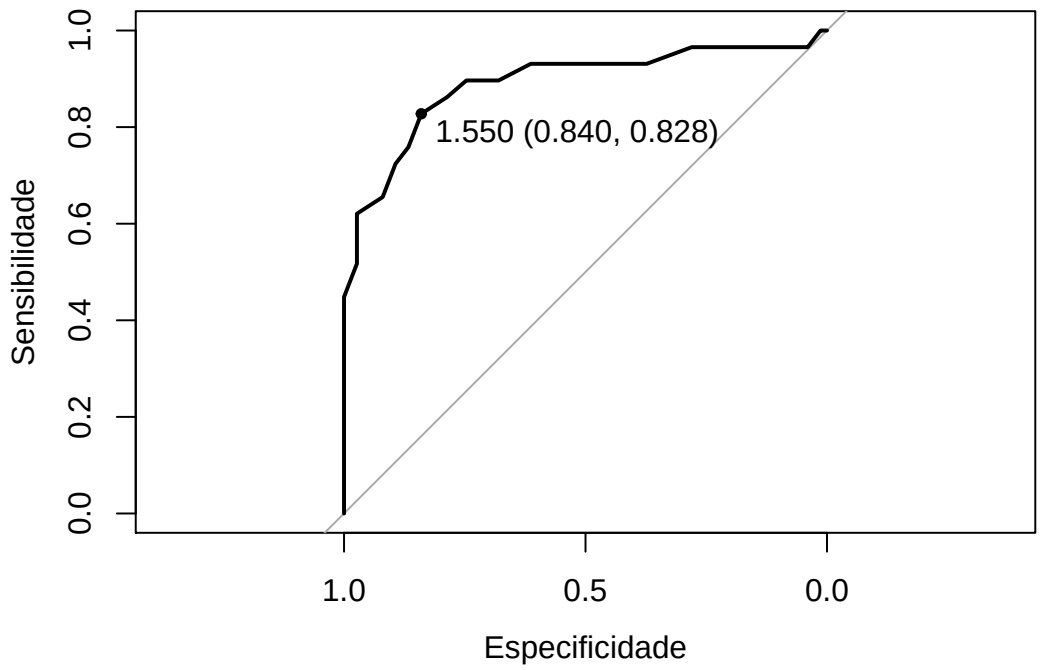
Lembrando que o objetivo do estudo é substituir o processo de identificação de deslocamento do disco realizado via ressonância magnética por aquele baseado na medida da distância aberta por meio de ultrassonografia, podemos estimar as probabilidades de sucesso para todas as articulações e identificar um **ponto de corte** d_0 segundo o qual, distâncias abertas com valores acima dele sugerem decidirmos pelo deslocamento do disco e distâncias abertas com valores abaixo dele sugerem a decisão oposta. Obviamente, não esperamos que todas as decisões tomadas dessa forma sejam corretas e consequentemente, a escolha do ponto de corte deve ser feita com o objetivo de minimizar os erros (decidir pelo deslocamento quando ele não existe ou *vice versa*). Nesse contexto, um contraste entre as decisões tomadas com base em um determinado ponto de corte d_0 e o padrão áureo definido pela ressonância magnética para todas as 104 articulações pode ser resumido por meio da Tabela C.7.2 em que as frequências da diagonal principal correspondem a decisões corretas e aquelas da diagonal secundária às decisões erradas. O quociente $n_{11}/(n_{11} + n_{21})$ é

Tabela C.7.2: Frequência de decisões para um ponto de corte d_0

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	n_{11}	n_{12}
distância aberta d_0	não	n_{21}	n_{22}

conhecido como **sensibilidade** do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco está realmente deslocado. O quociente $n_{22}/(n_{12} + n_{22})$ é é conhecido como **especificidade** do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco está realmente não está deslocado. A situação ideal é aquela em que tanto a sensibilidade quanto a especificidade do processo de decisão são iguais a 100%. O problema a resolver é determinar o ponto de corte d_{max} que gere o melhor equilíbrio entre sensibilidade e especificidade. Com essa finalidade, podemos construir tabelas com o mesmo formato da Tabela C.7.2) para diferentes pontos de corte e um gráfico cartesiano entre a sensibilidade e especificidade obtida de cada uma delas. Esse gráfico, conhecido como **curva ROC** (do termo inglês *Receiver Operating Characteristic*) gerado para os dados da Tabela C.7.1 está apresentado na Figura C.7.1.

Figura C.7.1: Curva ROC para os dados da Tabela C.7.1 baseada no modelo (C.7.8 com distância aberta como variável explicativa



O ponto de corte ótimo é aquele mais próximo do vértice superior esquerdo (em que tanto a sensibilidade quanto a especificidade seriam iguais a 100%. Para o exemplo, esse ponto está salientado na Figura C.7.1 e corresponde à distância aberta com valor $d_{max} = 2.05 (= 1.55 + 0.50)$. A sensibilidade e a especificidade associadas à decisão baseada nesse ponto de corte, são, respectivamente, 83% e 84% e as frequências de decisões corretas estão indicadas na Tabela C.7.3. Com

Tabela C.7.3: Frequência de decisões para um ponto de corte para distância aberta $d_{max} = 2.05$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	24	12
distância aberta $d_{max} = 2.05$	não	5	63

esse procedimento de decisão a porcentagem de acertos (**acurácia**) é 84% [= (24 +

63)/104]. A porcentagem de **falsos positivos** é 17% [= 5/(5+29)] e a porcentagem de **falsos negativos** é 16% [= 12/(12 + 63)].

Uma análise similar, baseada na distância fechada (transformada por meio da subtração de deu valor mínimo (0,4) gera a curva ROC apresentada na Figura C.7.2 e frequências de decisões apresentada na Tabela C.7.4.

Figura C.7.2: Curva ROC para os dados da Tabela C.7.1 baseada no modelo (C.7.8 com distância fechada como variável explicativa

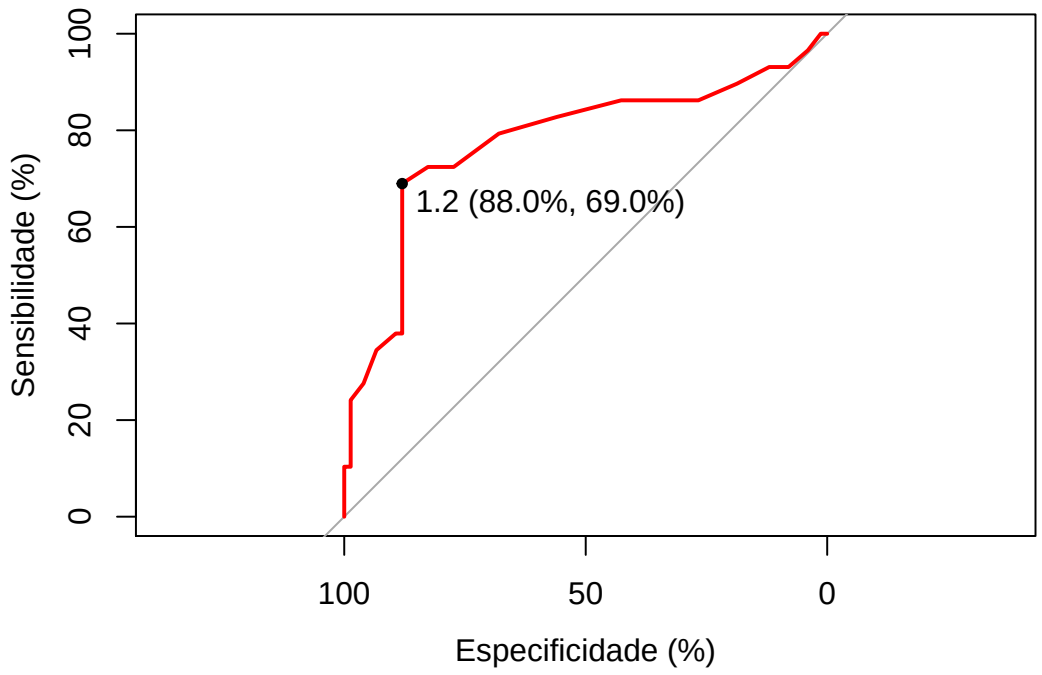


Tabela C.7.4: Frequência de decisões para um ponto de corte para distância fechada $d_{max} = 1,60$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	20	9
distância fechada $d_{max} = 1.60$	não	9	66

A acurácia associada a processo de decisão baseado apenas na distância fechada, 83% [= (20 + 66)/104] é praticamente igual àquela obtida com base apenas na

distância aberta; no entanto aquele processo apresenta um melhor equilíbrio entre sensibilidade e especificidade (83% e 84%, respectivamente, *versus* 88% e 69%).

Se quisermos avaliar o processo de decisão com base nas observações das distâncias aberta e fechada simultaneamente, podemos considerar o modelo

$$\log[\theta(x_i;\alpha,\beta,\gamma)]/[1-\theta(x_i;\alpha,\beta,\gamma)] = \alpha + x_i\beta + w_i\gamma \qquad \text{(C.7.9)}$$

$i = 1, \ldots, 104$ em que w_i corresponde à distância fechada observada na i-ésima articulação. Neste caso, γ corresponde à razão entre a chance de deslocamento do disco para articulações com distância fechada $w+1$ e a chance de deslocamento do disco para articulações com distância fechada w para aquelas com mesmo valor da distância aberta; uma interpretação similar vale para o parâmetro β. Estimativas dos parâmetros (com erros padrões entre parênteses) do modelo (C.7.9) obtidas após a transformação das variáveis explicativas segundo o mesmo figurino adotado nas análises univariadas são $\widehat{\alpha} = -6.38$ (1.19), $\widehat{\beta} = 2.83$ (0.67) e $\widehat{\gamma} = 0.98$ (0.54). A estimativa do parâmetro γ é apenas marginalmente significativa, ou seja a inclusão da variável explicativa distância fechada não acrescenta muito poder de discriminação além daquele correspondente à distância aberta. Uma das razões para isso é que as duas variáveis são correlacionadas (com coeficiente de correlação de Pearson igual a 0.46. A determinação de pontos de corte para modelos com duas ou mais variáveis explicativas é bem mais complexa do que no caso univariado e não será abordada neste texto. Para efeito de comparação com as análises anteriores, as frequências de decisões obtidas com os pontos de corte utilizados naquelas estão dispostas na Tabela C.7.5, e correspondem a uma sensibilidade de 62%, especificidade de 97% e acurácia de 88%.

Tabela C.7.5: Frequência de decisões correspondentes a pontos de corte $d_{max} = 2.05$ para distância aberta e $d_{max} = 1.60$ para distância fechada

		Deslocamento real do disco	
		sim	não
Decisão baseada em	sim	18	2
ambas as distâncias	não	11	73

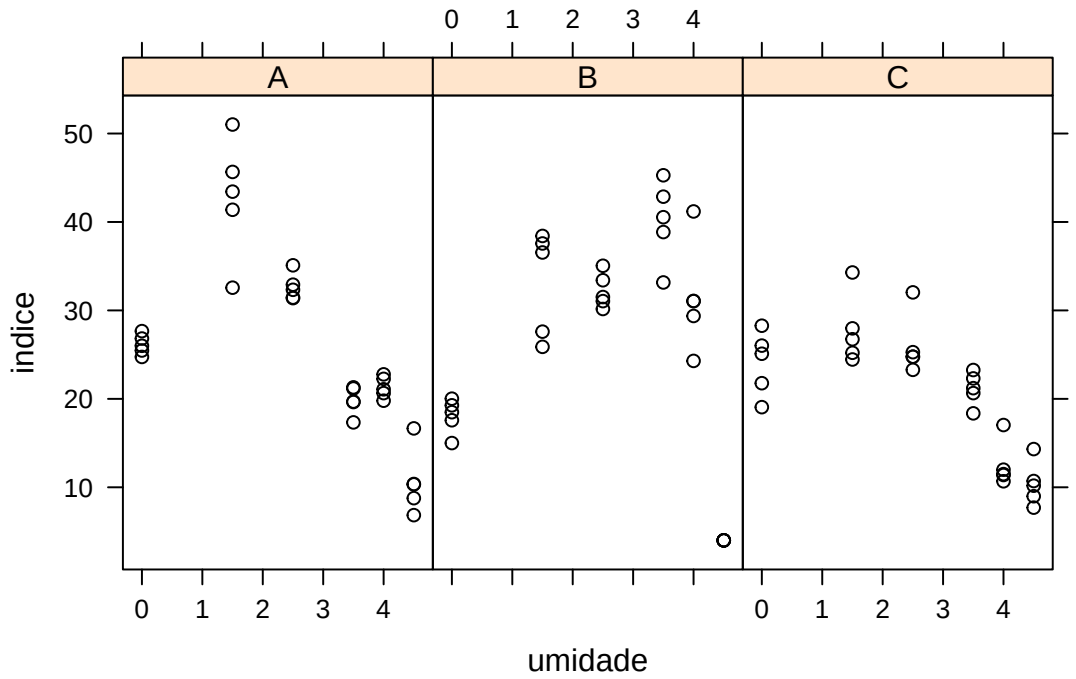
C.8 Exemplos

Exemplo C.8.1: Os dados representados na Tabela C.8.1 e disponíveis em

www.ime.usp.br/∼jmsinger/Dados/Singer&Nobre&Rocha2018exempc81.xls

são provenientes de um estudo conduzido na Faculdade de Odontologia da Universidade de São Paulo com o objetivo de avaliar o efeito do nível de umidade na resistência de união (medida por meio de um índice) de três adesivos dentários. Cada um de três conjuntos de trinta molares extraídos foi tratado com um dos três adesivos (A, B ou C), com 5 dentes submetidos a cada nível de umidade (0.00, 1.50, 2.50, 3.50, 4.00 e 4.50). Depois de um certo tempo, a resistência de união foi avaliada em cada um dos 90 dentes. Detalhes podem ser obtidos em Reis, Loguercio, Azevedo, Carvalho, Singer & Grande (2003). Médias e desvios padrões correspondentes aos índices de resistência de união estão dispostos na Tabela C.8.2 e gráficos de dispersão correspondentes estão apresentados na Figura C.8.1.

Figura C.8.1: Gráficos de dispersão para os dados da Tabela C.8.1



Ambos sugerem um erro de observação ou transcrição para os dados associados às observações dos dentes submetidos ao adesivo B com nível de umidade 4.50. Esses valores serão eliminados na análise subsequente. O gráfico de dispersão sugere um

Tabela C.8.1: Índice de resistência de adesivos dentários

adesivo	umid	indice	adesivo	umid	indice	adesivo	umid	indice
A	0.0	24.74	B	0.0	15.00	C	0.0	28.27
A	0.0	27.66	B	0.0	20.02	C	0.0	19.06
A	0.0	26.00	B	0.0	19.27	C	0.0	26.02
A	0.0	25.47	B	0.0	18.49	C	0.0	25.10
A	0.0	26.82	B	0.0	17.58	C	0.0	21.77
A	1.5	45.66	B	1.5	25.88	C	1.5	24.45
A	1.5	51.02	B	1.5	36.55	C	1.5	34.29
A	1.5	32.57	B	1.5	38.42	C	1.5	25.21
A	1.5	41.38	B	1.5	37.57	C	1.5	26.74
A	1.5	43.43	B	1.5	27.60	C	1.5	27.97
A	2.5	32.91	B	2.5	31.05	C	2.5	24.75
A	2.5	35.10	B	2.5	35.05	C	2.5	24.77
A	2.5	31.39	B	2.5	30.16	C	2.5	32.04
A	2.5	32.33	B	2.5	33.41	C	2.5	25.28
A	2.5	31.45	B	2.5	31.51	C	2.5	23.28
A	3.5	19.69	B	3.5	45.28	C	3.5	18.36
A	3.5	21.17	B	3.5	38.86	C	3.5	23.26
A	3.5	19.64	B	3.5	40.54	C	3.5	22.31
A	3.5	21.30	B	3.5	42.86	C	3.5	20.65
A	3.5	17.34	B	3.5	33.16	C	3.5	21.21
A	4.0	21.08	B	4.0	24.29	C	4.0	11.39
A	4.0	22.77	B	4.0	41.19	C	4.0	10.68
A	4.0	19.79	B	4.0	29.37	C	4.0	17.04
A	4.0	22.26	B	4.0	31.07	C	4.0	11.49
A	4.0	20.65	B	4.0	31.05	C	4.0	11.99
A	4.5	6.85	B	4.5	4.00	C	4.5	8.99
A	4.5	8.77	B	4.5	4.00	C	4.5	10.71
A	4.5	16.66	B	4.5	4.00	C	4.5	10.18
A	4.5	10.33	B	4.5	4.00	C	4.5	14.32
A	4.5	10.37	B	4.5	4.00	C	4.5	7.70

Tabela C.8.2: Médias (desvios padrões) do índice de resistência de união

	Umidade					
Adesivo	0.00	1.50	2.50	3.50	4.00	4.50
A	26.1 (1.1)	42.8 (6.8)	32.6 (1.5)	19.8 (1.6)	21.3 (1.2)	10.6 (3.7)
B	18.1 (1.9)	33.2 (6.0)	32.2 (2.0)	40.1 (4.6)	31.4 (6.1)	4.0 (0.0)
C	24.0 (3.6)	27.7 (3.9)	26.0 (3.4)	21.2 (1.9)	12.5 (2.6)	10.4 (2.5)

efeito quadrático dos níveis de umidade. Uma análise inicial que incorpora essa sugestão pode ser concretizada por meio do ajuste do seguinte modelo quadrático

$$y_{ijk} = \alpha_i + \beta_i x_k + \gamma_i x_k^2 + e_{ijk}, \tag{C.8.1}$$

$i = 1, \ldots, 3,\ j = 1, \ldots, 5,\ k = 1, \ldots, 6$ em que y_{ijk} representa o índice de resistência para o j-ésimo dente submetido ao i-ésimo tratamento ($i = 1$ correspondendo ao adesivo A, $i = 2$, ao adesivo B e $i = 3$, ao adesivo C) sob o k-ésimo nível de umidade, x_k (lembrando que para o adesivo B, $k = 1, \ldots, 5$ dada a eliminação dos valores correspondentes ao nível de umidade 4.50). Supomos que $e_{ijk} \sim N(0, \sigma^2)$ são erros aleatórios independentes.

Os resultados do ajuste desse modelo aos dados estão dispostos na Tabela C.8.3. O desvio padrão residual é $S = 4.55$ e o coeficiente de determinação ajustado,

Tabela C.8.3: Estimativas (erros padrões) dos coeficientes do modelo (C.8.1) aos dados da Tabela C.8.1

	Coeficiente		
Adesivo	linear	angular	quadrático
A	27.94 (1.97)	11.35 (1.96)	-3.43 (0.42)
B	18.18 (2.00)	12.25 (2.27)	-2.09 (0.54)
C	23.98 (1.97)	5.71 (1.96)	-1.99 (0.42)

$R^2_{aj} = 0.972$. O excelente ajuste obtido sob essa ótica é confirmado por intermédio do gráfico de resíduos padronizados e do correspondente histograma, apresentados na Figura C.8.2. A sugestão de uma leve assimetria não deve ser importante no que tange à estimação dos parâmetros.

Figura C.8.2: Gráfico de resíduos padronizados e o correspondente histograma referentes ao ajuste do modelo (C.8.1)

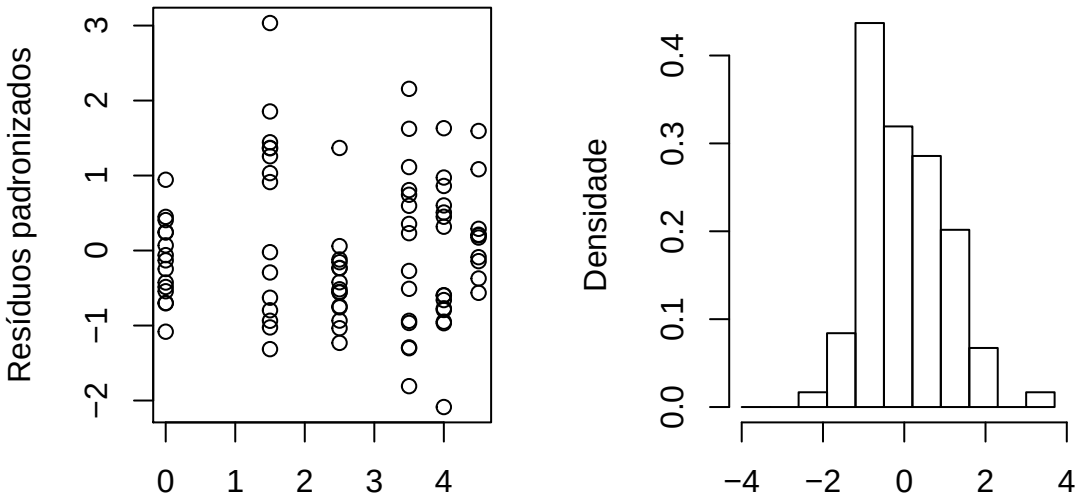
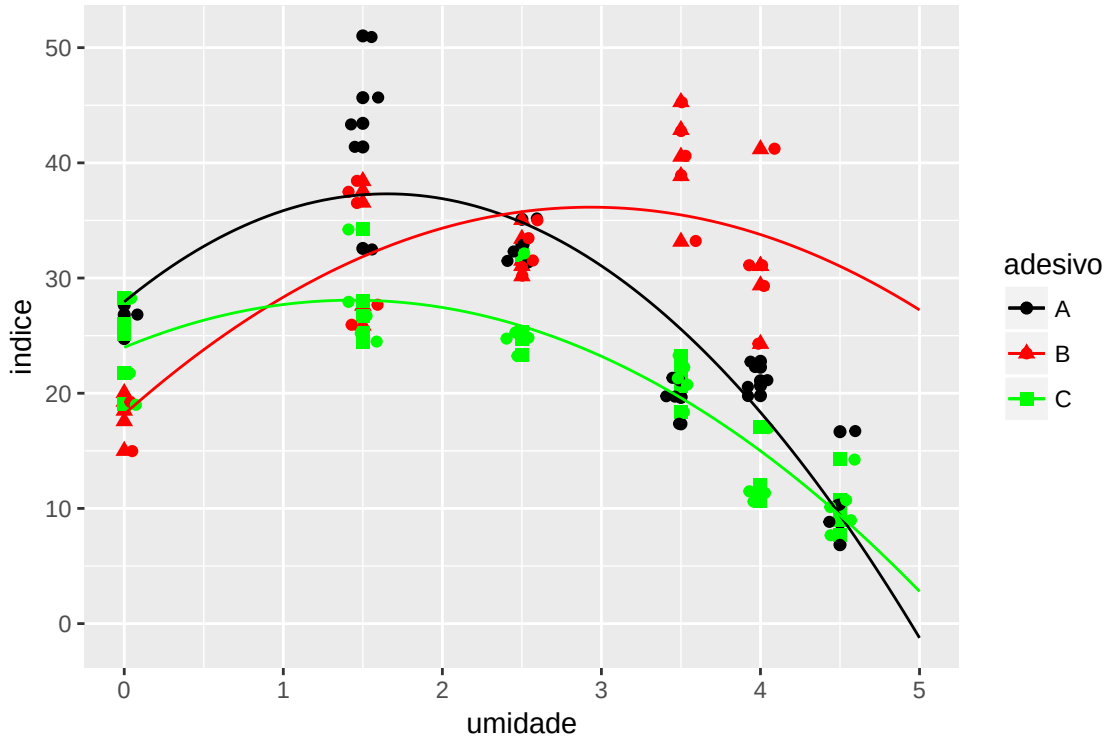


Figura C.8.3: Gráfico com curvas ajustadas pelo modelo (C.8.1)



Na Figura C.8.3 apresentamos as curvas ajustadas para os três adesivos.

A comparação entre as curvas que representam o efeito da umidade no índice de resistência correspondentes aos três adesivos pode ser realizada por meio de testes de hipóteses sobre seus parâmetros. A hipótese de que as três curvas são coincidentes pode ser expressa na forma $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ com

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}$$

e $\boldsymbol{\beta} = (\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2, \alpha_3, \beta_3, \gamma_3)^\top$. O resultado do teste, baseado numa estatística F com 6 graus de liberdade no numerador e 76 graus de liberdade no denominador mostra forte evidência de diferença entre as curvas ($p < 0.001$). Comparações múltiplas podem ser concretizadas de forma semelhante. Por exemplo, para comparar as curvas associadas aos adesivos A e B, basta testar a hipótese utilizando apenas a primeira, terceira e quinta linhas da matriz $\mathbf{C}$. Nesse caso, a estatística F tem 3 graus de liberdade no numerador. Todas as comparações entre as curvas duas a duas sugerem diferenças altamente significativas ($p < 0.001$).

Os níveis ótimos de umidade (correspondentes ao máximo índice de resistência de união) são dados por $g(\beta_i, \gamma_i) = -\beta_i/(2\gamma_i)$, $i = 1, \ldots, 3$ e suas estimativas obtidas por meio dos valores apresentados na Tabela C.8.3. Estimativas de suas variâncias podem ser obtidas via Método Delta, observando que

$$\mathbf{u}_i = \left[\frac{\partial g(\beta_i, \gamma_i)}{\partial \beta_i}, \frac{\partial g(\beta_i, \gamma_i)}{\partial \gamma_i}\right]^\top = \left[-\frac{1}{2\gamma_i}, \frac{\beta_i}{2\gamma_i^2}\right]^\top$$

de forma que $\mathbb{V}[g(\widehat{\beta}_i, \widehat{\gamma}_i)] = \mathbf{u}_i(\widehat{\boldsymbol{\beta}})^\top \mathbb{V}_i(\widehat{\boldsymbol{\beta}})\mathbf{u}_i(\widehat{\boldsymbol{\beta}})$ com $\mathbb{V}_i(\widehat{\boldsymbol{\beta}})$ representando a submatriz (com dimensão 2×2) de $\mathbb{V}(\widehat{\boldsymbol{\beta}})$ correspondente aos parâmetros β_i, γ_i. Estimativas dos pontos de resistência máxima e intervalos de confiança (aproximados) com coeficientes de confiança de 95% estão dispostos na Tabela C.8.4.

Tabela C.8.4: Estimativas e intervalos de confiança (95%) para os pontos de resistência máxima

	Ponto de	Limites do IC(95%)	
Adesivo	máximo	inferior	superior
A	1.65	1.44	1.86
B	2.93	1.26	4.60
C	1.44	0.52	2.35

C.9 Exercícios

C.9.1. Seja $x_1, \ldots, x_n$ uma amostra aleatória de uma variável X cuja distribuição é Normal.

a) Mostre que a média e a variância amostrais são independentes.

b) Enuncie um resultado semelhante no contexto de regressão linear simples.

c) Qual a utilidade desse resultado?

C.9.2. Mostre que se $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{V})$ então $(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2_p$.

C.9.3. Considere um vetor aleatório (X, Y) com distribuição normal bivariada. Mostre que a esperança condicional de Y dado $X = x$ é da forma $E(Y|X = x) = \alpha + \beta x$ explicitando os parâmetros α e β em termos dos parâmetros da distribuição normal bivariada adotada.

C.9.4. Obtenha os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão

$$y_i = x_i^\beta e_i,$$

$i = 1, \ldots, n$ em que os e_i representam erros aleatórios independentes log-normais com média $\exp(\sigma^2/2)$ e variância $\exp(\sigma^2)[\exp(\sigma^2) - 1]$.

i) Obtenha a distribuição do estimador de β e construa um intervalo de confiança.

ii) Obtenha um intervalo de confiança aproximado para o valor esperado de y dado x_0.

Sugestão: linearize o modelo e lembre-se que se $\log X \sim N(\mu, \sigma^2)$ então X tem distribuição log-normal com média $\exp(\mu + \sigma^2/2)$ e variância $[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$.

C.9.5. Suponha que as variáveis x_i e y_i estejam relacionadas de acordo com o modelo de regressão linear simples

$$y_i = \alpha + \beta x_i + e_i,$$

$i = 1, \ldots, n$ em que os e_i representam erros aleatórios independentes de média zero e variância σ^2. Mostre que o estimador

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2$$

em que $\widehat{\alpha}$ e $\widehat{\beta}$ denotam os estimadores de mínimos quadrados de α e β, respectivamente, é não enviesado para a variância σ^2.

C.9.6. Expresse o modelo de regressão linear simples $y_i = \alpha + \beta x_i + e_i$, $i = 1, \ldots, n$ em que os e_i representam erros aleatórios independentes de média zero e variância σ^2 na forma $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ e calcule $\widehat{\boldsymbol{\beta}}$ e $\mathbb{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma_2$.

C.9.7. Para avaliar a associação entre a pressão arterial sistólica e idade, colheram-se os dados dispostos na Tabela C.1.1. Utilize esses dados para avaliar se essa associação pode ser representada por um modelo de regressão linear simples. Com essa finalidade,

a) Especifique o modelo, interpretando os parâmetros;

b) Construa um diagrama de dispersão (rotulando os eixos convenientemente);

c) Estime os parâmetros e os correspondentes erros padrões;

d) Construa intervalos de confiança (com coeficiente de confiança de 95%) para os parâmetros;

e) Obtenha o valor-p correspondente ao teste da hipótese de que o coeficiente angular é nulo.

C.9.8. Considere os modelos

$$y_i = \alpha + \beta x_i + e_i \quad \text{e} \quad y_i = \beta x_i + e_i$$

$i = 1, \ldots, n$ em que os e_i representam erros aleatórios independentes de média zero e variância σ^2. Para ambos os modelos, expresse o coeficiente de determinação R^2 em termos de x_i e y_i e discuta a diferença entre eles.

C.9.9. Para investigar a associação entre tipo de escola (particular ou pública), cursada por calouros de uma universidade e a média no curso de Cálculo I, obtiveram-se os seguintes dados:

Escola	Média no curso de Cálculo I								
Particular	8.6	8.6	7.8	6.5	7.2	6.6	5.6	5.5	8.2
Pública	5.8	7.6	8.0	6.2	7.6	6.5	5.6	5.7	5.8

Seja y_i a nota obtida pelo i-ésimo aluno, $x_i = 1$ se o aluno cursou escola particular e $x_i = -1$ se o aluno cursou escola pública, $i = 1, \ldots, 18$. Considere o modelo $y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, 18$ em que os e_i são erros aleatórios não correlacionados com $E(e_i) = 0$ e $Var(e_i) = \sigma^2$.

i) Interprete os parâmetros α e β.

ii) Estime α e β pelo método de mínimos quadrados. Obtenha também uma estimativa de σ^2.

iii) Construa intervalos de confiança para α e β.

iv) Com base nas estimativas obtidas no item ii), construa intervalos de confiança para os valores esperados das notas dos alunos das escolas particulares e públicas.

v) Ainda utilizando o modelo proposto, especifique e teste a hipótese de que ambos os valores esperados são iguais.

vi) Repita os itens i)-v) definindo $x_i = 1$ se o aluno cursou escola particular e $x_i = 0$ se o aluno cursou escola pública, $i = 1, \ldots, 18$.

C.9.10. Num estudo realizado na Faculdade de Medicina da Universidade de São Paulo foram colhidos dados de 16 pacientes submetidos a transplante intervivos e em cada um deles obtiveram-se medidas tanto do peso (g) real do lobo direito do fígado quanto de seu volume (cm^3) previsto pré operatoriamente por métodos ultrassonográficos. O objetivo é estimar o peso real por meio do volume previsto. Os dados estão dispostos na tabela abaixo.

i) Proponha um modelo de regressão linear simples para analisar os dados e interprete seus parâmetros.

ii) Construa um gráfico de dispersão apropriado.

iii) Ajuste o modelo e construa intervalos de confiança para seus parâmetros.

iv) Avalie o ajuste do modelo por meio de medidas descritivas, de testes de hipóteses convenientes e de uma análise de resíduos.

v) Construa uma tabela com intervalos de confiança para o peso esperado do lobo direito do fígado correspondentes a volumes (estimados ultrassonograficamente) de 600, 700, 800, 900 e 1000 cm^3.

vi) Repita os itens anteriores considerando um modelo linear simples sem intercepto.

Volume USG (cm^3)	Peso real (g)	Volume USG (cm^3)	Peso real (g)
656	630	737	705
692	745	921	955
588	690	923	990
799	890	945	725
766	825	816	840
800	960	584	640
693	835	642	740
602	570	970	945

C.9.11. Os dados abaixo são provenientes de uma pesquisa cujo objetivo é propor um modelo para a relação entre a área construída de um determinado tipo de imóvel e o seu preço.

Imóvel	Área (m^2)	Preço (R$)
1	128	10 000
2	125	9 000
3	200	17 000
4	4.000	200 000
5	258	25 000
6	360	40 000
7	896	70 000
8	400	25 000
9	352	35 000
10	250	27 000
11	135	11 000
12	6.492	120 000
13	1.040	35 000
14	3.000	300 000

i) Construa um gráfico de dispersão apropriado para o problema.

ii) Ajuste um modelo de regressão linear simples e avalie a qualidade do ajuste (obtenha estimativas dos parâmetros e de seus erros padrões, calcule o coeficiente de determinação e construa gráficos de resíduos e um gráfico do tipo QQ).

iii) Ajuste o modelo linearizável

$$y = \beta x^{\gamma} e$$

em que y representa o preço e x representa a área e avalie a qualidade do ajuste comparativamente ao modelo linear ajustado no item ii); construa um gráfico de dispersão com os dados transformados.

iv) Utilizando o modelo com o melhor ajuste, construa intervalos de confiança com coeficiente de confiança (aproximado) de 95% para os preços esperados de imóveis com $200m^2$, $500m^2$ e $1000m^2$.

C.9.12. O arquivo Bosco (1998), disponível na forma de uma planilha Excel no sítio www.ime.usp.br∼jmsinger contém dados provenientes de um estudo observacional baseado numa amostra de 66 pacientes matriculadas na Clínica Ginecológica do Departamento de Obstetrícia e Ginecologia do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo e no Setor de Mamas do Centro de Referência da Saúde da Mulher e de Nutrição, Alimentação e Desenvolvimento Infantil entre novembro de 1995 e outubro de 1997. Um dos objetivos é estudar a relação entre o tamanho clínico de tumores de mama e seu tamanho obtido ultrassonograficamente. O tamanho clínico do tumor é definido como a média dos valores encontrados nas colunas N e O da planilha mencionada; as três medidas ultrassonográficas (altura, comprimento e largura) estão disponíveis nas colunas U, V e W.

i) Construa gráficos de dispersão apropriados para o problema.

ii) Ajuste um modelo de regressão linear múltipla tendo como variável resposta o tamanho clínico do tumor e como variáveis explicativas as três medidas ultrassonográficas.

iii) Avalie a qualidade do ajuste por meio do coeficiente de determinação, testes de hipóteses convenientes, gráficos de resíduos e um gráfico do tipo QQ.

iv) Com base nas conclusões obtidas dos itens acima, verifique se é possível reduzir o modelo (eliminando uma ou duas variáveis explicativas).

v) Repita a análise utilizando a raiz cúbica do volume ultrassonográfico do tumor, definido como

$$Volume = \frac{\pi}{6}(altura \times comprimento \times largura)$$

como única variável explicativa.

vi) Compare o modelo que você julgou mais adequado por meio da análise realizada nos itens i)-iv) com aquele obtido no item v).

C.9.13. Os dados abaixo são provenientes de um estudo cujo objetivo era avaliar a eficácia de um tipo de escova de dentes na remoção de placa bacteriana. Para isso foram observados índices de placa bacteriana (maiores valores do índice correspondendo a maiores quantidades de placa) antes (X) e após (Y) a escovação numa amostra de 26 crianças.

a) Assuma que o par (X, Y) segue uma distribuição normal bivariada. Proponha um modelo que permita a comparação das médias de X e Y, expresse-o em

notação matricial e interprete todos os seus parâmetros. Teste a hipótese de que as médias de X e Y são iguais e construa um intervalo de confiança para a sua diferença. Indique precisamente como devem ser realizados os cálculos.

b) Construa um gráfico de dispersão tendo o índice de placa bacteriana pré-escovação (X) no eixo das abscissas e o índice de placa bacteriana pós-escovação (Y) no eixo das ordenadas.

c) Proponha um modelo de regressão para explicar a variação do índice pós-escovação como função do índice pré-escovação, levando em conta o fato de que índices pré-escovação nulos implicam índices pós-escovação nulos (em média). Explicite as suposições e interprete os parâmetros.

d) Ajuste o modelo proposto e apresente os resultados de forma não técnica.

e) Utilize técnicas de diagnóstico para avaliar o ajuste do modelo.

f) Qual dos modelos você usaria para analisar os dados? Por que?

Índice de placa bacteriana			
pré-escovação	pós-escovação	pré-escovação	pós-escovação
2.18	0.43	1.40	0.24
2.05	0.08	0.90	0.15
1.05	0.18	0.58	0.10
1.95	0.78	2.50	0.33
0.28	0.03	2.25	0.33
2.63	0.23	1.53	0.53
1.50	0.20	1.43	0.43
0.45	0.00	3.48	0.65
0.70	0.05	1.80	0.20
1.30	0.30	1.50	0.25
1.25	0.33	2.55	0.15
0.18	0.00	1.30	0.05
3.30	0.90	2.65	0.25

C.9.14. Considere o modelo $y_i = \alpha + \beta x_i + e_i$, $i = 1, \ldots, 20$ em que $e_i = \rho e_{i-1} + u_i$ com $u_i \sim N(0, 1)$ e assuma que $\alpha = 2$, $\beta = 0.5$, $e_0 = 3$ e $\rho = 0.9$.

a) Utilizando um gerador de números aleatórios obtenha 20 valores de u_i e construa os valores correspondentes de e_i, $i = 1, \ldots, 20$.

b) Construa um gráfico de e_i em função de i.

c) Obtenha os valores de y_i para $x_i = i, \quad i = 1, \ldots, 20$ e construa o gráfico de dispersão correspondente incluindo nele a reta $E(y_i|x_i) = \alpha + \beta x_i$.

d) Obtenha os estimadores de mínimos quadrados de α e β a partir dos dados gerados no item c) e inclua a reta estimada no gráfico do item c).

e) Calcule a estatística de Durbin-Watson e discuta os resultados, interpretando o efeito da autocorrelação dos erros.

f) Repita os itens c) - e) com $\rho = 0.5$ e $\rho = 0.1$

g) Compare os resultados obtidos com os diferentes valores de ρ e comente as diferenças encontradas.

C.9.15. Os dados do arquivo intitulado Singer&Andrade (1997), disponíveis na forma de uma planilha Excel no sítio www.ime.usp.br/∼jmsinger são provenientes de um estudo cujo objetivo era avaliar a eficácia de dois tipo de escova de dentes (Hugger e Convencional) na remoção de placa bacteriana. Para isso foram observados índices de placa bacteriana (maiores valores do índice correspondendo a maiores quantidades de placa) antes (X) e após (Y) a escovação com cada tipo de escova numa amostra de $n_1 = 14$ crianças do gênero feminino (F) e $n_2 = 12$ do gênero masculino (M).

a) Construa um gráfico de dispersão para cada tipo de escova, tendo o índice de placa bacteriana pré-escovação (X) no eixo das abscissas e o índice de placa bacteriana pós-escovação (Y) no eixo das ordenadas. Use símbolos diferentes para identificar crianças de cada gênero.

b) Utilize métodos de regressão linear simples para ajustar modelos do tipo

$$Y_{ij} = \beta_i X_{ij}^{\gamma_i} e_{ij},$$

$i = 1, 2,\ j = 1, \ldots, n_i$ para cada tipo de escova separadamente, indicando as suposições adotadas.

c) Teste a hipótese de que $\gamma_1 = \gamma_2$ e no caso de não rejeição, reajuste o modelo com $\gamma_1 = \gamma_2 = \gamma$.

d) No modelo reduzido, teste a hipótese de que $\gamma = 1$ e em caso de não rejeição, ajuste um novo modelo reduzido que incorpore esse resultado.

e) No terceiro modelo reduzido, teste a hipótese de que $\beta_1 = \beta_2 = \beta$ e, se for o caso, ajuste um novo modelo que incorpore o resultado.

f) Avalie a qualidade do ajuste do último modelo ajustado por intermédio de técnicas de diagnóstico.

g) Compare os resultados das análises dos dados das duas escovas.

h) Escreva sua conclusão, interpretando as hipóteses testadas e construindo um tabela com valores esperados para os índices de placa pós-escovação com diferentes níveis de índices de placa pré-escovação.

C.9.16. Considere o modelo

$$y_{ij} = \alpha + \beta x_{ij} + a_i + e_{ij},$$

$i = 1, \ldots, n, j = 1, \ldots, m$ em que y_{ij}, α, β e x_{ij} têm as interpretações usuais e as variáveis aleatórias $a_i \sim N(0, \sigma_a^2)$, $e_{ij} \sim N(0, \sigma^2)$ são independentes.

i) Mostre que $Var(y_{ij}) = \sigma_a^2 + \sigma^2$, $Cov(y_{ij}, y_{ij'}) = \sigma_a^2$, $j \neq j'$ e que $Cov(y_{ij}, y_{i'j'}) = 0$, $i \neq i'$.

ii) Escreva o modelo descrito acima na forma matricial, especificando a matriz de covariâncias do vetor de erros, *i.e.*, do vetor cujos componentes são $a_i + e_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

iii) Assuma que tanto σ_a^2 quanto σ^2 são conhecidas e obtenha os estimadores de mínimos quadrados generalizados e máxima verossimilhança dos demais parâmetros do modelo e suas respectivas distribuições.

C.9.17. A Tabela C.1.2 contém dados de capacidade instalada (ton), potência instalada ($1000\ kW$) e área construída ($100\ m^2$) de 10 empresas de uma certa indústria. Com o objetivo de estimar a capacidade instalada (Y) a partir das informações sobre potência instalada (X_1) e área construída (X_2),

i) Construa gráficos de dispersão apropriados.

ii) Especifique um modelo de regressão linear e interprete seus parâmetros.

iii) Obtenha estimativas dos parâmetros e de seus erros padrões e calcule o coeficiente de determinação múltiplo.

iv) Avalie o ajuste do modelo por meio de gráficos de resíduos e gráficos QQ.

v) Avalie a perda de precisão dos estimadores decorrente do uso de cada uma das variáveis explicativas em modelos de regressão linear simples construídos com o mesmo propósito do modelo descrito no item ii).

vi) Com base no modelo mais adequado dentre aqueles que você analisou, construa uma tabela com intervalos de confiança (coeficiente de confiança = 95%) para as capacidades de produção esperadas de empresas com todas as combinações de potências instaladas de 1.0, 2.5 e 5.0 ($\times$ 1000 kW) e áreas construídas 8.0, 10.0 e 12.0 ($\times$ 100 m^2).

C.9.18. Os dados do arquivo intitulado Braga (1998), disponível na forma de uma planilha Excel no sítio www.ime.usp.br/∼jmsinger são oriundos de um estudo realizado na Faculdade de Medicina da Universidade de São Paulo para avaliar pacientes com insuficiência cardíaca. Foram estudados 87 pacientes com algum nível de insuficiência cardíaca, além de 40 pacientes controle (coluna K). Para cada paciente foram registradas algumas características físicas (altura, peso, superfície corporal, idade, sexo). Eles foram submetidos a um teste de esforço cardiopulmonar em cicloergômetro em que foram medidos a frequência cardíaca, o consumo de oxigênio, o equivalente ventilatório de oxigênio, o equivalente ventilatório de dióxido de carbono, o pulso de oxigênio e a pressão parcial de dióxido de carbono ao final da expiração, em três momentos diferentes: no limiar anaeróbio, no ponto de compensação respiratória e no pico do exercício.

Ajuste um modelo linear que permita comparar a relação entre o consumo de oxigênio no limiar anaeróbio do exercício (coluna X) e a carga na esteira ergométrica (coluna U) para pacientes com diferentes níveis de insuficiência cardíaca (medida segundo a classificação NYHA - coluna K). Com essa finalidade, você deve:

a) Construir gráficos de dispersão convenientes.

b) Interpretar os diferentes parâmetros do modelo.

c) Estimar os parâmetros do modelo e apresentar os respectivos erros padrões.

d) Avaliar a qualidade de ajuste do modelo por meio de gráficos de resíduos e gráficos QQ.

e) Identificar possíveis valores discrepantes (*outliers*) e reajustar o modelo sem esses pontos.

f) Comparar os ajustes dos modelos obtidos com e sem os *outliers*.

g) Definir e testar hipóteses adequadas para avaliar se a relação entre consumo de oxigênio no limiar anaeróbio do exercício e carga na esteira ergométrica depende da classificação NYHA.

h) Reajustar o modelo com base nas conclusões do item (g) e avaliar o seu ajuste.

i) Apresentar conclusões que evitem jargão técnico.

C.9.19. As tabelas abaixo foram obtidas da análise de um conjunto de dados.

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	4.1003	2.8355	1.446	0.1862
w	-0.8334	0.2640	-3.157	0.0135 *
x	2.5496	0.3151	8.093	4.02e-05 ***

Residual standard error: 2.153 on 8 degrees of freedom Multiple R-Squared: 0.905, Adjusted R-squared: 0.8813 F-statistic: 38.11 on 2 and 8 DF, p-value: 8.141e-05

	(Intercept)	w	x
(Intercept)	1.46702786	-0.08622291	-0.11625387
w	-0.08622291	0.03978328	-0.01996904
x	-0.11625387	-0.01996904	0.03142415

a) Quantas variáveis explicativas e quantas observações (n) foram utilizadas na análise?

b) Especifique o modelo adotado.

c) Há alguma evidência de que o modelo se ajusta bem aos dados? Justifique.

d) Fixe um valor (entre 1 e 10) para cada variável explicativa e construa um intervalo de confiança com coeficiente de confiança de 95% para o valor esperado da resposta correspondente aos valores fixados para as variáveis explicativas.

C.9.20. Para estudar a associação entre gênero (1=Masc, 0=Fem) e idade (anos) e a preferência (1=sim, 0=não) pelo refrigerante Kcola, o seguinte modelo de regressão logística foi ajustado aos dados de 50 crianças escolhidas ao acaso:

$$\log\left\{\frac{\pi_i(x_i, w_i)}{1-\pi_i(x_i, w_i)}\right\} = \alpha + \beta x_i + \gamma(w_i - 5),$$

em que x_i (w_i) representa o gênero (idade) da i-ésima criança e $\pi_i(x_i, w_i)$ a probabilidade de uma criança do gênero x_i e idade w_i preferir Kcola. As seguintes estimativas para os parâmetros foram obtidas:

Parâmetro	Estimativa	Erro-padrão	Valor p
α	0.69	0.12	< 0.01
β	0.33	0.10	< 0.01
γ	-0.03	0.005	< 0.01

a) Interprete os parâmetros do modelo por intermédio de chances e razões de chances.

b) Com as informações acima, estime a razão de chances de preferência por Kcola correspondente à comparação de crianças do mesmo gênero com 10 e 15 anos.

c) Construa intervalos de confiança (com coeficiente de confiança aproximado de 95%) para $\exp(\beta)$ e $\exp(\gamma)$ e traduza o resultado em linguagem não técnica.

d) Estime a probabilidade de meninos com 15 anos preferirem Kcola.

C.9.21. No arquivo Singer&Ikeda (1996) disponível em www.ime.usp.br/∼jmsinger você encontra dados provenientes de um estudo cuja finalidade é identificar fatores de risco para a doença aterosclerótica coronariana (definida como obstrução de mais de 50% de pelo menos uma coronária).

a) Utilize modelos de regressão logística para verificar se a presença de angina estável (ANGEST), antecedentes hereditários (AH), infarto do miocárdio prévio (IMP), nível de triglicérides para pacientes sem medicamento (TRIGS), nível de colesterol para pacientes sem medicamento (COLS), idade (IDADE1) e sexo (SEXO) podem ser consideradas como fatores de risco para a doença aterosclerótica coronariana (LO3). Considere somente os participantes com dados completos para as variáveis indicadas.

b) Com base no modelo selecionado, construa uma tabela com limites inferiores e superiores de intervalos de confiança (com coeficiente de confiança de 95%) para razões de chances definidas a partir de um conjunto de valores pré-especificados das variáveis explicativas.

C.9.22. Para avaliar a relação entre o tempo de uso (X) e o número de defeitos (Y) de um determinado componente eletrônico, obteve-se uma amostra de n itens desse tipo e em cada um observaram-se as duas variáveis. O estatístico encarregado da análise concluiu que o número de defeitos segue uma distribuição de Poisson, *i.e.,* que $P(Y = y) = \exp(-\mu)\mu^y/y!$, $y = 0, 1, \ldots$ e propôs um modelo linear sem intercepto para a relação entre o número esperado de defeitos e o tempo de uso, *i.e.*, $\mu = \beta x$.

a) Obtenha o estimador de máxima verossimilhança do parâmetro do modelo.

b) Mostre que esse estimador é não enviesado.

c) Obtenha a variância desse estimador.

d) Construa um intervalo de confiança com coeficiente de confiança aproximado de 95% para o valor esperado do número de defeitos de componentes com tempo de uso igual a x_0.

C.9.23. Os dados abaixo correspondem ao faturamento de empresas similares de um mesmo setor industrial nos últimos 15 meses.

mês	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez	jan	fev	mar
vendas	1.0	1.6	1.8	2.0	1.8	2.2	3.6	3.4	3.3	3.7	4.0	6.4	5.7	6.0	6.8

Utilize técnicas de análise de regressão para quantificar o crescimento do faturamento de empresas desse setor ao longo do período observado. Com essa finalidade:

a) Proponha um modelo adequado, interpretando todos os parâmetros e especificando as suposições.

b) Estime os parâmetros do modelo e apresente os resultados numa linguagem não técnica.

c) Utilize técnicas de diagnóstico para avaliar o ajuste do modelo.

C.9.24. Obtenha os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão

$$y_i = \beta x_i + e_i,$$

$i = 1, \ldots, n$ em que os e_i representam erros aleatórios independentes normais com média zero e variância σ^2.

i) Obtenha a distribuição do estimador de β e construa um intervalo de confiança.

ii) Repita o item i) com a suposição adicional de que $Var(e_i) = x_i^2\sigma^2$.

C.9.25. Utilizando a notação usual, considere o modelo linear

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} E(y_{11}) \\ E(y_{12}) \\ E(y_{21}) \\ E(y_{22}) \end{pmatrix}$$

a) Expresse os parâmetros do modelo em termos dos valores esperados $E(y_{ij})$.

b) Repita o procedimento do item anterior sob as restrições $\alpha_1 = 0$ e $\alpha_1 + \alpha_2 = 0$.

c) Repita o procedimento agora sob a reparametrização $\beta_1 = \mu + \alpha_1$ e $\beta_2 = \mu + \alpha_2$.

d) Repita o procedimento agora sob a reparametrização $\beta_1 = \mu$ e $\beta_2 = \mu + \alpha_1$.

Interprete os parâmetros em cada caso.

C.9.26. A tabela abaixo contém dados obtidos de diferentes institutos de pesquisa coletados entre fevereiro de 2008 e março de 2010 e correspondem às porcentagens de eleitores favoráveis a cada um dos dois principais candidatos à presidência do Brasil.

a) Construa um diagrama de dispersão apropriado, evidenciando os pontos correspondentes a cada um dos candidatos.

b) Especifique um modelo polinomial de segundo grau, homocedástico, que represente a variação da preferência eleitoral de cada candidato ao longo do tempo.

c) Ajuste o modelo especificado no item anterior.

d) Avalie o ajuste do modelo e verifique, por meio de testes de hipóteses adequadas, se ele pode ser simplificado; em caso afirmativo, ajuste o modelo mais simples.

e) Com base no modelo escolhido, estime a porcentagem esperada de eleitores favoráveis a cada um dos candidatos em 3 de outubro de 2010 e construa um intervalo de confiança para a diferença entre essas porcentagens esperadas.

f) Faça uma crítica da análise e indique o que poderia ser feito para melhorá-la (mesmo não que não saiba implementar suas sugestões).

Porcentagem de eleitores favoráveis

Fonte	Data	Dilma	Serra	Fonte	Data	Dilma	Serra
sensus	16/02/2008	4.5	38.2	sensus	13/08/2009	19	39.5
dataf	27/03/2008	3	38	ibope	04/09/2009	14	34
sensus	25/04/2008	6.2	36.4	sensus	14/09/2009	21.7	31.8
sensus	19/09/2008	8.4	38.1	ibope	20/11/2009	17	38
dataf	28/11/2008	8	41	vox	30/11/2009	17	39
sensus	30/11/2008	10.4	46.5	vox	07/12/2009	18	39
ibope	12/12/2008	5	42	dataf	14/12/2009	23	37
sensus	14/12/2008	13.3	42.8	vox	18/12/2009	27	34
dataf	30/01/2009	11	41	sensus	17/01/2010	27	33.2
sensus	19/03/2009	16.3	45.7	ibope	29/01/2010	25	36
dataf	27/03/2009	16	38	dataf	06/02/2010	28	32
sensus	28/05/2009	23.5	40.4	ibope	25/02/2010	30	35
ibobe	29/05/2009	18	38	dataf	27/03/2010	27	36
dataf	01/06/2009	17	36	vox	31/03/2010	31	34

C.9.27. Os dados da tabela abaixo foram obtidos de um estudo cujo objetivo era avaliar a relação entre a quantidade de um certo aditivo (X) e o tempo de vida (Y) de um determinado alimento. Os valores substituídos por ? ficaram ilegíveis depois que o responsável pelo estudo derramou café sobre eles.

X (g/kg)	5	10	15	20	30
Y (dias)	3.2	?	?	?	?

Um modelo de regressão linear simples (com a suposição de normalidade e independência dos erros) foi ajustado aos dados gerando os seguintes resultados:

Tabela de ANOVA

Fonte de variação	*gl*	*SQ*	*QM*	*F*	*Valor p*
Regressão	1	42,30	42,30	156,53	0,001
Resíduo	3	0,81	0,27		
Total	4	43,11			

Intervalos de confiança (95%)

Parâmetro	*Limite inferior*	*Limite superior*
Intercepto	-0,19	2,93
X	0,25	0,42

Resíduos

Observação	*Resíduos*
1	0,14
2	-0,55
3	0,66
4	-0,23
5	-0,01

$$(\mathbf{X}^t\mathbf{X})^{-1} = \begin{pmatrix} 0.892 & -0.043 \\ -0.043 & 0.003 \end{pmatrix}$$

a) Escreva o modelo na forma matricial e interprete seus parâmetros.

b) Construa um intervalo de confiança para o valor esperado do tempo de vida do produto quando a quantidade de aditivo utilizada é de 25 g/kg.

c) Construa um intervalo de previsão para o valor do tempo de vida do produto quando a quantidade de aditivo utilizada é de 25 g/kg.

d) Reconstrua a tabela dos dados, *i.e.*, calcule os valores de Y substituídos por ?.

Observação: O quantil de ordem $97,5\%$ da distribuição t com 3 graus de liberdade é 3.182.

C.9.28. Os dados abaixo são provenientes de uma pesquisa cujo objetivo é avaliar o efeito da dosagem de uma certa droga (X) na redução de pressão arterial (Y) de pacientes hipertensos.

Homens		**Mulheres**	
Dose	Redução de pressão	Dose	Redução de pressão
1	3	2	4
3	5	3	7
4	9	5	11
6	15	6	14
		6	13

O pesquisador sugeriu o seguinte modelo para a análise dos dados

$$y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}) + e_{ij}$$

$i = 1, 2,\ j = 1, \ldots, n_i$ em que os erros e_{ij} seguem distribuições $N(0, \sigma^2)$ independentes e $\bar{x}$ denota a dose média empregada no estudo.

a) Interprete os parâmetros do modelo.

b) Escreva o modelo na forma matricial.

C.9.29. Uma determinada empresa de transportes deseja saber se o número de passageiros está relacionado com o preço da gasolina. Com essa finalidade, adotaram o modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

em que y corresponde ao número médio de passageiros no mês, x_1 representa a variação no preço médio mensal da gasolina relativamente à média do ano anterior, x_2 é uma variável indicadora do tipo de percurso do ônibus (0 = expresso e 1 = usual) e e é um erro aleatório de média zero.

a) Interprete os parâmetros do modelo.

b) Que hipótese você testaria para avaliar se a relação entre o número médio mensal de passageiros (y) e a variação no preço da gasolina (x_1) é igual para ônibus com os dois tipos de percurso.

c) Suponha que as estimativas de mínimos quadrados dos parâmetros sejam $\widehat{\beta}_0 = 500$, $\widehat{\beta}_1 = 50$, $\widehat{\beta}_2 = 5$ e $\widehat{\beta}_3 = -10$. Esboce um gráfico que represente a relação entre y e x_1 para ônibus com cada tipo de percurso.

C.9.30. Num estudo cujo objetivo era avaliar o efeito do sexo (X) e da idade (Y) de indivíduos no envolvimento em acidentes automobilísticos foram coletadas informações de sobre essas três variáveis para uma amostra de clientes de uma empresa de seguros.

a) Proponha um modelo de regressão logística para representar a associação entre as três variáveis e interprete os seus parâmetros.

b) Calcule a chance de um homem com 40 anos se envolver num acidente.

c) Por que valor fica multiplicada essa chance (de envolvimento em acidentes) para uma mulher com 50 anos?

C.9.31. Uma fábrica de cadeiras dispõe dos seguintes dados sobre sua produção mensal:

Número de cadeiras produzidas	105	130	141	159	160	172
Custos fixos e variáveis (R$)	1700	1850	1872	1922	1951	1970

a) Proponha um modelo de regressão linear simples para a relação entre o custo e o número de cadeiras produzidas e interprete os parâmetros;

b) Utilize um intervalo de confiança com coeficiente de confiança de 95% para estimar o custo esperado de produção para 200 cadeiras;

c) Admitindo que o preço de venda é de R$ 20.00 por unidade, qual a menor quantidade de cadeiras que deve ser produzida para que o lucro seja positivo.

C.9.32. Considere o seguinte modelo de regressão

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + e_i, & x_i \leq x_0, \quad i = 1, \ldots, j \\ \beta_0 + \beta_1 x_0 + \beta_2 (x_i - x_0) + e_i, & x_i \geq x_0, \quad i = j+1, \ldots, n. \end{cases}$$

em que os termos y_i, x_i, e_i, têm a interpretação usual e x_0 é uma constante positiva. Escreva o modelo com notação matricial e interprete os parâmetros β_0, β_1, e β_2.

C.9.33. Numa pesquisa realizada na Faculdade de Medicina da Universidade de São Paulo, foi observada uma amostra de 32 recém-nascidos pré-termo (RNPT), vulgarmente chamados de prematuros. Em cada um deles foram medidos o diâmetro da aorta (em mm) e a idade (em semanas) desde a concepção (instante em que houve a fecundação). O objetivo era verificar se 39 semanas após a concepção, o diâmetro médio da aorta dos RNPT era equivalente ao diâmetro médio da aorta de recém-nascidos a termo (RNT), vulgarmente conhecidos como normais. Para isso, foi ajustado um modelo de regressão linear simples, cujos resultados estão indicados abaixo.

Coeficiente	Estimativa	Erro padrão
Intercepto	8.42	0.20
Angular	-0.12	0.02

$$(\mathbf{X}^t\mathbf{X})^{-1} = \begin{pmatrix} 0.25 & -0.0125 \\ -0.0125 & 0.0025 \end{pmatrix}$$

em que $\mathbf{X}$ denota a matriz com os valores das variáveis explicativas.

a) Obtenha uma estimativa da variância dos erros (σ^2).

b) Estime o valor esperado do diâmetro da aorta para 39 semanas pós-concepção e estime o seu erro padrão.

c) Sabendo que o diâmetro médio (populacional) da aorta de RNT (39 semanas após a concepção) é de 3.23 mm, responda a pergunta que originou o estudo, justificando sua resposta.

d) Repita a análise do item anterior no caso de o valor do diâmetro médio da aorta de RNT (3.23 mm) ser proveniente de uma amostra de tamanho 30 e que o erro padrão correspondente é de 0.50 mm.

Observações:

i) Utilize um nível de significância de 5% em todas as suas análises.

ii) Utilize uma aproximação normal para as distribuições envolvidas nos métodos inferenciais que você utilizar.

iii) Indique claramente como os cálculos foram realizados.

C.9.34. Os dados da tabela abaixo foram obtidos de empresas de duas indústrias (Alimentos e Produtos de Limpeza) com o objetivo de avaliar a relação entre o valor investido em propaganda (X) num certo trimestre e o aumento de faturamento no trimestre subseqüente (Y).

Indústria	Valor investido em propaganda	Aumento de faturamento
Alimentos	5	20
Alimentos	7	25
Alimentos	9	28
Alimentos	10	27
Alimentos	8	26
Alimentos	6	22
Limpeza	4	16
Limpeza	11	19
Limpeza	7	17
Limpeza	9	16
Limpeza	6	15

a) Especifique um modelo em que a relação entre X e Y é quadrática, com coeficientes possivelmente diferentes para cada indústria. Interprete os parâmetros do modelo.

b) Escreva o modelo na forma matricial.

c) Utilizando notação matricial, especifique a hipótese de que as curvas correspondentes à relação entre X e Y são iguais para as duas indústrias.

C.9.35. Com a finalidade de comparar homens (H) e mulheres (M) quanto à relação entre o tempo gasto entre a chegada e a saída em um centro de compras, (X) e o valor dispendido (Y), considerou-se o modelo:

$$y_{ij} = \mu + \alpha_i + \gamma(x_{ij} - x_0) + e_{ij},$$

$\sum_{i=1}^{2} \alpha_i = 0, \quad i = 1\ (M), 2(H),\ j = 1, \ldots, n_i$ em que as variáveis aleatórias e_{ij} têm distribuições $N(0, \sigma^2)$ independentes.

a) Descreva todos os símbolos utilizados, interpretando-os.

b) Especifique o modelo na forma matricial.

c) Utilizando notação matricial, especifique a hipótese de que a relação entre Y e X é igual para homens e mulheres.

d) Esboce um gráfico que represente o modelo.

C.9.36. Num estudo cujo objetivo era avaliar a associação entre atividade (técnica ou administrativa), tempo de serviço (anos) e salário (R$ 1000.00) e a participação em um programa de demissão voluntária (sim ou não), foram coletadas informações sobre essas quatro variáveis para uma amostra de empregados de uma grande empresa.

a) Proponha um modelo de regressão logística para representar a associação entre as variáveis e interprete os seus parâmetros.

b) Com base no modelo proposto, obtenha uma expressão para cálculo da chance de um empregado do setor administrativo com 5 anos de serviço e ganhando um salário de R$ 3000.00 participar do programa de demissão voluntária.

c) Segundo o modelo proposto, por que valor fica multiplicada essa chance para um funcionário da área técnica com salário de R$ 2000.00 e de mesma idade?

C.9.37. A tabela abaixo contem dados de uma investigação cujo objetivo era estudar a relação entre a duração de diabete e a ocorrência de retinoplastia (uma moléstia dos olhos).

a) Considere um modelo linear para avaliar a intensidade dessa relação e utilize o método de mínimos quadrados generalizados para ajustá-lo. **Sugestão**: Considere o ponto médio de cada intervalo como valor da variável explicativa e use as frequências relativas observadas para estimação das variâncias.

b) Utilize o método de máxima verossimilhança para ajustar um modelo linear e um modelo logístico aos dados.

c) Compare os resultados do ajuste do modelos lineares obtidos pelos dois métodos por meio de uma tabela com as estimativas e erros padrões dos parâmetros.

Duração da	Retinoplastia	
Diabete (anos)	Sim	Não
0 - 2	17	215
3 - 5	26	218
6 - 8	39	137
9 - 11	27	62
12 - 14	35	36
15 - 17	37	16
18 - 20	26	13
21 - 23	23	15

C.9.38. Num estudo em que se desejava comparar um tipo de pneu experimental ($i = 2$) com um tipo de pneu convencional ($i = 1$) com relação ao desgaste Y (profundidade do sulco em mm) em função da distância percorrida X (em 1000 km), foram observados $n_1 = 4$ pneus convencionais e $n_2 = 5$ pneus convencionais. Os dados estão dispostos abaixo.

Tipo de pneu	Distância percorrida	Profundidade do sulco
Convencional	10	99
Convencional	20	95
Convencional	30	72
Convencional	40	56
Experimental	10	93
Experimental	20	86
Experimental	25	83
Experimental	30	77
Experimental	40	68

O estatístico responsável pela análise propôs o seguinte modelo

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + e_{ij},$$

$i = 1, 2,\ j = 1, \ldots, m_i$ com $\alpha_1 = 0$ e $e_{ij} \sim N(0, \sigma^2)$, independentes.

a) Escreva o modelo na forma matricial e interprete seus parâmetros.

b) Represente o modelo graficamente indicando claramente o significado dos parâmetros.

c) Expresse em termos matriciais, a hipótese de que os dois tipos de pneus têm desgaste esperado equivalente. Especifique a distribuição da estatística que você utilizaria para testar essa hipótese.

C.9.39. Considere o modelo

$$y_i = \alpha x_i + e_i,$$

$i = 1, \ldots, n$ em que $e_i \sim N(0, \sigma^2)$ são variáveis aleatórias independentes.

a) Obtenha o estimador de máxima verossimilhança de α e proponha um estimador não enviesado para σ^2.

b) Especifique a distribuição do estimador de α.

c) Especifique um intervalo de confiança para o parâmetro α com coeficiente de confiança γ, $0 < \gamma < 1$.

Observação: a função densidade da distribuição $N(\mu, \sigma^2)$ é

$$f(x) = (\sqrt{2\pi}\sigma)^{-1} exp(-2^{-1}[(x-\mu)/\sigma]^2).$$

C.9.40. Considere o modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ em que $\mathbf{y}$ é um vetor de respostas com dimensão $(n_1 + n_2) \times 1$, $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_1+n_2})$, $\boldsymbol{\beta}$ é um vetor de parâmetros,

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{bmatrix},$$

com $\mathbf{I}_m$ representando uma matriz identidade dimensão m, $\mathbf{1}_m$, um vetor de dimensão m com todos os elementos iguais a 1 e $\mathbf{0}$ um vetor com todos os elementos iguais a zero.

a) Interprete os parâmetros do modelo.

b) Expresse o estimador de mínimos quadrados dos elementos de $\boldsymbol{\beta}$ em termos de somatórios.

c) Expresse um estimador não enviesado de σ^2 em termos de somatórios .

d) Dê um exemplo de situação prática em que esse modelo poderia ser aplicado.

C.9.41. As tabelas abaixo foram obtidas da análise de um conjunto de dados.

Ajuste do Modelo

Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	13.32	6.93	1.92	0.084
x_1	4.37	0.81	5.42	0.000
x_2	-22.60	5.46	-4.14	0.002
x_3	-7.36	5.46	-1.35	0.208

Multiple R-Squared: 81.02%, Adjusted R-squared: 75.06%

ANOVA

Source	DF	SS	MS	F	p
Regression	3	2807.90	935.97	14.42	0.001
Error	10	649.09	64.91		
Total	13	3456.99			

a) Quantas variáveis explicativas e quantas observações (n) foram utilizadas na análise?

b) Especifique o modelo adotado e interprete seus parâmetros.

c) Apresente uma estimativa para a variância dos erros baseada num estimador não enviesado.

d) Indique como se calcula o coeficiente R^2 a partir das tabelas acima.

e) Especifique (em termos dos parâmetros do modelo) a hipótese testada por meio da estatística F da tabela de ANOVA e indique sua distribuição.

f) Especifique (em termos dos parâmetros do modelo) as hipóteses testadas por meio das estatísticas t da tabela de ajuste do modelo e indique suas distribuições.

g) Esclareça as diferenças entre as hipóteses consideradas nos itens e) e f).

h) Com base nos resultados da análise, proponha um modelo mais simples para os dados, justificando sua resposta.

C.9.42. Reproduza todas as análises do Exemplo C.8.1 com e sem os pontos associados ao adesivo B e nível de umidade 4.50 adotando as parametrizações correspondentes aos seguintes comandos da função lm() do pacote R:

a) lm(indice $\sim$ adesivo + umidade + umidade2 + adesivo:umidade + adesivo:umidade2)

b) lm(indice $\sim$ adesivo + umidade + umidade2 + adesivo:umidade + adesivo:umidade2 -1)

com umidade2 = umidade^2. Em cada caso,

i) especifique e interprete os parâmetros do modelo;

ii) mostre como você obteve os resultados apresentados no Exemplo C.8.1;

iii) avalie o impacto da eliminação dos dados mencionados nas estimativas dos parâmetros;

iv) compare os coeficientes de determinação dos dois modelos e explique possíveis diferenças.

C.9.43. Os dados disponíveis em

www.ime.usp.br/$\sim$jmsinger/Dados/Singer&Nobre&Rocha2018exercc943.xls

são provenientes de um estudo conduzido no Instituto de Biociências da Universidade de São Paulo com o objetivo de avaliar o efeito do número de malárias contraídas durante a gestação em algumas características de recém nascidos. Avalie o efeito de quantidade de malárias (qntmal, coluna C), idade da mãe [idade (anos), coluna E], peso da mãe na triagem [pesotriag (kg), coluna H], peso da mãe no parto [pesoparto (kg), coluna I]), altura da mãe [altura (cm), coluna J] e idade gestacional [ig (sem), coluna K] no peso do recém nascido [peso (g), coluna M] por meio de modelos de regressão múltipla. Com essa finalidade,

a) construa gráficos de dispersão e *boxplots* para descrever o comportamento da variável peso;

b) explicite a estratégia de análise empregada;

c) ajuste os modelos adotados para concretizar a estratégia adotada;

d) utilize técnicas de diagnóstico para avaliar a qualidade dos ajustes;

e) apresente os resultados na forma de um relatório descrevendo cada passo da análise e a sua conclusão, quantificando-a.

C.9.44. Considere o modelo

$$y_{ij} = \alpha + \beta t_j + a_i + b_i t_j + e_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m_i,$$

em que y_{ij} é a resposta da i-ésima unidade amostral no j-ésimo tempo, α e β são efeitos fixos, $\mathbf{b}_i = (a_i, b_i)^\top$ são efeitos aleatórios independentes com distribuição $N(\mathbf{0}, \mathbf{G})$ e e_{ij} são erros aleatórios independentes com distribuição $N(0, \sigma^2)$. Suponha que $\mathbf{b}_i$ e e_{ij} são independentes. Obtenha expressões para a variância de y_{ij} e para a covariância entre y_{ij} e y_{il}, $j \neq l$ nos seguintes casos:

a) $\mathbf{G} = diag[\sigma_a^2, \sigma_b^2]$

b) $\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$

Bibliografia

Andrews, D. F. & Pregibon, D. (1978). Finding outliers that matter, *Journal of the Royal Statistical Society B* **40**: 85–93.

Anscombe, F. J. (1973). Graphs in statistical analysis, *The American Statistician* **27**: 17–21.

Atkinson, A. C. (1981). Two graphical display for outlying and influential observations in regression, *Biometrika* **68**: 13–20.

Atkinson, A. C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*, Oxford: Oxford University Press.

Atkinson, A. C. & Riani, M. (2000). *Robust diagnostic regression analysis*, New York: Springer.

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*, New York: Wiley.

Bickel, P. J. & Doksum, K. A. (2001). *Mathematical Statistics, Volume 1*, 2 edn, Upper Saddle River, NJ: Prentice-Hall.

Chatterjee, S. & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression (with discussion), *Statistical Science* **1**: 379–393.

Chatterjee, S. & Hadi, A. S. (1988). *Sensitivity analysis in linear regression*, New York: Wiley.

Chesher, A. (1991). The effect of measurement error, *Biometrika* **78**: 451–462.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics* **19**: 15–18.

Cook, R. D. (1986). Assessment of local influence (with discussion), *Journal of the Royal Statistical Society B* **48**: 133–169.

Cook, R. D. (1996). Added variables plots and curvature in linear regression, *Technometrics* **38**: 275–278.

Cook, R. D., Peña, D. & Weisberg, S. (1988). The likelihood: a unifying principle for influence measures, *Communications in Statistics, Theory and Methods* **17**: 623–640.

Cook, R. D. & Weisberg, S. (1982). *Residuals and influence regression*, New York: Chapman and Hall.

Cook, R. D. & Weisberg, S. (1989). Regression diagnostics with dynamic graphics, *Technometrics* **31**: 277–311.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals (with discussion), *Journal of the Royal Statistical Society B* **30**: 248–275.

Diggle, P. J., Heagerty, P., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of longitudinal data*, 2 edn, Oxford: Oxford University Press.

Durbin, J. & Watson, G. S. (1950). Testing for serial correlation in least squares regression, I, *Biometrika* **37**: 409–428.

Durbin, J. & Watson, G. S. (1951). Testing for serial correlation in least squ17-21ares regression, II, *Biometrika* **38**: 159–178.

Durbin, J. & Watson, G. S. (1971). Testing for serial correlation in least squares regression, III, *Biometrika* **58**: 1–19.

Elias, F. M., Birman, E. G., Matsuda, C. K., Oliveira, I. R. S. & Jorge, W. A. (2006). Ultrasonographic findings in normal temporomandibular joints, *Brazilian Oral Research* **20**: 25–32.

Freedman, D. (2005). *Statistical Models: Theory and Practice*, Cambridge: Cambridge University Press.

Fuller, W. A. (1987). *Measurement error models*, New York: Wiley.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*, New York: Springer.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2 edn, New York: John Wiley.

Kutner, M., Nachtsheim, C. J., Neter, J. & Li, W. (2005). *Applied Linear Statistical Models*, 5 edn, Boston: McGraw-Hill.

Lencina, V. B., Singer, J. M. & Stanek III, E. J. (2005). Much ado about nothing: the mixed models controversy revisited, *International Statistical Review* **73**: 9–20.

Lencina, V. B., Singer, J. M. & Stanek III, E. J. (2008). Response to J.A. Nelder: What is the mixed models controversy?, *International Statistical Review* **76**: 134–139.

Magnus, J. R. & Neudecker, H. (1988). *Matrix differential calculus with applications in Statistics and Econometrics*, New York: Wiley.

Nelder, J. A. (1998). The great mixed-model muddle is alive and flourishing, alas!, *Food quality and preference* **9**: 157–159.

Paula, G. A. (2004). *Modelos de regressão com apoio computacional*, São Paulo: IME/USP.

Paulino, C. D. & Singer, J. M. (2006). *Análise de dados categorizados*, São Paulo: Blucher.

Reis, A., Loguercio, A. D., Azevedo, C. L., Carvalho, R. M., Singer, J. M. & Grande, R. H. (2003). Moisture spectrum of demineralized dentin for adhesive systems with different solvent bases, *Journal of Adhesive Dentistry* **5**: 183–192.

Searle, S. R. (1971). *Linear models*, New York: Wiley.

Searle, S. R. (1982). *Matrix algebra useful for Statistics*, New York: Wiley.

Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear regression*, New York: Wiley.

Sen, P. K., Singer, J. M. & Pedroso-de Lima, A. C. (2009). *From finite sample to asymptotic methods in Statistics*, Cambridge: Cambridge University Press.

Souza, G. S. (1998). *Introdução aos modelos de regressão linear e não-linear*, Brasília: EMBRAPA.

Wei, B. C., Hu, Y. Q. & Fung, W. K. (1998). Generalized leverage and its applications, *Scandinavian Journal of Statistics* **25**: 25–37.

Índice