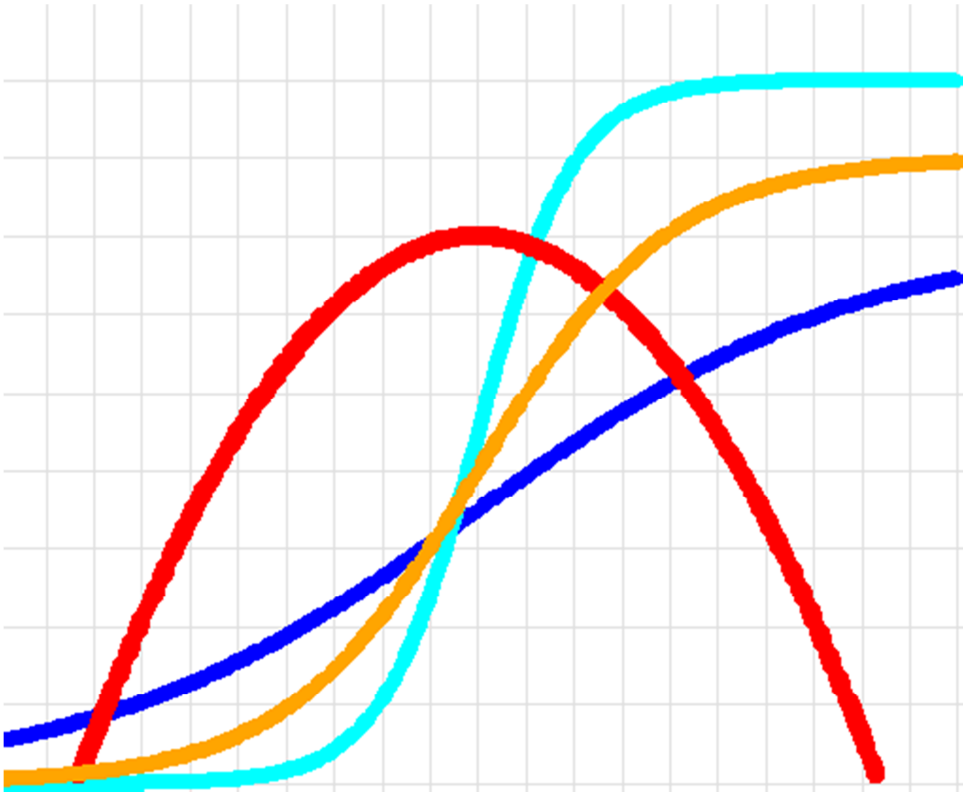
RODOLFO HOFFMANN



ANÁLISE ESTATÍSTICA DE RELAÇÕES LINEARES E NÃO LINEARES

Portal de Livros Abertos

da USP

RODOLFO HOFFMANN

ANÁLISE ESTATÍSTICA DE RELAÇÕES LINEARES E NÃO LINEARES

1ª edição digital

Piracicaba, ESALQ-USP

Edição do Autor

2016

DOI: 10.11606/9788592105716

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Hoffmann, Rodolfo
Análise estatística de relações lineares e não lineares [recurso eletrônico] / Rodolfo Hoffmann. - - Piracicaba: O Autor, 2016.
246 p. : il.

ISBN: 978-85-921057-1-6

1. Análise estatística 2. Relações lineares 3. Relações não lineares

I. Título

CDD 519.5
H711a

DOI: 10.11606/9788592105716

Sumário

INTRODUÇÃO

Objetivo e conteúdo do livro

Serão apresentadas várias técnicas estatísticas para analisar relações lineares e não lineares entre variáveis: regressão linear múltipla, incluindo as maneiras de detectar observações discrepantes ou muito influentes, regressão não linear, modelos para variáveis dependentes binárias (lógite e próbite), análise de componentes principais e análise fatorial.

O objetivo é fazer uma apresentação dessas técnicas, de maneira que o estudante compreenda seus fundamentos estatísticos, podendo avaliar seu potencial de aplicação em vários tipos de pesquisa, e também suas limitações. O texto tem finalidade didática, incluindo exercícios com respostas no final de cada capítulo.

Sempre que possível, são usados exemplos numéricos simples que permitem ao estudante acompanhar, com relativa facilidade, todas as etapas do procedimento estatístico.

Pré-requisitos

Admite-se que o estudante tenha conhecimentos básicos de estatística, incluindo os conceitos de distribuição, função de densidade de probabilidade, esperança matemática, variância e covariância, correlação e teste de hipóteses com base na variável normal reduzida (Z), qui-quadrado, t ou F. É conveniente, também, que o leitor tenha alguma familiaridade com a análise de regressão.

Será utilizada a álgebra matricial, pressupondo-se que o leitor saiba somar, subtrair, multiplicar e transpor matrizes e inverter matrizes quadradas não-singulares. Para algumas demonstrações é necessário conhecer os conceitos de traço de uma matriz quadrada e característica (ou posto) de uma matriz qualquer.

Agradecimentos

É importante reconhecer que as boas condições de trabalho no Instituto de Economia da UNICAMP e na ESALQ-USP foram essenciais para a elaboração desse livro didático. Foi

fundamental, também, o apoio financeiro recebido do CNPq. O autor agradece, ainda, a colaboração de alunos do curso de pós-graduação do IE-UNICAMP e da Profa. Angela Kageyama, que fizeram sugestões para o aperfeiçoamento do texto.

Versão publicada

Uma versão anterior deste livro foi publicada pela LP-Books em julho de 2011 (ISBN 978-85-7869-288-9).

Correções

Se o leitor tiver sugestões para corrigir ou aperfeiçoar o texto ou dúvidas que o autor possa esclarecer, favor escrever para hoffmannr@usp.br.

1. REGRESSÃO LINEAR

1.1. Origem

Neste capítulo é feita uma apresentação concisa da análise de regressão linear[1], que é, certamente, a técnica estatística mais usada em estudos que exigem a análise de relações entre duas ou mais variáveis.

A expressão "análise de regressão" se originou de um artigo em que Sir Francis Galton estuda a relação entre estatura de pais e filhos, publicado em 1886 no *Journal of the Anthropological Institute of Great Britain and Ireland*. O artigo é intitulado "*Regression towards mediocrity in hereditary stature*", refletindo o esnobismo do autor ao constatar que filhos de pais muito altos ou muito baixos tendem a diferir menos da média do que seus genitores.

1.2. O modelo e o estimador de mínimos quadrados

Admitindo que uma variável Y_j é linearmente dependente de *k* variáveis explanatórias $(X_{1j}, X_{2j}, \ldots, X_{kj})$, temos o seguinte modelo de regressão linear múltipla:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + u_j \ , \qquad (1.1)$$

com *j* = 1, 2, 3, ..., *n*. O índice *j* indica uma das *n* observações de uma amostra. O erro u_j é que torna a equação (1.1) um modelo estatístico. O erro u_j pode ser interpretado como o efeito de todas as demais variáveis, com importância secundária, que não foram incluídas no modelo.

É conveniente definir as seguintes matrizes:

a) os vetores-coluna com os *n* valores da variável dependente e com os *n* valores do erro:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \text{e} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

[1] Uma apresentação mais extensa, começando com o caso particular da regressão linear simples (com apenas duas variáveis), pode ser encontrada em livro do mesmo autor (**Análise de Regressão**: uma introdução à econometria).

b) a matriz $n \times (k+1)$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}$$

c) o vetor-coluna com os $p = k+1$ parâmetros:

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Assim, o modelo de uma regressão linear múltipla fica

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1.2}$$

Na versão mais simples do modelo, pressupomos que a matriz **X** é fixa e que os erros u_j têm as seguintes propriedades:

a) $E(u_j) = 0$, ou

$$E(\mathbf{u}) = \mathbf{0} \tag{1.3}$$

b) $V(u_j) = E(u_j^2) = \sigma^2$, variância constante ou homocedasticia.

c) $E(u_j u_h) = 0$ para $h \neq j$, ausência de covariância entre erros de diferentes observações.

Esses dois últimos pressupostos podem ser sintetizados na expressão

$$E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2 \tag{1.4}$$

na qual **I é** uma matriz identidade de ordem *n*. Essa expressão mostra que a matriz de variâncias e covariâncias do vetor **u** é uma matriz diagonal com todos os termos da diagonal iguais a σ^2.

Para fazer os testes de hipóteses com base nos valores de *t* ou *F* é necessário pressupor que os erros u_j têm distribuição normal, ou seja,

$$u_j \sim N(0, \sigma^2)$$

Note-se que é a existência de um termo constante no modelo (representado por α) que torna necessário que a 1ª coluna de **X** tenha todos os seus elementos iguais a 1. Um vetor-coluna de uns será representado por $\boldsymbol{\iota}$ (letra grega iota):

$$\boldsymbol{\iota} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{1.5}$$

Seja **b** um vetor-coluna com as estimativas dos $p = k + 1$ parâmetros da equação (1.1):

$$\mathbf{b} = \begin{bmatrix} a \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

Então as estimativas $(\hat{Y}_j)$ dos valores de Y_j são dadas por

$$\hat{\mathbf{y}} = \mathbf{Xb} \tag{1.6}$$

e o vetor dos desvios (ou resíduos) é

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{Xb} \tag{1.7}$$

Verifica-se que a soma de quadrados dos desvios é

$$S = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}$$

O segundo e o terceiro termo são iguais, pois se trata de matrizes 1×1 e uma é a transposta da outra. Então, podemos escrever

$$S = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \tag{1.8}$$

Dadas as matrizes **X** e **y**, essa expressão mostra como a soma de quadrados dos desvios depende do vetor **b**. Diferenciando, obtemos

$$dS = -2(d\mathbf{b})'\mathbf{X}'\mathbf{y} + (d\mathbf{b})'\mathbf{X}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{X}(d\mathbf{b})$$

Como os dois últimos termos são matrizes 1×1 e uma é a transposta da outra, podemos escrever

$$dS = -2(d\mathbf{b})'\mathbf{X}'\mathbf{y} + 2(d\mathbf{b})'\mathbf{X}'\mathbf{X}\mathbf{b} = 2(d\mathbf{b})'(\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y}) \tag{1.9}$$

De acordo com o método de mínimos quadrados, o estimador **b** deve ser aquele que minimiza o valor da soma de quadrados dos desvios. No ponto de mínimo de *S* como função de **b**, o diferencial *dS* será nulo para qualquer vetor *d***b**. Então, de acordo com (1.9), devemos ter

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y} \tag{1.10}$$

Deixamos de verificar a condição de segunda ordem para mínimo. Cabe assinalar que, por ser uma soma de quadrados de desvios, *S* não pode ter um valor máximo finito.

Se $\mathbf{X}'\mathbf{X}$ for não-singular, de (1.10) obtemos

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.11}$$

que é o estimador de mínimos quadrados ordinários de $\boldsymbol{\beta}$.

A expressão (1.10) é um sistema de equações lineares cujas incógnitas são as estimativas dos parâmetros. Ele é denominado *sistema de equações normais* e a sua solução, quando existe, é dada por (1.11). De (1.10) segue-se que

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$

ou

$$\mathbf{X}'\mathbf{e} = \mathbf{0} \tag{1.12}$$

Se o modelo tiver um termo constante (α), a 1ª linha de $\mathbf{X}'$ é $\boldsymbol{\iota}'$ e, de acordo com (1.12), temos

$$\boldsymbol{\iota}'\mathbf{e} = \sum_{j=1}^{n} e_j = 0 \tag{1.13}$$

Cabe ressaltar que a expressão (1.12) é uma propriedade geral dos resíduos obtidos pelo método de mínimos quadrados, mas o resultado (1.13) é válido apenas quando o modelo de regressão tem um termo constante.

Substituindo (1.10) em (1.8) obtemos

$$S = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} \tag{1.14}$$

Esse resultado será usado adiante na análise de variância da regressão.

1.3. Propriedades do estimador de mínimos quadrados

Se dispusermos de uma amostra de dados, isto é, se tivermos as matrizes **y** e **X**, e a matriz $\mathbf{X}'\mathbf{X}$ for não-singular, podemos calcular o vetor (**b**) de estimativas dos parâmetros, conforme a expressão (1.11). Para interpretar **b** como um estimador de $\boldsymbol{\beta}$, obviamente é necessário pressupor que há uma relação entre as variáveis, conforme (1.1) ou (1.2).

A expressão (1.11) mostra que cada elemento de **b** é uma combinação linear dos elementos de **y** (os valores Y_j da variável independente). Diz-se, então, que **b** é um estimador *linear* de $\boldsymbol{\beta}$.

Vejamos a demonstração de que, em certas condições, **b** é um estimador linear não-tendencioso (não-viesado).

Substituindo (1.2) em (1.11), obtemos

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})$$

ou

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{1.15}$$

Se a matriz **X** for fixa, de acordo com as propriedades do operador de esperança matemática, segue-se que

$$E(\mathbf{b}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u})$$

Se $E(\mathbf{u}) = \mathbf{0}$, obtemos

$$E(\mathbf{b}) = \boldsymbol{\beta}, \tag{1.16}$$

mostrando que **b** é um estimador linear não-tendencioso de $\boldsymbol{\beta}$.

Pode-se mostrar que, em geral, existem infinitos estimadores lineares não-tendenciosos de $\boldsymbol{\beta}$. É desejável que se use o estimador linear não-tendencioso de variância mínima, também denominado o *melhor* estimador linear não-tendencioso.[2]

Utilizando a pressuposição (1.4) (homocedasticia e ausência de correlação entre os erros u_j) é possível demonstrar que o estimador de mínimos quadrados (**b**) é o estimador

[2] A sigla da expressão em inglês é BLUE (best linear unbiased estimator).

linear não-tendencioso de variância mínima. Esse resultado é conhecido como “teorema de Gauss-Markov”.

De (1.15) segue-se que

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{1.17}$$

Por definição, a matriz de variâncias e covariâncias das estimativas dos parâmetros é

$$V(\mathbf{b}) = E(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' \tag{1.18}$$

Lembrando (1.17), segue-se que

$$V(\mathbf{b}) = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$$

Com $\mathbf{X}$ fixa e $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$, obtemos

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \tag{1.19}$$

Para obter as estimativas dessas variâncias e covariâncias, é necessário obter uma estimativa de σ^2, como se mostra na próxima seção.

1.4. Análise de variância da regressão linear

Substituindo (1.11) em (1.6), obtém-se

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

ou

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}\,, \tag{1.20}$$

com

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{1.21}$$

Pode-se verificar que $\mathbf{H}$ é uma matriz $n \times n$, simétrica e idempotente, isto é, $\mathbf{HH} = \mathbf{H}$. É curioso notar que há apenas dois números idempotentes (zero e 1), mas há infinitas matrizes idempotentes. Notar que uma matriz quadrada de zeros e uma matriz identidade também são idempotentes.

O vetor de desvios é

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}\,, \tag{1.22}$$

com

$$\mathbf{M} = \mathbf{I} - \mathbf{H} \tag{1.23}$$

É fácil verificar que **M** também é uma matriz $n \times n$, simétrica e idempotente.

Lembrando a definição de $\boldsymbol{\iota}$ em (1.5), define-se a matriz $n \times n$

$$\mathbf{A} = \mathbf{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}' \tag{1.24}$$

Pode-se verificar que **A** também é uma matriz simétrica e idempotente. Note-se que

$$\mathbf{Ay} = \mathbf{y} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'\mathbf{y}$$

Com $\boldsymbol{\iota}'\mathbf{y} = \sum Y_j$, a média dos Y_j é

$$\overline{Y} = \frac{1}{n}\boldsymbol{\iota}'\mathbf{y} \tag{1.25}$$

Então

$$\mathbf{Ay} = \mathbf{y} - \boldsymbol{\iota}\overline{Y}\,, \tag{1.26}$$

que é um vetor-coluna com os valores de $Y_j - \overline{Y}$, denominados *valores centrados* de Y_j. Verifica-se, portanto, que a pré-multiplicação de um vetor-coluna pela matriz **A** faz com que a variável correspondente se torne centrada.

Analogamente, **AX** é uma matriz com as mesmas dimensões que **X** e com todas as suas variáveis centradas. Se a primeira coluna de **X** for o vetor $\boldsymbol{\iota}$, a primeira coluna de **AX** será um vetor de zeros.

É usual denominar de *soma de quadrados total* (S.Q.Total) a soma dos quadrados dos valores centrados da variável dependente:

$$\text{S.Q.Total} = (\mathbf{Ay})'\mathbf{Ay}$$

Como **A** é uma matriz simétrica e idempotente, obtemos

$$\text{S.Q.Total} = \mathbf{y}'\mathbf{Ay} =$$
$$= \mathbf{y}'\left(\mathbf{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'\right)\mathbf{y} =$$

$$= \mathbf{y}'\mathbf{y} - \frac{1}{n}(\boldsymbol{\iota}'\mathbf{y})^2 = \mathbf{y}'\mathbf{y} - C\,, \tag{1.27}$$

com

$$C = \frac{1}{n}(\boldsymbol{\iota}'\mathbf{y})^2 \tag{1.28}$$

A expressão (1.28) é a *correção* que deve ser subtraída de $\mathbf{y}'\mathbf{y}$ para obter a S.Q.Total.

Analogamente, denomina-se de *soma de quadrados de regressão* (S.Q.Regr.) a soma dos quadrados dos valores centrados da variável dependente estimada:

$$\begin{aligned} \text{S.Q.Regr.} &= (\mathbf{A}\hat{\mathbf{y}})'\mathbf{A}\hat{\mathbf{y}} = \\ &= \hat{\mathbf{y}}'\mathbf{A}\hat{\mathbf{y}} = \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} - \frac{1}{n}(\boldsymbol{\iota}'\hat{\mathbf{y}})^2 \end{aligned} \tag{1.29}$$

Pré-multiplicando por $\boldsymbol{\iota}'$ a relação $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, obtemos

$$\boldsymbol{\iota}'\mathbf{e} = \boldsymbol{\iota}'\mathbf{y} - \boldsymbol{\iota}'\hat{\mathbf{y}}$$

Se a equação estimada tiver um termo constante, de acordo com (1.13) segue-se que

$$\boldsymbol{\iota}'\mathbf{y} = \boldsymbol{\iota}'\hat{\mathbf{y}}$$

Substituindo esse resultado em (1.29), obtemos

$$\text{S.Q.Regr.} = \hat{\mathbf{y}}'\hat{\mathbf{y}} - \frac{1}{n}(\boldsymbol{\iota}'\mathbf{y})^2 \tag{1.30}$$

Lembrando que $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ e utilizando a relação (1.10), verifica-se que

$$\text{S.Q.Regr.} = \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{1}{n}(\boldsymbol{\iota}'\mathbf{y})^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - C \tag{1.31}$$

De (1.14), (1.27) e (1.31) conclui-se que a soma de quadrados dos desvios, ou soma de quadrados dos resíduos (S.Q.Res.), é igual à diferença entre a S.Q.Total e a S.Q.Regr.:

$$S = \mathbf{e}'\mathbf{e} = \text{S.Q.Total} - \text{S.Q.Regr.} \tag{1.32}$$

ou

$$S = \mathbf{e}'\mathbf{e} = [\,\mathbf{y}'\mathbf{y} - C] - [\,\mathbf{b}'\mathbf{X}'\mathbf{y} - C]$$

A seguir será deduzida a esperança matemática de cada uma das somas de quadrados, iniciando com $E(\text{S.Q.Res.})$. De (1.22) segue-se que

$$\mathbf{e} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{u}$$

É fácil verificar que $\mathbf{H}\mathbf{X} = \mathbf{X}$ e $\mathbf{M}\mathbf{X} = \mathbf{0}$ (uma matriz $n \times p$ de zeros). Então

$$\mathbf{e} = \mathbf{M}\mathbf{u} \tag{1.33}$$

e, como **M** é simétrica e idempotente,

$$S = \mathbf{e}'\mathbf{e} = \mathbf{u}'\mathbf{M}\mathbf{u} \tag{1.34}$$

Antes de aplicar o operador de esperança matemática, é necessário usar um artifício. Como $\mathbf{u}'\mathbf{M}\mathbf{u}$ é uma matriz com um único elemento, ela é igual ao seu traço:

$$S = \mathbf{u}'\mathbf{M}\mathbf{u} = \text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})$$

Uma vez que o traço de um produto matricial não muda se for alterada a ordem dos fatores de maneira apropriada, temos

$$S = \text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u}) = \text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')$$

Aplicando esperança matemática, lembrando a pressuposição (1.4), considerando que **M** só depende de **X**, que é considerada fixa, e tendo em vista que a esperança do traço é igual ao traço da esperança, obtemos

$$E(S) = [\text{tr}(\mathbf{M})]\sigma^2 \tag{1.35}$$

Temos

$$\text{tr}(\mathbf{H}) = \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{tr}(\mathbf{I}_p) = p$$

$$\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n - \mathbf{H}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{H}) = n - p$$

Então

$$E(S) = (n - p)\sigma^2 \tag{1.36}$$

A seguir vamos deduzir a expressão para a esperança matemática da S.Q.Total. Temos

$$\begin{aligned}\text{S.Q.Total} &= \mathbf{y}'\mathbf{A}\mathbf{y} = \\ &= (\boldsymbol{\beta}'\mathbf{X}' + \mathbf{u}')\mathbf{A}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \\ &= \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{u} + \mathbf{u}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{u}'\mathbf{A}\mathbf{u}\end{aligned}$$

Com $\mathbf{X}$ fixa e $E(\mathbf{u}) = \mathbf{0}$, obtemos

$$E(\text{S.Q.Total}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + E(\mathbf{u}'\mathbf{A}\mathbf{u}) \tag{1.37}$$

Repetindo o artifício utilizado na dedução da esperança da S.Q.Res., temos

$$E(\mathbf{u}'\mathbf{A}\mathbf{u}) = E[\text{tr}(\mathbf{u}'\mathbf{A}\mathbf{u})] = E[\text{tr}(\mathbf{A}\mathbf{u}\mathbf{u}')] =$$

$$= [\text{tr}(\mathbf{A})]\sigma^2 \tag{1.38}$$

Mas

$$\text{tr}(\mathbf{A}) = \text{tr}\left(\mathbf{I}_n - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'\right) = n - 1$$

Então

$$E(\mathbf{u}'\mathbf{A}\mathbf{u}) = (n-1)\sigma^2$$

Substituindo esse resultado em (1.37), obtemos

$$E(\text{S.Q.Total}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + (n-1)\sigma^2 \tag{1.39}$$

De acordo com (1.32), temos

$$E(\text{S.Q.Regr.}) = E(\text{S.Q.Total}) - E(S)$$

Utilizando (1.36) e (1.39), obtemos

$$E(\text{S.Q.Regr.}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + (p-1)\sigma^2 \tag{1.40}$$

Para um modelo de regressão linear com termo constante, $p - 1 = k$ é o número de variáveis explanatórias.

Os resultados expressos em (1.14), (1.27), (1.31), (1.36), (1.39) e (1.40) podem ser resumidos no esquema a seguir:

Causa de variação	S.Q.	E(S.Q.)
Regressão	$\hat{\mathbf{y}}'\mathbf{A}\hat{\mathbf{y}} = \mathbf{b}'\mathbf{X}'\mathbf{y} - C$	$\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + (p-1)\sigma^2$
Resíduo	$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$	$(n-p)\sigma^2$
Total	$\mathbf{y}'\mathbf{A}\mathbf{y} = \mathbf{y}'\mathbf{y} - C$	$\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + (n-1)\sigma^2$

O coeficiente de σ^2 na expressão da *E*(S.Q.) é o número de *graus de liberdade* (G.L.) associado à respectiva soma de quadrados. O *quadrado médio* (Q.M.) é, por definição, a razão entre a soma de quadrados e o respectivo número de graus de liberdade. Assim

$$\text{Q.M.Res.} = \frac{1}{n-p}\mathbf{e}'\mathbf{e} = \frac{1}{n-p}(\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}) \tag{1.41}$$

Lembrando (1.36), segue-se que

$$E(\text{Q.M.Res.}) = \sigma^2, \tag{1.42}$$

mostrando que o Q.M.Res. é um estimador não-tendencioso da variância do erro. Por isso é usual representar o Q.M.Res. por s^2.

Lembrando (1.40), verifica-se que

$$E(\text{Q.M.Regr.}) = \frac{1}{k}\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + \sigma^2 \tag{1.43}$$

É importante notar que a existência do termo constante (α) faz com que todos os elementos da 1ª coluna de **AX** sejam iguais a zero, o mesmo ocorrendo com todos os elementos da 1ª coluna e da 1ª linha de $\mathbf{X}'\mathbf{A}\mathbf{X} = (\mathbf{A}\mathbf{X})'\mathbf{A}\mathbf{X}$. Consequentemente, o valor $\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}$ não é afetado por α.

Sob a hipótese de que $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, $\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} = 0$ e $E(\text{Q.M.Regr.}) = \sigma^2$. Nessa situação a relação

$$F = \frac{\text{Q.M.Regr.}}{\text{Q.M.Res.}} \tag{1.44}$$

tende a ficar próxima de 1. Se aquela hipótese não for verdadeira, teremos $\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} > 0$ (pois $\mathbf{X}'\mathbf{A}\mathbf{X}$ é uma matriz quadrada definida positiva) e a razão (1.44) tende a assumir valores maiores do que 1. Na seção 1.8 será apresentada a fundamentação teórica para o uso da relação (1.44) para testar a hipótese $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, mostrando que, sob essa hipótese, e sendo válidas as pressuposições de que $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$ e $u_j \sim N(0, \sigma^2)$, a relação (1.44) tem distribuição de *F* com *k* e *n* – *p* graus de liberdade.

Se, na expressão (1.19), substituirmos σ^2 pelo seu estimador não-tendencioso ($s^2 = \text{Q.M.Res.}$), obtemos a matriz das estimativas das variâncias e covariâncias das estimativas dos parâmetros:

$$\hat{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} s^2 \tag{1.45}$$

A relação entre a S.Q.Regr. e a S.Q.Total é denominada *coeficiente de determinação* e é indicada por R^2:

$$R^2 = \frac{\text{S.Q.Regr.}}{\text{S.Q.Total}} = 1 - \frac{\text{S.Q.Res.}}{\text{S.Q.Total}} \tag{1.46}$$

Lembrando (1.32) e tendo em vista que uma soma de quadrados não pode ser negativa, conclui-se que

$$0 \le R^2 \le 1.$$

O coeficiente de determinação é uma medida da qualidade do ajustamento da equação aos dados. Ele pode ser interpretado como a fração da variabilidade dos Y_j em torno da sua média (medida pela S.Q.Total) que é estatisticamente explicada por meio das variáveis $X_{1j}, X_{2j}, \ldots, X_{kj}$. Temos $R^2 = 1$ apenas quando todos os desvios forem iguais a zero, isto é, quando $\hat{Y}_j = Y_j$ para todo *j*.

É óbvio que a análise de regressão só poderá ser feita se $n > p$. Se $n = p$, o resíduo fica com zero graus de liberdade e não se pode fazer nenhuma análise estatística. Com $n = p$, o sistema

$$\mathbf{Xb} = \mathbf{y}$$

tem *p* equações e *p* incógnitas e a determinação de **b** é um problema de geometria analítica (determinar a reta que passa por 2 pontos, determinar o plano que passa por 3 pontos ou, com $p > 3$, determinar o hiperplano que passa pelos *p* pontos) e não existem desvios. Há, então, uma tendência de R^2 se aproximar de 1 quando a amostra tem pouco mais que *p* observações. Para evitar essa falsa indicação de "boa qualidade" do ajustamento para amostras pequenas, pode ser usado o coeficiente de determinação corrigido para graus de liberdade, definido como

$$\overline{R}^2 = 1 - \frac{\text{S.Q.Res.}}{\text{S.Q.Total}} \cdot \frac{n-1}{n-p} \tag{1.47}$$

De (1.46) e (1.47), após algumas passagens algébricas, obtemos

$$\overline{R}^2 = R^2 - (1-R^2)\frac{p-1}{n-p}, \tag{1.48}$$

mostrando que $\overline{R}^2 \leq R^2$. O símbolo $\overline{R}^2$ está consagrado, mas a rigor é inapropriado, pois o coeficiente de determinação corrigido para graus de liberdade pode ser negativo.

1.5. Inversa de uma matriz decomposta

Esta seção se destina exclusivamente à apresentação de expressões referentes à inversão de matrizes quadradas decompostas de maneira apropriada, que serão utilizadas nas próximas seções.

Consideremos uma matriz quadrada simétrica e não-singular decomposta da seguinte maneira:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix},$$

onde $\mathbf{A}$ e $\mathbf{C}$ são matrizes quadradas. Verifica-se que

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix}^{-1} =$$

$$= \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}'\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\left(\mathbf{C} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}'\mathbf{A}^{-1} & \left(\mathbf{C} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{bmatrix} \tag{1.49}$$

$$= \begin{bmatrix} \left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'\right)^{-1} & -\left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'\right)^{-1}\mathbf{B}\mathbf{C}^{-1} \\ -\mathbf{C}^{-1}\mathbf{B}'\left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'\right)^{-1} & \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{B}'\left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'\right)^{-1}\mathbf{B}\mathbf{C}^{-1} \end{bmatrix} \tag{1.50}$$

A validade das expressões (1.49) e (1.50) pode ser comprovada multiplicando-as pela matriz original e verificando que o resultado é uma matriz identidade.

É interessante, também, verificar que as expressões (1.49) e (1.50) levam ao resultado correto na inversão de uma simples matriz 2×2 (quando as matrizes **A**, **B** e **C** são constituídas por um único elemento) como, por exemplo,

$$\mathbf{M} = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}$$

Note que no caso particular em que **B** = **0**, isto é, quando a matriz original é bloco-diagonal, temos

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} \end{bmatrix} \tag{1.51}$$

1.6. Exemplo numérico de uma regressão com duas variáveis explanatórias

O modelo de uma regressão linear com duas variáveis explanatórias é

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j \tag{1.52}$$

com $E(\mathbf{u}) = \mathbf{0}$, $E(\mathbf{uu}') = \mathbf{I}\sigma^2$ e $u_j \sim N(0, \sigma^2)$.

Os cálculos ficam mais fáceis se usarmos as variáveis explanatórias centradas, que serão representadas por letras minúsculas, isto é,

$$x_{ij} = X_{ij} - \overline{X}_i,$$

com $$\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

Então o modelo de regressão fica

$$Y_j = \gamma + \beta_1 x_{1j} + \beta_2 x_{2j} + u_j, \tag{1.53}$$

com $$\gamma = \alpha + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 \tag{1.54}$$

Uma vez que a soma dos valores de uma variável centrada é igual a zero, para o modelo (1.53) obtemos

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & 0 \\ 0 & \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ 0 & \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix} \tag{1.55}$$

Como essa matriz é bloco-diagonal, podemos inverter separadamente o *n* e a matriz 2×2, obtendo

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{D}\sum x_{2j}^2 & -\frac{1}{D}\sum x_{1j}x_{2j} \\ 0 & -\frac{1}{D}\sum x_{1j}x_{2j} & \frac{1}{D}\sum x_{1j}^2 \end{bmatrix}, \tag{1.56}$$

com $D = \sum x_{1j}^2 \sum x_{2j}^2 - \left(\sum x_{1j} x_{2j}\right)^2$

Temos

$$\mathbf{X'y} = \begin{bmatrix} \sum Y_j \\ \sum x_{1j} Y_j \\ \sum x_{2j} Y_j \end{bmatrix} \tag{1.57}$$

Verifica-se que o primeiro elemento de $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$, que é a estimativa de γ, é igual a $\overline{Y}$.

Então a estimativa da equação (1.53) é

$$\hat{Y}_j = \overline{Y} + b_1 x_{1j} + b_2 x_{2j}$$

Como $x_{ij} = X_{ij} - \overline{X}_i$, obtemos

$$\hat{Y}_j = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 + b_1 X_{1j} + b_2 X_{2j}$$

ou, ainda,

$$\hat{Y}_j = a + b_1 X_{1j} + b_2 X_{2j}$$

com $$a = \overline{Y} - \sum_{i=1}^{k} b_i \overline{X}_i, \tag{1.58}$$

que é a estimativa de α no modelo (1.52).

De acordo com (1.31),

$$\text{S.Q.Regr.} = \mathbf{b'X'y} - C$$

Mas para o modelo (1.53), verifica-se que

$$\mathbf{b'X'y} = \overline{Y} \sum Y_j + b_1 \sum x_{1j} Y_j + b_2 \sum x_{2j} Y_j$$

Como o primeiro elemento dessa expressão é exatamente a correção *C*, conclui-se que

$$\text{S.Q.Regr.} = b_1 \sum x_{1j} Y_j + b_2 \sum x_{2j} Y_j .$$

Em geral, para um modelo com termo constante,

$$\text{S.Q.Regr.} = \sum_{i=1}^{k} b_i \sum_{j=1}^{n} x_{ij} Y_j \tag{1.59}$$

A tabela 1.1 mostra os valores de X_{1j}, X_{2j} e Y_j em uma amostra com 5 observações. Trata-se de dados artificiais e, para facilitar os cálculos, considera-se uma amostra muito pequena.

É fácil verificar que as médias são $\overline{X}_1 = 5$, $\overline{X}_2 = 9$ e $\overline{Y} = 21$. Os valores das variáveis centradas foram calculados e são apresentados na mesma tabela.

Tabela 1.1. Amostra artificial com 5 observações das variáveis X_{1j}, X_{2j} e Y_j e respectivas variáveis centradas.

X_{1j}	X_{2j}	Y_j	x_{1j}	x_{2j}	y_j
5	5	32	0	–4	11
8	15	10	3	6	–11
3	9	16	–2	0	–5
6	11	16	1	2	–5
3	5	31	–2	–4	10

Considerando o modelo (1.53), obtemos

$$\mathbf{X'X} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 18 & 28 \\ 0 & 28 & 72 \end{bmatrix}, \quad \mathbf{X'y} = \begin{bmatrix} 105 \\ -48 \\ -160 \end{bmatrix},$$

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{36}{256} & -\frac{14}{256} \\ 0 & -\frac{14}{256} & \frac{9}{256} \end{bmatrix},$$

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y} = \begin{bmatrix} 21 \\ 2 \\ -3 \end{bmatrix} \quad \text{e} \quad \text{S.Q.Total} = \sum y_j^2 = 392$$

Por meio de (1.58) obtemos *a* = 38. Então, para o modelo (1.52), a equação estimada é

$$\hat{Y} = 38 + 2X_1 - 3X_2$$

De acordo com (1.59), temos

$$\text{S.Q.Regr.} = 2(-48) + (-3)(-160) = 384$$

Segue-se que

$$\text{S.Q.Res.} = 392 - 384 = 8$$

A tabela 1.2 mostra a análise de variância dessa regressão.

Tabela 1.2. Análise de variância.

Causa de variação	G.L.	S.Q.	Q.M.	*F*
Regressão	2	384	192	48
Resíduo	2	8	4	
Total	4	392		

A estimativa não-tendenciosa de σ^2 é

$$s^2 = \text{Q.M.Res.} = 4$$

Ao nível de significância de 5%, o valor crítico de *F* com 2 e 2 graus de liberdade é 19,00. Então, o valor calculado (*F* = 48) é significativo. Ao nível de significância de 5%, rejeita-se a hipótese de que $\beta_1 = \beta_2 = 0$.

Se os cálculos forem feitos por meio de um computador, pode ser obtida a probabilidade caudal associada a *F* = 48, isto é, a probabilidade de obter um valor maior do que 48 em uma distribuição de *F* com 2 e 2 graus de liberdade, também denominada "valor-*p*" do teste. Verifica-se que a probabilidade de obtermos *F* > 48, em uma distribuição de *F* com 2 e 2 graus de liberdade (sob a hipótese de que $\beta_1 = \beta_2 = 0$), é 2,04%. Como esse valor é menor do que o nível de significância adotado (5%), rejeita-se a hipótese de que $\beta_1 = \beta_2 = 0$.

De acordo com (1.45), obtemos

$$\hat{V}(b_1) = \frac{36}{256} s^2 = \frac{36}{256} \cdot 4 = \frac{9}{16}$$

e

$$\hat{V}(b_2) = \frac{9}{256} s^2 = \frac{9}{256} \cdot 4 = \frac{9}{64}$$

Segue-se que os respectivos desvios padrões são

$$s(b_1) = \frac{3}{4} = 0{,}75 \quad \text{e} \quad s(b_2) = \frac{3}{8} = 0{,}375 .$$

Vamos admitir que se deseja testar a hipótese $H_0 : \beta_2 = 0$, contra a hipótese alternativa $H_A : \beta_2 < 0$. Ao nível de significância de 1%. Então calculamos

$$t = \frac{b_2}{s(b_2)} = -8$$

Trata-se de uma distribuição de *t* com 2 graus de liberdade, que é o número de graus de liberdade associado a s^2. A região de rejeição para esse teste unilateral é $t \leq -6{,}965$. Portanto, o valor calculado ($t = -8$) é significativo, rejeitando-se $H_0 : \beta_2 = 0$ em favor de $H_A : \beta_2 < 0$.

Os programas de computador usualmente fornecem a probabilidade caudal para um teste *t* bilateral, que nesse caso é 1,53%. Como a metade desse valor é inferior a 1%, a conclusão é a mesma: o teste é significativo ao nível de 1%.

1.7. Regressão múltipla com decomposição da matriz **X**

Para o modelo de regressão linear múltipla

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$$

sabemos que o estimador de mínimos quadrados ordinários para o vetor de parâmetros $\boldsymbol{\beta}$ é

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

o vetor dos valores estimados de **y** é

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{Hy},$$

com $$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

e o vetor dos desvios é

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{My} \ ,$$

com

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Vimos que tanto **H** como **M** são matrizes simétricas e idempotentes.

Vamos, agora, agrupar as variáveis explanatórias em dois grupos, o que corresponde a decompor a matriz **X** da seguinte maneira:

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2] \tag{1.60}$$

A decomposição correspondente no vetor **b** é

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \tag{1.61}$$

Considerando a decomposição da matriz **X**, obtemos

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} \quad \text{e} \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \tag{1.62}$$

Fazendo $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$, de acordo com (1.50) obtemos

$$[\mathbf{X}'\mathbf{X}]^{-1} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \\ -(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{bmatrix} \tag{1.63}$$

Como $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, segue-se que

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$$

ou

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y} \tag{1.64}$$

Mas $\mathbf{M}_2\mathbf{y}$ é o vetor dos desvios de uma regressão de **y** contra $\mathbf{X}_2$ e $\mathbf{M}_2\mathbf{X}_1$ é a matriz dos desvios de regressões de cada coluna de $\mathbf{X}_1$ contra $\mathbf{X}_2$. Lembrando que $\mathbf{M}_2$ é uma matriz idempotente, a expressão (1.64) mostra que, em uma regressão múltipla de **y** contra **X**, os coeficientes das variáveis incluídas em $\mathbf{X}_1$ podem ser obtidos fazendo a regressão de $\mathbf{M}_2\mathbf{y}$ contra $\mathbf{M}_2\mathbf{X}_1$ (ou fazendo a regressão de **y** contra $\mathbf{M}_2\mathbf{X}_1$).

Esse resultado é conhecido como teorema de Frisch-Waugh ou teorema de Frisch-Waugh-Lovell.

O vetor-coluna $\mathbf{M}_2\mathbf{y}$ corresponde ao vetor-coluna **y** depois que ele foi depurado das variações de Y_j que podiam ser linearmente associadas às variáveis em $\mathbf{X}_2$. Analogamente, as colunas de $\mathbf{M}_2\mathbf{X}_1$ correspondem às variáveis em $\mathbf{X}_1$ depois que elas foram depuradas das suas variações linearmente associadas às variáveis em $\mathbf{X}_2$. E o teorema mostra que os coeficientes de $\mathbf{b}_1$ na regressão de **y** contra toda a matriz **X** podem ser obtidos fazendo a

regressão de $\mathbf{M}_2\mathbf{y}$ contra $\mathbf{M}_2\mathbf{X}_1$, isto é, usando as variáveis previamente depuradas dos efeitos lineares das variáveis em $\mathbf{X}_2$. E o mesmo vetor $\mathbf{b}_1$ pode ser obtido fazendo a regressão de $\mathbf{y}$ contra $\mathbf{M}_2\mathbf{X}_1$, ou seja, basta fazer aquela depuração nas variáveis explanatórias.

A seguir demonstra-se, ainda, que o vetor-coluna de desvios da regressão de $\mathbf{M}_2\mathbf{y}$ contra $\mathbf{M}_2\mathbf{X}_1$, dado por

$$\mathbf{d} = \mathbf{M}_2\mathbf{y} - \mathbf{M}_2\mathbf{X}_1\mathbf{b}_1 \tag{1.65}$$

é igual ao vetor de desvios da regressão de $\mathbf{y}$ contra $\mathbf{X}$, que é

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \mathbf{X}_1\mathbf{b}_1 - \mathbf{X}_2\mathbf{b}_2 \tag{1.66}$$

Utilizando a segunda linha da matriz (1.63), obtemos

$$\begin{aligned}\mathbf{b}_2 &= -(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} + \\ &\quad + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} = \\ &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}\end{aligned}$$

Lembrando (1.64), segue-se que

$$\mathbf{b}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1 \tag{1.67}$$

Substituindo esse resultado em (1.66), obtemos

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{X}_1\mathbf{b}_1 - \mathbf{H}_2\mathbf{y} + \mathbf{H}_2\mathbf{X}_1\mathbf{b}_1 = \\ &= (\mathbf{I} - \mathbf{H}_2)\mathbf{y} - (\mathbf{I} - \mathbf{H}_2)\mathbf{X}_1\mathbf{b}_1 = \mathbf{M}_2\mathbf{y} - \mathbf{M}_2\mathbf{X}_1\mathbf{b}_1\end{aligned}$$

Comparando esse resultado com (1.65), conclui-se que $\mathbf{d} = \mathbf{e}$, c.q.d. Cabe ressaltar que a regressão de $\mathbf{y}$ contra $\mathbf{M}_2\mathbf{X}_1$, que gera o mesmo vetor de estimativas de parâmetros $\mathbf{b}_1$, não produz, em geral, o mesmo vetor de desvios que a regressão de $\mathbf{M}_2\mathbf{y}$ contra $\mathbf{M}_2\mathbf{X}_1$.

É importante interpretar o teorema de Frisch-Waugh-Lowell quando a matriz $\mathbf{X}_1$ tem apenas uma coluna, e todas as demais variáveis explanatórias e $\boldsymbol{\iota}$ estão na matriz $\mathbf{X}_2$. Vamos admitir, sem perda de generalidade, que a única coluna de $\mathbf{X}_1$ seja formada pelos valores de X_{1j}. Para destacar que se trata de um vetor-coluna, passamos a indicar a matriz $\mathbf{X}_1$ por $\mathbf{x}_1$. De acordo com o teorema de Frisch-Waugh-Lowell, o coeficiente b_1 de X_{1j}, na regressão

múltipla de **y** contra **X**, pode ser obtido fazendo a regressão linear simples de $\mathbf{M}_2\mathbf{y}$ contra $\mathbf{M}_2\mathbf{x}_1$. Podemos dizer, então, que o coeficiente b_1 na regressão múltipla é igual ao coeficiente de uma regressão linear simples de Y_j contra X_{1j} depois que essas duas variáveis foram depuradas dos efeitos lineares de todas as demais variáveis explanatórias incluídas no modelo de regressão linear múltipla.

Em uma ciência experimental, para analisar como uma variável X_1 afeta uma variável dependente *Y*, é usual manter constantes as demais variáveis que afetam *Y*. Em uma ciência tipicamente não-experimental, como economia ou sociologia, o pesquisador interessado no efeito de X_1 sobre *Y* não tem a opção de "manter constantes as demais variáveis que afetam *Y*". Nessa situação, a técnica estatística de regressão múltipla é uma tentativa de obter o mesmo resultado a partir de um conjunto de dados com variações simultâneas em todas as variáveis relevantes. Como os resultados dependem do modelo de regressão adotado, é claro que eles não são tão confiáveis como os obtidos de um experimento no qual as variáveis podem ser controladas, conforme os objetivos do pesquisador.

Quando a matriz $\mathbf{X}_1 = \mathbf{x}_1$ é constituída por apenas uma coluna com os valores de X_{1j}, a correlação simples entre $\mathbf{M}_2\mathbf{y}$ e $\mathbf{M}_2\mathbf{x}_1$ é denominada correlação *parcial* entre Y_j e X_{1j}, dados os valores de X_{2j}, X_{3j}, ..., X_{kj}. Sendo **A** a matriz que centra as variáveis, definida em (1.24), a correlação simples entre Y_j e X_{1j} é dada por

$$r_{Y1} = \frac{\mathbf{y}'\mathbf{A}\mathbf{x}_1}{\sqrt{(\mathbf{y}'\mathbf{A}\mathbf{y})(\mathbf{x}_1'\,\mathbf{A}\mathbf{x}_1)}}$$

A correlação parcial entre Y_j e X_{1j}, dados os valores de X_{2j}, X_{3j}, ..., X_{kj}, é

$$\pi_{Y1} = \frac{\mathbf{y}'\mathbf{M}_2\mathbf{x}_1}{\sqrt{(\mathbf{y}'\mathbf{M}_2\mathbf{y})(\mathbf{x}_1'\,\mathbf{M}_2\mathbf{x}_1)}} \qquad (1.68)$$

Sabemos que os resíduos da regressão de **y** contra **X** são iguais aos resíduos da regressão de $\mathbf{M}_2\mathbf{y}$ contra $\mathbf{M}_2\mathbf{x}_1$, para a qual temos

$$\text{S.Q.Total} = \mathbf{y}'\mathbf{M}_2\mathbf{y},$$

$$b_1 = \frac{\mathbf{x}_1'\mathbf{M}_2\mathbf{y}}{\mathbf{x}_1'\mathbf{M}_2\mathbf{x}},$$

$$\text{S.Q.Regr.} = b_1 \mathbf{x}_1' \mathbf{M}_2 \mathbf{y} = \frac{(\mathbf{x}_1' \mathbf{M}_2 \mathbf{y})^2}{\mathbf{x}_1' \mathbf{M}_2 \mathbf{x}_1}$$

e

$$\text{S.Q.Res.} = \mathbf{y}' \mathbf{M}_2 \mathbf{y} - \frac{(\mathbf{x}_1' \mathbf{M}_2 \mathbf{y})^2}{\mathbf{x}_1' \mathbf{M}_2 \mathbf{x}_1}.$$

Como $\mathbf{y}'\mathbf{M}_2\mathbf{y}$ é a soma de quadrados dos resíduos da regressão de $\mathbf{y}$ contra $\mathbf{X}_2$, verifica-se que a redução da soma de quadrados residual devida à inclusão de X_{1j} como variável explanatória é

$$\text{S.Q.Contribuição de } X_{1j} = \frac{(\mathbf{x}_1' \mathbf{M}_2 \mathbf{y})^2}{\mathbf{x}_1' \mathbf{M}_2 \mathbf{x}_1}.$$

Dada a matriz $\mathbf{X}_2$, o valor máximo da "S.Q.Contribuição de X_{1j}" é igual à soma de quadrados residual da regressão de $\mathbf{y}$ contra $\mathbf{X}_2$, que é igual a $\mathbf{y}'\mathbf{M}_2\mathbf{y}$. Então a razão entre a "S.Q.Contribuição de X_{1j}" e seu valor máximo é

$$\frac{(\mathbf{x}_1' \mathbf{M}_2 \mathbf{y})^2}{(\mathbf{x}_1' \mathbf{M}_2 \mathbf{x}_1)(\mathbf{y}' \mathbf{M}_2 \mathbf{y})},$$

que é denominada *coeficiente de determinação parcial* entre Y_j e X_{1j}, dados os valores de X_{2j}, X_{3j}, ..., X_{kj}. Comparando essa expressão com (1.68), verifica-se que o coeficiente de determinação parcial entre Y_j e X_{1j} é igual ao quadrado do coeficiente de correlação parcial entre Y_j e X_{1j}, dados os valores das demais variáveis explanatórias.

Uma aplicação específica do teorema de Frisch-Waugh-Lowell se refere ao uso de variáveis centradas. Considerando um modelo de regressão linear com termo constante no final da equação, decompomos a matriz $\mathbf{X}$ em uma matriz $\mathbf{W}$ com os valores das variáveis explanatórias e o vetor-coluna $\boldsymbol{\iota}$, isto é,

$$\mathbf{X} = [\mathbf{W} \quad \boldsymbol{\iota}]$$

Conforme a notação usada no início desta seção, neste caso temos $\mathbf{X}_2 = \boldsymbol{\iota}$ e

$$\mathbf{M}_2 = \mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}' = \mathbf{I} - \frac{1}{n}\boldsymbol{\iota}\,\boldsymbol{\iota}' = \mathbf{A},$$

isto é, $\mathbf{M}_2$ é a matriz $\mathbf{A}$ que faz com que $\mathbf{Ay}$ e $\mathbf{AW}$ sejam matrizes de variáveis centradas. O teorema de Frisch-Waugh-Lowell garante, então, que a regressão com as variáveis centradas

(de $\mathbf{Ay}$ contra $\mathbf{AW}$) produz o mesmo vetor de coeficientes de regressão ($\mathbf{b}_1$) que a regressão de **y** contra **X**. Além disso, o vetor de resíduos da regressão de $\mathbf{Ay}$ contra $\mathbf{AW}$ é idêntico ao vetor de resíduos da regressão de **y** contra **X**.

1.8. Teste de hipóteses no modelo linear[3]

Nesta seção vamos delinear a dedução de uma expressão geral para testar hipóteses a respeito dos parâmetros de uma regressão linear cujo modelo é

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

onde **X** é uma matriz de dimensão $n \times p$ com valores fixos e característica igual a *p* (as *p* colunas de **X** são linearmente independentes), $\boldsymbol{\beta}$ é um vetor com os *p* parâmetros a serem estimados e **y** e **u** são vetores *n*-dimensionais de variáveis aleatórias. Admite-se que **u** tem distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2)$.

Consideremos que a hipótese de nulidade a respeito dos valores dos parâmetros seja constituída por *m* relações lineares independentes, isto é,

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}$$

onde **C** é uma matriz com *m* linhas e *p* colunas e $\boldsymbol{\theta}$ é um vetor *m* dimensional de constantes conhecidas. Se as *m* relações lineares assim definidas são independentes, a característica de **C** é igual a *m*. Note que devemos ter $m \leq p < n$.

Desenvolveremos duas maneiras de testar $H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}$ contra $H_A : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\theta}$. Uma delas se baseia no estimador de mínimos quadrados e a outra na soma de quadrados residual. Mostraremos, a seguir, que as duas maneiras de fazer o teste são equivalentes.

Consideremos, inicialmente, o teste baseado em $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, que é o estimador de mínimos quadrados, sem considerar a hipótese de nulidade.

Seja

$$\mathbf{g} = \mathbf{C}\mathbf{b} - \boldsymbol{\theta}$$

o vetor que mostra em quanto os valores estimados diferem dos valores estabelecidos pela hipótese de nulidade. No caso de um modelo com um único parâmetro, é lógico que se **g** for

[3] Esta seção foi desenvolvida com base em anotações de um curso de econometria ministrado pelo Prof. T. Rothenberg, da Universidade da Califórnia, Berkeley, em 1974.

bastante grande devemos rejeitar H_0. Mas **g** é, em geral, um vetor. Podemos avaliar se é "pequeno" ou "grande" por meio do valor da forma quadrática $\mathbf{g}'\mathbf{A}\mathbf{g}$, onde **A** é uma matriz definida positiva.[4] Uma vez que o valor de $\mathbf{g}'\mathbf{A}\mathbf{g}$ tende a crescer com os valores absolutos dos elementos de **g**, é razoável rejeitar H_0 se o valor de $\mathbf{g}'\mathbf{A}\mathbf{g}$ for bastante grande. Mas como escolher a matriz **A**? Ela é escolhida de maneira que a distribuição de $\mathbf{g}'\mathbf{A}\mathbf{g}$ seja conveniente e o poder do teste seja elevado.

De acordo com (1.15), temos

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},$$

mostrando que a parte aleatória de cada estimativa de parâmetro é uma combinação linear dos erros u_j. Então, se esses erros têm distribuição normal, as estimativas dos parâmetros têm distribuição normal. Lembrando (1.19), conclui-se que **b** tem distribuição $N[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2]$. Considerando a hipótese da nulidade, verifica-se que **g** tem distribuição normal com média

$$E(\mathbf{g}) = \mathbf{C}\boldsymbol{\beta} - \boldsymbol{\theta} = \mathbf{0}$$

e variância

$$\begin{aligned} V(\mathbf{g}) &= E[(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})'] = \\ &= E[\mathbf{C}(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'\mathbf{C}'] = \\ &= \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\sigma^2 \end{aligned}$$

Então devemos escolher $\mathbf{A} = \frac{1}{\sigma^2}[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}$

Segue-se que

$$\mathbf{g}'\mathbf{A}\mathbf{g} = \frac{(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})}{\sigma^2}$$

É possível demonstrar, como veremos adiante, que, sendo verdadeira a hipótese da nulidade, $\mathbf{g}'\mathbf{A}\mathbf{g}$ tem distribuição de qui-quadrado com *m* graus de liberdade.

De acordo com (1.34) temos que

$$\text{S.Q.Res.} = (n - p)s^2 = \mathbf{e}'\mathbf{e} = \mathbf{u}'\mathbf{M}\mathbf{u}, \tag{1.69}$$

onde

[4] Embora se use o mesmo símbolo **A**, não se trata, aqui, da matriz usada para centrar as variáveis.

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

É possível demonstrar, também, que

$$\frac{(n-p)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{\sigma^2}$$

tem distribuição de qui-quadrado com $n - p$ graus de liberdade.

Então, se H_0 for verdadeira,

$$T_1 = \frac{(\mathbf{Cb}-\boldsymbol{\theta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb}-\boldsymbol{\theta})}{ms^2} \tag{1.70}$$

tem distribuição de F com m e $n - p$ graus de liberdade.

Veremos, agora, o teste baseado no valor da soma de quadrados dos desvios. Para isso, consideremos as seguintes etapas:

a) Calculamos a soma de quadrados de desvios da regressão de **y** em relação a **X**,

$$S = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}), \tag{1.71}$$

à qual se associam $n - p$ graus de liberdade;

b) Determinamos, de acordo com o método de mínimos quadrados, as estimativas dos parâmetros ($\mathbf{b}_*$) da equação de regressão de **y** em relação a **X**, sujeitas à restrição $\mathbf{Cb}_* = \boldsymbol{\theta}$, isto é, determinamos o valor de $\mathbf{b}_*$ que minimiza $(\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*)$, condicionado a $\mathbf{Cb}_* = \boldsymbol{\theta}$.

A correspondente soma de quadrados de desvios é

$$S_* = (\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*), \tag{1.72}$$

à qual se associam $n - (p - m)$ graus de liberdade;

c) Calculamos a relação

$$T_2 = \frac{S_* - S}{S} \cdot \frac{n-p}{m} \tag{1.73}$$

Essa relação mede o crescimento relativo da soma de quadrados residual devido à restrição imposta pela hipótese de nulidade. Se H_0 é falsa, T_2 tende a assumir valores grandes. Pode-se demonstrar que, se H_0 é verdadeira, T_2 tem distribuição de F com m e $n - p$

graus de liberdade; o valor obtido pode, então, ser utilizado para testar H_0 ao nível de significância escolhido.

Passemos à demonstração de que $T_2 = T_1$.

Como $\frac{S}{n-p} = s^2$, de (1.73) segue-se que

$$T_2 = \frac{S_* - S}{ms^2} \tag{1.74}$$

Lembremos que S_* é a soma de quadrados dos desvios de uma regressão de **y** contra **X**, condicionada a $\mathbf{Cb}_* = \boldsymbol{\theta}$. Utilizando o método do multiplicador de Lagrange, formamos a função

$$(\mathbf{y} - \mathbf{Xb}_*)'(\mathbf{y} - \mathbf{Xb}_*) + 2\boldsymbol{\lambda}'(\mathbf{Cb}_* - \boldsymbol{\theta}),$$

onde $\boldsymbol{\lambda}$ é um vetor com *m* elementos.

As condições de primeira ordem para mínimo são

$$-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\mathbf{b}_* + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0} \tag{1.75}$$

e

$$\mathbf{Cb}_* = \boldsymbol{\theta} \tag{1.76}$$

Pré-multiplicando (1.75) por $\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}$ obtemos

$$-\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + \mathbf{Cb}_* + \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\boldsymbol{\lambda} = \mathbf{0}$$

ou

$$\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\boldsymbol{\lambda} = \mathbf{Cb} - \mathbf{Cb}_*$$

Considerando (1.76), segue-se que

$$\boldsymbol{\lambda} = [\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb} - \boldsymbol{\theta})$$

Substituindo esse resultado em (1.75), obtemos

$$\mathbf{X}'\mathbf{X}\mathbf{b}_* = \mathbf{X}'\mathbf{y} - \mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb} - \boldsymbol{\theta}),$$

Pré-multiplicando por $(\mathbf{X}'\mathbf{X})^{-1}$ e lembrando que $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, segue-se que

$$\mathbf{b}_* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\mathbf{b} - \boldsymbol{\theta}), \tag{1.77}$$

Isto é, o estimador de mínimos quadrados condicionado ($\mathbf{b}_*$) é igual ao estimador não-condicionado (**b**), mais uma combinação linear das diferenças $\mathbf{C}\mathbf{b} - \boldsymbol{\theta}$.

De acordo com (1.71) e (1.72), temos

$$\begin{aligned} S_* - S &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}_* + \mathbf{b}_*'\mathbf{X}'\mathbf{X}\mathbf{b}_* - \mathbf{y}'\mathbf{y} + 2\mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \\ &= -2\mathbf{y}'\mathbf{X}\mathbf{b}_* + \mathbf{b}_*'\mathbf{X}'\mathbf{X}\mathbf{b}_* + 2\mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \end{aligned}$$

Considerando (1.77) e lembrando que $\mathbf{b}' = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, obtemos, após várias simplificações,

$$S_* - S = (\mathbf{C}\mathbf{b} - \boldsymbol{\theta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})$$

Substituindo esse resultado em (1.74), obtemos

$$T_2 = \frac{(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\mathbf{b} - \boldsymbol{\theta})}{ms^2} \tag{1.78}$$

Comparando (1.70) e (1.78) concluímos que $T_2 = T_1$, c.q.d.

Resta mostrar que $T = T_1 = T_2$ tem distribuição de *F* com *m* e $n - p$ graus de liberdade.

Determinemos, inicialmente, a distribuição de $\mathbf{u}'\mathbf{A}\mathbf{u}$, onde **A** é uma matriz simétrica e idempotente qualquer e **u** é um vetor de *n* variáveis aleatórias com distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2)$.

Do estudo da álgebra de matrizes sabemos que, dada uma matriz simétrica **A**, existe uma matriz ortogonal **P** tal que

$$\mathbf{P}'\mathbf{A}\mathbf{P} = \boldsymbol{\Lambda},$$

onde $\boldsymbol{\Lambda}$ é a matriz diagonal cujos elementos são as raízes características (λ_i) de **A**.

Uma vez que a inversa de uma matriz ortogonal é igual à sua transposta, segue-se que

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$$

Então

$$\mathbf{u}'\mathbf{A}\mathbf{u} = \mathbf{u}'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{u} = \mathbf{v}'\boldsymbol{\Lambda}\mathbf{v} = \sum_{i=1}^{n} \lambda_i v_i^2 \tag{1.79}$$

onde

$$\mathbf{v} = \mathbf{P}'\mathbf{u}$$

Como $E(\mathbf{u}) = \mathbf{0}$ e $E(\mathbf{uu}') = \mathbf{I}\sigma^2$, temos $E(\mathbf{v}) = \mathbf{0}$ e $E(\mathbf{vv}') = E(\mathbf{P}'\mathbf{uu}'\mathbf{P}) = \mathbf{I}\sigma^2$, pois $\mathbf{P}'\mathbf{P} = \mathbf{I}$. Portanto **v** tem distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2)$.

Do estudo da álgebra de matrizes sabemos que as raízes características de uma matriz simétrica e idempotente são ou iguais a um ou iguais a zero. Sabemos também que o número de raízes características iguais a um, a característica da matriz e o traço da matriz são, neste caso, iguais entre si. Seja $h = \text{tr}(\mathbf{A})$. Se considerarmos, sem perda de generalidade, que as *h* primeiras raízes características de **A** são iguais a um, de (1.79) obtemos

$$\mathbf{u}'\mathbf{A}\mathbf{u} = \sum_{i=1}^{n} v_i^2$$

ou

$$\frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\sigma^2} = \sum_{i=1}^{h} \left(\frac{v_i}{\sigma} \right)^2$$

Uma vez que **v** tem distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2)$, v_i/σ são variáveis normais reduzidas independentes. Concluímos que $\mathbf{u}'\mathbf{A}\mathbf{u}/\sigma^2$ tem distribuição de qui-quadrado com $h = \text{tr}(\mathbf{A})$ graus de liberdade.

Consideremos, agora, duas formas quadráticas: $\mathbf{u}'\mathbf{A}_1\mathbf{u}$ e $\mathbf{u}'\mathbf{A}_2\mathbf{u}$, onde $\mathbf{A}_1$ e $\mathbf{A}_2$ são matrizes simétricas idempotentes, $\text{tr}(\mathbf{A}_1) = h_1$ e $\text{tr}(\mathbf{A}_2) = h_2$. Sabemos que $\mathbf{u}'\mathbf{A}_1\mathbf{u}/\sigma^2$ tem distribuição de qui-quadrado com h_1 graus de liberdade e $\mathbf{u}'\mathbf{A}_2\mathbf{u}/\sigma^2$ tem distribuição de qui-quadrado com h_2 graus de liberdade. É possível demonstrar que, se $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$, essas duas distribuições são independentes e então

$$\frac{\dfrac{\mathbf{u}'\mathbf{A}_1\mathbf{u}}{h_1}}{\dfrac{\mathbf{u}'\mathbf{A}_2\mathbf{u}}{h_2}} = \frac{\mathbf{u}'\mathbf{A}_1\mathbf{u}}{\mathbf{u}'\mathbf{A}_2\mathbf{u}} \cdot \frac{h_2}{h_1}$$

tem distribuição de *F* com h_1 e h_2 graus de liberdade.

Temos, portanto, o seguinte teorema:

Se

1º) **u** é um vetor de variáveis aleatórias com distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2)$

$2^{\underline{o}}$) $\mathbf{A}_1$ e $\mathbf{A}_2$ são matrizes simétricas idempotentes, isto é, $\mathbf{A}_1 = \mathbf{A}_1' = \mathbf{A}_1^2$ e $\mathbf{A}_2 = \mathbf{A}_2' = \mathbf{A}_2^2$, com $\text{tr}(\mathbf{A}_1) = h_1$ e $\text{tr}(\mathbf{A}_2) = h_2$, e

$3^{\underline{o}}$) $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$,

então

$$\frac{\mathbf{u}'\mathbf{A}_1\mathbf{u}}{\mathbf{u}'\mathbf{A}_2\mathbf{u}} \cdot \frac{h_2}{h_1}$$

tem distribuição de *F* com h_1 e h_2 graus de liberdade.

Para o modelo linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, se admitirmos que $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{\theta}$ é verdadeira, temos que

$$\mathbf{Cb} - \mathbf{\theta} = \mathbf{Cb} - \mathbf{C}\boldsymbol{\beta} = \mathbf{C}(\mathbf{b} - \boldsymbol{\beta})$$

Lembrando (1.17), obtemos

$$\mathbf{Cb} - \mathbf{\theta} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{1.80}$$

Considerando (1.69) e (1.80), a expressão de *T*, dada em (1.70) ou (1.78), pode ser colocada na seguinte forma:

$$T = \frac{\mathbf{u}'\mathbf{Q}\mathbf{u}}{\mathbf{u}'\mathbf{M}\mathbf{u}} \cdot \frac{n-p}{m}$$

onde

$$\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Verifica-se que

$$\mathbf{M} = \mathbf{M}' = \mathbf{M}^2$$

$$\text{tr}(\mathbf{M}) = n - p$$

$$\mathbf{Q} = \mathbf{Q}' = \mathbf{Q}^2$$

$\text{tr}(\mathbf{Q}) = m$ = característica de $\mathbf{C}$

$$\mathbf{MQ} = \mathbf{0}$$

Concluímos que, se $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{\theta}$ for verdadeira,

$$T = \frac{(\mathbf{Cb} - \boldsymbol{\theta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb} - \boldsymbol{\theta})}{ms^2} \tag{1.81}$$

tem distribuição de F com m e $n– p$ graus de liberdade.

Admitamos que um mesmo conjunto de dados ($\mathbf{X}$, $\mathbf{y}$) seja utilizado para testar várias hipóteses $H_{0i} : \mathbf{C}_i\boldsymbol{\beta} = \boldsymbol{\theta}_i$ (i = 1, 2, ...). A aplicação sucessiva de (1.81) para efetuar esses testes só é válida, a rigor, se as diferentes distribuições de qui-quadrado, associadas aos numeradores dos diferentes valores de F calculados, forem independentes entre si. Para isso é necessário que tenhamos $\mathbf{Q}_i\mathbf{Q}_k = 0$ para $i \neq k$, onde $\mathbf{Q}_i = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_i'[\mathbf{C}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_i']^{-1}\mathbf{C}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Isto é conseguido se tivermos

$$\mathbf{C}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_k' = \mathbf{0} \text{ para } i \neq k$$

No caso do modelo

$$Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + u_j, \qquad j = 1, ..., n$$

por exemplo, as hipóteses $H_{01} : \beta_1 = 0$ e $H_{02} : \beta_2 = 0$ só são independentes se $\sum X_{1j}X_{2j} = 0$. Isso porque temos

$$\mathbf{C}_1 = [1 \quad 0], \qquad \mathbf{C}_2 = [0 \quad 1],$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum X_{1j}^2 & \sum X_{1j}X_{2j} \\ \sum X_{1j}X_{2j} & \sum X_{2j}^2 \end{bmatrix}$$

e

$$\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_2' = \frac{-\sum X_{1j}X_{2j}}{\sum X_{1j}^2 X_{2j}^2 - (\sum X_{1j}X_{2j})^2}$$

Então, a rigor, só podemos utilizar o teste F ou o teste t para testar essas duas hipóteses, nesse modelo, se as colunas da matriz $\mathbf{X}$ forem ortogonais entre si.

Admitamos que $\sum X_{1j}X_{2j} \neq 0$ e, apesar disso, os dois testes são efetuados, comprando-se o valor de F (ou de t) calculado com o valor crítico ao nível de significância de 5%, obtido na tabela. Nesse caso, como os testes não são independentes, o nível de significância verdadeiro é maior do que 5%. O erro depende, obviamente, do grau de multicolinearidade existente.

Tendo em vista suas aplicações, é conveniente escrever a expressão (1.81) como um teste *F* para a hipótese $H_0 : \mathbf{C\beta} = \boldsymbol{\theta}$, ou seja,

$$F = \frac{1}{ms^2}(\mathbf{Cb} - \boldsymbol{\theta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb} - \boldsymbol{\theta}) \tag{1.82}$$

De acordo com o que foi demonstrado, o mesmo valor é obtido por meio de (1.73), isto é,

$$F = \frac{\dfrac{S_* - S}{m}}{\dfrac{S}{n-p}}, \tag{1.83}$$

sendo *S* a S.Q.Res. do modelo sem restrição e S_* a S.Q.Res. do modelo restrito (com $\mathbf{C\beta} = \boldsymbol{\theta}$).

Uma adaptação da relação (1.82) permite obter a *região de confiança* para um conjunto de combinações lineares dos parâmetros definido por $\mathbf{C\beta}$.

Seja F_0 o valor crítico de *F*, com *m* e $n-p$ graus de liberdade, ao nível de significância de $(100-\varphi)\%$, que corresponde a um nível de confiança de $\varphi\%$. Os pontos da região de $\varphi\%$ de confiança para o conjunto de combinações lineares $\mathbf{C\beta}$ são aqueles que satisfazem a condição

$$\frac{1}{ms^2}(\mathbf{Cb} - \mathbf{C\beta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb} - \mathbf{C\beta}) < F_0$$

ou

$$(\mathbf{C\beta} - \mathbf{Cb})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C\beta} - \mathbf{Cb}) < ms^2 F_0 \tag{1.84}$$

Note-se que, conforme (1.82), essa condição implica não rejeitar, ao nível de significância de $(1-\varphi)\%$, a hipótese $\mathbf{C\beta} = \boldsymbol{\theta}$.

Se **C** for uma matriz identidade, a expressão (1.84) fornecerá a região de confiança para todos os parâmetros do modelo. Para o caso particular de apenas dois parâmetros, verifica-se que essa região de confiança é a área interna de uma elipse.

No caso particular em que **C** é igual a uma linha de uma matriz identidade, a expressão (1.84) fornece o *intervalo* de confiança para um único parâmetro.

Como exemplo de aplicação da expressão (1.84), vamos determinar a região de 95% de confiança para os parâmetros β_1 e β_2 de um modelo de regressão linear de *Y* contra X_1 e X_2, para a amostra de 5 observações apresentada na seção 1.6. Para o modelo (1.53), utilizado nos cálculos, o vetor de parâmetros é

$$\boldsymbol{\beta} = \begin{bmatrix} \gamma \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Como desejamos obter a região de confiança para β_1 e β_2, adotamos a matriz

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

de maneira que

$$\mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

A característica (ou posto) da matriz **C** é $m = 2$.

Obtemos

$$\mathbf{C}\boldsymbol{\beta} - \mathbf{C}\mathbf{b} = \begin{bmatrix} \beta_1 - 2 \\ \beta_2 + 3 \end{bmatrix}$$

e, tendo em vista a matriz $\mathbf{X}'\mathbf{X}$ apresentada logo após a tabela 1.1,

$$[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1} = \begin{bmatrix} 18 & 28 \\ 28 & 72 \end{bmatrix}$$

Conforme resultados apresentados na tabela 1.2, temos $s^2 = 4$, com 2 graus de liberdade. O valor crítico de *F*, ao nível de significância de 5%, com 2 e 2 graus de liberdade, é $F_o = 19{,}00$.

Substituindo esses resultados em (1.84), obtemos

$$\begin{bmatrix} \beta_1 - 2 & \beta_2 + 3 \end{bmatrix} \begin{bmatrix} 18 & 28 \\ 28 & 72 \end{bmatrix} \begin{bmatrix} \beta_1 - 2 \\ \beta_2 + 3 \end{bmatrix} < 152$$

Podemos definir $g_1 = \beta_1 - 2$ e $g_2 = \beta_2 + 3$, o que corresponde a fazer uma translação do sistema de eixos, cuja origem passa a ser o ponto com coordenadas iguais às estimativas dos parâmetros, obtendo a equação

$$[g_1 \quad g_2]\begin{bmatrix} 18 & 28 \\ 28 & 72 \end{bmatrix}\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} < 152$$

Essa região de confiança é delimitada pela elipse traçada na figura 1.1 e cuja equação é

$$[g_1 \quad g_2]\begin{bmatrix} 18 & 28 \\ 28 & 72 \end{bmatrix}\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = 152$$

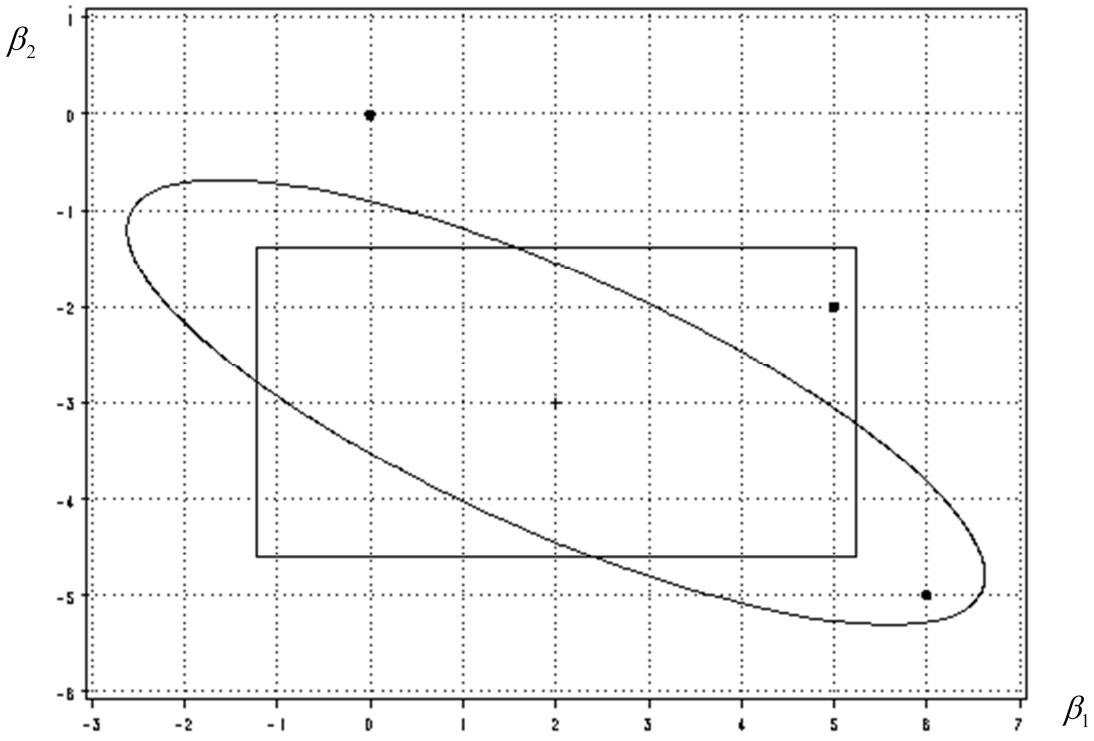


Figura 1.1. A região de 95% de confiança para β_1 e β_2 e o retângulo cujos lados são os intervalos de 95% de confiança para β_1 e para β_2.

A mesma figura mostra o retângulo cujos lados correspondem aos intervalos de 95% de confiança para β_1 e para β_2.

Os limites do intervalo de 95% de confiança para β_1 são

$$b_1 \pm t_0 \sqrt{\hat{V}(b_1)}$$

Considerando os resultados obtidos na seção 1.6, os limites do intervalo de 95% de confiança para β_1 são

$$2 \pm 4{,}303 \cdot 0{,}75$$

ou $$-1{,}23 < \beta_1 < 5{,}23$$

Analogamente, o intervalo de 95% de confiança para β_2 é

$$-4{,}61 < \beta_2 < -1{,}39$$

Consideremos, na figura 1.1, o ponto $\beta_1 = \beta_2 = 0$. Como ele está fora da região de 95% de confiança para β_1 e β_2, conclui-se que a hipótese $H_0 : \beta_1 = \beta_2 = 0$ é rejeitada ao nível de significância de 5%, confirmando resultado já obtido na seção 1.6. Esse ponto também está fora do retângulo, pois $\beta_2 = 0$ não pertence ao intervalo de 95% de confiança para β_2, mostrando que a hipótese $H_0 : \beta_2 = 0$ é rejeitada ao nível de significância de 5% (considerando um teste bilateral).

A figura 1.1 mostra que há pontos que levam a resultados aparentemente contraditórios. Consideremos, por exemplo, o ponto $\beta_1 = 5$ e $\beta_2 = -2$. Como esse ponto está fora da elipse, a hipótese $H_0 : \beta_1 = 5$ e $\beta_2 = -2$ é rejeitada ao nível de significância de 5%. Mas, como o ponto está dentro do retângulo, não se rejeita, ao nível de 5%, nem a hipótese $H_0 : \beta_1 = 5$, nem a hipótese $H_0 : \beta_2 = -2$. Isso ocorre porque ao testar a hipótese $H_0 : \beta_1 = 5$ e $\beta_2 = -2$ estamos considerando a distribuição *conjunta* de b_1 e b_2 e o teste das hipótese separadas para β_1 e para β_2 leva em consideração as respectivas distribuições marginais. Pode-se dizer que a possibilidade de resultados aparentemente contraditórios como esses se deve ao fato de que um retângulo não pode coincidir com uma elipse.

Devido à inclinação da elipse (associada à covariância entre b_1 e b_2), o ponto $\beta_1 = 6$ e $\beta_2 = -5$ fica dentro da região de confiança para β_1 e β_2, mas fora dos dois intervalos de confiança. Mantendo sempre o nível de significância de 5%, não se rejeita a hipótese $H_0 : \beta_1 = 6$ e $\beta_2 = -5$, mas rejeita-se tanto a hipótese $H_0 : \beta_1 = 6$ como a hipótese $H_0 : \beta_2 = -5$.

1.9. Teste de mudança estrutural

Vamos admitir que a variável Y_j está relacionada com as variáveis explanatórias de acordo com modelo

$$Y_j = \sum_{i=0}^{k} \beta_i X_{ij} + u_j \tag{1.85}$$

ou

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$$

Se $X_{0j} = 1$ para todo *j*, β_0 é o termo constante ou intercepto.

Dispomos de $n = n_1 + n_2$ observações, sendo n_1 observações referentes a uma situação (categoria, região ou período) e n_2 observações referentes a outra situação. Seja $\mathbf{y}_1$ o vetor-coluna com os n_1 valores de Y_j na primeira situação e seja $\mathbf{y}_2$ o vetor-coluna com os n_2 valores de Y_j na segunda situação. Analogamente, seja $\mathbf{X}_1$ a matriz $n_1 \times p$ (com $p = k+1$) referente à primeira situação e seja $\mathbf{X}_2$ a matriz $n_2 \times p$ referente à segunda situação.

Desejamos saber se a relação linear entre as variáveis é a mesma nas duas situações, ou seja, se a "estrutura" caracterizada pelo vetor $\boldsymbol{\beta}$ é a mesma nas duas situações. Uma maneira de formular o teste de hipóteses é definir uma variável binária Z_j, com $Z_j = 0$ para todas as observações da primeira situação e $Z_j = 1$ para todas as observações da segunda situação, e considerar o modelo

$$Y_j = \sum_{i=0}^{k} \beta_i X_{ij} + \sum_{i=0}^{k} \gamma_i Z_j X_{ij} + u_j \tag{1.86}$$

Na primeira situação, com $Z_j = 0$, o coeficiente de X_{ij} é β_i. Na segunda situação, com $Z_j = 1$, o coeficiente de X_{ij} é $\beta_i + \gamma_i$. Os parâmetros γ_i (com $i = 0, 1, ..., k$) são as mudanças nos coeficientes entre as duas situações. A hipótese de que a "estrutura" é a mesma nas duas situações corresponde à hipótese

$$H_0 : \gamma_o = \gamma_1 = \ldots = \gamma_k = 0$$

O modelo (1.86) é equivalente a

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}\,, \tag{1.87}$$

com

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \qquad \text{e} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{X}_2 \end{bmatrix}$$

Note-se que **X** é uma matriz $n \times 2p$. O vetor das estimativas dos parâmetros, com 2*p* elementos, é composto por um vetor-coluna **b** com as estimativas dos β_i e um vetor-coluna **c** com as estimativas dos γ_i (com $i = 0, 1, ..., k$).

Fazendo regressões separadas para cada situação, obtemos

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 \qquad \text{e} \qquad \mathbf{b}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2$$

As respectivas somas de quadrados de resíduos são

$$S_1 = \mathbf{y}_1'\mathbf{y}_1 - \mathbf{b}_1'\mathbf{X}_1'\mathbf{y}_1,$$

com $n_1 - p$ graus de liberdade, e

$$S_2 = \mathbf{y}_2'\mathbf{y}_2 - \mathbf{b}_2'\mathbf{X}_2'\mathbf{y}_2$$

com $n_2 - p$ graus de liberdade.

Pode-se demonstrar que o modelo (1.86) produz os mesmos valores para a estimativa de Y_j que as duas regressões separadas, sendo $\mathbf{b} = \mathbf{b}_1$ e $\mathbf{b} + \mathbf{c} = \mathbf{b}_2$. Assim, a soma de quadrados dos resíduos do modelo (1.87) é

$$S = S_1 + S_2, \tag{1.88}$$

com $n_1 - p + n_2 - p = n - 2p$ graus de liberdade.

Considerando a hipótese de que $\gamma_0 = \gamma_1 = \ldots = \gamma_k = 0$ como uma restrição, fazemos a regressão de

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \quad \text{contra} \quad \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix},$$

obtendo $\mathbf{b}_* = (\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-1}(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2)$

e a soma de quadrados residual

$$S_* = \mathbf{y}_1'\mathbf{y}_1 + \mathbf{y}_2'\mathbf{y}_2 - \mathbf{b}_*'(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2),$$

com $n - p$ graus de liberdade.

Tendo em vista a expressão (1.83) e o resultado (1.88), o teste *F* para a hipótese de que não há mudança estrutural ($H_0 : \gamma_0 = \gamma_1 = \ldots = \gamma_k = 0$) é dado por

$$F = \frac{\dfrac{S_* - (S_1 + S_2)}{p}}{\dfrac{S_1 + S_2}{n - 2p}} \tag{1.89}$$

Esse procedimento é conhecido como "teste de Chow". Em resumo, ele consiste nas seguintes etapas:

a) Fazer uma regressão com os dados referentes à primeira situação ($\mathbf{y}_1$ contra $\mathbf{X}_1$), obtendo a soma de quadrados residual S_1.

b) Fazer uma regressão com os dados referentes à segunda situação ($\mathbf{y}_2$ contra $\mathbf{X}_2$), obtendo a soma de quadrados residual S_2.

c) Usando o mesmo modelo, fazer uma regressão com os dados de todas as $n = n_1 + n_2$ observações, obtendo a soma de quadrados residual S_*.

d) Calcular o teste *F* com *p* e $n - 2p$ graus de liberdade, de acordo com a expressão (1.89).

1.10. Erros de especificação

Retomando a análise desenvolvida na seção 1.7, vamos admitir que Y_j seja uma função linear das variáveis incluídas na matriz **X** e que esta é decomposta como indicado em (1.60). A correspondente decomposição de **b** é dada em (1.61) e a respectiva decomposição do vetor de parâmetros é

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

Sabemos que

$$E(\mathbf{b}) = E\begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

Segue-se, obviamente, que

$$E(\mathbf{b}_1)=\boldsymbol{\beta}_1 \qquad \text{e} \qquad E(\mathbf{b}_2)=\boldsymbol{\beta}_2 \tag{1.90}$$

Se for cometido um erro de especificação do modelo de regressão, admitindo que Y_j seja uma função linear apenas das variáveis contidas em $\mathbf{X}_2$, seriam calculadas as estimativas de parâmetros dadas por

$$\mathbf{b}_2^* = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} \tag{1.91}$$

De acordo com (1.67), temos

$$\mathbf{b}_2 = \mathbf{b}_2^* - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1$$

ou

$$\mathbf{b}_2^* = \mathbf{b}_2 + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1$$

Tendo em vista (1.90), se a matriz **X** for considerada fixa, segue-se que

$$E(\mathbf{b}_2^*)=\boldsymbol{\beta}_2 + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\boldsymbol{\beta}_1 \tag{1.92}$$

Note-se que cada coluna da matriz $(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1$ é formada pelos coeficientes de regressão de uma coluna de $\mathbf{X}_1$ contra $\mathbf{X}_2$.

Admite-se, a seguir, que $\mathbf{X}_1$ tem apenas uma coluna, com os valores da variável X_{1j}. Nesse caso a matriz $(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1$ é o vetor-coluna dos coeficientes de regressão de X_{1j} contra $\mathbf{X}_2$ e $\boldsymbol{\beta}_1$ tem apenas um elemento, o escalar β_1. Seja b_h^* o *h*-ésimo elemento de $\mathbf{b}_2^*$ e seja $\hat{\theta}_5$ o *h*-ésimo coeficiente da regressão de X_{1j} contra $\mathbf{X}_2$. Então

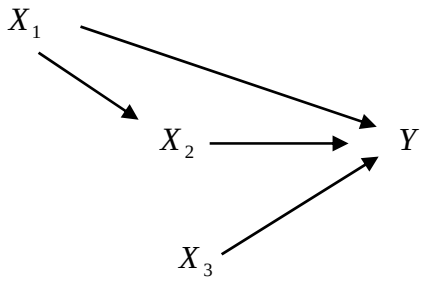
$$E\left(b_h^*\right) = \beta_h + \hat{\theta}_h \beta_1 \qquad (1.93)$$

Essa expressão mostra que o coeficiente estimado com a regressão incompleta, sem considerar a variável X_{1j}, será um estimador não-tendencioso de β_h apenas se $\hat{\theta}_h = 0$ e/ou $\beta_1 = 0$. Daí decorre a preocupação dos econometristas em incluir no modelo todas as variáveis (controles) relevantes. Se uma variável relevante for omitida, o coeficiente estimado é, em geral, tendencioso, com sua esperança matemática incluindo não só o efeito direto da variável (β_h), mas também o efeito associado à ausência do controle relevante $\left(\hat{\theta}_h \beta_1\right)$.

Os livros-texto de econometria costumam mostrar que a inclusão de controles irrelevantes não prejudica a não-tendenciosidade dos estimadores dos parâmetros β_h, embora tenda a torná-los menos precisos. Estabelece-se, assim, a ideia de que o perigo maior reside na omissão de controles, e não no uso excessivo de controles.

Mas há, sim, situações em que a inclusão de controles indevidos pode levar a conclusões erradas[5].

Vamos imaginar que as variáveis X_1, X_2, X_3 e Y estão relacionadas conforme indica o esquema a seguir, no qual a seta indica a existência de efeito de uma variável sobre outra.



Se for feita a regressão múltipla de Y contra X_1, X_2 e X_3, o coeficiente de regressão de X_1 capta apenas o efeito *direto* de X_1 sobre Y. Se o pesquisador estiver interessado em avaliar a influência de X_1 sobre Y, incluindo a que ocorre via X_2, ele não deve incluir X_2 como variável explanatória (ou controle). É claro que poderia ser interessante ajustar um

[5] O problema é analisado por Angrist e Pischke (2009, p. 64-68) em uma seção intitulada "*Bad Control*".

sistema de equações, incluindo uma equação com X_2 como função de X_1 e uma outra equação de Y como função de X_1, X_2 e X_3. O sistema de equações permitiria estimar tanto o efeito direto como o efeito indireto de X_1 sobre Y. Mas se o pesquisador está interessado no efeito total de X_1 sobre *Y*, incluindo tanto o efeito direto como o indireto (via X_2), estimar o sistema de equações é uma complicação desnecessária. O exercício 31 apresenta dados numéricos artificiais que ilustram a questão.

Um pesquisador deseja avaliar se as transferências de renda do programa Bolsa Família (X_1) afetam a pobreza (Y) nas Unidades da Federação. Ele estima um modelo de regressão múltipla de Y e inclui como controles a renda média (X_2) e o índice de Gini (X_3) da distribuição da renda em cada Unidade da Federação. Tal modelo de regressão seria claramente inapropriado, pois o efeito de X_1 sobre Y se dá, essencialmente via X_2 e X_3. É aumentando a renda dos pobres que as transferências reduzem a pobreza e o crescimento da renda dos pobres se reflete no crescimento da renda média e na redução da desigualdade. Incluindo a renda média e o índice de Gini como controles fica quase impossível captar o efeito das transferências do programa Bolsa Família sobre a pobreza[6].

Além da escolha das variáveis explanatórias e dos controles a serem incluídos, a especificação apropriada de um modelo de regressão envolve a decisão sobre a forma matemática: linear nas variáveis, linear nos logaritmos, polinômio (de que grau?) etc.

Desnecessário dizer que a especificação de um modelo apropriado depende de integrar o conhecimento do fenômeno analisado e da metodologia estatística mais conveniente.

Exercícios

1. Dispomos das 6 observações das variáveis X_1, X_2 e *Y*, apresentadas na tabela a seguir:

X_1	X_2	Y
2	5	6
3	16	25
4	13	2
5	26	25
6	25	6
4	17	8

[6] Devo reconhecer que esse exemplo de uso de controles inapropriados é inspirado em artigos publicados na Revista Brasileira de Economia (Marinho, Linhares e Campelo, 2011, e Marinho e Araújo, 2010).

Pode-se verificar que

$\sum X_1 = 24$	$\sum X_2 = 102$	$\sum Y = 72$
$\sum X_1^2 = 106$	$\sum X_2^2 = 2040$	$\sum Y^2 = 1390$
$\overline{X}_1 = 4$	$\overline{X}_2 = 17$	$\overline{Y} = 12$
$\sum X_1 X_2 = 458$	$\sum X_1 Y = 288$	$\sum X_2 Y = 1392$

a) Obtenha as estimativas dos parâmetros da regressão linear múltipla de *Y* contra X_1 e X_2.

b) Para modelo $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u$ e adotando as pressuposições usuais a respeito do erro (*u*), teste a hipótese $H_0 : \beta_1 = \beta_2 = 0$ ao nível de significância de 1%.

c) Teste, ao nível de significância de 1%, a hipótese $H_0 : \beta_1 = 0$ contra $H_A : \beta_1 < 0$.

d) Calcule o coeficiente de determinação e o coeficiente de determinação corrigido para graus de liberdade.

e) Calcule os três coeficientes de correlação simples entre as 3 variáveis. Note que o coeficiente de correlação entre *Y* e X_1 é zero. É correto afirmar, então, que "como X_1 não tem relação linear com *Y*, é apenas X_2 que ajuda a explicar as variações de *Y*"? Discuta (explique).

2. Considere o modelo estatístico de uma função de produção tipo Cobb-Douglas

$$Z_j = \theta W_{1j}^{\beta_1} W_{2j}^{\beta_2} \varepsilon_j \quad ,$$

onde Z_j é o valor da produção, W_{ij} é a quantidade de mão de obra empregada e W_{2j} é o valor do capital (inclusive terra) na *j*-ésima unidade de produção. Admite-se que $u_j = \ln \varepsilon_j$ são erros aleatórios não correlacionados entre si, com distribuição normal de média zero e variância σ^2.

São dados os seguintes valores obtidos de uma amostra de 9 unidades de produção:

$\ln W_1$	$\ln W_2$	$\ln Z = Y$
2	3	4,4
0	2	1,6
1	3	3,4
1	4	4,2
3	4	5,6
2	4	4,6
3	5	5,8
4	6	6,4
2	5	5,4

Verifica-se que $\sum Y = 41{,}4$, $\sum Y^2 = 207{,}4$ e $\sum y^2 = 16{,}96$.

a) Determine as estimativas de $\alpha = \ln\theta$, β_1 e β_2 de acordo com o método de mínimos quadrados.

b) Determine a estimativa não-tendenciosa de σ^2.

c) Calcule o coeficiente de determinação corrigido para graus de liberdade.

d) Teste, ao nível de significância de 1%, a hipótese de que a elasticidade parcial do valor da produção em relação a W_1 é igual a zero.

e) Teste, ao nível de significância de 1%, a hipótese de que a elasticidade parcial do valor da produção em relação a W_2 é igual a zero.

f) Teste, ao nível de significância de 1%, a hipótese de que as duas elasticidades parciais são (simultaneamente) iguais a zero.

g) Discuta a existência, nesses dados, de um problema de multicolinearidade elevada (indicando como medir o grau de multicolinearidade e avaliar suas consequências, eventualmente utilizando resultados dos itens anteriores).

h) Teste, ao nível de significância de 10%, a hipótese de que a função de produção é linearmente homogênea (os rendimentos médios não são afetados pela escala de produção).

3. Admita que em uma regressão linear múltipla com *p* parâmetros, o teste *t* referente à hipótese $H_0 : \beta_h = 0$ tenha valor absoluto menor do que 1, isto é,

$$| t_h | = \frac{| b_h |}{s(b_h)} < 1$$

Demonstre que, nesse caso, a exclusão da variável X_h da regressão faz com que diminua o valor do quadrado médio do resíduo, isto é, que o quadrado médio do resíduo da regressão sem X_h (com $p - 1$ parâmetros) é menor do que o quadrado médio do resíduo da regressão completa (com p parâmetros).

4. Admite-se que as variáveis X_1, X_2 e Y estão relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j \ ,$$

onde os u_j são erros aleatórios, com as pressuposições usuais.

A partir de uma amostra com 10 observações foram obtidos os seguintes valores:

$\sum X_1 = 60$	$\sum X_2 = 30$	$\sum Y = 120$
$\sum X_1^2 = 440$	$\sum X_2^2 = 110$	$\sum Y^2 = 1936$
$\sum X_1 X_2 = 212$	$\sum X_1 Y = 888$	$\sum X_2 Y = 456$
$\sum x_1^2 = 80$	$\sum x_2^2 = 20$	$\sum x_1 x_2 = 32$

a) Determine as estimativas de α, β_1 e β_2 de acordo com o método de mínimos quadrados ordinários.

b) Obtenha a estimativa não-tendenciosa da variância do erro *u*.

c) Calcule o coeficiente de determinação e o coeficiente de determinação corrigido para graus de liberdade.

d) Teste, ao nível de significância de 10%, a hipótese de que $\beta_1 = 0$.

e) Teste, ao nível de significância de 5%, a hipótese de que $\beta_1 + \beta_2 = 3$.

f) Teste, ao nível de significância de 5%, a hipótese de que $\beta_1 = 1$ e $\beta_2 = 2$.

5. É dada uma série de 9 valores anuais da variável *Y*. Admite-se que *Y* varia linearmente em função do tempo (em anos), mas acredita-se que ocorreu uma mudança estrutural entre a 4ª e a 5ª observação, de maneira que haveria uma tendência linear durante os 4 primeiros anos da série e uma tendência linear distinta durante os 5 últimos anos.

Verifica-se que $\sum Y = 237$, $\sum Y^2 = 7229$ e $\sum y^2 = 988$.

a) Estime as taxas aritméticas de crescimento anual de *Y* nos dois períodos.
b) Verifique se a mudança estrutural é estatisticamente significativa. Sugere-se fazer o teste com base nas regressões simples, como indicado por Chow.
c) Há diferença estatisticamente significativa entre as taxas aritméticas de crescimento de *Y* nos dois períodos?

Ano	*Y*
1º	10
2º	15
3º	18
4º	19
5º	29
6º	33
7º	34
8º	37
9º	42

Adote um nível de significância de 5% em todos os testes de hipóteses desta questão.

6. É dada uma amostra com 9 pares de valores de *X* e *Y*. Admite-se que *Y* é função linear de *X*, mas há razões para acreditar que ocorreu uma mudança estrutural entre a 3ª e a 4ª observação e uma outra mudança estrutural entre a 6ª e a 7ª observação. Estenda o método de Chow e faça um teste *F* para a hipótese de que os 9 pares pertencem a uma mesma relação linear, contra a hipótese alternativa de que há 3 "estruturas" distintas (cada uma incluindo uma sequência de 3 pares de valores).

X	*Y*
0	8
1	4
2	6
3	32
4	32
5	44
6	30
7	40
8	56

7. Com base em uma amostra com 40 observações foi estimada a equação de regressão de *Y* contra X_1, X_2 e X_3, considerando o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + u_j$$

O coeficiente de determinação parcial entre *Y* e X_1, dados X_2 e X_3, é igual a 0,1.

Teste, ao nível de significância de 5%, a hipótese de que $\beta_1 = 0$.

8. Admite-se que as variáveis W_i, Z_{1i} e Z_{2i} estão relacionadas de acordo com o modelo

$$W_i = \theta Z_{1i}^{\beta_1} Z_{2i}^{\beta_2} \varepsilon_i \qquad (1)$$

Admite-se, também, que $u_i = \ln \varepsilon_i$ são erros com distribuição normal com média igual a zero, variância σ^2 e independentes entre si.

É dada uma amostra de 8 valores das variáveis $X_{1i} = \ln Z_{1i}$, $X_{2i} = \ln Z_{2i}$ e $Y_i = \ln W_i$:

X_{1i}	X_{2i}	Y_i
2	4	4,4
3	5	5,2
4	5	5,3
3	6	5,1
4	7	5,9
5	6	6,1
5	7	6,8
6	8	7,6

a) Como o modelo (1) é transformado em um modelo de regressão linear?

b) Estime os parâmetros θ, β_1 e β_2.

c) Determine a estimativa de σ^2.

d) Calcule o coeficiente de determinação da regressão linear.

e) Teste, ao nível de significância de 1%, a hipótese $H_0 : \beta_1 = 0$ contra $H_A : \beta_1 > 0$.

f) Idem, $H_0 : \beta_2 = 0$ contra $H_A : \beta_2 > 0$.

g) Qual é a estimativa da elasticidade parcial de *W* em relação a Z_2?

h) Admitindo que (1) é o modelo de uma função de produção, teste a hipótese de que os rendimentos à escala são constantes, isto é, que a função é linearmente homogênea, adotando um nível de significância de 5%.

9. Dispomos de valores de *X* e *Y* para uma amostra de 5 empresas da categoria I e para uma amostra de 5 empresas da categoria II.

Categoria I	
X	*Y*
2	7
4	11
6	11
8	17
10	19

Categoria II	
X	*Y*
2	6
4	8
6	13
8	16
10	22

Admite-se que dentro de uma categoria, *Y* é uma função linear de *X*, podendo haver diferenças tanto no intercepto como no coeficiente angular das retas das duas categorias.

a) Estime a reta para cada categoria.

b) Faça o teste de Chow para a diferença estrutural entre categorias, adotando um nível de significância de 10%.

c) Considere que seja definida uma variável binária (*Z*) para distinguir as duas categorias e, utilizando as 10 observações, seja estimado o modelo

$$Y_i = \alpha + \beta X_i + \delta Z_i + \gamma Z_i X_i + u_i$$

Qual é a S.Q.Res. dessa regressão? Qual é o valor do seu coeficiente de determinação?

10. Sejam b_1 e b_2 as estimativas de mínimos quadrados dos coeficientes angulares de uma regressão linear de $\ln Y_i$ contra $\ln X_{1i}$ e $\ln X_{2i}$.

Sejam c_1 e c_2 as estimativas de mínimos quadrados dos coeficientes angulares de uma regressão linear de $\ln \frac{Y_i}{X_{2i}}$ contra $\ln \frac{X_{1i}}{X_{2i}}$ e $\ln X_{2i}$.

Deduza as expressões de c_1 e c_2 em função de b_1 e b_2.

Observação: as expressões de c_1 e c_2 em função de b_1 e b_2 podem ser obtidas com base nas relações matemáticas correspondentes aos dois modelos de regressão, mas a demonstração deve, necessariamente, ser feita com base nos estimadores de mínimos quadrados.

11. São dados os valores de X_1, X_2 e *Y* observados em uma amostra aleatória:

X_1	X_2	Y
2	8	12
3	5	4
4	6	10
5	7	20
6	4	4

Verifica-se que:

$\sum X_1 = 20$ $\quad \sum X_2 = 30$ $\quad \sum Y = 50$

$\sum X_1^2 = 90$ $\quad \sum X_2^2 = 190$ $\quad \sum Y^2 = 676$

$\sum X_1 X_2 = 114$ $\quad \sum X_1 Y = 200$ $\quad \sum X_2 Y = 332$

Admitimos que X_1, X_2 e *Y* estão relacionados de acordo com o modelo $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$, onde os u_j são variáveis aleatórias independentes com distribuição normal de média zero e variância σ^2.

a) Obtenha as estimativas lineares não-tendenciosas de variância mínima de α, β_1 e β_2.

b) Teste (sempre ao nível de significância de 5%) a hipótese $H_0 : \beta_1 = \beta_2 = 0$

c) Teste $H_0 : \beta_1 = 0$ contra $H_A : \beta_1 > 0$.

d) Teste $H_0 : \beta_1 = -1$ e $\beta_2 = 1$.

e) Calcule os coeficientes de correlação parcial $r_{Y2.1}$ e $r_{Y1.2}$.

f) Seja δ a variação do valor de $E(Y)$ quando X_1 e X_2 aumentam de uma unidade cada um $(\Delta X_1 = \Delta X_2 = 1)$. Qual é a estimativa de δ? Determine o intervalo de 95% de confiança para δ.

12. Considere o modelo $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$ ou, em notação matricial, $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$. Admita que os valores de X_1 e X_2 dados sejam tais que $|\mathbf{X'X}| \neq 0$ e que os valores de Y_j sejam todos iguais a 1, isto é, $\mathbf{y} = \boldsymbol{\iota}$. Qual será o vetor das estimativas dos parâmetros? O que se pode dizer dos vetores $\hat{\mathbf{y}}$ e $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$?

13. Dispõe-se das 8 observações das variáveis X_1, X_2 e Y apresentados na tabela a seguir:

X_1	X_2	Y
1	7	8
11	7	32
6	12	27
6	2	13
1	7	11
6	12	24
11	7	27
6	2	18

a) Obtenha as estimativas dos parâmetros da regressão linear múltipla de Y contra X_1 e X_2 (incluindo um termo constante).

b) Estime a variância residual.

c) Teste, ao nível de significância de 5%, a hipótese $H_0 : \beta_1 = \beta_2$.

d) Note que há apenas 4 pares de valores distintos para X_1 e X_2. Considere cada um desses pares como um "tratamento" distinto para Y. Estabeleça um modelo em que os efeitos dos 4 tratamentos são captados por variáveis binárias. Para facilitar as contas, é aconselhável considerar um modelo com 4 variáveis binárias e sem termo constante. Estime os parâmetros do modelo e estime a variância residual.

e) Na regressão ajustada inicialmente admite-se que há uma relação *linear* entre Y e X_1 e X_2. No modelo do item (d) não se impõe essa linearidade na relação entre Y e as variáveis explanatórias. Esse último é, portanto, um modelo menos restritivo. Verifique se há razões para rejeitar a hipótese de que Y é uma função linear de X_1 e X_2, adotando um nível de significância de 10%.

14. Considere o modelo $Y_i = \beta X_i + u_i$, com X_i fixos, $E(u_i) = 0$, $E(u_i^2) = \sigma^2$ e $E(u_i u_j) = 0$ para $i \neq j$.

Sabe-se que o estimador de mínimos quadrados para β é $b = \dfrac{\sum X_i Y_i}{\sum X_i^2}$, não-tendencioso, com $V(b) = \dfrac{\sigma^2}{\sum X_i^2}$.

Um estimador alternativo para β é $\hat{\beta} = \overline{Y} / \overline{X}$, que é a inclinação da reta unindo a origem do sistema de eixos ao ponto $\overline{X}, \overline{Y}$.

a) Prove que $\hat{\beta}$ é um estimador linear não-tendencioso.

b) Deduza a expressão que dá $V(\hat{\beta})$ em função de σ^2 e dos valores de X.

c) Prove (sem utilizar o teorema de Gauss-Markov) que $V(\hat{\beta}) \geq V(b)$. Em que condições tem-se $V(\hat{\beta}) = V(b)$?

15. Uma equação de regressão linear múltipla é ajustada a uma amostra de n_1 observações. Seja $\mathbf{X}_1$ a matriz $n_1 \times p$ das variáveis explanatórias e seja $\mathbf{y}_1$ o vetor dos valores da variável dependente. O vetor-coluna das estimativas dos parâmetros é $\mathbf{b}_1 = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{y}$ e a soma de quadrados dos resíduos dessa regressão é

$$Q_1 = \mathbf{y}_1' \mathbf{M}_1 \mathbf{y}_1, \text{ onde } \mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$$

Vamos admitir que sejam acrescentadas n_2 observações e que, ao mesmo tempo, se crie uma variável binária para cada nova observação. Então, o novo vetor dos valores da variável dependente será

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \text{ com } n_1 + n_2 \text{ elementos,}$$

e, sendo **0** uma matriz $n_1 \times n_2$ de zeros, a matriz das variáveis explanatórias fica

$$\mathbf{W} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{I} \end{bmatrix} \text{ , com } n_1 + n_2 \text{ linhas e } p + n_2 \text{ colunas.}$$

O vetor-coluna das estimativas dos parâmetros é

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} \text{ , sendo que } \mathbf{d}_1 \text{ tem } p \text{ elementos e } \mathbf{d}_2 \text{ tem } n_2 \text{ elementos.}$$

a) Obtenha a relação entre $\mathbf{d}_1$ e $\mathbf{b}_1$.

b) Prove que a soma de quadrados dos resíduos dessa nova regressão linear múltipla é a mesma (igual a Q_1).

16. Em uma regressão múltipla o teste da hipótese $H_0 : \beta_i = 0$ pode ser feito através de

$$t = \frac{b_i}{s(b_i)} .$$

A "contribuição" da *i*-ésima variável pode ser testada através de

$$F = \frac{\text{S.Q."Contribuição de } X_i\text{"}}{s^2},$$

onde s^2 é o Q.M.Res. da regressão completa e S.Q."Contribuição de X_i" é a diferença entre a soma de quadrados de regressão com toda as *k* variáveis explanatórias e a soma de quadrados de regressão da equação estimada sem a variável X_i.

Prove que $t^2 = F$.

SUGESTÃO: Sem perda de generalidade, admita que o coeficiente a ser testado é o da última variável explanatória. O vetor-coluna com os valores dessa variável será indicado por **x**. Seja **X** a matriz com as demais variáveis explanatórias e seja **y** o vetor-coluna com os valores da variável dependente. Na regressão completa a matriz das variáveis explanatórias é $[\mathbf{X} \quad \mathbf{x}]$. Verifica-se que o vetor das estimativas dos parâmetros dessa regressão é dado por

$$\begin{bmatrix} \mathbf{c} \\ d \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{x} \\ \mathbf{x}'\mathbf{X} & \mathbf{x}'\mathbf{x} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{x}'\mathbf{y} \end{bmatrix} \quad ,$$

onde **c** é um vetor-coluna com *k* elementos e *d* é um escalar.

a) Obtenha as expressões para **c** e *d*.

Verifica-se que

$$d = (\mathbf{x}'\mathbf{M}\mathbf{x})^{-1}\mathbf{x}'\mathbf{M}\mathbf{y} \text{ , onde } \mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

b) Mostre que a introdução da variável **x** reduz a S.Q.Res. de

$$\frac{(\mathbf{x}'\mathbf{M}\mathbf{y})^2}{\mathbf{x}'\mathbf{M}\mathbf{x}}$$

c) Mostre que $s^2(d) = s^2(b_i) = (\mathbf{x}'\mathbf{M}\mathbf{x})^{-1}s^2$

d) Finalmente, mostre que

$$t^2 = F = \frac{(\mathbf{x}'\mathbf{M}\mathbf{y})^2}{(\mathbf{x}'\mathbf{M}\mathbf{x})\,s^2}$$

17. (Greene, 1990, p. 308) Com *n* observações temos $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Considere-se uma observação adicional $\mathbf{x}_a'$, Y_a. Seja $\mathbf{b}_*$ o vetor das estimativas dos parâmetros com as *n* + 1 observações.

Prove que $\mathbf{b}_* = \mathbf{b} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_a}{1+\mathbf{x}_a'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_a}(Y_a - \mathbf{x}_a'\mathbf{b})$

Essa relação mostra que a nova observação não modifica as estimativas dos parâmetros se o novo valor de *Y* pudesse ser exatamente previsto com a regressão baseada nas *n* observações iniciais.

Sugestão: considere as duas formas para $\begin{bmatrix} -\mathbf{X}'\mathbf{X} & \mathbf{x}_a \\ \mathbf{x}_a' & 1 \end{bmatrix}^{-1}$

18. Em artigo de José W. Rossi publicado em *Pesq. Plan. Econ. 12*(2), de agosto de 1982, o autor enfrentou o problema de ajustar uma regressão múltipla sem intercepto. Ele afirma que “como muitos dos pacotes computacionais em uso corrente não dispõem de opção “regressão pela origem” então nas suas utilizações ter-se-á automaticamente tal interseção, conduzindo, portanto, a uma estimação inapropriada. Entretanto, através de um simples método proposto recentemente por Hawkins (1980), pode-se estimar o modelo sem interseção, com os pacotes em uso, sem qualquer dificuldade, bastando para tal utilizar,

além das *n* observações de $X_i (i = 1,..., k)$ e *Y*, também estes mesmos valores com o sinal trocado e proceder à regressão com 2*n* observações, o que garantirá efetivamente que a linha ajustada passe pela origem, tendo os estimadores assim obtidos os valores apropriados (há, entretanto, que se proceder a uma ligeira correção nos valores dos desvios-padrões produzidos na estimação)"

a) Demonstre que o método proposto é correto.

b) Qual é a correção a ser feita nas estimativas dos desvios padrões das estimativas dos parâmetros?

19. Admite-se que *Y* varia linearmente com X_1, X_2, ..., X_k. Dispõe-se de n_1 observações para o período I (matrizes $\mathbf{X}_1$ e $\mathbf{y}_1$) e n_2 observações para o período II (matrizes $\mathbf{X}_2$ e $\mathbf{y}_2$). Se forem ajustadas regressões para cada período, obtém-se

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 \qquad s_1^2 = \frac{1}{n_1 - p}(\mathbf{y}_1'\mathbf{y}_1 - \mathbf{b}_1'\mathbf{X}_1'\mathbf{y}_1)$$

$$\mathbf{b}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2 \qquad s_2^2 = \frac{1}{n_2 - p}(\mathbf{y}_2'\mathbf{y}_2 - \mathbf{b}_2'\mathbf{X}_2'\mathbf{y}_2)$$

com $p = k + 1$.

Considere-se, agora, que é ajustada a seguinte equação (com $X_0 = 1$ para toda observação)

$$\hat{Y}_j = \sum_{i=0}^{k} c_i X_{ij} + \sum_{i=0}^{k} d_i Z_j X_{ij} \qquad (j = 1, ..., n_1, n_1 + 1, ..., n_1 + n_2)$$

onde Z_j é uma variável binária que assume valor zero para as observações do período I e valor 1 para as observações do período II. Seja s^2 o Q.M.Res. dessa regressão.

Define-se

$$\mathbf{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_k \end{bmatrix} \qquad \text{e} \qquad \mathbf{d} = \begin{bmatrix} d_0 \\ \vdots \\ d_k \end{bmatrix}$$

Prove que $\mathbf{c} = \mathbf{b}_1$, $\mathbf{c} + \mathbf{d} = \mathbf{b}_2$ e

$$s^2 = \frac{(n_1 - p)s_1^2 + (n_2 - p)s_2^2}{n_1 + n_2 - 2p}$$

20. Considere-se o modelo de regressão linear múltipla $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, onde **u** é um vetor-coluna de erros com as propriedades usuais, isto é, $E(\mathbf{u}) = 0$ e $E(\mathbf{uu}') = \mathbf{I}\sigma^2$.

Dispõe-se das matrizes **X** e **y**, que permitem obter

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

e

$$s^2 = \frac{1}{n-p}(\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y})$$

Deseja-se estimar o valor da variável dependente para o vetor-linha de valores das variáveis explanatórias $\mathbf{x}_h'$ e obter a estimativa da variância do erro de previsão ($\hat{V}_h$, que é necessária para determinar um intervalo de previsão $\hat{Y}_h \pm t_0\sqrt{\hat{V}_h}$). Para isso formam-se as matrizes:

$$\mathbf{W} = \begin{bmatrix} \mathbf{X} & 0 \\ \mathbf{x}_h' & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{z} = \begin{bmatrix} \mathbf{y} \\ \overline{Y} \end{bmatrix}$$

onde $\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$

Fazendo uma regressão de **z** contra **W** obtém-se as estimativas

$$\begin{bmatrix} \mathbf{b}_* \\ c \end{bmatrix} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{z}$$

Demonstre que $\mathbf{b}_* = \mathbf{b}$, $\hat{Y}_h = \overline{Y} - c$ e $\hat{V}(c) = \hat{V}_h$

Note que, caso se esteja utilizando um programa de computação para regressão múltipla que não tem comandos específicos para o cálculo de $\hat{Y}_h$ e da respectiva estimativa da variância do erro de previsão, o método apresentado permite obter esses resultados com relativa facilidade.

21. Vamos admitir que $\mathbf{y}_1 = \alpha\mathbf{y}_2 + \mathbf{X_1\beta} + \mathbf{u}$ seja uma equação de um sistema de equações simultâneas. A matriz das variáveis exógenas é

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$$

($\mathbf{y}_1$, $\mathbf{y}_2$ e $\mathbf{u}$ são vetores-coluna com *n* elementos, $\mathbf{X}_1$ é uma matriz $n \times p_1$ e $\mathbf{X}_2$ é uma matriz $n \times p_2$).

Seja *a* o estimador de mínimos quadrados ordinários de α e seja $\hat{\alpha}$ o estimador de mínimos quadrados em 2 estágios. Mostre que

$$a = \frac{\mathbf{y}_2'(\mathbf{I} - \mathbf{H}_1)\mathbf{y}_1}{\mathbf{y}_2'(\mathbf{I} - \mathbf{H}_1)\mathbf{y}_2}$$

e $$\hat{\alpha} = \frac{\mathbf{y}_2'(\mathbf{H} - \mathbf{H}_1)\mathbf{y}_1}{\mathbf{y}_2'(\mathbf{H} - \mathbf{H}_1)\mathbf{y}_2} \quad ,$$

onde $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ e $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$

22. Consideremos o modelo de regressão múltipla $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, onde $\mathbf{X}$ é uma matriz $n \times p$.

Seja $\mathbf{x}$ uma coluna qualquer de $\mathbf{X}$ e seja $\mathbf{W}$ a matriz formada pelas demais colunas de $\mathbf{X}$. Podemos reordenar as colunas de $\mathbf{X}$ fazendo com que a variável destacada passe a ser a primeira. Passamos, então, a considerar a regressão múltipla de $\mathbf{y}$ contra $[\mathbf{x} \quad \mathbf{W}]$. Seja β_1 o coeficiente de regressão correspondente a $\mathbf{x}$. Para testar $H_0 : \beta_1 = 0$ devemos calcular

$$t_1 = \frac{b_1}{\sqrt{\hat{V}(b_1)}} \quad ,$$

onde b_1 é a estimativa de β_1.

Seja γ_1 o coeficiente correspondente a $\mathbf{y}$ em uma regressão múltipla de $\mathbf{x}$ contra $[\mathbf{y} \quad \mathbf{W}]$. Para testar $H_0 : \gamma_1 = 0$ devemos calcular

$$t_2 = \frac{c_1}{\sqrt{\hat{V}(c_1)}} \quad ,$$

onde c_1 é a estimativa de γ_1.

a) Demonstre que $t_1 = \dfrac{\mathbf{x}'\mathbf{M}\mathbf{y}}{\sqrt{\dfrac{1}{n-p}\left[(\mathbf{x}'\mathbf{M}\mathbf{x})(\mathbf{y}'\mathbf{M}\mathbf{y}) - (\mathbf{x}'\mathbf{M}\mathbf{y})^2\right]}}$, onde $\mathbf{M} = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$.

b) Obtenha, por analogia, a expressão para t_2 e mostre que $t_1 = t_2$.

23. No contexto da seção 1.7, tem-se que:

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2] \quad , \quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ ,$$

$$\mathbf{H}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' \ , \quad \mathbf{M} = \mathbf{I} - \mathbf{H} \quad \text{e} \quad \mathbf{M}_2 = \mathbf{I} - \mathbf{H}_2$$

a) Demonstre que $\mathbf{M}\mathbf{M}_2 = \mathbf{M}$

b) Demonstre que $\mathbf{H} - \mathbf{H}_2 = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2$

Observação: esses resultados são importantes na dedução e interpretação da expressão

$$\frac{\dfrac{\mathbf{y}'\mathbf{M}_2\mathbf{y} - \mathbf{y}'\mathbf{M}\mathbf{y}}{k_1}}{\dfrac{\mathbf{y}'\mathbf{M}\mathbf{y}}{n-p}} = F$$

onde k_1 é o número de colunas em $\mathbf{X}_1$. O valor de *F* permite testar a hipótese de que os coeficientes das variáveis em $\mathbf{X}_1$ são iguais a zero $(H_0 : \boldsymbol{\beta}_1 = \mathbf{0})$.

24. A tabela abaixo mostra os valores da idade (X_1), da escolaridade (X_2) e da renda mensal (*Y*) para uma amostra de 8 pessoas.

X_1	X_2	Y
20	12	16
20	8	10
26	10	13
32	12	20
32	4	10
38	6	19
44	8	20
44	4	20

Considere o modelo $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

onde u_i são erros aleatórios com as propriedades usuais (independentes, com distribuição normal com média zero e variância constante).

a) Obtenha estimativas não-tendenciosas de variância mínima para α , β_1 e β_2 .

Adotando um nível de significância de 1%, teste as seguintes hipóteses:

b) $H_o : \beta_1 = \beta_2 = 0$.

c) $H_o : \beta_1 = 0$

d) $H_o : \beta_2 = 0$

e) Determine o intervalo de 90% de confiança para a esperança do crescimento da renda mensal quando a escolaridade aumenta de 1 ano $(\Delta X_2 = 1)$.

f) Determine o intervalo de 90% de confiança para a esperança do crescimento quando há acréscimos de 10 anos na idade e 2 anos na escolaridade ($\Delta X_1 = 10$ e $\Delta X_2 = 2$, simultaneamente)

25. A tabela ao lado mostra 5 valores de *X* e *Y* em uma amostra. Admite-se que *Y* varia linearmente em função de *X*, mas há razões para acreditar que ocorreu uma "mudança estrutural" entre a 4ª e a 5ª observação.

X	*Y*
3	12
5	14
7	18
9	24
11	17

a) Ajuste aos dados um modelo de regressão linear múltipla, utilizando uma variável binária para captar a "mudança estrutural".

b) Com base nessa regressão, faça um teste para verificar a significância estatística da "mudança estrutural" (adotando um nível de significância de 5%).

c) Descreva uma outra maneira de fazer o teste, sem usar variável binária.

26. Admite-se que as variáveis X_1, X_2 e Y estão relacionadas conforme o modelo: $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$.

Admite-se, ainda, que os u_j são erros aleatórios independentes entre si, com distribuição normal com média zero e variância σ^2.

Dispomos de uma amostra com 6 observações, apresentada na tabela ao lado. Verifica-se que $\sum X_{1j} = 30$, $\sum X_{2j} = 36$, $\sum Y_j = 474$, $\bar{Y} = 79$, $\sum Y_j^2 = 50.590$ e $\sum y_j^2 = 13.144$.

X_1	X_2	Y
9	6	13
5	10	145
5	6	69
1	6	125
5	2	33
5	6	89

a) Obtenha estimativas de α, β_1 e β_2.

b) Determine a estimativa não-tendenciosa de σ^2.

c) Determine a região de 99% de confiança para β_1 e β_2.

Em todos os testes de hipótese solicitados a seguir, adote o nível de significância de 1%. No caso de teste de hipóteses múltiplas envolvendo β_1 e β_2, a conclusão pode ser

estabelecida utilizando a resposta do item (c) e conhecimentos de geometria analítica, sendo dispensável o cálculo do valor do teste *F*.

d) Teste a hipótese $H_0 : \beta_1 = \beta_2 = 0$.

e) Teste, separadamente, as hipóteses $H_0 : \beta_1 = 0$ e $H_0 : \beta_2 = 0$. Comente, sumariamente, se há contradição com o resultado do item anterior.

f) Teste a hipótese $H_0 : \beta_1 = 0$ e $\beta_2 = 35$.

g) Teste a hipótese $H_0 : \beta_1 = -15$ e $\beta_2 = 30$.

h) Teste a hipótese $H_0 : \beta_2 = 30$.

27. A tabela a seguir mostra a escolaridade (*X*) e o rendimento (*Y*) de 5 pessoas ocupadas na agricultura e 5 pessoas ocupadas nos setores "urbanos" (indústria ou serviços). Define-se uma variável binária *Z* que é igual a zero para pessoas ocupadas na agricultura e é igual a 1 nos demais casos.

Z	*X*	*Y*
0	2	25
0	4	29
0	6	45
0	8	53
0	10	73
1	4	47
1	6	73
1	8	87
1	10	109
1	12	119

Para o modelo $Y_j = \alpha + \beta X_j + \gamma Z_j + \delta Z_j X_j + u_j$ obteve-se

$$\mathbf{X'X} = \begin{bmatrix} 10 & 70 & 5 & 40 \\ 70 & 580 & 40 & 360 \\ 5 & 40 & 5 & 40 \\ 40 & 360 & 40 & 360 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} 660 \\ 5430 \\ 435 \\ 3840 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 1{,}10 & -0{,}15 & -1{,}10 & 0{,}15 \\ -0{,}15 & 0{,}025 & 0{,}15 & -0{,}025 \\ -1{,}10 & 0{,}15 & 2{,}90 & -0{,}35 \\ 0{,}15 & -0{,}025 & -0{,}35 & 0{,}05 \end{bmatrix}$$

$\hat{Y} = 9 + 6X + 6Z + 3ZX$ S.Q.Res. = 128 e $s^2 = \frac{128}{6} = \frac{64}{3}$

a) Determine a equação de regressão de *Y* contra *X* e a respectiva S.Q.Res. para as 5 pessoas do setor agrícola.

b) Idem, para as 5 pessoas do setor "urbano".

c) Teste $H_0 : \gamma = 0$ ao nível de significância de 5%.

d) Teste $H_0 : \delta = 0$ contra $H_A : \delta > 0$, ao nível de significância de 5%.

e) Ao nível de significância de 1%, há diferença estrutural entre setor agrícola e setor "urbano" no que se refere à relação linear entre escolaridade e rendimento?

f) Teste ao nível de significância de 5% a hipótese de que para o nível de escolaridade médio (*X* = 7) não há diferença no rendimento esperado para pessoas ocupadas no setor agrícola e no setor "urbano".

g) Com base no modelo mais geral (com 4 parâmetros), determine o intervalo de previsão, ao nível de 95% de confiança, para o rendimento de uma pessoa com 11 anos de escolaridade ocupada no setor "urbano".

28. Na tabela ao lado está uma amostra de 5 pares de valores das variáveis *X* e *Y*. Define-se uma variável binária Z_1 de maneira que $Z_1 = 0$ se $X \leq 3$ e $Z_1 = 1$ se $X > 3$. Define-se, também, a variável binária $Z_2 = 1 - Z_1$.

X_i	Y_i
2	13
4	19
4	25
2	7
4	22

Considere 3 diferentes modelos para analisar as variações em *Y*:

$$Y_i = \alpha + \beta X_i + u_i \qquad (1)$$

$$Y_i = \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + u_i \qquad (2)$$

$$Y_i = \delta_0 + \delta_1 X_i + \delta_2 X_i^2 + u_i \qquad (3)$$

Nos 3 modelos admite-se que os erros u_i têm as propriedades usuais.

a) Para cada modelo obtenha, se possível, as estimativas dos parâmetros com base na amostra dada. Explique eventuais impossibilidades.

b) Calcule a estimativa da variância do erro para cada modelo estimado. Se houver igualdade de duas estimativas, explique qual é a característica básica dessa amostra que causa essa igualdade.

c) Qual é a hipótese sobre os parâmetros γ_1 e γ_2 que, nesse caso, é equivalente à hipótese de que $\beta = 0$? Calcule a estatística de teste nos dois casos, mostrando sua equivalência.

29. A tabela ao lado mostra os valores de X e Y observados em uma amostra com 4 observações

X_i	Y_i
1	32
2	10
5	9
10	2

Considere o modelo

$$Y_i = \frac{\alpha}{X_i} + \frac{\beta}{X_1} + u_i$$

Admite-se que os u_i são erros independentes, com $u_i \sim N(0, \sigma^2)$.

a) É possível estimar α e β? Explique porque não ou obtenha as estimativas.

b) É possível estimar $\alpha + \beta$? Explique por que não ou obtenha a estimativa e teste, ao nível de significância de 1%, a hipótese de que $\alpha + \beta = 0$, contra a alternativa de que $\alpha + \beta > 0$.

c) Teste, ao nível de significância de 5%, a hipótese de que $\alpha + \beta = 20$.

30. A tabela ao lado mostra os valores de X_1, X_2 e Y em uma amostra com 10 observações. Para distinguir as duas situações (ou dois períodos), cria-se uma variável binária *Z*, com *Z* = 0 na situação *A* e *Z* = 1 na situação *B*. Admitindo que possa haver diferença estrutural entre as duas situações, considera-se o modelo de regressão

Situação	X_1	X_2	Y
A	3	6	32
A	4	9	45
A	5	8	42
A	6	7	45
A	7	10	56
B	3	6	40
B	4	9	43
B	5	8	46
B	6	7	43
B	7	10	48

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma Z_i + \delta_1 Z_i X_{1i} + \delta_1 Z_i X_{1i} + \delta_2 Z_i X_{2i} + u_i$$

A equação estimada é

$$\hat{Y} = 5 + 3X_1 + 3X_2 + 26Z - 2ZX_1 - 2ZX_2$$

com

$$\mathbf{X'X} = \begin{bmatrix} 10 & 50 & 80 & 5 & 25 & 40 \\ 50 & 270 & 412 & 25 & 135 & 206 \\ 80 & 412 & 660 & 40 & 206 & 330 \\ 5 & 25 & 40 & 5 & 25 & 40 \\ 25 & 135 & 206 & 25 & 135 & 206 \\ 40 & 206 & 330 & 40 & 206 & 330 \end{bmatrix}$$

e

$$(\mathbf{X'X})^{-1} = \frac{1}{16}\begin{bmatrix} 105{,}7 & -0{,}5 & -12{,}5 & -105{,}7 & 0{,}5 & 12{,}5 \\ -0{,}5 & 2{,}5 & -1{,}5 & 0{,}5 & -2{,}5 & 1{,}5 \\ -12{,}5 & -1{,}5 & 2{,}5 & 12{,}5 & 1{,}5 & -2{,}5 \\ -105{,}7 & 0{,}5 & 12{,}5 & 211{,}4 & -1 & -2{,}5 \\ 0{,}5 & -2{,}5 & 1{,}5 & -1 & 5 & -3 \\ 12{,}5 & 1{,}5 & -2{,}5 & -2{,}5 & -3 & 5 \end{bmatrix}$$

Sem a variável binária, obtém-se a equação de regressão

$$\hat{Y} = 18 + 2Z_1 + 2X_2 \ ,$$

com S.Q.Res. = 76

Teste, ao nível de significância de 5%, a hipótese de que não há diferença estrutural entre as duas situações (equivalente à hipótese de que $\gamma = \delta_1 = \delta_2 = 0$).

31. A tabela ao lado mostra os valores de X_1, X_2, X_3 e Y em uma amostra com 20 observações. Sabe-se que X_1 afeta X_2 e que X_1, X_2 e X_3 tem efeito sobre Y. Usando um computador, faça a regressão de *Y* contra X_1, X_2, X_3, verificando que os respectivos coeficientes de regressão são iguais a 1, 2 e 4 e que não há efeito estatisticamente significativo de X_1 sobre *Y*. Em seguida, faça a regressão de *Y* contra X_1 e X_3 verificando que os respectivos coeficientes de regressão são iguais a 7 e 4 e que agora o efeito de X_1 sobre *Y* é fortemente significativo. Na primeira regressão o coeficiente de X_1 capta apenas o efeito *direto* de X_1 sobre *Y*, ao passo que na segunda regressão esse coeficiente capta, também, o efeito de X_1 sobre *Y* via X_2 (já que existem efeitos estatisticamente fortes de X_1 sobre X_2 e de X_2 sobre *Y*).

X_1	X_2	X_3	Y
13	37	9	94
13	35	5	72
8	21	7	48
3	7	5	6
13	37	5	78
13	35	9	90
3	7	9	24
3	5	5	2
8	21	7	48
13	37	5	76
13	35	9	88
3	5	9	18
3	7	5	8
3	5	5	4
8	21	7	48
3	7	9	22
13	35	5	74
13	37	9	92
3	5	9	20
8	21	7	48

Respostas

1. a) $\hat{Y} = 21 - 15X_1 + 3X_2$

   b) $F = 34{,}36$, significativo $(F_0 = 30{,}82)$

   c) $t = -7{,}493$, significativo (a região de rejeição é $t < -4{,}541$)

   d) $R^2 = 0{,}958$ e $\overline{R}^2 = 0{,}930$

   e) $r_{12} = 0{,}904$, $r_{Y1} = 0$ e $r_{Y2} = 0{,}419$

A afirmativa não é correta. A variável X_1 tem efeito negativo e estatisticamente significativo sobre *Y* (ver item *c*).

O valor de $r_{Y1} = 0$ resulta da combinação de um efeito direto negativo com um efeito indireto positivo, através de X_2, devido à forte correlação entre X_1 e X_2.

2. a) Com $X_1 = \ln W_1$ e $X_2 = \ln W_2$, obtemos $\hat{Y} = 1{,}2 + 0{,}7X_1 + 0{,}5X_2$

   b) $s^2 = 0{,}18$

   c) $\overline{R}^2 = 0{,}9151$

   d) $t = 3{,}159$, não-significativo $(t_0 = 3{,}707)$

   e) $t = 2{,}257$, não-significativo $(t_0 = 3{,}707)$

   f) $F = 44{,}11$, significativo $(F_0 = 10{,}9)$

   g) A correlação entre $X_1 = \ln W_1$ e $X_2 = \ln W_2$ é forte: $r_{12} = 0{,}833$. As consequencias da multicolinearidade não são mais graves porque a função se ajusta muito bem aos dados ($R^2 = 0{,}936$). Os resultados obtidos nos itens (d), (e) e (f) estão associados com a forte covariância negativa entre b_1 e b_2.

   h) $t = 1{,}563$, não-significativo $(t_0 = 1{,}943)$.

3. Temos $t_h^2 = F_h < 1$ ou $\dfrac{\text{S.Q.Contribuição de } X_h}{s_c^2} < 1$, (1)

onde s_c^2 é o Q.M.Res. da regressão completa, com $n - p + 1$ graus de liberdade.

Seja s_i^2 o Q.M.Res. da regressão sem X_h, com $n - p + 1$ graus de liberdade.

Então

$$\text{S.Q. Contribuição de } X_h = (n-p+1)s_i^2 - (n-p)s_c^2$$

Substituindo em (1) obtemos

$$\frac{(n-p+1)s_i^2 - (n-p)s_c^2}{s_c^2} < 1$$

$$(n-p+1)s_i^2 - (n-p)s_c^2 < s_c^2$$

$$(n-p+1)s_i^2 < (n-p+1)s_c^2$$

$$s_i^2 < s_c^2 \quad \text{, c.q.d.}$$

4. a) $\hat{Y} = -3 + 0{,}5X_1 + 4X_2$

b) $s^2 = 4$

c) $R^2 = 0{,}9435$, $\overline{R}^2 = 0{,}9274$

d) t = 1,342, não-significativo ($t_0 = 1{,}895$)

e) t = 3, significativo ($t_0 = 2{,}365$)

f) F = 4,5, não-significativo ($F_0 = 4,74$)

5. a) $b_1 = b_2 = 3$

b) F = 6,25, significativo ($F_0 = 5{,}79$)

c) t = 0, obviamente não-significativo.

6. Utilizando as 9 observações, obtemos $\hat{Y} = 4 + 6X$, com soma de quadrados residual S = 520.

Com as 3 primeiras observações obtemos $b_1 = -1$ e soma de quadrados residual $S_1 = 6$. Com as 3 observações seguintes obtemos $b_2 = 6$ e $S_2 = 24$. Com as 3 últimas observações obtemos $b_3 = 13$ e $S_3 = 6$.

O teste F para a hipótese de que a "estrutura" linear é a mesma nos 3 períodos é

$$F = \frac{\dfrac{S - (S_1 + S_2 + S_3)}{7 - 3}}{\dfrac{S_1 + S_2 + S_3}{3}} = 10{,}08$$

O resultado é significativo, pois o valor crítico, ao nível de significância de 5%, é $F_0 = 9{,}12$.

7. F = 4, não-significativo $(F_0 = 4{,}12)$.

8. a) Aplicando logaritmos:

$$\ln W_i = \ln \theta + \beta_1 \ln Z_{1i} + \beta_2 \ln Z_{2i} + \ln \varepsilon_i$$

ou $$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

b) $b_1 = 0{,}5$, $b_2 = 0{,}3$ e $\hat{\theta} = \exp(2) = 7{,}389$

c) $s^2 = 0{,}064$

d) $R^2 = 0{,}9568$

e) t = 3,785, significativo (a região de rejeição é $t \geq 3{,}365$).

f) t = 2,271, não-significativo.

g) $b_2 = 0{,}3$

h) $t = \dfrac{b_1 + b_2 - 1}{\sqrt{\hat{V}(b_1 + b_2)}} = -2{,}622$, significativo ($t_0 = 2{,}571$).

9. a) $\hat{Y}_1 = 1 + 2X$, com S.Q.Res. = $S_1 = 4$

$\hat{Y}_2 = 4 + 1{,}5X$, com S.Q.Res. = $S_2 = 6$

b) Utilizando as 10 observações, obtemos $\hat{Y} = 2{,}5 + 1{,}75X$, com S.Q.Res. = $S_R = 15$

$$F = \frac{\dfrac{15 - (4+6)}{2}}{\dfrac{4+6}{6}} = 1{,}5, \text{ não-significativo } (F_0 = 3{,}46)$$

c) S.Q.Res. = 10

$$R^2 = \frac{260 - 10}{260} = 0{,}9615$$

10. $c_1 = b_1$ e $c_2 = b_1 + b_2 - 1$

Indicando por z_i, w_{1i} e w_{2i} as variáveis centradas correspondentes, respectivamente, a $\ln Y_{1i}$, $\ln X_{1i}$ e $\ln X_{2i}$, a demonstração das relações acima é feita a partir de

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum w_{1i}^2 & \sum w_{1i} w_{2i} \\ \sum w_{1i} w_{2i} & \sum w_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum w_{1i} z_i \\ \sum w_{2i} z_i \end{bmatrix}$$

e

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \sum (w_{1i} - w_{2i})^2 & \sum (w_{1i} - w_{2i}) w_{2i} \\ \sum (w_{1i} - w_{2i}) w_{2i} & \sum w_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum (w_{1i} - w_{2i})(z_i - w_{2i}) \\ \sum w_{2i} (z_i - w_{2i}) \end{bmatrix}$$

11. a) $\hat{Y} = -32 + 3X_1 + 5X_2$

b) F =10, não-significativo ($F_0 = 19{,}00$)

c) t = 2,683, não-significativo (a região de rejeição é $t \geq 2{,}920$)

d) F = 8, não-significativo ($F_0 = 19{,}00$)

e) $r_{Y1.2} = 0{,}8847$ e $r_{Y2.1} = 0{,}9535$

f) $d = b_1 + b_2 = 8$, $-0{,}6 < \delta < 16{,}6$.

12. Como $\mathbf{y} = \boldsymbol{\iota}$, $\mathbf{X}'\mathbf{y} = \mathbf{X}'\boldsymbol{\iota} = \begin{bmatrix} n \\ \sum X_1 \\ \sum X_2 \end{bmatrix}$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{bmatrix}^{-1} \begin{bmatrix} n \\ \sum X_1 \\ \sum X_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

igual à 1ª coluna de $\mathbf{I}_3$, pois se está multiplicando a inversa de uma matriz pela sua própria 1ª coluna.

$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \boldsymbol{\iota}$

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \boldsymbol{\iota} - \boldsymbol{\iota} = \mathbf{0}$ (vetor de zeros)

13. Com todas as variáveis centradas, obtemos:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 200 \\ 100 \end{bmatrix}$$

a) $\hat{Y} = 1 + 2X_1 + X_2$

b) $s^2 = 36/5 = 7{,}2$

c) t = 2,635, significativo ($t_0 = 2{,}571$)

d) $\hat{Y} = 9{,}5Z_1 + 29{,}5Z_2 + 25{,}5Z_3 + 15{,}5Z_4$, com $s^2 = 34/4 = 8{,}5$

e) $F = 2/8{,}5 = 0{,}235$, não-significativo ($F_0 = t_0^2 = 4{,}54$)

14. $\hat{\beta} = \dfrac{\overline{Y}}{\overline{X}} = \dfrac{\sum Y}{\sum X} = \dfrac{\sum(\beta X + u)}{\sum X} = \beta + \dfrac{\sum u}{\sum X}$

Então, $E(\hat{\beta}) = \beta$ e $V(\hat{\beta}) = E\left(\dfrac{\sum u}{\sum X}\right)^2 = \dfrac{n\sigma^2}{(\sum X)^2}$ ou $V(\hat{\beta}) = \dfrac{\sigma^2}{\dfrac{(\sum X)^2}{n}}$

Temos $V(b) = \dfrac{\sigma^2}{\sum X^2}$

Mas $\sum x^2 = \sum X^2 - \dfrac{(\sum X)^2}{n} \geq 0$, donde $\sum X^2 \geq \dfrac{(\sum X)^2}{n}$

Então $V(\hat{\beta}) \geq V(b)$, com igualdade apenas quando todos os X forem iguais.

15. $\mathbf{W'W} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2' \\ \mathbf{X}_2 & \mathbf{I} \end{bmatrix} \quad \mathbf{W'y} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2 \\ \mathbf{y}_2 \end{bmatrix}$

Utilizando a expressão (1.50), obtemos

$$(\mathbf{W'W})^{-1} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \\ -\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{I} + \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \end{bmatrix}$$

Segue-se que

$\mathbf{d}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 = \mathbf{b}_1$

$\mathbf{d}_2 = -\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 + \mathbf{y}_2$

S.Q.Res. = $\mathbf{y}_1'\mathbf{y}_1 + \mathbf{y}_2'\mathbf{y}_2 - [\mathbf{d}_1' \quad \mathbf{d}_2']\mathbf{W'y} =$

$= \mathbf{y}_1'\mathbf{y}_1 - \mathbf{y}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} = \mathbf{y_1'M}_1\mathbf{y}_1$, c.q.d.

18. b) As estimativas dos desvios padrões das estimativas dos parâmetros devem ser multiplicadas por

$$\sqrt{\frac{2n-k-1}{n-k}}$$

24. a) $\hat{Y} = -8 + 0{,}5X_1 + X_2$

b) F = 11,88, não-significativo ($F_0 = 13{,}3$).

c) t = 4,830, significativo ($t_0 = 4{,}032$).

d) t = 3,220, não-significativo ($t_0 = 4{,}032$).

e) $1 \pm 0{,}626 \rightarrow 0{,}374 < \beta_2 < 1{,}626$.

f) $7 \pm 2{,}969 \rightarrow 4{,}031 < 10\beta_1 + 2\beta_2 < 9{,}969$.

25. a) $Y = \alpha + \beta X + \gamma Z + u$, com $Z = 0$ para as 4 primeiras observações e $Z = 1$ para a 5ª observação.

$\hat{Y} = 5 + 2X - 10Z$, com S.Q.Res. = 4 e $s^2 = 2$

b) $H_0 : \gamma = 0$, t = – 4, 472, significativo ($t_0 = 4{,}303$). Notar que $t^2 = 20$.

c) Seja S_U a S.Q.Res. de uma regressão linear simples com as 4 primeiras observações e seja S_R a S.Q.Res. de uma regressão linear simples com as 5 observações. Então,

$$F = \frac{S_R - S_U}{\frac{S_U}{2}},$$ com 1 e 2 graus de liberdade.

Para esses dados obtemos $F = \frac{44-4}{2} = 20$, significativo ($F_0 = 18{,}51$).

26. Com as 3 variáveis centradas, obtemos:

$$\mathbf{X'X} = \begin{bmatrix} 32 & 0 \\ 0 & 32 \end{bmatrix}, \quad \mathbf{X'y} = \begin{bmatrix} -448 \\ 448 \end{bmatrix} \quad \text{e} \quad \mathbf{b} = \begin{bmatrix} -14 \\ 14 \end{bmatrix}$$

a) $a = 65$, $b_1 = -14$ e $b_2 = 14$

b) $s^2 = 200$

c) $(\beta_1 + 14)^2 + (\beta_2 - 14)^2 < 12{,}5 \cdot 30{,}8 = 385$, com $\sqrt{385} = 19{,}62$

Essa região de 99% de confiança para β_1 e β_2 é a área delimitada por uma circunferência com centro no ponto (–14; 14) e raio 19,62.

d) $F = 31{,}36$, significativo ($F_0 = 30{,}8$).

Alternativa: a distância do ponto (0; 0) ao ponto (–14; 14) é 19,80. Como esse valor é maior do que o raio da circunferência, o resultado do teste é significativo.

e) Para as duas hipóteses o valor absoluto de *t* é 5,60, não-significativo ($t_0 = 5{,}841$). Só há contradição se as conclusões forem consideradas como afirmativas "matemáticas", esquecendo seu caráter estatístico. No item (d) considera-se a distribuição *conjunta* de b_1 e b_2, ao passo que os resultados dos testes para as hipóteses separadas são baseadas nas respectivas distribuições marginais.

f) A distância entre os pontos (–14; 14) e (0; 35) é 25,24, maior do que o raio. Rejeita-se $H_0 : \beta_1 = 0$ e $\beta_2 = 35$.

g) A distância entre os pontos (–14; 14) e (–15; 30) é 16,03, menor do que o raio. Não se rejeita $H_0 : \beta_1 = -15$ e $\beta_2 = 30$.

h) $t = \frac{14-30}{2{,}5} = -6{,}40$, significativo ($t_0 = 5{,}841$).

Novamente, uma aparente contradição com o resultado do item anterior.

27. a) $\hat{Y} = 9 + 6X$, com S.Q.Res. = 64.

b) $\hat{Y} = 15 + 9X$, com S.Q.Res. = 64.

c) t = 0,763, não-significativo ($t_0 = 2{,}447$).

d) t = 2,905, significativo (a região de rejeição é $t \geq 1{,}943$).

e) F = 42,19, significativo ($F_0 = 10{,}92$).

f) $H_0 : \gamma + 7\delta = 0$, t = 8,714, significativo ($t_0 = 2{,}447$).

g) Para $X_h = 11$ obtemos $\hat{Y}_h = 114$ e 100,5 < Y_h < 127,5.

28. a) Modelo (1): $\hat{Y} = -2 + 6X$;

Modelo (2): $\hat{Y} = 22Z_1 + 10Z_2$

O modelo (3) não pode ser estimado porque há apenas 2 valores distintos de X. A matriz $\mathbf{X}$ tem apenas duas linhas distintas e, consequentemente, sua característica é igual a 2. A matriz $\mathbf{X'X}$ é 3×3, mas sua característica também é igual a 2. Então o determinante de $\mathbf{X'X}$ é igual a zero.

b) Tanto para o modelo (1) como para o modelo (2) obtemos $s^2 = 12$, com 3 graus de liberdade. Como há apenas 2 valores distintos, a variável X já é binária, o que torna os modelos (1) e (2) equivalentes.

c) As hipóteses $H_0 : \beta = 0$ e $H_0 : \gamma_1 = \gamma_2$ são equivalentes. Para a primeira hipótese calculamos

$$t = \frac{b}{s(b)} = \frac{6}{\sqrt{12/4{,}8}} = 3{,}795$$

Para a segunda hipótese calculamos

$$t = \frac{22 - 10}{\sqrt{\left(\frac{1}{2} + \frac{1}{3}\right)12}} = \frac{12}{\sqrt{10}} = 3{,}795$$

29. a) O modelo pode ser escrito como $Y_i = (\alpha + \beta)W_i) + u_i$ com $W_i = \frac{1}{X_i}$

Trata-se de uma regressão linear simples de Y_i contra W_i sem intercepto. Pode-se estimar $\alpha + \beta$, mas é totalmente arbitrário separar o total em 2 partes.

b) $t = 9{,}487$, significativo $(t_0 = 4{,}541)$

c) $t = 3{,}162$, não-significativo $(t_0 = 3{,}182)$

30. $F = \frac{64}{9} = 7{,}11$, significativo $(F_0 = 6{,}59)$

Há diferença estrutural entre as duas situações.

2. A INFLUÊNCIA DE UMA OBSERVAÇÃO EM ANÁLISE DE REGRESSÃO

2.1. A matriz **H**

Considere-se o modelo de regressão

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

onde $E(\mathbf{u}) = 0$ e $E(\mathbf{uu}') = \mathbf{I}\sigma^2$

Sabemos que o estimador de mínimos quadrados para $\boldsymbol{\beta}$ é

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Então o vetor-coluna dos valores estimados de Y é

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \ ,$$

onde $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é a matriz que faz a projeção ortogonal do vetor **y** no sub-espaço que pode ser gerado pelas colunas de **X**.

O vetor-coluna dos desvios (ou resíduos) é

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y} \quad ,$$

onde $\mathbf{M} = \mathbf{I} - \mathbf{H}$.

Tanto **H** como **M** são matrizes simétricas e idempotentes. Além disso, pode-se verificar que

$$\mathbf{HX} = \mathbf{X}, \quad \mathbf{X}'\mathbf{H} = \mathbf{X}' \quad , \quad \mathbf{MX} = \mathbf{0} \quad \text{e} \quad \mathbf{X}'\mathbf{M} = \mathbf{0}$$

De $\mathbf{e} = \mathbf{M}\mathbf{y}$ segue-se, então, que

$$\mathbf{e} = \mathbf{M}\mathbf{u}$$

e
$$V(\mathbf{e}) = E(\mathbf{ee}') = E(\mathbf{Muu}'\mathbf{M}) = \mathbf{M}\sigma^2 \qquad (2.1)$$

Indicando o *i*-ésimo elemento da diagonal de **H** por $h_i = h_{ii}$, tem-se

$$V(e_i) = (1 - h_i)\sigma^2 \qquad (2.2)$$

A seguir são demonstradas algumas propriedades dos h_i.

Como $\mathbf{H}'\mathbf{H} = \mathbf{H}$, tem-se $\sum_{j=1}^{n} h_{ij}^2 = h_i$

ou
$$h_i = h_i^2 + \sum_{j \neq i} h_{ij}^2$$

Conclui-se que

$$0 \leq h_i \leq 1 \qquad (2.3)$$

Tem-se $\text{tr}(\mathbf{H}) = \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = p$

Então $\sum_{i=1}^{n} h_i = p$ e o valor médio dos h_i é igual a p/n.

Uma vez que $\mathbf{HX} = \mathbf{X}$, se o modelo tem um termo constante e a primeira coluna de **X** é **ι,** tem-se

$$\mathbf{H\iota} = \boldsymbol{\iota}$$

ou

$$\sum_{j=1}^{n} h_{ij} = 1$$

Se **W** é a matriz com as variáveis explanatórias, exclusive a coluna correspondente ao termo constante, temos

$$\mathbf{X} = [\boldsymbol{\iota} \quad \mathbf{W}]$$

e

$$\mathbf{X'X} = \begin{bmatrix} \boldsymbol{\iota}'\boldsymbol{\iota} & \boldsymbol{\iota}'\mathbf{W} \\ \mathbf{W}'\boldsymbol{\iota} & \mathbf{W}'\mathbf{W} \end{bmatrix} = \begin{bmatrix} n & \boldsymbol{\iota}'\mathbf{W} \\ \mathbf{W}'\boldsymbol{\iota} & \mathbf{W}'\mathbf{W} \end{bmatrix}$$

De acordo com (1.49), o elemento inferior-direito de $(\mathbf{X'X})^{-1}$ é

$$\left(\mathbf{W'W} - \mathbf{W'\iota}\frac{1}{n}\boldsymbol{\iota}'\mathbf{W}\right)^{-1} = (\mathbf{W'AW})^{-1}$$

onde $\mathbf{A} = \mathbf{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'$ é uma matriz idempotente que torna centradas as colunas de qualquer matriz que ela pré-multiplica.

Ainda de acordo com (1.49), obtemos

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{1}{n^2}\boldsymbol{\iota}'\mathbf{W}(\mathbf{W'AW})^{-1}\mathbf{W'\iota} & -\frac{1}{n}\boldsymbol{\iota}'\mathbf{W}(\mathbf{W'AW})^{-1} \\ -\frac{1}{n}(\mathbf{W'AW})^{-1}\mathbf{W'\iota} & (\mathbf{W'AW})^{-1} \end{bmatrix} \tag{2.4}$$

Substituindo esse resultado em

$$\mathbf{H} = [\boldsymbol{\iota} \quad \mathbf{W}](\mathbf{X'X})^{-1}\begin{bmatrix} \boldsymbol{\iota}' \\ \mathbf{W}' \end{bmatrix}$$

e fazendo algumas manipulações algébricas obtemos

$$\mathbf{H} = \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}' + \mathbf{AW}(\mathbf{W'AW})^{-1}\mathbf{W'A} \tag{2.5}$$

ou

$$\mathbf{W}_*(\mathbf{W}_*'\mathbf{W}_*)^{-1}\mathbf{W}_*' = \mathbf{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}' \quad ,$$

com $\mathbf{W}_* = \mathbf{AW}$

Por analogia com **H**, os elementos da diagonal de $\mathbf{W}_*(\mathbf{W}_*'\mathbf{W}_*)^{-1}\mathbf{W}_*'$ também pertencem ao intervalo [0, 1]. Conclui-se que, para uma regressão com termo constante, temos

$$\frac{1}{n} \le h_i \le 1 \tag{2.6}$$

Como regra prática, recomenda-se examinar com especial atenção as observações para as quais o valor de h_i supera 2 ou 3 vezes a média, isto é, supera

$$\frac{2p}{n} \quad \text{ou} \quad \frac{3p}{n}$$

Veremos que tais observações tendem a ter forte influência nos resultados da regressão.

2.2. Inclusão de uma variável binária para captar a influência de uma observação

Seja $\mathbf{1}_i$ a *i*-ésima coluna de uma matriz $\mathbf{I}_n$. Introduzindo a variável binária $\mathbf{1}_i$ entre as variáveis explanatórias, ela vai "roubar" a influência da *i*-ésima observação. A matriz das variáveis explanatórias passa a ser $\mathbf{Z} = [\mathbf{X} \quad \mathbf{1}_i]$

Vamos reservar os símbolos $\mathbf{b}$, $\hat{\mathbf{y}}$, $\mathbf{e}$, s^2 e h_i para os resultados referentes à regressão original, isto é, a regressão de **y** contra **X**.

Para a regressão de **y** contra **Z** obtemos

$$\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{x}_i \\ \mathbf{x}_i' & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{Z}'\mathbf{y} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ Y_i \end{bmatrix}$$

onde $\mathbf{x}_i'$ é a *i*-ésima linha de **X** e Y_i é o correspondente valor observado da variável dependente. Notando que $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = h_i$, de acordo com (1.49), obtém-se

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} + \dfrac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_i} & -\dfrac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1-h_i} \\ -\dfrac{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_i} & \dfrac{1}{1-h_i} \end{bmatrix}$$

As estimativas dos parâmetros da regressão de **y** contra $\mathbf{Z} = [\mathbf{X} \quad \mathbf{1}_i]$ são

$$\begin{bmatrix} \mathbf{b}_* \\ d \end{bmatrix} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \begin{bmatrix} \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \dfrac{e_i}{1-h_i} \\ \dfrac{e_i}{1-h_i} \end{bmatrix}, \tag{2.7}$$

onde $e_i = Y_i - \mathbf{x}_i'\mathbf{b}$

Notar que $\mathbf{b}$ e $\mathbf{b}_*$ são vetores com o mesmo número de elementos, os quais são os coeficientes de regressão das variáveis explanatórias em $\mathbf{X}$.
Os valores de Y estimados por meio da regressão incluindo a variável binária $\mathbf{1}_i$ são

$$\hat{\hat{\mathbf{y}}} = \begin{bmatrix}\mathbf{X} & \mathbf{1}_i\end{bmatrix}\begin{bmatrix}\mathbf{b}_* \\ d\end{bmatrix} = \mathbf{Xb} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_i} + \mathbf{1}_i \frac{e_i}{1-h_i}$$

Então, para a i-ésima observação, obtemos

$$\hat{\hat{Y}}_i = \mathbf{x}_i'\mathbf{b} - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_i} + \frac{e_i}{1-h_i} = \mathbf{x}_i'\mathbf{b} + e_i = Y_i \quad ,$$

mostrando que a introdução de $\mathbf{1}_i$ "elimina" o desvio referente à i-ésima observação.
A soma de quadrados residual é dada por

$$\mathbf{y}'\mathbf{y} - \begin{bmatrix}\mathbf{b}_*' & d\end{bmatrix}\mathbf{Z}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \begin{bmatrix}\mathbf{b}' - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\frac{e_i}{1-h_i} & \frac{e_i}{1-h_i}\end{bmatrix}\begin{bmatrix}\mathbf{X}'\mathbf{y} \\ Y_i\end{bmatrix} =$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{e_i^2}{1-h_i}$$

Verifica-se que a redução da soma de quadrados residual devida à introdução de $\mathbf{1}_i$ é

$$\frac{e_i^2}{1-h_i} \tag{2.8}$$

Se s^2 é o quadrado médio residual da regressão de $\mathbf{y}$ contra $\mathbf{X}$, então o quadrado médio residual da regressão de $\mathbf{y}$ contra $\begin{bmatrix}\mathbf{X} & \mathbf{1}_i\end{bmatrix}$ é

$$s_*^2 = \frac{1}{n-p-1}\left[(n-p)s^2 - \frac{e_i^2}{1-h_i}\right] \tag{2.9}$$

Admitindo que os erros u_j têm distribuição normal, o valor de F para testar a contribuição de $\mathbf{1}_i$ é

$$F = \frac{e_i^2}{s_*^2(1-h_i)}$$

Como o numerador está associado a 1 grau de liberdade, a raiz quadrada do F é uma variável com distribuição de t de Student com $n - p - 1$ graus de liberdade:

$$t = \frac{e_i}{\sqrt{s_*^2(1-h_i)}} \tag{2.10}$$

Se esse valor de t for estatisticamente significativo, conclui-se que a contribuição da variável binária $\mathbf{1}_i$ (que "rouba" a influência da i-ésima observação) para a soma de quadrados de regressão é estatisticamente significativa, o que eqüivale a dizer que a i-ésima observação é *discrepante*. Isso deverá ficar mais claro com as deduções apresentadas nas três próximas seções.

2.3. Eliminando uma linha da matriz X

Preliminarmente, será obtida uma relação matricial básica para as deduções que se seguem. Sendo $\mathbf{x}_i'$ a *i*-ésima linha da matriz **X**, considere-se a matriz

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{x}_i \\ \mathbf{x}_i' & 1 \end{bmatrix}^{-1}$$

De acordo com as expressões para o primeiro elemento da primeira linha em (1.49) e (1.50), temos

$$(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i')^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_i} \tag{2.11}$$

Obtida essa relação, passamos à análise de como a eliminação da *i*-ésima observação afeta os resultados da regressão.
Notar que

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$$

e que na regressão sem a *i*-ésima observação a matriz $\mathbf{X}'\mathbf{X}$ é substituída por $\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'$. Analogamente, a matriz $\mathbf{X}'\mathbf{y}$ é substituída por $\mathbf{X}'\mathbf{y} - \mathbf{x}_i Y_i$. Seja $\mathbf{b}_{(i)}$ o vetor das estimativas dos parâmetros sem a *i*-ésima observação. Tem-se

$$\mathbf{b}_{(i)} = (\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i')^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{x}_i Y_i)$$

Lembrando (2.11), segue-se que

$$\mathbf{b}_{(i)} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i Y_i + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_i}(\mathbf{X}'\mathbf{y} - \mathbf{x}_i Y_i)$$

ou

$$\mathbf{b}_{(i)} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_i} \tag{2.12}$$

Indicando o Q.M.Res. por $s_{(i)}^2$, a S.Q.Res. fica:

$$(n-p-1)s_{(i)}^2 = \mathbf{y}'\mathbf{y} - Y_i^2 - \mathbf{b}_{(i)}'(\mathbf{X}'\mathbf{y} - \mathbf{x}_i Y_i) =$$

$$= \mathbf{y}'\mathbf{y} - Y_i^2 - \left[\mathbf{b}' - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\frac{e_i}{1-h_i}\right](\mathbf{X}'\mathbf{y} - \mathbf{x}_i Y_i) =$$

$$= \mathbf{y}'\mathbf{y} - Y_i^2 - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{x}_i Y_i + \mathbf{x}_i'\mathbf{b}\frac{e_i}{1-h_i} - \frac{h_i e_i Y_i}{1-h_i} =$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{e_i^2}{1-h_i}$$

Segue-se que

$$s^2_{(i)} = \frac{1}{n-p-1}\left[(n-p)s^2 - \frac{e_i^2}{1-h_i}\right] \tag{2.13}$$

Comparando (2.7) e (2.12) verifica-se que $\mathbf{b}_* = \mathbf{b}_{(i)}$, isto é, a alteração nas estimativas dos parâmetros (**b**) devida à introdução de $\mathbf{1}_i$ é igual à alteração devida à eliminação de $\mathbf{x}_i'$. Comparando (2.9) e (2.13) verifica-se que o efeito da eliminação de $\mathbf{x}_i'$ sobre a S.Q.Res. (e sobre o Q.M.Res.) é idêntico ao efeito da introdução de $\mathbf{1}_i$.

Utilizando $\mathbf{b}_{(i)}$, a estimativa da variável dependente para $\mathbf{x}_i'$ é

$$\mathbf{x}_i'\mathbf{b}_{(i)} = \mathbf{x}_i'\left[\mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\frac{e_i}{1-h_i}\right] = \mathbf{x}_i'\mathbf{b} - \frac{h_i e_i}{1-h_i}$$

Portanto, a alteração no valor dessa estimativa devida à eliminação de $\mathbf{x}_i'$ é

$$\hat{Y}_i - \mathbf{x}_i'\mathbf{b}_{(i)} = \frac{h_i e_i}{1-h_i} \tag{2.14}$$

2.4. Resíduos

Foi visto que $V(e_i) = (1-h_i)\sigma^2$

Se os erros tem distribuição normal, então $\dfrac{e_i}{\sqrt{(1-h_i)\sigma^2}}$

tem distribuição normal reduzida.

Entretanto,

$$\frac{e_i}{\sqrt{(1-h_i)s^2}} \tag{2.15}$$

não tem distribuição de *t* porque numerador e denominador não são independentes. Os valores de (2.15) são denominados *resíduos estudentizados internamente*. Às vezes, por simplicidade, são utilizados os valores de $\dfrac{e_i}{s}$ ou resíduos padronizados.

Os valores dados por (2.10), isto é,

$$t = \frac{e_i}{\sqrt{s^2_{(i)}(1-h_i)}} = e_i^* \tag{2.16}$$

são denominados *resíduos estudentizados externamente*. É claro que estes são os mais apropriados para detectar uma observação discrepante.

2.5. Outra maneira de interpretar o resíduo estudentizado externamente

Fazendo a regressão *sem* a *i*-ésima observação, o valor de *Y* estimado para essa *i*-ésima observação será dado por

$\mathbf{x}_i'\mathbf{b}_{(i)}$

O erro de previsão é, então,

$$f_i = Y_i - \mathbf{x}_i'\mathbf{b}_{(i)} \tag{2.17}$$

e a estimativa da respectiva variância é

$$\hat{V}(f) = \left[1 + \mathbf{x}_i'(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i')^{-1}\mathbf{x}_i\right]s^2_{(i)} \tag{2.18}$$

Temos, então, o seguinte teste:

$$t(f_i) = \frac{f_i}{\sqrt{\hat{V}(f_i)}} \tag{2.19}$$

Vamos demonstrar, em seguida, que esse teste é igual ao resíduo estudentizado externamente, dado por (2.16).
Substituindo (2.12) em (2.17) obtemos

$$f_i = Y_i - \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\frac{e_i}{1-h_i}$$

$$f_i = e_i + h_i\frac{e_i}{1-h_i}$$

$$f_i = \frac{e_i - h_ie_i + h_ie_i}{1-h_i}$$

ou $$f_i = \frac{e_i}{1-h_i} \tag{2.20}$$

Substituindo (2.11) em (2.18) obtemos

$$\hat{V}(f_i) = \left\{1 + \mathbf{x}_i'\left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_i}\right]\mathbf{x}_i\right\}s^2_{(i)}$$

$$\hat{V}(f_i) = \left(1 + h_i + \frac{h_i^2}{1-h_i}\right)s^2_{(i)}$$

$$\hat{V}(f_i) = \frac{1 - h_i^2 + h_i^2}{1-h_i}s^2_{(i)}$$

ou $$\hat{V}(f_i) = \frac{s^2_{(i)}}{1-h_i} \tag{2.21}$$

Finalmente, substituindo (2.20) e (2.21) em (2.19) segue-se que

$$t(f_i) = \frac{\frac{e_i}{1-h_i}}{\sqrt{\frac{s^2_{(i)}}{1-h_i}}}$$

ou

$$t(f_i) = \frac{e_i}{\sqrt{(1-h_i)s^2_{(i)}}}$$

Comparando esse resultado com (2.16) conclui-se que

$$t(f_i) = e_i^* \quad , \quad \text{c.q.d.}$$

2.6. DFBETAS

De acordo com (2.12), tem-se

$$\mathbf{b} - \mathbf{b}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_i} \tag{2.22}$$

Sejam c_{ki} os elementos da matriz $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Note-se que o índice *k* é utilizado, nesta seção, para indicar qualquer uma das *p* linhas da matriz **C**. Então

$$\mathbf{b} - \mathbf{b}_{(i)} = \begin{bmatrix} c_{1i} \\ c_{2i} \\ \vdots \\ c_{pi} \end{bmatrix} \frac{e_i}{1-h_i}$$

A alteração em b_k devida à inclusão da *i*-ésima observação é

$$\Delta b_k = \frac{c_{ki}e_i}{1-h_i} \tag{2.23}$$

Tendo em vista que $V(b_k) = \omega_{kk}\sigma^2$, onde ω_{kk} é o *k*-ésimo elemento da diagonal principal de $(\mathbf{X}'\mathbf{X})^{-1}$, uma medida padronizada da alteração em b_k é dada por

$$\text{DFBETAS}_{ki} = \frac{c_{ki}e_i}{(1-h_i)\sqrt{\omega_{kk}s^2_{(i)}}} \tag{2.24}$$

Lembrando (2.16), segue-se que

$$\text{DFBETAS}_{ki} = \frac{c_{ki}e_i^*}{\sqrt{\omega_{kk}(1-h_i)}} \tag{2.25}$$

Levando em consideração que o efeito de uma única observação diminui quando aumenta *n*, Belsley, Kuh e Welsh (1980, p. 28) recomendam o exame das observações com

$$\left|\text{DFBETAS}_{ki}\right| > \frac{2}{\sqrt{n}} \tag{2.26}$$

Antes de encerrar esta seção, vamos mostrar que é indiferente obter os coeficientes c_{ki} referentes aos coeficientes de regressão b_k utilizando a matriz X com as variáveis originais ou com as variáveis centradas. Conforme a notação já utilizada na *seção 2.1, se*

o modelo de regressão linear incluir uma constante, a matriz X pode ser decomposta da seguinte maneira:

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{\iota} & \mathbf{W} \end{bmatrix}$$

Utilizando (2.4), verifica-se que

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} \frac{1}{n} + \frac{1}{n^2}\boldsymbol{\iota}'\mathbf{W}(\mathbf{W}'\mathbf{A}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\iota} & -\frac{1}{n}\boldsymbol{\iota}'\mathbf{W}(\mathbf{W}'\mathbf{A}\mathbf{W})^{-1} \\ -\frac{1}{n}(\mathbf{W}'\mathbf{A}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\iota} & (\mathbf{W}'\mathbf{A}\mathbf{W})^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\iota}' \\ \mathbf{W}' \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{1}{n}\boldsymbol{\iota}' - \frac{1}{n}\boldsymbol{\iota}'\mathbf{W}(\mathbf{W}'\mathbf{A}\mathbf{W})^{-1}\mathbf{W}'\mathbf{A} \\ (\mathbf{W}'\mathbf{A}\mathbf{W})^{-1}\mathbf{W}'\mathbf{A} \end{bmatrix}$$

Lembrando que **AW** é a matriz com as variáveis explanatórias centradas, verifica-se que as últimas $k = p - 1$ linhas de **C** são idênticas às linhas da matriz correspondente obtida com $\mathbf{W}_* = \mathbf{A}\mathbf{W}$:

$$(\mathbf{W}_*'\mathbf{W}_*)^{-1}\mathbf{W}_*' = (\mathbf{W}'\mathbf{A}\mathbf{W})^{-1}\mathbf{W}'\mathbf{A}$$

Então os DFBETAS$_{ki}$ para os coeficientes de regressão podem ser calculados usando o elemento c_{ki} correspondente nessa matriz obtida com as variáveis centradas.

2.7. DFFITS

De acordo com (2.14), e tendo em vista que $V(\hat{Y}_i) = h_i\sigma^2$, uma medida padronizada da alteração em $\hat{Y}_i$ devida à exclusão de $\mathbf{x}_i'$ é dada por

$$\text{DFFITS}_i = \frac{h_i e_i}{(1-h_i)\sqrt{h_i s_{(i)}^2}} = \frac{e_i\sqrt{h_i}}{s_{(i)}(1-h_i)} \tag{2.27}$$

ou, lembrando (2.16),

$$\text{DFFITS}_i = e_i^*\sqrt{\frac{h_i}{(1-h_i)}} \tag{2.28}$$

Belsley, Kuh eWelsch (1980) recomendam que se dê atenção às observações para as quais

$$\left|\text{DFFITS}_i\right| > 2\sqrt{\frac{p}{n}}$$

Valor elevado do $\left|\text{DFFITS}_i\right|$ indica que a observação é influente. Posteriormente os DFBETAS$_{ki}$ podem ser utilizados para verificar se o fenômeno está associado a determinadas variáveis explanatórias.

A expressão (2.28) mostra que uma observação discrepante (valor absoluto de e_i^* elevado) tende a ser, também, uma observação influente $\left(\left|\text{DFFITS}_i\right| \text{ elevado}\right)$. É claro, também, que uma observação discrepante pode ser pouco influente, se o respectivo h_i for bastante pequeno. Vice-versa, se h_i for bastante grande, uma observação pode ser muito influente mesmo que o respectivo desvio seja pequeno (desde que não seja nulo).

2.8. O D de COOK

Uma outra estatística para diagnóstico é o *D* de Cook, definido por

$$D_i = \frac{(\mathbf{b}-\mathbf{b}_{(i)})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\mathbf{b}_{(i)})}{ps^2} \tag{2.29}$$

Essa expressão é formalmente semelhante ao *F* para testar uma hipótese do tipo $H_0 : \boldsymbol{\beta} = \boldsymbol{\theta}$ (ou para delimitar uma região de confiança para os *p* parâmetros). Trata-se de uma medida padronizada da influência da exclusão de $\mathbf{x}_i'$ sobre as estimativas dos *p* parâmetros.
Lembrando (2.12), verifica-se que

$$D_i = \frac{e_i^2 h_i}{ps^2(1-h_i)^2} \tag{2.30}$$

Comparando (2.27) e (2.30) verifica-se que o D_i de Cook é semelhante ao quadrado do DFFITS_i. A principal diferença é que no cálculo do D_i utiliza-se o Q.M.Res. obtido com as *n* observações (s^2), ao passo que no cálculo do DFFITS_i utiliza-se o Q.M.Res. obtido excluindo a *i*-ésima observação $(s_{(i)}^2)$. O DFFITS_i é uma medida mais sensível da influência da *i*-ésima observação, podendo-se recomendar o abandono do D_i de Cook.

2.9. Exemplos

Consideremos dois exemplos numéricos de regressão linear simples apresentados por Dachs e Carvalho (1984). Os dados estão na tabela 2.1.

Tabela 2.1. Valores de *X* e *Y* em dois conjuntos.

Exemplo 1		Exemplo 2	
X	*Y*	*X*	*Y*
1	4	1	1
3	2	1	2
5	3	11	3
7	1	11	4
14	6	6	6

Para ambos os exemplos temos $n = 5$ e a análise de regressão linear simples produz os seguintes resultados:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 30 \\ 30 & 280 \end{bmatrix} \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0{,}56 & -0{,}06 \\ -0{,}06 & 0{,}01 \end{bmatrix}$$

$\mathbf{X'y} = \begin{bmatrix} 16 \\ 116 \end{bmatrix}$ $\mathbf{b} = \begin{bmatrix} 2 \\ 0{,}2 \end{bmatrix}$ $\hat{Y} = 2 + 0{,}2X$

S.Q.Res. = 10,8 $s^2 = 3{,}6$ $r^2 = 0{,}2703$

As figuras 2.1 e 2.2 mostram os pontos observados e a reta ajustada, para os exemplos 1 e 2, respectivamente.

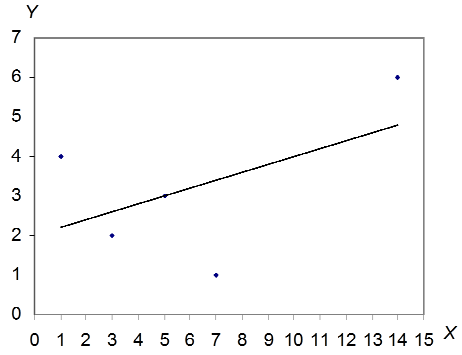


Figura 2.1. Exemplo 1

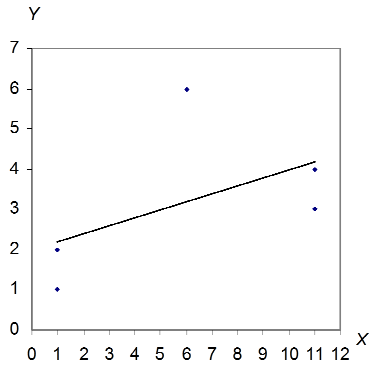


Figura 2.2. Exemplo 2

As tabelas 2.2 e 2.3 mostram os principais indicadores estatísticos para análise dos resíduos e avaliação da influência de cada observação.

Tabela 2.2. Análise dos resíduos e da influência de cada observação para o exemplo 1.

Y_i	$\hat{Y}_i$	e_i	h_i	$s^2_{(i)}$	e^*_i	DFFITS$_i$
4	2,2	1,8	0,45	2,4545	1,5492	1,4013
2	2,6	–0,6	0,29	5,1465	–0,3139	–0,2006
3	3,0	0	0,21	5,4000	0	0
1	3,4	–2,4	0,21	1,7544	–2,0386	–1,0511
6	4,8	1,2	0,84	0,9000	3,1623	7,2457

Tabela 2.3. Análise dos resíduos e da influência de cada observação para o exemplo 2.

Y_i	$\hat{Y}_i$	e_i	h_i	$s^2_{(i)}$	e^*_i	DFFITS$_i$
1	2,2	–1,2	0,45	4,0909	–0,8000	–0,7236
2	2,2	–0,2	0,45	5,3636	–0,1164	–0,1053
3	4,2	–1,2	0,45	4,0909	–0,8000	–0,7236
4	4,2	–0,2	0,45	5,3636	–0,1164	–0,1053
6	3,2	2,8	0,20	0,5000	4,4272	2,2136

Ao examinar os valores de h_i devemos ter em vista que, nesse caso, $2\,p/n = 0{,}8$.

Ao examinar os valores do resíduo estudentizado externamente devemos lembrar que ao nível de significância de 5% o valor crítico de *t*, com 2 graus de liberdade, é 4,303.

Ao examinar os valores de DFFITS$_i$ devemos ter em vista que o valor crítico sugerido por Belsley, Kuh e Welsch (1980) é

$$2\sqrt{\frac{p}{n}} = 1{,}2649$$

No caso do exemplo 1 a última observação se destaca pelo valor elevado do DFFITS$_i$ (superior a 7). Trata-se, claramente, de uma observação muito influente, com $h_i = 0{,}84$ (ver tabela 2.2 e figura 2.1). O valor do resíduo estudentizado externamente é elevado (3,1623), mas não alcança o valor crítico ao nível de significância de 5%. Note-se que o desvio (e_i) dessa observação é relativamente pequeno. Para verificar que a quinta observação é muito influente nesse exemplo, é interessante ajustar a regressão utilizando apenas as 4 primeiras observações, obtendo $\hat{Y} = 4{,}1 - 0{,}4X$. Note-se que há uma mudança radical nos valores dos coeficientes, com inversão do sinal do coeficiente de regressão. Verifica-se que DFBETAS$_{15}$ = –2,958 e DFBETAS$_{25}$ = 6,325, mostrando que a quinta observação tem influência muito forte sobre as estimativas dos parâmetros. A inclusão dessa observação reduz a estimativa de α (de 4,1 para 2) e tem um fortíssimo efeito positivo sobre a estimativa de β (que passa de $-0{,}4$ para 0,2).

No caso de exemplo 2 a única observação com DFFITS$_i$ elevado é a última (ver tabela 2.3 e figura 2.2). Não são os valores da variável explanatória que tornam essa observação influente, pois o valor de h_i é relativamente baixo. O valor do resíduo estudentizado externamente é bastante elevado, superando o valor crítico ao nível de significância de 5%. Trata-se, portanto, de uma observação discrepante. Verifica-se que nesse exemplo DFBETAS$_{15}$ = 1,323 e DEFBETAS$_{25}$ = 0, mostrando que a quinta observação tem uma influência substancial no sentido de aumentar a estimativa de α, mas não afeta a estimativa de β.

Exercícios

1. Admite-se que as variáveis *X* e *Y* estão relacionadas de acordo com o modelo $Y_i = \alpha + \beta X_i + u_i$, onde os u_i são erros aleatórios independentes com distribuição normal, média zero e variância σ^2.

   A tabela a seguir mostra os valores de *X* e *Y* em uma amostra com 6 observações.

Observação	*X*	*Y*
1ª	10	57,6
2ª	12	57,6
3ª	14	45,6
4ª	16	45,6
5ª	2	2,4
6ª	18	33,6

   a) Obtenha a equação de regressão linear simples de *Y* contra *X* com base nas 6 observações e calcule os desvios. Qual é o maior desvio, em valor absoluto?
   b) Determine o valor de h_i para a 1ª e para a 5ª observação. [Sugere-se a utilização da relação logo após (2.5), pois neste caso a matriz de variáveis centradas $\mathbf{W}_*$ é constituída por uma única coluna, facilitando os cálculos].
   c) Calcule o valor do resíduo estudentizado externamente para a 1ª e para a 5ª observação.
   d) Calcule o DFFITS para a 1ª e para a 5ª observação.

e) Alguma dessas duas observações pode ser considerada discrepante e/ou muito influente? (explique).

2. Admite-se que as variáveis *X* e *Y* estão relacionadas de acordo com o modelo $Y_i = \alpha + \beta X_i + u_i$, onde os u_i são erros aleatórios independentes com distribuição normal, média zero e variância σ^2.

A tabela ao lado mostra os valores de *X* e *Y* em uma amostra com 5 observações.

Observação	*X*	*Y*
1ª	1	7
2ª	1	9
3ª	2	7
4ª	2	11
5ª	3	3

a) Obtenha a equação de regressão linear simples de *Y* contra *X* com base nas 5 observações e calcule os desvios. Qual é, em valor absoluto, o maior desvio?
b) Para as 3 últimas observações (3ª, 4ª e 5ª), determine o valor de h_i, do resíduo estudentizado externamente e do DFFITS.
c) Alguma dessas três observações pode ser considerada discrepante e/ou muito influente? (explique)

Se fizer algum teste estatístico, adote um nível de significância de 10%.

3. Admite-se que *Y*, X_1 e X_2 estão relacionadas de acordo com o modelo

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

onde os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2.

A tabela ao lado mostra os valores das três variáveis em uma amostra com 4 observações.

Observação	X_1	X_2	*Y*
1ª	3	1	36
2ª	12	4	114
3ª	6	0	39
4ª	0	2	39

a) Estime a equação de regressão e calcule os desvios.
b) Determine o valor de h_i, do resíduo estudentizado externamente e do DFFITS para cada observação.
c) Qual é a observação mais influente? (Justifique).
d) O resíduo estudentizado externamente dessa observação é estatisticamente significativo ao nível de 10%?
e) Com base nos valores dos DFBETAS, qual dos dois coeficientes de regressão é mais afetado pela observação mais influente?

4. Admite-se que *Y*, X_1 e X_2 estão relacionadas de acordo com o modelo

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

onde os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2.

A tabela ao lado mostra os valores das três variáveis em uma amostra com 6 observações.

Observação	Y	X_1	X_2
1ª	70	3	1
2ª	120	12	4
3ª	20	6	0
4ª	20	0	2
5ª	55	12	0
6ª	55	0	4

a) Estime a equação de regressão e calcule os desvios.

b) Determine o valor de h_i, do resíduo estudentizado externamente e do DFFITS para cada observação.

c) Discuta os resultados, verificando se há alguma observação discrepante e/ou muito influente (Adote um nível de significância de 5%).

5. Admite-se que as variáveis X_i e Y_i estão relacionadas de acordo com o modelo

$$Y_i = \alpha + \beta X_i + u_i,$$

onde os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2.

A tabela ao lado mostra os valores das duas variáveis em uma amostra com 4 observações.

Obs.	X_i	Y_i
1ª	1	19
2ª	1	21
3ª	7	14
4ª	7	44

a) Estime a equação de regressão.

b) Faça uma tabela com os valores de *Y* estimado, desvio, h_i, resíduo padronizado, resíduo estudentizado externamente e $DFFITS_i$ para cada observação.

c) Identifique as observações discrepantes, adotando um nível de significância de 5%.

d) Há observações muito influentes? (comente).

e) Calcule os $DFBETAS_{ki}$ para a última observação. Qual dos dois coeficientes é mais afetado por essa observação?

6. Admite-se que as variáveis X e Y estão relacionadas de acordo com o modelo

$$Y_i = \alpha + \beta X_i + u_i,$$

no qual os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2.

A tabela mostra os valores das duas variáveis em uma amostra com 5 observações.

Obs.	X	Y
1ª	1	6
2ª	2	7
3ª	3	10
4ª	4	10
5ª	5	12

a) Estime a equação e teste, ao nível de significância de 1%, a hipótese de que $\beta = 0$, contra a hipótese de que $\beta > 0$.

b) Adotando um nível de significância de 10%, verifique se a 5ª observação é discrepante. Ela pode ser considerada muito influente?

c) Idem, para a 3ª observação.

7. Admite-se que as variáveis X e Y estão relacionadas de acordo com o modelo

$$Y_i = \alpha + \beta X_i + u_i,$$

no qual os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2.

Obs.	X	Y
1ª	1	15
2ª	2	8
3ª	2	12
4ª	3	5
5ª	8	54
6ª	8	56

A tabela mostra os valores das duas variáveis em uma amostra com 6 observações. Adotando um nível de significância de 5%, verifique se o conjunto das duas últimas observações é discrepante das demais.

8. São dados 5 valores da variável aleatória *Y*: 5, 9, 8, 6 e 2. Adotando um nível de significância de 10%, a quinta observação é discrepante? Justifique.

9. Admite-se que as variáveis X e Y estão relacionadas de acordo com o modelo

$$Y_i = \alpha + \beta X_i + u_i,$$

no qual os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2. A tabela ao lado mostra os valores das duas variáveis em uma amostra com 7 observações.

Obs.	X	Y
1ª	2	12
2ª	3	11
3ª	3	12
4ª	3	13
5ª	4	12
6ª	9	22
7ª	11	30

a) Estime a equação.

b) Teste, ao nível de significância de 1%, a hipótese de que $\beta = 1$, contra a hipótese alternativa de que $\beta > 1$.

c) Verifique se a 7ª observação é discrepante, adotando um nível de significância de 5%.

d) Verifique se o conjunto formado pelas duas últimas observações (a 6ª e a 7ª) é discrepante, adotando um nível de significância de 5%.

10. Admite-se que as variáveis X_1, X_2 e Y estão relacionadas de acordo com o modelo

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

onde os u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2. A tabela a seguir mostra os valores das três variáveis em uma amostra com 5 observações.

Observação	X_1	X_2	Y	x_1	x_2	y
1ª	2	3	27	–5	0	11
2ª	11	7	26	4	4	10
3ª	5	1	5	–2	–2	–11
4ª	7	3	21	0	0	5
5ª	10	1	1	3	–2	–15

a) Estime a equação de regressão e calcule os desvios.

b) Teste, ao nível de significância de 10%, a hipótese $H_o : \beta_1 = \beta_2 = 0$.

c) Teste, ao nível de significância de 10%, a hipótese $H_o : \beta_1 = 0$ contra a hipótese alternativa $H_a : \beta_1 < 0$.

d) Determine o valor de h_i, do resíduo estudentizado externamente e do DFFITS$_i$ para a 2ª e a 3ª observação. Qual dessas duas observações tem desvio ($e_i = Y_i - \hat{Y}_i$) maior? Verifique se alguma dessas duas observações é discrepante e/ou muito influente. Adote um nível de significância de 10%.

e) Se uma das duas observações analisadas for muito influente, calcule os correspondentes DFBETAS$_{ki}$ para os dois coeficientes de regressão e interprete os resultados.

Observação: Se fizer os cálculos usando variáveis centradas, lembre da relação (2.5).

11. A tabela a seguir mostra os valores de *X* e *Y* em uma amostra com 8 observações.

a) Ajuste uma regressão linear simples de *Y* contra *X*, por mínimos quadrados ordinários, aos dados de toda a amostra, e determine a soma de quadrados residual.

b) Faça um teste para verificar se o conjunto das duas últimas observações (a 7ª e a 8ª) pode ser considerado discrepante das demais, adotando um nível de significância de 5%.

Obs.	*X*	*Y*
1ª	1	14
2ª	3	16
3ª	2	12
4ª	3	14
5ª	4	20
6ª	3	16
7ª	1	3
8ª	5	25

12. Admite-se que as variáveis X_1, X_2 e Y estão relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j,$$

no qual os u_j são erros aleatórios independentes com distribuição normal com média zero e variância σ^2. A tabela a seguir mostra os valores das três variáveis e das respectivas variáveis centradas em uma amostra com 9 observações.

Observação	X_1	X_2	Y	x_1	x_2	y
1ª	8	9	102	2	0	42
2ª	6	15	102	0	6	42
3ª	4	15	54	–2	6	–6
4ª	4	9	18	–2	0	–42
5ª	6	3	18	0	–6	–42
6ª	8	3	54	2	–6	–6
7ª	7	9	96	1	0	36
8ª	6	12	96	0	3	36
9ª	5	6	0	–1	–3	–60

Para o modelo com todas as variáveis centradas obtemos

$$\mathbf{X'X} = \begin{bmatrix} 18 & -21 \\ -21 & 162 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} 264 \\ 792 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} = \frac{1}{825}\begin{bmatrix} 54 & 7 \\ 7 & 6 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 24 \\ 8 \end{bmatrix}$$

a) Determine a soma de quadrados dos desvios dessa regressão.

b) Estime a equação de regressão considerando apenas as 6 primeiras observações e obtenha a respectiva soma de quadrados residual.

c) Adotando um nível de significância de 5%, verifique se o conjunto das 3 últimas observações é *discrepante*.

13. Para uma regressão linear múltipla de **y** contra **X** (com p colunas e n linhas) temos

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y} \quad , \quad \mathbf{e} = \mathbf{y} - \mathbf{Xb} \qquad \text{e} \qquad s^2 = \frac{1}{n-p}(\mathbf{y'y} - \mathbf{b'X'y})$$

Vamos admitir que haja desconfiança de que as duas últimas observações são discrepantes. Então são introduzidas, na regressão, duas variáveis binárias para captar a influência dessas duas últimas observações. A nova matriz de variáveis explanatórias fica

$$\mathbf{W} = [\mathbf{X} \quad \mathbf{L}]$$

onde $\mathbf{L}$ é uma matriz $n \times 2$ cujas $n - 2$ primeiras linhas são constituídas por zeros e cujas duas últimas linhas formam uma matriz identidade de ordem 2. Note que $\mathbf{L'L} = \mathbf{I}_2$.

a) Demonstre que o novo vetor de estimativas dos parâmetros, com p+2 elementos, é

$$\begin{bmatrix} \mathbf{b}_* \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{b} - (\mathbf{X'X})^{-1}\mathbf{X'L}(\mathbf{I}_2 - \mathbf{L'HL})^{-1}\mathbf{L'e} \\ (\mathbf{I}_2 - \mathbf{L'HL})^{-1}\mathbf{L'e} \end{bmatrix} \qquad (2.31)$$

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

Note que a expressão (2.7) é um caso particular de (2.31).

b) Demonstre que a diminuição da S.Q.Res. devida à introdução das duas variáveis binárias é

$$\mathbf{e'L(I_2 - L'HL)^{-1}L'e} \qquad (2.32)$$

Note que a expressão (2.8) é um caso particular de (2.32).

c) Mostre como obter um valor de F com 2 e $n - p - 2$ graus de liberdade para testar se as duas últimas observações são ou não discrepantes.

d) Se, em lugar de incluirmos as duas variáveis binárias, excluirmos as duas últimas observações, o vetor das estimativas dos parâmetros fica

$$\mathbf{b}_L = (\mathbf{X'X - X'LL'X})^{-1}(\mathbf{X'y - X'LL'y})$$

e a S.Q.Res. fica

$$Q_L = \mathbf{y'y - y'LL'y - b}'_L(\mathbf{X'y - X'LL'y})$$

Demonstre que $\mathbf{b}_L = \mathbf{b}_*$

e) Demonstre que Q_L é igual à S.Q.Res. da regressão incluindo as duas variáveis binárias (ambas com $n - p - 2$ graus de liberdade).

f) Considere a amostra de valores de X e Y na tabela ao lado e mostre que a S.Q.Res de uma regressão linear simples de Y contra X com as 6 observações é 124.

g) Verifique que a S.Q.Res de uma regressão linear simples com as 4 primeiras observações é 4.

h) Mostre que o teste F para verificar se o conjunto das duas últimas observações é discrepante é $F = 30$, significativo a 5%, apesar de os resíduos estudentizados externamente dessas observações serem baixos (0 e 0,976).

X	Y
1	10
1	12
3	2
3	4
7	13
7	17

i) Faça uma regressão múltipla com as 6 observações, incluindo duas variáveis binárias para captar o efeito das duas últimas observações e verifique que o teste F para a contribuição dessas duas variáveis também é igual a 30.

j) Faça o teste para "mudança estrutural" entre as 4 primeiras e as duas últimas observações.

14. Uma variável Y é observada em 3 diferentes situações (tratamentos). Com base em uma amostra com 5 observações para cada tratamento (total de 15 observações) foram obtidas as seguintes médias de tratamento: $\overline{Y}_1 = 6$, $\overline{Y}_2 = 8$ e $\overline{Y}_3 = 13$. Considera-se o modelo de regressão

$$Y_i = \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + u_i) \ ,$$

onde Z_{hi} (com h = 1, 2, 3) são variáveis binárias, com $Z_{hi} = 1$ para observações do h-ésimo tratamento, e $Z_{hi} = 0$ nos demais casos, e u_i são erros aleatórios independentes com distribuição normal com média zero e variância σ^2. Estimando os 3 parâmetros por mínimos quadrados ordinários, obtém-se o seguinte vetor de desvios:

$$e' = [-1 \quad 1 \quad 0 \quad 1 \quad -1 \quad -4 \quad 1 \quad 2 \quad 0 \quad 1 \quad -2 \quad 0 \quad 2 \quad -1 \quad 1]$$

a) Verifique se a 6ª observação é discrepante, adotando um nível de significância de 1%.

b) Teste, ao nível de significância de 5%, a hipótese de que $\beta_1 = \beta_2$.

c) Teste, ao nível de significância de 1%, a hipótese de que $\beta_1 + \beta_2 = 2\beta_3$.

15. Admite-se que as variáveis *X* e *Y* estão relacionadas de acordo com o modelo usual de regressão linear simples. Uma amostra de 6 observações é fornecida na tabela ao lado. Note que a média dos valores de X para as 6 observações e para as 4 primeiras observações é a mesma $(\overline{X} = 7)$. Usando letras minúsculas para indicar as variáveis centradas, obtém-se

X_i	Y_i
5	15
5	13
9	21
9	23
6	25
8	29

$$\sum_{i=1}^{6} x_i^2 = 18 \ , \ \sum_{i=1}^{4} x_i^2 = 16$$

$$\sum_{i=1}^{6} x_i y_i = 36 \ , \ \sum_{i=1}^{4} x_i y_i = 32$$

Adote um nível de significância de 5%.

a) A 6ª observação é discrepante?

b) O conjunto formado pelas 2 últimas observações é discrepante?

Respostas

1. Equação estimada: $\hat{Y} = 15{,}2 + 2{,}1X$

X_i	Y_i	$\hat{Y}_i$	e_i	h_i	$t = e_i^*$	DFFITS$_i$
10	57,6	36,2	21,4	0,1917	1,4027	0,6830
12	57,6	40,4	17,2	0,1667	0,9952	0,4451
14	45,6	44,6	1,0	0,1917	0,0510	0,0248
16	45,6	48,8	–3,2	0,2667	–0,1720	–0,1037
2	2,4	19,4	–17,0	0,7917	–9,8150	–19,1329
18	33,6	53,0	–19,4	0,3917	–1,5121	–1,2133

O valor médio dos h_i é $\frac{p}{n} = \frac{2}{6} = 0{,}3333$.

Recomenda-se dar atenção às observações com $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}} = 1{,}1547$

O valor crítico de *t*, com 3 graus de liberdade, ao nível de significância de 5%, é 3,182

A 5ª observação é influente (h_i elevado) e discrepante (valor absoluto de e_i^* elevado). Note-se que o correspondente valor de DFFITS$_i$ é extremamente elevado, em valor absoluto.

A 1ª observação, apesar de apresentar o maior desvio (em valor absoluto), não pode ser considerada discrepante.

2. a) $\hat{Y} = 11 - 2X$

O maior desvio, em valor absoluto, é o da 4ª observação (e_4 = 4).

b) $\sum x^2 = 2{,}8$
$\sum e^2 = 24$
$s^2 = 8$

Obs.	h	$s^2_{(i)}$	$t = e^*$	DFFITS
3ª	0,214	12	0	0
4ª	0,214	1,8182	3,347	1,748
5ª	0,714	5	–1,673	–2,646

c) O valor médio dos h_i é 0,4.
O valor crítico de *t* com 2 graus de liberdade, ao nível de significância de 10%, é 2,92.
São considerados elevados os valores de DFFITS maiores do que

$$2\sqrt{\frac{p}{n}} = 2\sqrt{0{,}4} = 1{,}265$$

Conclui-se que a 4ª observação é discrepante e que a 5ª observação é muito influente.

3. $\mathbf{X'X} = \begin{bmatrix} 189 & 51 \\ 51 & 21 \end{bmatrix}$ $\mathbf{X'y} = \begin{bmatrix} 1710 \\ 570 \end{bmatrix}$ $(\mathbf{X'X})^{-1} = \frac{1}{456}\begin{bmatrix} 7 & -17 \\ -17 & 63 \end{bmatrix}$

$\mathbf{b} = \begin{bmatrix} 5 \\ 15 \end{bmatrix}$ $\hat{Y} = 5X_i + 15X_2$

	Y_i	$\hat{Y}_i$	e_i	h_i	$s^2_{(i)}$	e^*_i	DFFITS$_i$
1ª	36	30	6	$\frac{1}{19} = 0{,}052$	196	0,440	0,104
2ª	114	120	–6	$\frac{16}{19} = 0{,}8421$	6	–6,164	–14,236
3ª	39	30	9	$\frac{21}{38} = 0{,}5526$	$\frac{900}{17}$	1,849	2,055
4ª	39	30	9	$\frac{21}{38} = 0{,}5526$	$\frac{900}{17}$	1,849	2,055

c) A observação mais influente é a 2ª, com o maior DFFITS
d) Ao nível de significância de 10%, com 1 grau de liberdade, o valor crítico de *t* é 6,314. O valor de e^*_i para a 2ª observação não chega a ser estatisticamente significativo ao nível de 10%.
e) Os dois coeficientes são igualmente afetados, pois

$$\text{DFBETAS}_{12} = \text{DFBETAS}_{22} = -4{,}393$$

4. $\hat{Y} = 5X_1 + 15X_2$, S.Q.Res. = 1.850 , s^2 = 462,5

Y	e	h	e^*	DFFITS
70	40	$\frac{1}{27}$	5,143	1,009
120	0	$\frac{16}{27}$	0	0
20	–10	$\frac{37}{270}$	–0,448	–0,178
20	–10	$\frac{37}{270}$	–0,448	–0,178
55	–5	$\frac{74}{135}$	–0,304	–0,335
55	–5	$\frac{74}{135}$	–0,304	–0,335

Com 3 graus de liberdade, ao nível de significância de 5%, $t_0 = 3{,}182$.

Valor "crítico" do DFFITS = $2\sqrt{\frac{p}{n}} = 1{,}155$.

Verifica-se que a 1ª observação é discrepante.
Os valores do DFFITS não indicam nenhuma observação muito influente.

5. $\hat{Y} = 18{,}5 + 1{,}5X$
$s^2 = 226$
$s = 15{,}033$
$\frac{2p}{n} = 1$ e $t_c = 12{,}706$

Obs.	$\hat{Y}_i$	e_i	h_i	$\frac{e_i}{s}$	$e_i^* = \text{DFFITS}_i$
1ª	20	–1	0,5	–0,0665	–0,0667
2ª	20	1	0,5	0,0665	0,0667
3ª	29	–15	0,5	–0,9978	–15
4ª	29	15	0,5	0,9978	15

O valor crítico para DFI

Tanto a 3ª como a 4ª observação são discrepantes e muito influentes.
$\text{DFBETAS}_{a4} = -2{,}12$ e $\text{DFBETAS}_{b4} = 10{,}61$.

6. a) $\hat{Y} = 4{,}5 + 1{,}5X$, t = 6,708, significativo ($t_0 = 4{,}541$).
b) $e_5 = 0$, não é discrepante ou influente.
c) $e_3 = 1$, $t_3 = 3{,}162$, significativo ($t_0 = 2{,}920$).
DFFITS$_3$ = 1,581 > 1,265. Pode ser considerada influente.

7. Regressão com os 6 pontos: $\hat{Y} = -3 + 7X$, com S.Q.Res. = 310.
Regressão com os 4 pontos: $\hat{Y} = 20 - 5X$, com S.Q.Res. = 8.
F = 37,75, significativo ($F_0 = 19{,}00$), mostrando que o conjunto das duas últimas observações é discrepante das demais. Pode-se verificar que, isoladamente, nenhuma das duas observações é discrepante.

8. A quinta observação é discrepante pois, adotando o modelo $Y_i = \alpha + u_i$, o valor de t estudentizado externamente, t = 2,449, é significativo ($t_0 = 2{,}353$).

9. a) $\hat{Y} = 6 + 2X$ b) t = 4,535, significativo ($t_0 = 3{,}365$).

c) $h_7 = 0{,}6293$, $s^2_{(7)} = \dfrac{173}{96}$, t = 2,447, não-significativo ($t_0 = 2{,}776$).

d) F = 12, significativo ($F_0 = 9{,}55$).

10. a) Com variáveis centradas obtemos

$$\mathbf{X'X} = \begin{bmatrix} 54 & 14 \\ 14 & 24 \end{bmatrix} \qquad (\mathbf{X'X})^{-1} = \frac{1}{550}\begin{bmatrix} 12 & 7 \\ 7 & 27 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} -38 \\ 92 \end{bmatrix}$$

$\overline{X}_1 = 7$, $\overline{X}_2 = 3$, $\overline{Y} = 16$

$\overline{Y} = 15 - 2X_1 + 5X_2$

Os desvios são 1, −2, −5, 5 e 1.

S.Q.Res = 56

$s^2 = 28$

b) $F = \dfrac{268}{28} = 9{,}57$, significativo ($F_0 = 9{,}00$)

c) t = –2,559, significativo (a região de rejeição é $t \leq -1{,}886$)

d)

Observação	h_i	$s^2_{(i)}$	e^*_i	DFFITS$_i$
2ª	$\dfrac{51}{55} = 0{,}9273$	1	$-\sqrt{55} = -7{,}416$	–26,48
3ª	$\dfrac{21}{55} = 0{,}3818$	15,56	–1,612	–1,27

Embora o valor absoluto do desvio da 3ª observação seja substancialmente maior, é a 2ª observação que é discrepante ($e^*_2 = -7{,}416$, com $t_0 = 6{,}314$). A 2ª observação também é muito influente.

e) $\text{DFBETAS}_{12} = -6{,}77$ e $\text{DFBETAS}_{22} = -18{,}05$

A inclusão da 2ª observação causa forte redução no valor de b_1 e (mais ainda) de b_2

11. a) $\hat{Y} = 4 + 4X$, com S.Q.Res. = 66.

b) F = 6,25, não-significativo ($F_0 = 6{,}94$).

Pode-se verificar que, ao nível de significância de 5%, a 7ª observação, isoladamente, é discrepante (t = –2,652).

12. a) S.Q.Res. = 648

b) $\hat{Y} = -131 + 21X_1 + 7X_2$

S.Q.Res. = 48

c) F = 12,50, significativo. Com 3 e 3 graus de liberdade, $F_0 = 9{,}28$. Conclui-se que o conjunto das 3 últimas observações é discrepante. Pode-se verificar que, isoladamente, nenhuma das 3 observações é discrepante.

13. f) $\hat{Y} = 6 + X$, com S.Q.Res. = 124

g) $\hat{Y} = 15 - 4X$, com S.Q.Res. = 4

h) $F = \dfrac{\dfrac{124-4}{2}}{\dfrac{4}{2}} = 30$, com 2 e 2 graus de liberdade. $F_0 = 19{,}00$.

j) Como nas duas últimas observações o valor de X é o mesmo, só é possível verificar "mudança estrutural" no parâmetro α (termo constante). Considerando o modelo $Y = \alpha + u$ para essas duas observações, a S.Q.Res é igual a 8, com 1 grau de liberdade. Então, o teste para mudança estrutural é

$F = \dfrac{\dfrac{124-(4+8)}{1}}{\dfrac{4+8}{3}} = 28$, com 1 e 3 graus de liberdade, significativo ao nível de 5%, pois $F_0 = 10{,}13$.

14. a) $t = -3{,}708$, com $t_0 = 3{,}106$. A 6ª observação é discrepante.

b) $t = -1{,}826$, não-significativo $(t_0 = 2{,}179)$

c) $t = -6{,}325$, significativo $(t_0 = 3{,}055)$

15. a) $t = 1{,}454$, não-significativo $(t_0 = 3{,}182)$. A 6ª observação não é discrepante.

b) $F = 27$, significativo $(F_0 = 19)$. O conjunto das duas últimas observações é discrepante. Observação: pode-se verificar que, isoladamente, a 5ª observação também não é discrepante (com t = 1,454, da mesma maneira que para a 6ª observação).

3. ANÁLISE HARMÔNICA

3.1. Introdução

Frequentemente a variável dependente em uma análise de regressão apresenta variações cíclicas. Esse é o caso, por exemplo, da variação estacional em uma série de preços mensais de um produto agropecuário. Essas variações podem ser captadas utilizando variáveis binárias. Se a equação de regressão tem um termo constante, são necessárias $T-1$ variáveis binárias para captar variações cíclicas com período igual a T termos da série. Assim, para uma série de preços mensais (T = 12) será necessário utilizar 11 variáveis binárias para captar a sua variação estacional por meio de uma equação de regressão incluindo um termo constante. Se a equação de regressão não tiver termo constante deverão ser utilizadas T variáveis binárias[7].

As variações cíclicas de uma variável também podem ser captadas, em análise de regressão, usando a função cosseno (cuja representação gráfica é denominada cossenóide).

Na próxima seção serão examinadas duas formas de representar uma cossenóide, ou seja, as duas formas de escrever um *componente harmônico*. Na seção seguinte mostra-se como uma soma de componentes harmônicos pode ser utilizada para representar variações cíclicas com formas mais complexas.

Admite-se que o período (T) da variação cíclica é conhecido. Para uma série de dados mensais o período da variação cíclica estacional é, obviamente, T = 12. No caso de dados trimestrais o período da variação estacional é T = 4.

Quando usamos variáveis binárias para captar uma variação cíclica, é necessário que o período do ciclo seja um número inteiro de unidades de tempo. Isso ocorre, por exemplo, quando queremos captar o ciclo estacional em séries de dados mensais ou trimestrais, pois o ano tem 12 meses ou 4 trimestres. Uma vantagem da análise harmônica é que ela não tem essa restrição, permitindo captar variações cíclicas cujo período não é um número inteiro de unidades de tempo, como é o caso da variação estacional em uma série de dados semanais (pois o ano não tem um número inteiro de semanas). Na análise harmônica o período T pode ser qualquer número real positivo[8].

---

[7] Sobre o uso de variáveis binárias em regressão ver, por exemplo, o capítulo 5 de "Análise de Regressão - Uma Introdução à Econometria", de R. Hoffmann.

[8] Como exemplo interessante de aplicação, ver Okawa (1985).

3.2. Componente Harmônico

As ordenadas de uma cossenóide com período T, amplitude A e fase inicial $-\psi$ são dados por

$$Y = A\cos\left(\frac{2\pi}{T}t - \psi\right) \tag{3.1}$$

ou

$$Y = A\cos(\omega t - \psi) \tag{3.2}$$

onde

$\omega = \frac{2\pi}{T}$ é a velocidade angular[9].

A abcissa (ou variável) explanatória é indicada por t pois nesse tipo de análise ela corresponde, quase sempre, a tempo.

Note-se que os ângulos são medidos em radianos, como é usual. Se os ângulos fossem medidos em graus a velocidade angular seria $\omega = 360°/T$.

Uma vez que o cosseno varia de -1 a $+1$, o valor de Y varia de $-A$ a A. Portanto, a diferença entre o valor máximo e o valor mínimo de Y em (3.1) é igual a duas vezes a amplitude do componente harmônico.

Verifica-se que o valor inicial de Y (quando $t = 0$) é $A\cos(-\psi)$, valor que se repete para $t = T$, $t = 2T$, $t = 3T$, etc.

A figura 3.1 mostra a cossenóide com amplitude 10, período 12 e fase inicial $\pi/3$, para $0 \leq t \leq 24$ (dois períodos completos).

[9] É claro que essa terminologia tem origem na Física. Há muita semelhança entre a análise harmônica, em Estatística, e o estudo do movimento harmônico simples, na Física.

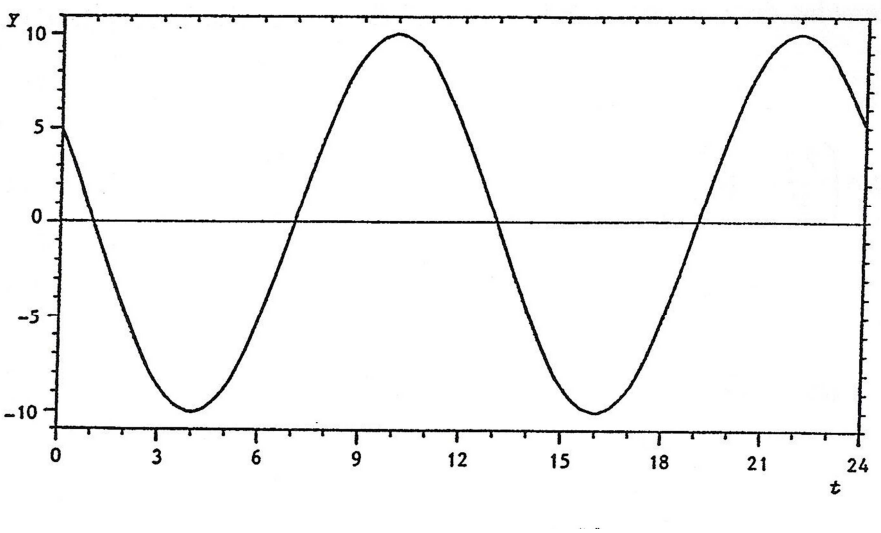


Figura 3.1. Cossenóide com período 12, amplitude 10 e fase inicial $\pi/3$

De (3.2) segue-se

$$Y = A(\cos\omega t \cos\psi \ + \text{sen}\,\omega t \,\text{sen}\,\psi)$$

ou

$$Y = A\cos\psi \ \cos\omega t + A\,\text{sen}\,\psi \ \text{sen}\,\omega t$$

Fazendo

$$A\cos\psi = \beta \tag{3.3}$$

e

$$A\,\text{sen}\,\psi = \gamma \tag{3.4}$$

obtém-se

$$Y = \beta\cos\omega t + \gamma \ \text{sen}\,\omega t \tag{3.5}$$

As expressões (3.2) e (3.5) são as duas formas alternativas básicas de representar um componente harmônico.

Se o período (*T*) é conhecido, a velocidade angular (ω) também é conhecida. Portanto, os parâmetros (desconhecidos) na equação (3.2) são A e ψ. Já na equação (3.5) os parâmetros são β e γ. As relações (3.3) e (3.4) mostram como os parâmetros A e ψ são transformados nos parâmetros β e γ.

Com ω conhecido, é fácil calcular os valores de

$$X_1 = \cos \omega t$$

e

$$X_2 = \text{sen } \omega t$$

para diferentes valores de *t*. Dada uma série temporal de valores de *Y*, é claro, pela expressão (3.5), que os parâmetros β e γ podem ser estimados fazendo-se uma regressão de *Y* contra X_1 e X_2.

De (3.3) e (3.4), elevando ao quadrado e somando, obtém-se

$$A^2 = \beta^2 + \gamma^2$$

ou

$$A = \sqrt{\beta^2 + \gamma^2} \tag{3.6}$$

Dividindo (3.4) por (3.3), membro-a-membro, tem-se

$$\text{tg } \psi = \frac{\gamma}{\beta} \tag{3.7}$$

As expressões (3.6) e (3.7) mostram com os parâmetros A e ψ podem ser obtidos a partir de β e γ. Analogamente, se b e c são estimativas de β e γ, respectivamente, as estimativas de A e ψ podem ser obtidas através das relações

$$\hat{A} = \sqrt{b^2 + c^2} \tag{3.8}$$

e

$$\text{tg } \hat{\psi} = \frac{c}{b} \tag{3.9}$$

De acordo com (3.3) e (3.4) tem-se

$$b = \hat{A} \cos \hat{\psi}$$

e

$$c = \hat{A} \text{ sen } \hat{\psi}$$

Então o sinal de b é igual ao sinal de $\cos\ \hat{\psi}$ e o sinal de c é igual ao sinal de $\text{sen}\ \hat{\psi}$.

Ao obter o valor de $\hat{\psi}$ a partir dos valores de b e c, utilizando a expressão (3.9), é necessário levar em consideração os sinais de b e c para determinar o quadrante em que se localiza o ângulo $\hat{\psi}$. Um determinado valor positivo de c/b, por exemplo, pode corresponder tanto a um ângulo $\hat{\psi}$ no primeiro quadrante como a um ângulo $\hat{\psi}$ no terceiro quadrante. Trata-se de um ângulo no primeiro quadrante se b e c forem positivos ($\cos\hat{\psi}$ e $\text{sen}\,\hat{\psi}$ positivos). Trata-se de um ângulo no terceiro quadrante se b e c forem negativos ($\cos\hat{\psi}$ e $\text{sen}\,\hat{\psi}$ negativos).

3.3. O Modelo Geral de Análise Harmônica

É claro que a equação (3.2) ou (3.5) dificilmente poderá ser utilizada para analisar o comportamento de variáveis econômicas, pois tipicamente essas variáveis não assumem valores negativos e geralmente suas variações cíclicas não tem a forma de uma cossenóide.

Para eliminar os valores negativos basta introduzir no modelo um termo constante $\alpha > A$. Então o modelo fica

$$Y = \alpha + A\cos(\omega t - \psi)$$

ou

$$Y = \alpha + \beta\cos\ \omega t + \gamma\ \text{sen}\ \omega t$$

Mas isso não altera a *forma* da curva. Formas mais complexas são obtidas *somando* a cossenóide com período T com cossenóides com período $T/2$, $T/3$, etc. A figura 3.2 ilustra esse procedimento mostrando o resultado que pode ser obtido somando uma cossenóide com período $T = 12$, uma cossenóide com período $T/2 = 6$ e uma cossenóide com período $T/3 = 4$. Note-se que os diversos componentes harmônicos também podem ter amplitudes e fase iniciais diferentes.

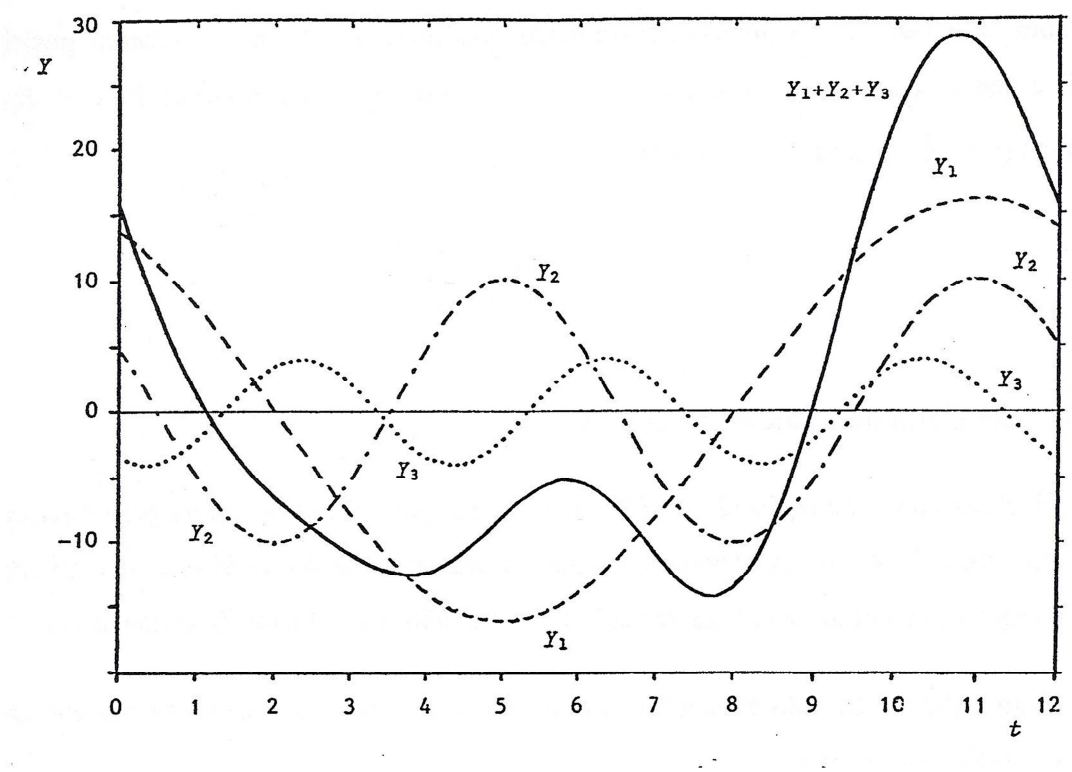


Figura 3.2. A soma de 3 cossenóides: $Y_{1t} = 16\cos\left(\frac{2\pi}{12}t + \frac{\pi}{6}\right)$,

$Y_{2t} = 10\cos\left(\frac{2\pi}{6}t + \frac{\pi}{3}\right)$ e $Y_{3t} = 4\cos\left(\frac{2\pi}{4}t + \frac{5\pi}{6}\right)$

O modelo geral de análise harmônica seria, então

$$Y_t = \alpha + \sum_{i=1}^{h} A_i \cos\left(\frac{2\pi\ i}{T}t - \psi_i\right) + u_t$$

ou

$$Y_t = \alpha + \sum_{i=1}^{h}\left(\beta_i \cos\frac{2\pi\, i}{T}t + \gamma_i \operatorname{sen}\frac{2\pi\, i}{T}t\right) + u_t$$

onde u_t é o erro associado à t-ésima observação de Y.

Entretanto, o número de termos com cosseno não é necessariamente igual ao número de termos com seno. Então, o modelo geral fica

$$Y_t = \alpha + \sum_{i=1}^{h} \beta_i \cos\frac{2\pi\, i}{T}t + \sum_{i=1}^{k} \gamma_i \operatorname{sen}\frac{2\pi\, i}{T}t + u_t \qquad (3.10)$$

ou

$$Y_t = \alpha + \sum_{i=1}^{h} \beta_i \cos \omega i t + \sum_{i=1}^{k} \gamma_i \operatorname{sen} \omega i t + u_t \qquad (3.11)$$

As estimativas dos parâmetros α, β_i e γ_i são obtidas fazendo-se a regressão de Y_t contra

$$X_i = \cos \frac{2\pi i}{T} t \qquad (\text{com } i = 1, \ldots, h)$$

e

$$W_i = \operatorname{sen} \frac{2\pi i}{T} t \qquad (\text{com } i = 1, \ldots, k)$$

Vamos admitir que o período (T) é um número inteiro e que cada termo da série corresponde a uma unidade de tempo.

Se o período (T) das variações cíclicas corresponde a um número *ímpar* de termos, isto é, se T for ímpar, então o número máximo de termos no segundo membro do modelo (3.10) é $h + k + 1$ com

$$h = k = \frac{T-1}{2} \qquad (3.12)$$

Se o período T for *par* então

$$h = \frac{T}{2} \text{ e } k = \frac{T}{2} - 1 \qquad (3.13)$$

Note-se que nos dois casos o número total de parâmetros (α, β_i e γ_i) no modelo (3.10) é igual a T.

O número total de parâmetros não pode exceder o número de posições distintas no ciclo. A tentativa de introduzir termos adicionais no modelo, fora dos limites estabelecidos por (3.12) ou (3.13), faria com que a matriz $\mathbf{X}$ tivesse característica menor do que o número de colunas e, consequentemente, a matriz $\mathbf{X}'\mathbf{X}$ seria singular.

Considere-se, por exemplo, o caso de $T = 12$. De acordo com (3.13) pode-se utilizar um modelo com no máximo 6 termos com cosseno e 5 termos com seno. A tentativa de introduzir um 6º termo com seno faria com que a correspondente coluna na matriz $\mathbf{X}$ fosse formada apenas por zeros pois, com $T = 12$ e $i = 6$ tem-se

$$\operatorname{sen} \frac{2\pi i}{T} t = \operatorname{sen} \pi t = 0 \quad \text{para todo } t.$$

Utilizando um modelo de análise harmônica incluindo todos os termos possíveis de acordo com (3.12) ou (3.13) obter-se-á um coeficiente de determinação exatamente igual ao de uma regressão com termo constante e $T-1$ variáveis binárias. Essa equivalência é óbvia quando $T = 2$: nesse caso a análise harmônica incluirá somente uma variável.

$$X = \cos \frac{2\pi}{T} t = \cos \pi t$$

que assume apenas dois valores distintos (1 e -1) .

Se, em uma série temporal de dados, cada observação corresponde a uma unidade de tempo, mas o período das variações cíclicas não é um número inteiro dessas unidades de tempo, então não se aplicam as condições (3.12) ou (3.13). Mas é sempre verdade que o número total de parâmetros do modelo (3.10) não pode exceder o número de posições distintas no ciclo para as quais há dados observados. Considere-se, por exemplo, o exercício 6, com observações a cada 8 meses. O período das variações estacionais é $T = 1{,}5$ unidades de 8 meses. Como os valores observados correspondem a apenas 3 posições distintas no ciclo anual, o número máximo de parâmetros no modelo (3.10) é 3, isto é, os valores máximos para h e k são $h = k = 1$. Como outro exemplo, considere uma variável que é observada a cada 5 meses, com variação estacional. Se a série de dados for bastante longa (mais de 5 anos), haverá valores observados para 12 posições distintas ao longo do ciclo anual e os valores máximos de h e k serão $h = 6$ e $k = 5$.

3.4. Exemplo

Para ilustrar o procedimento de estimação do parâmetros de um modelo de análise harmônica serão utilizados os dados artificiais apresentados na tabela 3.1. Trata-se de uma série de 12 observações trimestrais, cobrindo um período de 3 anos, de uma variável Y. Admite-se que essa variável apresente variação estacional. Adotando o trimestre como unidade de medida do tempo, o período das variações cíclicas é $T = 4$. Verifica-se que $\omega = \frac{2\pi}{T} = \frac{\pi}{2}$.

Tabela 3.1. Série de 12 observações trimestrais da variável Y

Trimestre (t)	Y
1	41
2	37
3	41
4	62
5	35
6	36
7	50
8	64
9	44
10	44
11	41
12	57

De acordo com (3.13) tem-se $h = 2$ e $k = 1$ e o modelo de análise harmônica a ser considerado terá no máximo 4 parâmetros. Esse modelo é

$$Y_t = \alpha + \beta_1 \cos \omega t + \gamma_1 \text{ sen } \omega t + \beta_2 \cos 2\,\omega t + u_t$$

As estimativas dos parâmetros α, β_1, γ_1 e β_2 são obtidas fazendo-se a regressão de Y_t contra $X_{1t} = \cos \omega t$, $X_{2t} = \text{sen } \omega t$ e $X_{3t} = \cos 2\,\omega t$. A tabela 3.2 mostra o valor dessas variáveis para as 12 observações. Note-se que os valores se repetem ciclicamente com período $T = 4$.

Tabela 3.2. Valores de $X_{1t} = \cos \omega t$, $X_{2t} = sen\, \omega t$ e $X_{3t} = \cos 2\omega t$

t	X_{1t}	X_{2t}	X_{3t}
1	0	1	−1
2	−1	0	1
3	0	−1	−1
4	1	0	1
5	0	1	−1
6	−1	0	1
7	0	−1	−1
8	1	0	1
9	0	1	−1
10	−1	0	1
11	0	−1	−1
12	1	0	1

É interessante notar que as variáveis X_{1t}, X_{2t} e X_{3t} são centradas e que os respectivos vetores-coluna são ortogonais entre si. Para a regressão múltipla de Y_t contra X_{1t}, X_{2t} e X_{3t} obtém-se

$$\mathbf{X'X} = \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 12 \end{bmatrix}, \quad \mathbf{X'y} = \begin{bmatrix} 552 \\ 66 \\ -12 \\ 48 \end{bmatrix} \quad \mathbf{b} = \left(\mathbf{X'X}\right)^{-1} \mathbf{X'y} = \begin{bmatrix} 46 \\ 11 \\ -2 \\ 4 \end{bmatrix}$$

A equação estimada é

$$\hat{Y}_t = 46 + 11 \cos \frac{\pi}{2} t - 2 \text{ sen } \frac{\pi}{2} t + 4 \cos \pi t$$

ou

$$\hat{Y}_t = 46 + \sqrt{125} \cos\left(\frac{\pi}{2} t + 0{,}179853\right) + 4 \cos \pi\ t$$

Os valores estimados de Y nos quatro trimestres de cada ano, em ordem cronológica, são 40, 39, 44 e 61.

A figura 3.3 mostra os 12 pontos observados e a curva ajustada. Note-se que a curva foi traçada admitindo que a variável t seja contínua.

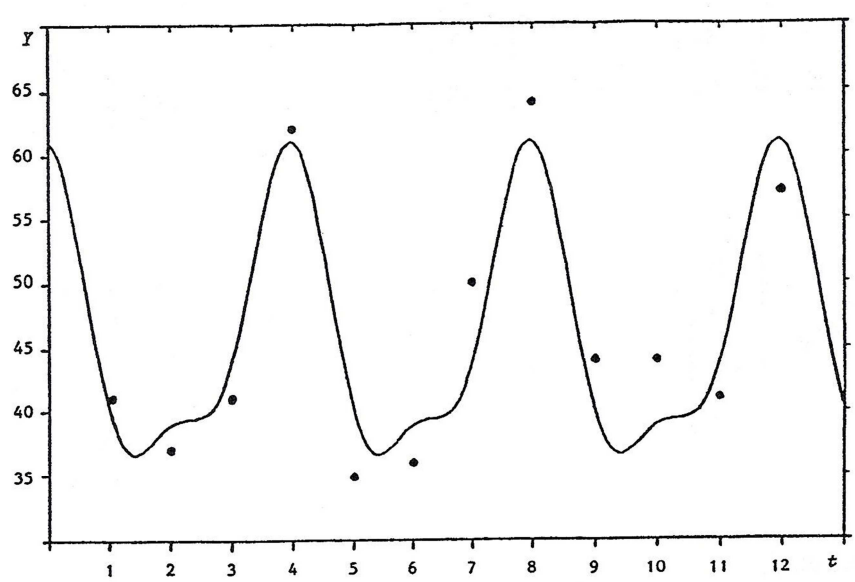


Figura 3.3. Os pontos observados e a curva ajustada

A tabela 3.3 mostra a análise de variância da regressão, separando as contribuições do primeiro componente harmônico (variáveis X_1 e X_2) e do segundo componente harmônico (variável X_3) para a Soma de Quadrados de Regressão.

Tabela 3.3. Análise de variância .

C.V.	G.L.	S.Q.	Q.M.	F
1º comp.harm.	2	750	375	$F_1 = 18{,}75$
2º comp.harm.	1	192	192	$F_2 = 9{,}60$
Regressão	3	942	314	$F_3 = 15{,}70$
Resíduo	8	160	20	
Total	11	1102		

O valor crítico de F para 3 e 8 graus de liberdade, ao nível de significância de 1%, é 7,59. Como $F_3 = 15{,}70 > 7{,}59$, rejeita-se, ao nível de significância de 1% , a hipótese de que $\beta_1 = \gamma_1 = \beta_2 = 0$. O valor de F para a contribuição do 1º componente harmônico ($F_1 = 18{,}75$) também é significativo ao nível de 1% (o respectivo valor crítico é 8,65). A

contribuição do 2^{o} componente harmônico não é significativa ao nível de 1% (valor crítico igual a 11,26), mas pode-se verificar que é significativa ao nível de 5%.

Considere-se que o pesquisador deseja prever o valor de Y para o segundo trimestre do próximo ano, quando *t* = 14. O valor estimado é $\hat{Y}_{14} = \mathbf{c}'\mathbf{b}$, onde $\mathbf{c}'$ é a linha da matriz **X** correspondente ao segundo trimestre, isto é,

$$\mathbf{c}' = \begin{bmatrix} 1 & -1 & 0 & 1 \end{bmatrix}$$

Verifica-se que $\hat{Y}_{14} = \mathbf{c}'\mathbf{b} = 39$.

A estimativa da variância do erro de previsão é

$$\left[1 + \mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c}\right]s^2 = \left(1 + \frac{1}{3}\right)20 = \frac{80}{3}$$

Então os limites do intervalo de previsão para Y_{14} , ao nível de confiança de 90% , são

$$39 \pm 1{,}86\sqrt{\frac{80}{3}} = 39 \pm 9{,}6$$

A variação estacional dos valores de Y também pode ser analisada através de uma regressão com variáveis binárias. Se a equação de regressão tiver um termo constante, para os dados da tabela 3.1 devem ser utilizadas $T - 1 = 3$ variáveis binárias. Seja $Z_{2t} = 1$ para o 2^{o} trimestre de cada ano e $Z_{2t} = 0$ para os demais trimestres. Seja $Z_{3t} = 1$ para o 3^{o} trimestre de cada ano e $Z_{3t} = 0$ para os demais trimestres. Finalmente, seja $Z_{4t} = 1$ para o 4^{o} trimestre de cada ano e $Z_{4t} = 0$ para os demais trimestres. O modelo de regressão fica

$$Y_t = \delta_1 + \delta_2 Z_{2t} + \delta_3 Z_{3t} + \delta_4 Z_{4t} + u_t$$

Pode-se verificar que, para os dados da tabela 3.1, a correspondente equação estimada é

$$\hat{Y}_t = 40 - Z_{2t} + 4 Z_{3t} + 21 Z_{4t}$$

com Soma de Quadrados Residual igual a 160, igual ao valor obtido com o modelo de análise harmônica. Os valores estimados de Y para os 4 trimestres também são os mesmos: 40, 39, 44 e 61.

A análise harmônica tem muitas aplicações em estudos econométricos. Vamos mencionar, como exemplos, apenas duas dissertações desenvolvidas na ESALQ/USP.

Okawa (1985) utilizou componentes harmônicos em uma análise das variações dos preços e das quantidades de sardinha fresca no mercado atacadista de São Paulo, nos anos de 1981 e 1982. Utilizando dados semanais ele mostrou a existência de dois tipos de variação cíclica nos preços e nas quantidades de sardinha: a estacional, com período de 52,14 semanas, e a lunar, com período de 29,53 dias.

Kassouf (1988) utilizou componentes harmônicos em modelos de análise de regressão para fazer previsões de preços na pecuária de corte do Estado de São Paulo. Foram utilizados preços mensais no período de janeiro de 1970 a dezembro de 1986. Os componentes harmônicos foram particularmente úteis para captar as variações cíclicas plurianuais, com período em torno de 6 anos, associadas às variações no estoque de matrizes.

Exercícios

1. A tabela abaixo mostra uma série temporal de valores trimestrais da variável econômica *Y*:

Trimestre	*Y*
1	29
2	18
3	13
4	22
5	27
6	21
7	11
8	11

a) Admitindo que a série apresente variações cíclicas estacionais, estime os parâmetros β_0 , β_1 e β_2 do modelo

$$Y_t = \beta_0 + \beta_1 \cos\frac{2\pi}{T}t + \beta_2\, sen\frac{2\pi}{T}t + u_t$$

onde $t = 1, 2, \ldots, 8$ indica o tempo (em trimestres) e T é o período do ciclo.

b) Coloque a equação estimada na forma

$$\hat{Y} = b_0 + C\cos\left(\frac{2\pi}{T}t - \psi\right)$$

c) Teste a hipótese $H_0 : \beta_1 = \beta_2 = 0$

d) Teste a hipótese $H_0 : \beta_2 = 0$.

2. A tabela a seguir mostra uma série de 8 valores trimestrais (2 anos) da variável *Y*. Admite-se que essa variável apresenta variações cíclicas estacionais.

Ano	Trimestre (*t*)	*Y*
1º	0	37
	1	32
	2	25
	3	30
2º	4	33
	5	32
	6	25
	7	34

a) Ajuste aos dados um modelo harmônico completo (com 3 coeficientes de regressão, além do termo constante).

b) Teste, ao nível de significância de 5%, a hipótese de que os 3 coeficientes de regressão são iguais a zero.

c) Determine o intervalo de previsão para o valor de *Y* no 3º trimestre do 3º ano, ao nível de confiança de 90%.

3. A tabela a seguir mostra uma série temporal de valores bimestrais da variável econômica *Y*.

Bimestre	Y
1	122
2	93
3	71
4	92
5	119
6	106
7	110
8	99
9	65
10	100
11	113
12	110

a) Admitindo que a série apresenta variações cíclicas estacionais, estime os parâmetros de um modelo com dois componentes harmônicos (com períodos de 6 e 3 bimestres).

b) Coloque a equação estimada na forma

$$\hat{Y} = b_0 + \sum_{i=1}^{2} C_i \cos\left(\frac{2\pi}{T_i} t - \psi_i\right) \text{ , com } C_i > 0$$

c) Verifique se a contribuição do 2º componente harmônico (com período de 3 bimestres) é estatisticamente significativa.

d) Teste a hipótese de que o verdadeiro valor dos 4 coeficientes de regressão é igual a zero ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$)

e) Determine a estimativa do desvio padrão das estimativas dos coeficientes de regressão.

f) Estime o valor de Y e determine o respectivo intervalo de previsão para $t = 13$ e para $t = 15$, ao nível de confiança de 95%.

4. A tabela ao lado mostra uma série temporal de valores trimestrais da variável econômica Y :

Trimestre	Y_t
1	28,5
2	19,0
3	12,5
4	21,0
5	27,5
6	21,0
7	11,5
8	19,0

a) Admitindo que a série apresente variações cíclicas estacionais, estime os parâmetros β_0 , β_1 e β_2 do modelo

$$Y_t = \beta_0 + \beta_1 \cos\frac{2\pi}{T} t + \beta_2 \operatorname{sen}\frac{2\pi}{T} t + u_t$$

onde $t = 1, 2, ..., 8$ indica o tempo (em trimestres) e T é o período do ciclo.

b) Coloque e equação estimada na forma $\hat{Y} = b_0 + A\cos\left(\frac{2\pi}{T} t - \psi\right)$.

c) Teste a hipótese $H_0 : \beta_1 = \beta_2 = 0$ ao nível de significância de 1%.

d) Estime o valor de Y e determine o respectivo intervalo de previsão para t = 11, ao nível de confiança de 90%.

5. Para os dados da questão anterior, estabeleça um modelo onde o efeito das variações estacionais é captado por variáveis binárias. Estime a equação e calcule a correspondente soma de quadrados dos desvios. Compare esse modelo com aquele analisado na questão anterior. Há razão para dar preferência a um dos dois modelos nesse caso? (Explique).

6. A tabela abaixo mostra uma série temporal de 6 valores da variável Y, observada de 8 em 8 meses (o intervalo entre duas observações consecutivas é 8 meses ou 2/3 de ano). Admite-se que Y apresenta variações cíclicas estacionais.

Y
36
42
12
42
36
18

a) Ajuste aos dados um modelo harmônico completo, com 2 coeficientes de regressão (além do termo constante).

b) Teste, ao nível de significância de 5%, a hipótese de que os dois coeficientes de regressão são iguais a zero.

c) Coloque a equação estimada na forma

$$\hat{Y} = a + \hat{A}\cos\left(\frac{2\pi}{T}t - \hat{\psi}\right)$$

d) Determine o intervalo de previsão para o próximo valor de Y, ao nível de confiança de 90%.

7. A tabela a seguir mostra uma série temporal de 12 valores trimestrais da variável *Y*. Admite-se que *Y* apresenta variações cíclicas estacionais.

Ano	Trimestre	*Y*
1	1º	57
	2º	14
	3º	18
	4º	55
2	1º	42
	2º	23
	3º	12
	4º	46
3	1º	42
	2º	26
	3º	21
	4º	40

a) Obtenha estimativas dos parâmetros do modelo harmônico

$$Y_t = \alpha + \delta \cos\left(\frac{2\pi}{T} t - \psi\right) + u_t \quad ,$$

onde u_t é um ruído branco.

b) Calcule o coeficiente de determinação da regressão ajustada e verifique se é estatisticamente diferente de zero ao nível de significância de 1%.

c) Determine o intervalo de previsão para *Y* no 3º trimestre do 4º ano, ao nível de confiança de 90%.

d) O modelo harmônico completo para dados trimestrais teria um termo adicional (ficando com um total de 4 parâmetros. Faça um teste (*t* ou *F*) para verificar se a contribuição desse termo adicional é significativa ao nível de 10%.

e) Descreva um modelo com variáveis binárias que levaria a um coeficiente de determinação igual ao do modelo harmônico completo. Tendo em vista o resultado obtido no item (d), na análise dessa série temporal você daria preferência a este modelo com variáveis binárias ou ao modelo dado no item (a)? (Justifique).

8. A tabela a seguir mostra uma série de 8 valores trimestrais (2 anos) da variável *Y*. Admite-se que as variações de *Y* podem ser explicadas pelo modelo

$$Y_t = \alpha + A \cos\left(\frac{2\pi}{T} t - \psi\right) + u_t \quad , \quad \text{com } T = 4 \text{ trimestres.}$$

Pressupõe-se que os u_t são erros aleatórios independentes com distribuição normal de média zero. Pressupõe-se, também, que a variância dos u_t é σ^2 para os dois primeiros trimestres de cada ano e $2\,\sigma^2$ para os dois últimos trimestres de cada ano.

Ano	Trimestre (t)	Y
1	1	8
	2	8
	3	16
	4	22
2	5	14
	6	2
	7	20
	8	14

a) Após colocar o modelo na forma de uma regressão linear múltipla com duas variáveis explanatórias, obtenha as estimativas lineares não-tendenciosas de variância mínima dos seus 3 parâmetros.

b) Estime σ^2 .

c) Teste, o nível de significância de 5%, a hipótese de que o coeficiente de $\cos\left(\frac{2\pi}{T}\right)$ é igual a zero.

d) Teste, ao nível de significância de 5%, a hipótese $H_0 : A = 0$ (que equivale a afirmar que os dois coeficientes de regressão são iguais a zero).

e) Determine a estimativa de Y no $4^{\underline{o}}$ trimestre do $3^{\underline{o}}$ ano.

f) Determine o respectivo intervalo de previsão, ao nível de confiança de 90%.

9. A tabela ao lado mostra uma série temporal de 12 valores trimestrais (3 anos) da variável *Y*.

Admite-se que as variações de Y podem ser explicadas pelo modelo

$$Y_t = \alpha + A\cos\left(\frac{2\pi}{T}t - \psi\right) + u_t \quad ,$$

Ano	Trimestre (t)	Y
1	1	5
	2	1
	3	16
	4	14
2	5	9
	6	6
	7	10
	8	18
3	9	6
	10	3
	11	14
	12	6

onde $T = 4$ trimestres e u_t são erros aleatórios independentes com distribuição normal de média zero. Admite-se, também, que a variância do erro é σ^2 para os dois primeiros trimestres de cada ano e é $2\sigma^2$ para os dois últimos trimestres de cada ano.

a) Após colocar o modelo na forma de uma regressão linear múltipla com duas variáveis explanatórias, obtenha as estimativas lineares não-tendenciosas de variância mínima dos seus 3 parâmetros.

b) Estime σ^2 .

c) Teste, ao nível de significância de 1% , a hipótese $H_0 : A = 0$ (que equivale a afirmar que os dois coeficientes de regressão são iguais a zero).

10. A tabela a seguir mostra uma série temporal de 12 valores trimestrais (3 anos) da variável *Y*.

Verifica-se que $\sum Y = 276$ e $\sum Y^2 = 6950$.

Trimestre (t)	Y
1	17
2	18
3	12
4	14
5	25
6	23
7	17
8	25
9	33
10	34
11	28
12	30

a) Estime os parâmetros do modelo

$$Y_t = \alpha + \beta\ t + A\cos\left(\frac{2\pi}{T}t - \psi\right) + u_t \qquad , \qquad \text{com } T = 4 \text{ trimestres.}$$

b) Calcule o coeficiente de determinação da regressão.

c) Verifique se a contribuição do componente harmônico é estatisticamente significativa a 1%.

d) Determine a estimativa de Y no 3º trimestre do 4º ano.

Sugestão: Faça a regressão múltipla de Y_t ou $y_t = Y_t - \overline{Y}$ contra $x_1 = 2(t - 6{,}5)$, $x_2 = \cos \frac{2\pi}{T} t$ e $x_3 = \text{sen} \frac{2\pi}{T} t$, notando que essas 3 últimas variáveis já são centradas.

11. A tabela abaixo mostra uma série temporal de valores bimestrais da variável econômica *Y*.

Bimestre	*Y*
1	103
2	119
3	121
4	105
5	85
6	79
7	83
8	95
9	105
10	101
11	85
12	71

a) Admitindo que a série apresente uma tendência linear e variações cíclicas estacionais, estime os parâmetros do modelo

$$Y_t = \alpha + \gamma\ t + \beta_1 \cos \frac{2\pi}{T} t + \beta_2\ \text{sen} \frac{2\pi}{T} t + u_t \quad ,$$

onde t = 1, 2, ..., 12 indica o tempo (em bimestres) e T é o período das variações cíclicas.

b) Determine a fase inicial e a amplitude das variações cíclicas.

c) Teste a hipótese $H_0 : \gamma = 0$, ao nível de significância de 5%.

d) Teste a hipótese $H_0 : \beta_1 = \beta_2 = 0$, ao nível de significância de 5%.

12. A tabela ao lado mostra uma série temporal de 9 valores quadrimestrais da variável *Y*. Admite-se que essa variável apresenta variações cíclicas estacionais.

a) Estabeleça um modelo de regressão para captar as variações estacionais de *Y*, utilizando variáveis binárias.

b) Estime os parâmetros do modelo.

c) Teste, ao nível de significância de 5%, a hipótese de que não há variações estacionais.

Ano	Quadrimestre	*Y*
1	1º	15
	2º	22
	3º	16
2	1º	11
	2º	18
	3º	17
3	1º	10
	2º	20
	3º	21

d) Estime um modelo harmônico completo, considerando uma variável *t* que varia de 0 a 8, e refaça o teste do item (c).

13. A tabela ao lado mostra uma série de 8 valores semestrais da variável *Y*, correspondendo a um período de 4 anos. Admite-se que essa variável apresenta apenas variações estacionais e um erro aleatório.

Semestre (*t*)	Y_t
1	15
2	21
3	13
4	18
5	12
6	23
7	16
8	18

a) Ajuste aos dados um modelo harmônico completo.

b) Teste, ao nível de significância de 1%, a hipótese de que não há variações estacionais.

c) Determine o intervalo de previsão para o valor de *Y* no segundo semestre do 5º ano, ao nível de confiança de 95%.

14. A tabela a seguir mostra uma série temporal de 16 valores trimestrais (4 anos) da variável Y_t. Admite-se que essa variável apresenta variações cíclicas estacionais, sem tendência de crescimento ou diminuição. Admite-se, também, que essa variável inclui um termo aleatório (u_t) cuja variância é σ^2.

Ano	Trimestre (*t*)	Y_t
1	0	29
	1	24
	2	27
	3	41
2	4	32
	5	28
	6	31
	7	38
3	8	34
	9	24
	10	25
	11	47
4	12	29
	13	28
	14	29
	15	38

a) Obtenha estimativas não-tendenciosas de variância mínima para os 4 coeficientes de um modelo harmônico completo.

b) Estime σ^2.

c) Teste, ao nível de significância de 1%, a hipótese de que não há variações estacionais.

d) Teste, ao nível de significância de 5%, a hipótse de que é nula a contribuição do segundo componente harmônico.

e) Determine o intervalo de previsão para o valor de Y_t no 4º trimestre do 5º ano, ao nível de confiança de 95%.

15. A tabela ao lado mostra uma série temporal de 16 valores trimestrais (4 anos) da variável Y_t. Admite-se que essa variável apresenta variações cíclicas estacionais. Admite-se, também que não há tendência de crescimento ou diminuição, mas que houve uma mudança de nível (patamar) na passagem do segundopara o terceiro ano. Admite-se, finalmente, que essa variável inclui um termo aleatório (u_t) com média zero, variância (σ^2) constante e sem autocorrelação.

Ano	Trimestre (*t*)	Y_t
1	0	19
	1	16
	2	10
	3	15
2	4	23
	5	14
	6	12
	7	11
3	8	29
	9	20
	10	14
	11	15
4	12	23
	13	20
	14	18
	15	21

   a) Obtenha estimativas não-tendenciosas de variância mínima para os 4 coeficientes de um modelo incluindo apenas um componente harmônico (seno e cosseno) e uma variável binária para captar a mudança de nível na passagem do segundo para o terceiro ano. Para facilitar as contas é aconselhável usar uma variável binária com valores –1 e 1.
   b) Estime σ^2 com base nesse modelo.
   c) Teste, ao nível de significância de 1%, a hipótese de que não há variações estacionais.
   d) Teste, ao nível de significância de 1%, a hipótese de que não há mudança de nível na passagem do segundo para o terceiro ano.
   e) Determine o intervalo de 95% de confiança para o valor esperado de Y_t no 3^{o} trimestre do 5^{o} ano.

16. A tabela ao lado mostra uma série temporal de 8 valores semestrais da variável *Y*. Admite-se que *Y* apresenta uma tendência de crescimento linear no tempo, combinada com uma variação cíclica causada pelo fato de o valor esperado de *Y* diminuir do primeiro para o segundo semestre de cada ano.

Ano	Semestre	*Y*
1	0	62
	1	61
2	2	70
	3	67
3	4	75
	5	70
4	6	77
	7	70

   a) Construa o modelo apropriado, incluindo a tendência linear e um componente harmônico, e estime seus parâmetros.

b) Verifique se a tendência linear de crescimento é estatisticamente significativa ao nível de 1% (teste bilateral).

c) Verifique se a variação entre semestres, dentro de cada ano, é estatisticamente significativa ao nível de 1% (teste unilateral).

17. Admite-se que o preço (Y_t) de um produto agrícola apresenta variações estacionais com padrão distinto em anos pares e anos ímpares. Tal fenômeno pode ser causado pelo fato de a produção apresentar oscilações bienais, sendo que o padrão de variação estacional é diferente em anos de produção relativamente abundante e em anos de produção mais escassa.

Dispomos de uma série de 16 preços médios trimestrais apresentada na tabela ao lado.

Ano	Trimestre (t)	Y_t
1	1	14
	2	6
	3	21
	4	20
2	5	11
	6	15
	7	24
	8	25
3	9	12
	10	8
	11	17
	12	24
4	13	13
	14	15
	15	18
	16	29

a) Com base nesses dados, estime um modelo harmônico completo, captando as variações cíclicas com período de 8 trimestres.

Lembre que $\operatorname{sen}\frac{\pi}{4} = \cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}$

São dados 6 dos oito elementos do vetor $\mathbf{X'y}$:

$$\mathbf{X'y} = \begin{bmatrix} \cdots \\ 10+3\sqrt{2} \\ -16-\sqrt{2} \\ 54 \\ \cdots \\ 10-3\sqrt{2} \\ 16-\sqrt{2} \\ 12 \end{bmatrix}$$

b) Obtenha a estimativa da variância do erro.

c) Teste a hipótese de que os 7 parâmetros que multiplicam variáveis definidas como senos ou cossenos são todos iguais a zero, adotando um nível de significância de 1%.

d) Obtenha o intervalo de previsão, ao nível de confiança de 95%, para o valor de Y no $4^{\underline{o}}$ trimestre do próximo ano (o ano 5).

e) Adotando um nível de significância de 10%, verifique se podemos desprezar os últimos 3 termos do modelo, passando a utilizar um modelo com apenas os dois primeiros componentes harmônicos, isto é, apenas as "ondas" com períodos 8 e 4.

18. Dispomos de uma série de 12 valores trimestrais de uma variável macroeconômica Y_t, como mostra a tabela ao lado. A série cobre um período de 3 anos, começando no primeiro trimestre de um ano. Admitimos que Y_t apresenta apenas variações estacionais e que a variância do erro de Y_t é constante (σ^2).

Trimestre (t)	Y_t
1	5
2	7
3	10
4	26
5	6
6	10
7	11
8	20
9	7
10	7
11	9
12	26

a) Apresente um modelo harmônico completo para representar as variações de Y_t em função de *t*, e estime seus parâmetros.

b) Estime σ^2.

c) Calcule o coeficiente de determinação múltipla da regressão.

d) Teste a hipótese de que não há variações estacionais, adotando um nível de significância de 1%.

e) Teste, ao nível de significância de 1%, a hipótese de que o coeficiente de $\cos \pi t$ é igual a zero.

f) Obtenha o intervalo de previsão, ao nível de confiança de 95%, para o valor de Y_t no 4º trimestre do 4º ano (quando $t = 16$).

19. Continuando a questão anterior, admita que, um ano mais tarde, foram observados os valores de Y_t ao longo do 4º ano, apresentados na tabela ao lado:

Trimestre (t)	Y_t
13	14
14	8
15	10
16	16

a) Reestime os parâmetros do modelo harmônico e reestime σ^2.

b) Obtenha o resíduo estudentizado externamente para o 1º trimestre do 4º ano e verifique se essa observação é discrepante, adotando um nível de significância de 1%.

20. Dispomos de uma série temporal de 8 valores trimestrais da variável Y_t, apresentada na tabela ao lado. Admite-se que Y_t apresenta variações cíclicas estacionais e inclui um erro aleatório (

t	Y_t	f_t
1	20,0	1
2	5,0	2
3	8,4	3
4	12,5	4
5	17,5	4
6	5,0	3
7	7,4	2
8	15,0	1

u_t) cuja variância diminui ao longo do primeiro ano e aumenta ao longo do segundo ano, de maneira que

$$V(u_t) = \frac{\sigma^2}{f_t},$$

considerando os valores de f_t na última coluna da tabela. Admite-se, também, que os erros u_t têm média igual a zero e são independentes.

a) Obtenha estimativas não-tendenciosas de variância mínima dos 4 parâmetros de um modelo harmônico completo.

b) Verifique se a contribuição do 2º componente harmônico é estatisticamente significativa ao nível de 1%.

21. O modelo $Y_t = \alpha + \beta_1 \cos\frac{\pi}{2}t + \beta_2 \text{sen}\frac{\pi}{2}t + \beta_3 \cos \pi t + u_t$ foi ajustado aos dados da tabela ao lado, por mínimos quadrados ordinários, obtendo-se $\hat{Y} = 15 + 7\cos\frac{\pi}{2}t + 4\text{sen}\frac{\pi}{2} + 3\cos\pi t$

Trimestre (t)	Y_t
1	15
2	12
3	10
4	27
5	17
6	10
7	6
8	23

$$\mathbf{X'X} = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}, \qquad \mathbf{X'y} = \begin{bmatrix} 120 \\ 28 \\ 16 \\ 24 \end{bmatrix}, \qquad \mathbf{y'y} = 2152, \quad \text{S.Q.Res.} = 20$$

a) Determine a estimativa de Y no trimestre seguinte ($t = 9$).

b) Determine o intervalo de 95% de confiança para $E(Y_9)$.

c) Determine o intervalo de previsão para Y_9, ao nível de confiança de 95%.

d) Se fosse observado o valor $Y_9 = 6$ no trimestre seguinte, isso seria considerado uma observação discrepante? (adotando um nível de significância de 5%). Qual seria o valor do resíduo estudentizado externamente para essa observação?

22. Dispomos de uma série de 12 valores trimestrais de uma variável macroeconômica Y_t, como mostra a tabela a seguir. A série cobre um período de 3 anos, começando no primeiro trimestre de um ano. Admitimos que Y_t apresenta apenas variações estacionais e que a variância do erro de Y_t é constante (σ^2).

Trimestre (t)	Y_t
1	21
2	22
3	24
4	33
5	19
6	22
7	28
8	31
9	23
10	22
11	23
12	44

a) Apresente um modelo harmônico completo para representar as variações de Y_t em função de *t*, e estime seus parâmetros.

b) Estime σ^2 e calcule o coeficiente de determinação múltipla da regressão.

c) Teste a hipótese de que não há variações estacionais, adotando um nível de significância de 1%.

d) Teste, ao nível de significância de 1%, a hipótese de que o coeficiente de $\cos \pi t$ é igual a zero.

e) Obtenha o intervalo de previsão, ao nível de confiança de 95%, para o valor de Y_t no 1º trimestre do 4º ano (quando $t = 13$).

f) Teste, ao nível de significância de 5%, a hipótese de que ocorreu uma mudança estrutural entre o 2º e o 3º ano (entre o 8º e o 9º trimestre).

Respostas

1. a) $b_0 = \overline{Y} = 19$, $b_1 = -1{,}5$, $b_2 = 8$

b) $\hat{Y} = 19 + 8{,}14 \cos\left(\frac{\pi}{2} t - 1{,}756\right)$

c) $F = 8{,}604$, significativo $(F_0 = 5{,}79)$

d) $t = 4{,}077$, significativo $(t_0 = 2{,}571)$

2. a) $Y_t = \alpha + \beta_1 \cos \frac{\pi}{2} t + \beta_2 \text{ sen } \frac{\pi}{2} t + \beta_3 \cos \pi t + u_t$

$$\mathbf{X'X} = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} 248 \\ 20 \\ 0 \\ -8 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 31 \\ 5 \\ 0 \\ -1 \end{bmatrix}$$

$\hat{Y} = 31 + 5 \cos \frac{\pi}{2} t - \cos \pi t$

b)

Análise de Variância

CV	GL	SQ	QM	*F*
Regr.	3	108	36	9
Res.	4	16	4	
Total	7	124		

O valor de F é significativo pois o valor crítico é 6,59.

c) $\mathbf{c'} = \begin{bmatrix} 1 & -1 & 0 & 1 \end{bmatrix}$ $\quad \mathbf{c'(X'X)^{-1}c} = 0{,}5$

Os limites do intervalo de previsão são

$25 \pm 2{,}132\sqrt{1{,}5 \cdot 4}$ ou $25 \pm 5{,}22$

Então $19{,}78 < Y_h < 30{,}22$

3. a) $b_0 = 100$, $b_1 = 20$, $b_2 = 0$, $b_3 = -12$, $b_4 = 0$

b) $\hat{Y} = 100 + 20 \cos \frac{\pi}{3} t + 12 \cos\left(\frac{2\pi}{3} t - \pi\right)$

c) $F = 18{,}22$, significativo $(F_0 = 4{,}74)$

d) $F = 34{,}41$, significativo $(F_0 = 4{,}12)$

e) $s(b_i) = 1{,}988 \quad (i = 1, 2, 3, 4)$

f) $\hat{Y}_{13} = 116$, $\hat{Y}_{15} = 68$, $102{,}29 < Y_{13} < 129{,}71$, $54{,}29 < Y_{15} < 81{,}71$

4. a) $\hat{Y} = 20 + 8 \text{ sen } \frac{\pi}{2} t$

b) $\hat{Y} = 20 + 8 \cos\left(\frac{\pi}{2} t - \frac{\pi}{2}\right)$

c) $F = 128$, significativo $(F_0 = 13{,}27)$

d) $\hat{Y}_{11} = 12$, $9,637 < Y_{11} < 14,363$

5. $Y = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + u$

com $Z_i = 1$ no i-ésimo trimestre e $Z_i = 0$ nos demais trimestres.

$\hat{Y} = 28 Z_1 + 20 Z_2 + 12 Z_3 + 20 Z_4$, com S.Q. Res. = 5 , igual à da questão anterior, mas com apenas 4 graus de liberdade. Nesse caso a análise harmônica é melhor.

6. a) Com $t = 0, 1, 2, 3, 4, 5$ unidades de 8 meses e $T = 1,5$ ou, alternativamente, $t = 0, 8, 16, \ldots, 40$ meses e $T = 12$, obtemos

$$\hat{Y} = 31 + 8 \cos \frac{2\pi}{T} t - 8\sqrt{3} \operatorname{sen} \frac{2\pi}{T} t$$

b) $F = 21,33$, significativo $(F_0 = 9,55)$

c) $\hat{Y} = 31 + 16 \cos\left(\frac{2\pi}{T} t + \frac{\pi}{3}\right)$ d) $26,77 < Y_h < 51,23$

7. a) Considerando uma variável t (tempo, em trimestres) variando de 1 a 12, obtemos

$$\hat{Y}_t = 33 + 13 \cos\frac{\pi}{2} t + 15 \operatorname{sen}\frac{\pi}{2} t \quad \text{ou} \quad \hat{Y}_t = 33 + \sqrt{394} \cos\left(\frac{\pi}{2} t - 0,8567\right)$$

b) $R^2 = 0,8565$; $F = 26,86$, significativo $(F_0 = 8,02)$

c) $4,4 < Y_{15} < 31,6$

d) $t = 0,5$, não - significativo $(t_o = 1,860)$

e) $Y_t = \beta_1 Z_{1t} + \beta_2 Z_{2t} + \beta_3 Z_{3t} + \beta_4 Z_{4t} + u_t$

onde Z_{1t} , Z_{2t} , Z_{3t} e Z_{4t} são variáveis binárias, com valor 0 ou 1, de maneira que $Z_{it} = 1$ apenas no i-ésimo trimestre.

Daria preferência ao modelo do item (a) porque é mais simples, deixando mais graus de liberdade no resíduo. Nesse caso o Q.M.Res. da regressão com 4 parâmetros (48) é até mesmo maior do que o Q.M.Res. da regressão com 3 parâmetros (44).

8. a) $\hat{Y}_t = 13 + 7 \cos \frac{\pi}{2} t - 3 \operatorname{sen} \frac{\pi}{2} t$

b) $s^2 = 13,6$ c) $t = 3,190$, significativo $(t_0 = 2,571)$

d) $F = 5,78$, não - significativo $(F_0 = 5,79)$

e) $\hat{Y}_{12} = 20$ f) $7,865 < Y_{12} < 32,135$

9. a) $\hat{Y}_t = 9 + 5 \cos \frac{\pi}{2} t - 3 \text{ sen } \frac{\pi}{2} t$

b) $s^2 = \frac{76}{9}$

c) $F = 8,11$, significativo $(F_0 = 8,02)$

10. a) $\hat{y}_t = 2(t - 6,5) - 3\cos\frac{\pi}{2}t + 5\,\text{sen}\,\frac{\pi}{2}t$

$\hat{Y}_t = 10 + 2t - 3 \cos \frac{\pi}{2} t + 5 \text{ sen } \frac{\pi}{2} t$

ou

$$\hat{Y}_t = 10 + 2t + \sqrt{34}\cos\left(\frac{\pi}{2}t - 2,111\right)$$

b) $R^2 = 584/602 = 0,9701$

c) $F = 41,75$, significativo $(F_0 = 8,65)$

d) $\hat{Y}_{15} = 35$

11. a) $a = 109$, $c = -2$, $b_1 = -16$ e $b_2 = 0$

b) $\hat{Y}_t = 109 - 2t + 16\cos\left(\frac{\pi}{3}t - \pi\right)$

com fase inicial $-\psi = -\pi$ e amplitude 16

c) teste $t = -4,123$, significativo $(t_0 = 2,306)$

d) $F = 26,28$, significativo $(F_0 = 4,46)$

12.a) Uma alternativa é $Y = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + u$, com $Z_1 = 1$ apenas no 1º quadrimestre de cada ano (e $Z_1 = 0$ nos outros dois quadrimestres), $Z_2 = 1$ apenas no 2º quadrimestre e $Z_3 = 1$ apenas no 3º quadrimestre.

b) $\hat{Y} = 12Z_1 + 20Z_2 + 18Z_3$

c) $F = 8,67$, significativo $(F_0 = 5,14)$

d) $\hat{Y} = 16,667 - 4,667\cos\left(\frac{2\pi t}{3}\right) + 1,155\,\text{sen}\left(\frac{2\pi t}{3}\right)$ e $F = 8,67$.

13.a) Modelo: $Y_t = \alpha + \beta \cos \pi t + u_t$

Equação estimada: $\hat{Y}_t = 17 + 3 \cos \pi t$

b) t = 3,928, significativo ($t_0 = 3{,}707$)

Rejeita-se a hipótese de que não há variações estacionais.

c) $14{,}09 < Y_{10} < 25{,}91$

14. a) $\hat{Y}_t = 31{,}5 + 1{,}5 \cos \frac{\pi}{2} t - 7{,}5 \operatorname{sen} \frac{\pi}{2} t - 2 \cos \pi t$

b) $s^2 = 9$

c) F = 19,704, significativo ($F_0 = 5{,}95$)

d) t = –2,667, signifivativo ($t_0 = 2{,}179$) ou F = 7,11, significativo ($F_0 = 4{,}75$)

e) $33{,}692 < Y_{19} < 48{,}308$

15. a) $\hat{Y} = 17{,}5 + 5 \cos \frac{\pi}{2} t + \operatorname{sen} \frac{\pi}{2} t + 2{,}5Z$

com Z = –1 para os 8 primeiros trimestres e Z = 1 para os trimestres seguintes.

b) $s^2 = \frac{20}{3} = 6{,}67$

c) F = 15,6, significativo ($F_0 = 6{,}93$)

d) t = 3,873, significativo ($t_0 = 3{,}055$)

e) $12{,}19 < E(Y_{18}) < 17{,}81$

16. a) $Y_t = \alpha + \beta t + \gamma \cos \pi t + u_t$ e $\hat{Y} = 62 + 2t + 3 \cos \pi t$

b) $s^2 = 5{,}6$, t = 5,345, significativo ($t_0 = 4{,}032$)

c) t = 3,499, significativo ($t_0 = 3{,}365$)

17. a) $\hat{Y}_t = 17 + 1{,}7803 \cos \frac{\pi t}{4} - 2{,}1768 \operatorname{sen} \frac{\pi t}{4} + 6{,}75 \cos \frac{\pi t}{2}$

$$- 3{,}75 \operatorname{sen} \frac{\pi t}{2} + 0{,}7197 \cos \frac{3\pi t}{4} + 1{,}8232 \operatorname{sen} \frac{3\pi t}{4} + 0{,}75 \cos \pi t$$

b) $s^2 = 6$

c) F = 13,81, significativo ($F_0 = 6{,}18$)

d) $15{,}08 < Y_{20} < 28{,}92$

e) F = 2,21, não-significativo ($F_0 = 2{,}92$)

18. a) $Y_t = \alpha + \beta_1 \cos\frac{\pi}{2}t + \beta_2 \operatorname{sen}\frac{\pi}{2}t + \beta_3 \cos \pi t + u_t$

$\hat{Y}_t = 12 + 8\cos\frac{\pi}{2}t - 2\operatorname{sen}\frac{\pi}{2}t + 4\cos\pi t$

b) $s^2 = 4{,}25$

c) $R^2 = 0{,}9464$

d) F = 47,06, significativo ($F_0 = 7{,}59$)

e) t = 6,72, significativo ($t_0 = 3{,}355$)

f) $18{,}51 < Y_{16} < 29{,}49$

19. a) $\hat{Y} = 12 + 7\cos\frac{\pi}{2}t - \operatorname{sen}\frac{\pi}{2}t + 3\cos\pi t$ e $s^2 = \frac{130}{12} = 10{,}833$

b) $h_{13} = 0{,}25$, $e_{13} = 6$, $s^2_{(13)} = \frac{1}{11}\left(130 - \frac{36}{0{,}75}\right) = \frac{82}{11}$

$t = \frac{6}{\sqrt{\frac{82}{11} \cdot 0{,}75}} = 2{,}538$, não-significativo ($t_0 = 3{,}106$)

20. a) $\hat{Y} = 11 + 4\cos\frac{\pi}{2}t + 5\operatorname{sen}\frac{\pi}{2}t - 2\cos\pi t$

b) t = –5,345, significativo ($t_0 = 4{,}604$)

21. a) $\hat{Y}_9 = 16$

b) $11{,}61 < E(Y_9) < 20{,}39$

c) $8{,}40 < Y_9 < 23{,}60$

d) Sim, pois está fora do intervalo de previsão. O resíduo estudentizado externamente é $t = -3{,}651$.

22. a) $Y_t = \alpha + \beta_1 \cos\frac{\pi}{2}t + \beta_2 \operatorname{sen}\frac{\pi}{2}t + \beta_3 \cos \pi t + u_t$

$\hat{Y}_t = 26 + 7\cos\frac{\pi}{2}t - 2\operatorname{sen}\frac{\pi}{2}t + 3\cos \pi t$

b) $s^2 = 15$, $R^2 = 0{,}7802$

c) F = 9,47, significativo ($F_0 = 7{,}59$)

d) t = 2,68, não-significativo ($t_0 = 3{,}355$)

e) $21 \pm 10{,}31$ ou $10{,}69 < Y_{13} < 31{,}31$

f) F = 9, significativo ($F_0 = 6{,}39$)

4. REGRESSÃO NÃO LINEAR

4.1. Introdução

Os métodos estatísticos apresentados nos capítulos anteriores se referem a modelos lineares ou a modelos que se tornam lineares por anamorfose.

Como introdução ao estudo das regressões não lineares, consideremos o modelo

$$Y_i = \alpha + \beta\rho^{X_i} + u_i, \tag{4.1}$$

onde α, β e ρ são parâmetros e os u_i são erros aleatórios independentes com distribuição normal de média zero e variância σ^2.

A relação

$$f(X) = \alpha + \beta\rho^{X}, \tag{4.2}$$

com $\alpha > 0$, $\beta < 0$ e $|\rho| < 1$, é conhecida como função de Spillman. Essa função também pode ser colocada na forma

$$f(X) = \alpha[1 - 10^{-\gamma(X+\theta)}], \tag{4.3}$$

sendo então denominada equação de Mitscherlich. Nessa forma, a função é especialmente usada no estudo da variação do crescimento de vegetais em função da quantidade de nutriente fornecida.

De (4.2) e (4.3) obtemos

$$\rho = 10^{-\gamma} \tag{4.4}$$

e

$$\beta = -\alpha 10^{-\gamma\theta} \tag{4.5}$$

Em (4.2) e (4.3), com $\gamma > 0$ e, portanto, $|\rho| < 1$, temos

$$\lim_{X \to \infty} f(X) = \alpha,$$

isto é, o valor da função se aproxima assintoticamente de α quando X tende para o infinito. Isso é ilustrado na figura 4.1, onde pode ser vista a curva correspondente à equação

$$f(X) = 8 - 4\left(\frac{1}{\sqrt{2}}\right)^X = 8[1 - 2^{-0,5(X+2)}]$$

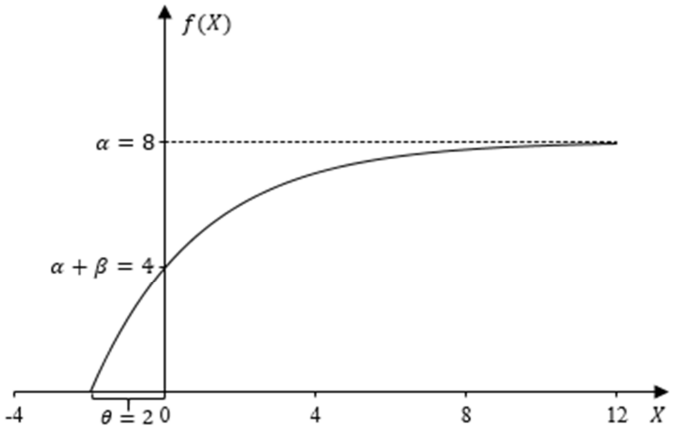


Figura 4.1. A função $f(X) = 8[1 - 2^{-0,5(X+2)}]$

Vejamos a interpretação dos parâmetros da função de Mitscherlich, quando usada para analisar a variação da produção de uma cultura em função da quantidade (X) de um nutriente. O parâmetro α corresponde a uma característica biológica da planta, ou seja, à produção máxima que pode ser alcançada com o fornecimento do nutriente em abundância. O parâmetro θ representa a quantidade de nutriente disponível no solo, sem que nada seja adicionado; note que $f(X) = 0$ se $X = -\theta$, isto é, a produção seria nula se pudéssemos retirar do solo a quantidade θ de nutriente. O parâmetro γ é denominado coeficiente de eficácia e está relacionado com a intensidade do efeito da unidade de nutriente no desenvolvimento da planta.

Analisemos, agora, algumas características matemáticas da função (4.2). Derivando, obtemos

$$\frac{d}{dX} f(X) = \beta\rho^X \ln\rho \tag{4.6}$$

Como $\beta\rho^X = f(X) - \alpha$, obtemos

$$\frac{d}{dX} f(X) = (\ln \rho)[f(X) - \alpha]$$

ou

$$\frac{d}{dX} f(X) = (-\ln \rho)[\alpha - f(X)] \tag{4.7}$$

Como $|\rho| < 1$, temos $(-\ln \rho) > 0$.

A expressão (4.7) mostra que, à medida que *X* cresce, a declividade da curva decresce porque o valor de $f(X)$ se aproxima de α. O valor $[\alpha - f(X)]$, que diminui com o aumento de *X*, é denominado "fator de contenção".

Outra característica da função (4.2) é a de que, se dermos acréscimos constantes a *X*, os valores das sucessivas variações no valor da função apresentam modificações porcentuais constantes, como demonstraremos a seguir. Temos

$$f(X_0) = \alpha + \beta\rho^{X_0},$$

$$f(X_0 + \Delta X) = \alpha + \beta\rho^{X_0 + \Delta X}$$

e

$$f(X_0 + 2\Delta X) = \alpha + \beta\rho^{X_0 + 2\Delta X}$$

Então

$$f(X_0 + \Delta X) - f(X_0) = \beta\rho^{X_0}(\rho^{\Delta X} - 1)$$

e

$$f(X_0 + 2\Delta X) - f(X_0 + \Delta X) = \beta\rho^{X_0 + \Delta X}(\rho^{\Delta X} - 1)$$

Segue-se que

$$\frac{f(X_0 + 2\Delta X) - f(X_0 + \Delta X)}{f(X_0 + \Delta X) - f(X_0)} = \rho^{\Delta X} \tag{4.8}$$

mostrando que, dado ΔX, a relação entre mudanças sucessivas em $f(X)$ é constante.

Um outro exemplo de regressão não linear é dado pela função logística

$$f(X) = \frac{\theta}{1 + \exp[-(\alpha + \beta X)]} \tag{4.9}$$

Essa função tem 3 parâmetros: θ, α e β. Usualmente $\theta > 0$ e $\beta > 0$, verificando-se que

$$\lim_{X \to \infty} f(X) = \theta \quad \text{e} \quad \lim_{X \to -\infty} f(X) = 0$$

Trata-se, nesse caso, de uma curva sigmoide compreendida entre duas assíntotas horizontais: uma reta horizontal com ordenada θ e o eixo das abscissas.

Tanto considerando um erro aditivo como considerando um erro multiplicativo, a função (4.9) dá origem a um modelo de regressão não linear.

A curva logística foi indicada para o estudo descritivo do crescimento de populações humanas por Verhulst (1845). Muitos anos mais tarde, Pearl e Reed (1920), sem conhecerem a contribuição de Verhulst (1845), obtiveram a mesma curva, que utilizaram para descrever o crescimento da população dos EUA, de 1790 a 1910, com base em dados censitários.

A partir daí, a curva logística tem sido bastante estudada quanto às suas características matemáticas e quanto ao método de estimar seus parâmetros. Ela tem sido largamente empregada para a representação de dados empíricos de crescimento de animais e vegetais, de crescimento de populações humanas e adoção de novos bens econômicos ou de novos métodos de produção.[10]

De (4.9) obtemos

$$\frac{d}{dX} f(X) = \frac{\theta\beta \exp[-(\alpha + \beta X)]}{\{1 + \exp[-(\alpha + \beta X)]\}^2} = \frac{\beta}{\theta} f(X)[\theta - f(X)] \tag{4.10}$$

Essa equação mostra que a taxa de crescimento da função logística é proporcional ao valor alcançado pela função e à diferença entre esse valor e o "nível de saturação" θ. O fator $f(X)$, cujo valor cresce com *X*, é denominado "fator de momento" e a diferença $\theta - f(x)$, cujo valor diminui à medida que *X* aumenta, é denominada "fator de contenção".

De (4.10) segue-se que

[10] Ver Lange (1967), p. 55-59, para discussão dos problemas relativos à aplicação da logística na análise do crescimento de grandezas econômicas e populações humanas.

$$\frac{1}{f(X)} \cdot \frac{d}{dX} f(X) = \frac{\beta}{\theta}[\alpha - f(X)],$$

isto é, a taxa de crescimento relativo de $f(X)$ decresce linearmente com o aumento do valor de $f(X)$.

A função (4.9) tem ponto de inflexão para a abscissa $X = -\alpha/\beta$, quando $f(X) = \theta/2$. Nesse ponto a derivada da função é máxima.

De (4.9) obtemos

$$f(X) - \frac{\theta}{2} = \frac{\theta}{2} \cdot \frac{1 - \exp(-\beta\upsilon)}{1 + \exp(-\beta\upsilon)}$$

ou

$$\Phi(\upsilon) = \frac{\theta}{2} \cdot \frac{1 - \exp(-\beta\upsilon)}{1 + \exp(-\beta\upsilon)}, \tag{4.11}$$

com $\upsilon = X + \frac{\alpha}{\beta}$ e $\Phi(\upsilon) = f(X) - \frac{\theta}{2}$. (4.12)

As transformações (4.12) correspondem, graficamente, a uma translação de eixos, de tal maneira que o ponto de inflexão da curva coincida com a origem do sistema de eixos.

É fácil verificar, em (4.11), que $\Phi(-\upsilon) = -\Phi(\upsilon)$. Isso mostra que a curva logística é radialmente simétrica em torno do seu ponto de inflexão.

Como ilustração, apresentamos, na figura 4.2, a curva correspondente à função $f(X) = 6/(1 + 2^{2-X})$, cujo ponto de inflexão tem coordenadas 2 e 3.

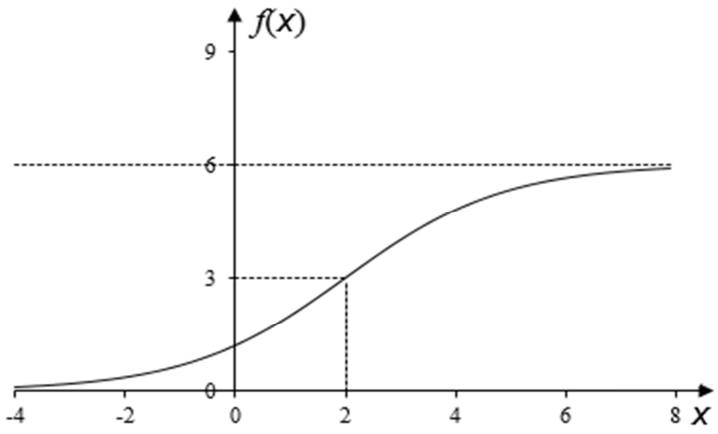


Figura 4.2. A função $f(X) = 6/(1 + 2^{2-X})$

Como terceiro exemplo de função que dá origem a uma regressão não linear, vamos considerar a função de Gompertz, que pode ser definida como

$$f(X) = \exp\{\alpha + \beta\rho^X\} \tag{4.13}$$

onde α, β e ρ são parâmetros, $\beta < 0$ e $0 < \rho < 1$.

De (4.13) obtemos

$$\frac{d}{dX} f(X) = \omega f(X)[\alpha - \ln f(X)], \tag{4.14}$$

onde $\omega = -\ln\rho$,

$$\lim_{X\to\infty} f(X) = e^{\alpha}$$

e

$$\lim_{X\to-\infty} f(X) = 0$$

A função de Gompertz é monotonicamente crescente e fica entre duas assíntotas horizontais: o eixo das abscissas e a reta de ordenada e^{α}.

A equação (4.14) mostra que a taxa de crescimento da função de Gompertz é proporcional ao valor alcançado pela função e à diferença entre o logaritmo desse valor e o logaritmo da ordenada da assíntota superior. Portanto, na função de Gompertz, da mesma maneira que na logística, se reconhece um "fator de momento", igual a $f(X)$, e um "fator de contenção", neste caso igual a $[\alpha - \ln f(X)]$.

De (4.14) segue-se que

$$\frac{1}{f(X)} \cdot \frac{d}{dX} f(X) = \omega[\alpha - \ln f(X)], \tag{4.15}$$

isto é, a taxa de crescimento relativo de *f*(*X*) decresce linearmente com o logaritmo de $f(X)$.

Pode-se verificar que a função de Gompertz tem ponto de inflexão quando $X = -\ln(-\beta)/\ln\rho$ e $f(X) = e^{\alpha-1}$.

De (4.13) obtemos

$$\ln f(X) = \alpha + \beta\rho^X$$

Comparando essa relação com (4.2), concluímos que o logaritmo do valor de uma função de Gompertz se comporta como uma função de Spillman. Então, de acordo com a expressão (4.8), se dermos acréscimos constantes a *X*, os valores das sucessivas variações causadas em $\ln f(X)$ apresentam modificações porcentuais constantes.

De (4.13), considerando um erro multiplicativo, obtemos o modelo

$$Y_i = (\exp\{\alpha + \beta\rho^{X_i}\})\varepsilon_i, \tag{4.16}$$

Aplicando logaritmos naturais, obtemos

$$\ln Y_i = \alpha + \beta\rho^{X_i} + u_i \tag{4.17}$$

onde $u_i = \ln\varepsilon_i$. Se admitirmos que os u_i são erros independentes com distribuição normal de média zero e variância σ^2, o modelo (4.17) equivale ao modelo (4.1).

4.2. O limite inferior de Cramér-Rao e as propriedades assintóticas dos estimadores de máxima verossimilhança

Para os modelos de regressão linear, sendo válidas as pressuposições usuais sobre os erros, o teorema de Gauss-Markov nos garante que o método de mínimos quadrados fornece estimadores lineares não-tendenciosos de variância mínima. Mas, esse teorema não se aplica a modelos não lineares. É necessário recorrer, então, a um teorema que, em condições bastante gerais, garante que os estimadores de máxima verossimilhança são consistentes e assintoticamente eficientes.

Vejamos, preliminarmente, como se determina o limite inferior de Cramér-Rao para a matriz de variâncias e covariâncias das estimativas dos parâmetros.

Consideremos uma amostra aleatória simples com *n* observações (X_j, com *j* = 1, ..., *n*) de uma variável cuja distribuição é caracterizada por um vetor-coluna $\boldsymbol{\alpha}$ de parâmetros (α_i, *i* = 1, ..., *p*). Se a função de densidade de *X* é $f(X)$, a função de verossimilhança para essa amostra é

$$L(X_1,\ldots,X_n;\alpha_1,\ldots,\alpha_p) = f(X_1)\cdot f(X_2)\cdot\ldots\cdot f(X_n)$$

Seja **W** a matriz $p\times p$, simétrica, cujos elementos são

$$w_{hi} = -E\left(\frac{\partial^2 \ln L}{\partial\alpha_h \partial\alpha_i}\right) \tag{4.18}$$

para h = 1, ..., p e i = 1, ..., p.

Essa matriz é denominada *matriz de informação*.

Seja **a** o vetor-coluna das estimativas dos parâmetros. Admitimos que $E(\mathbf{a}) = \boldsymbol{\alpha}$, isto é, que $a_1,...,a_p$, são estimadores não-tendenciosos de $\alpha_1,\ldots,\ \alpha_p$. Seja **V** a matriz de variâncias e covariâncias de **a**. Se a função de densidade obedecer a certas condições de regularidade relativas à integração e diferenciação, pode-se demonstrar que a diferença $\mathbf{V}-\mathbf{W}^{-1}$ é igual a uma matriz semidefinida positiva, ou seja, a matriz de variâncias e covariâncias das estimativas não-tendenciosas dos parâmetros excede a inversa da matriz de informação em uma matriz semidefinida positiva[11].

Para exemplificar, consideremos uma variável X com distribuição normal de média $\mu = E(X)$ e variância $\sigma^2 = E(X-\mu)^2$. A função de verossimilhança para uma amostra aleatória com n observações da variável ($X_1, X_2,\ldots, X_n$) é

$$L(X_1,\ldots,X_n;\mu,\sigma^2) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{\sum(X_i-\mu)^2}{2\sigma^2}\right\}$$

Então

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum(X_i-\mu)^2$$

Segue-se

[11] A demonstração pode ser encontrada em Theil (1971), p. 384-387. Devemos ressaltar que tais condições de regularidade são satisfeitas pela distribuição normal.

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum(X_i - \mu) \tag{4.19}$$

e

$$\frac{\partial^2 \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(X_i - \mu)^2 \tag{4.20}$$

Igualando a zero as expressões (4.19) e (4.20), obtemos um sistema de equações cuja solução é constituída pelos estimadores de máxima verossimilhança de μ e σ^2, que são

$$\overline{X} = \frac{1}{n}\sum X_i \tag{4.21}$$

e

$$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \overline{X})^2 \tag{4.22}$$

De (4.19) e (4.20), obtemos

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2},$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum(X_i - \mu)$$

e

$$\frac{\partial^2 \ln L}{(\partial \sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum(X_i - \mu)^2$$

De acordo com (4.18), segue-se que a matriz de informação é

$$\mathbf{W} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Então

$$\mathbf{W}^{-1} = \begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{bmatrix} \tag{4.23}$$

Sabemos que $\overline{X} = (\sum X_i)/n$ e $s^2 = \sum (X_i - \overline{X})^2/(n-1)$ são estimadores não-tendenciosos de μ e σ^2. Pode-se demonstrar que a matriz de variâncias e covariâncias desses estimadores é

$$\mathbf{V} = \begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{2\sigma^4}{n-1} \end{bmatrix} \tag{4.24}$$

Comparando (4.23) e (4.24) verifica-se que a variância de $\overline{X}$ é igual ao respectivo limite inferior de Cramér-Rao. Entretanto, a variância de s^2 é maior do que o elemento correspondente na matriz $\mathbf{W}^{-1}$. Pode-se demonstrar que não existe estimador não-tendencioso de σ^2 com variância inferior a $2\sigma^4/(n-1)$; este é, portanto, um caso em que o limite inferior de Cramér-Rao não é atingido, isto é, não existe estimador não-tendencioso com variância igual ao limite dado pelo elemento correspondente na inversa da matriz de informação.

Demonstra-se, em condições bastante gerais, que, se $\hat{\boldsymbol{\alpha}}$ é o vetor dos estimadores de máxima verossimilhança ($\hat{\alpha}_i$, i = 1, ..., p) do vetor $\boldsymbol{\alpha}$ de parâmetros (α_i, i = 1, ..., p), então $\hat{\boldsymbol{\alpha}}$ tem distribuição assintoticamente normal multidimensional com vetor de médias $\boldsymbol{\alpha}$ e matriz de variâncias e covariâncias igual à inversa ($\mathbf{W}^{-1}$) da matriz de informação, isto é, os estimadores de máxima verossimilhança são consistentes e assintoticamente eficientes[12].

4.3. Determinação das estimativas dos parâmetros

Para ilustrar o procedimento de obtenção das estimativas dos parâmetros e das respectivas variâncias vamos considerar um modelo de regressão não linear com apenas 2 parâmetros:

[12] Ver Theil (1971), PP. 392-396.

$$Y_j = \beta\rho^{X_j} + u_j \tag{4.25}$$

Admite-se que os erros u_j (com j = 1, ..., n) não são correlacionados entre si e têm distribuição normal com média zero e variância σ^2.

Note-se que o modelo (4.25) corresponde ao modelo (4.1), excluindo-se o termo constante.

Se, ao invés de um erro aditivo, fosse considerado um erro multiplicativo, teríamos

$$Y_j = \beta\rho^{X_j}\varepsilon_j$$

Esse é um modelo linearizável, pois aplicando logaritmos obtemos:

$$\ln Y_j = \ln\beta + (\ln\rho)X_j + \ln\varepsilon_j,$$

mostrando que, se $\ln\varepsilon_j$ for um erro com as propriedades usuais, a análise estatística apropriada consiste em fazer uma regressão linear simples de $\ln Y_j$ contra X_j.

Mas o modelo (4.25) não é linearizável. Vejamos como obter as estimativas de máxima verossimilhança dos parâmetros desse modelo.

Admitindo que dispomos dos valores de X_j e Y_j em uma amostra com n observações, e tendo em vista que os erros u_j são independentes e têm distribuição normal com $E(u_j) = 0$ e $E(u_j^2) = \sigma^2$, a verossimilhança da amostra é

$$L = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j - \beta\rho^{X_j})^2\right\} \tag{4.26}$$

Nas equações a seguir, por simplicidade, vamos omitir o índice j nas variáveis X_j e Y_j.

De (4.26) segue-se que

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum(Y - \beta\rho^X)^2$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\sum(Y - \beta\rho^X)\rho^X \tag{4.27}$$

$$\frac{\partial \ln L}{\partial \rho} = \frac{\beta}{\sigma^2}\sum(Y - \beta\rho^X)X\rho^{X-1} \tag{4.28}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma(Y - \beta\rho^X)^2 \tag{4.29}$$

As estimativas de máxima verossimilhança de β e ρ, indicadas por *b* e *r*, são obtidas resolvendo o sistema de equações obtido de (4.27) e (4.28) lembrando a condição de 1ª ordem para ponto de máximo:

$$\begin{cases} \Sigma(Y - br^X)r^X = 0 & (4.30) \\ \Sigma(Y - br^X)Xr^{X-1} = 0 & (4.31) \end{cases}$$

Cabe ressaltar que o método de mínimos quadrados leva a esse mesmo sistema de equações.

De acordo com o método de Newton (ou Newton-Raphson), os primeiros membros de (4.30) e (4.31) são considerados como funções de *b* e *r* e desenvolvidos pela série de Taylor (até o termo envolvendo a primeira derivada), obtendo-se

$$\begin{cases} \Sigma(Y - b_o r_o^X)r_o^X - \Sigma r_o^{2X}\Delta b + \Sigma[-b_o X r_o^{2X-1} + (Y - b_o r_o^X)Xr_o^{X-1}]\Delta r = 0 \\ \Sigma(Y - b_o r_o^X)Xr_o^{X-1} - \Sigma Xr_o^{2X-1}\Delta b + \Sigma[-b_o X^2 r_o^{2X-2} + (Y - b_o r_o^X)X(X-1)r_o^{X-2}]\Delta r = 0 \end{cases}$$

ou

$$\begin{cases} \Sigma r_o^{2X}\Delta b + \Sigma[b_o X r_o^{2X-1} - (Y - b_o r_o^X)Xr_o^{X-1}]\Delta r = \Sigma(y - b_o r_o^X)r_o^X \\ \Sigma Xr_o^{2X-1}\Delta b + \Sigma[b_o X^2 r_o^{2X-2} - (Y - b_o r_o^X)X(X-1)r_o^{X-2}]\Delta r = \Sigma(Y - b_o r_o^X)Xr_o^{X-1} \end{cases} \tag{4.32}$$

Inicialmente é preciso obter estimativas preliminares de β e ρ, indicadas por b_o e r_o. Uma maneira de obter essas estimativas preliminares consiste em considerar a relação funcional entre *Y* e *X*, sem o erro, escolher dois dos *n* pontos da amostra (*j* = *k* e *j* = *h*) e resolver o sistema de duas equações com duas incógnitas dado por

$$\begin{cases} Y_k = b_o r_o^{X_k} \\ Y_h = b_o r_o^{X_h} \end{cases} \tag{4.33}$$

Dispondo de estimativas preliminares dos parâmetros (b_o e r_o) e dos valores de X_j e Y_j (com *j* = 1, ..., *n*), (4.32) é um sistema de equações lineares em Δb e Δr. Uma vez obtidos os valores de Δb e Δr, as estimativas preliminares são corrigidas, calculando-se

$$b_o + \Delta b$$

e $r_o + \Delta r$

Esses valores são considerados como estimativas preliminares, recalculando-se os coeficientes do sistema (4.32) e obtendo novas correções Δb e Δr. Admitindo que o processo seja convergente, o ciclo de cálculos é repetido até que as correções Δb e Δr sejam consideradas desprezíveis.

Há um procedimento um pouco mais simples, para obter as estimativas dos parâmetros do modelo (4.25), denominado *método de Gauss-Newton*. Sendo *b* e *r* as estimativas de β e ρ, a equação a ser estimada é

$$\hat{Y} = br^X \tag{4.34}$$

Considerando $\hat{Y}$ como uma função de *r* e desenvolvendo-a pela série de Taylor (até o termo que envolve a primeira derivada), obtemos

$$\hat{Y} = br_o^{\ X} + Xr_o^{X-1}b\Delta r\,, \tag{4.35}$$

onde r_o é uma estimativa preliminar de ρ.

De acordo com (4.35), fazemos a regressão múltipla de *Y* contra r_o^X e Xr_o^{X-1}, obtendo as estimativas *b* e

$$c = b\Delta r \tag{4.36}$$

O respectivo sistema de equações normais é

$$\begin{bmatrix} \sum r_o^{2X} & \sum Xr_o^{2X-1} \\ \sum Xr_o^{2X-1} & \sum X^2 r_o^{2X-2} \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} \sum Yr_o^X \\ \sum XYr_o^{X-1} \end{bmatrix} \tag{4.37}$$

Após obter os valores de *b* e *c*, de acordo com (4.36) calculamos

$$\Delta r = \frac{c}{b}$$

Se Δr não for desprezível, obtemos o valor corrigido

$$r = r_o + \Delta r$$

e os cálculos indicados em (4.37) são refeitos utilizando esse valor corrigido. Admitindo que o processo seja convergente, o ciclo de cálculos é repetido até que Δr seja considerado desprezível. Na última iteração, já com as estimativas definitivas, temos a matriz

$$\mathbf{Q} = \begin{bmatrix} \sum r^{2X} & \sum Xr^{2X-1} \\ \sum Xr^{2X-1} & \sum X^2 r^{2X-2} \end{bmatrix} \tag{4.38}$$

A seguir vamos mostrar que (4.37) corresponde a uma simplificação do sistema (4.32), desprezando, nos primeiros membros, os termos que são somas ponderadas dos desvios. Desprezando tais termos o sistema fica

$$\begin{cases} \sum r_o^{2X}(b-b_o) + \sum Xr_o^{2X-1}(b_o \Delta r) = \sum (Y - b_o r_o^X) r_o^X \\ \sum Xr_o^{2X-1}(b-b_o) + \sum X^2 r_o^{2X-2}(b_o \Delta r) = \sum (Y - b_o r_o^X) Xr_o^{X-1} \end{cases}$$

Simplificando, obtemos

$$\begin{cases} (\sum r_o^{2X})b + (\sum Xr_o^{2X-1})b_o \Delta r = \sum Yr_o^X \\ (\sum Xr_o^{2X-1})b + (\sum X^2 r_o^{2X-2})b_o \Delta r = \sum YXr_o^{X-1} \end{cases} \tag{4.39}$$

que é equivalente a (4.37).

Podemos dizer que o método de Gauss-Newton corresponde a uma simplificação do método de Newton, desprezando, no sistema de equações em Δb e Δr, os termos que são somas ponderadas dos desvios. Isso faz com que, no método de Gauss-Newton o número de iterações necessárias para obter um Δr desprezível seja, em geral, maior do que no método de Newton. Por outro lado, os cálculos exigidos em cada iteração são mais simples no método de Gauss-Newton.

4.4. Determinação da matriz de variâncias e covariâncias assintóticas das estimativas dos parâmetros

De acordo com o teorema apresentado na seção 4.2, a matriz de variâncias e covariâncias assintóticas dos estimadores de máxima verossimilhança dos parâmetros do modelo (4.25) são iguais ao limite inferior de Cramér-Rao.

De (4.27), (4.28) e (4.29) segue-se que

$$\frac{\partial^2 \ln L}{\partial \beta^2} = -\frac{1}{\sigma^2} \sum \rho^{2X}$$

$$\frac{\partial^2 \ln L}{\partial \rho^2} = -\frac{\beta}{\sigma^2} \sum [\beta X^2 \rho^{2X-2} - (Y - \beta \rho^X) X (X-1) \rho^{X-2}]$$

$$\frac{\partial^2 \ln L}{\partial\beta\partial\rho} = -\frac{1}{\sigma^2}\sum[\beta X\rho^{2X-1} - (Y-\beta\rho^X)X\rho^{X-1}]$$

$$\frac{\partial^2 \ln L}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum(Y-\beta\rho^X)^2$$

$$\frac{\partial^2 \ln L}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}\sum(Y-\beta\rho^X)\rho^X$$

$$\frac{\partial^2 \ln L}{\partial\rho\partial\sigma^2} = -\frac{\beta}{\sigma^4}\sum(Y-\beta\rho^X)X\rho^{X-1}$$

Como $E(Y) = \beta\rho^X$, verifica-se que a matriz de informação é

$$\mathbf{W} = \begin{bmatrix} \frac{1}{\sigma^2}\sum\rho^{2X} & \frac{\beta}{\sigma^2}\sum X\rho^{2X-1} & 0 \\ \frac{\beta}{\sigma^2}\sum X\rho^{2X-1} & \frac{\beta^2}{\sigma^2}\sum X^2\rho^{2X-2} & 0 \\ 0 & 0 & \frac{n}{2\sigma^4} \end{bmatrix} \tag{4.40}$$

e a matriz das estimativas das variâncias e covariância assintóticas de *b* e *r* é

$$\hat{\mathbf{V}} = \mathbf{G}^{-1}s^2, \tag{4.41}$$

onde

$$\mathbf{G} = \begin{bmatrix} \sum r^{2X} & b\sum Xr^{2X-1} \\ b\sum Xr^{2X-1} & b^2\sum X^2r^{2X-2} \end{bmatrix} \tag{4.42}$$

e

$$s^2 = \frac{1}{n-2}\sum(Y-br^X)^2 \tag{4.43}$$

Nessa estimativa de σ^2 a soma de quadrados dos desvios é dividida por $n-2$ (em geral $n-p$, onde p é o número de parâmetros do modelo), por analogia com os modelos lineares.

Fazendo

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & b \end{bmatrix}, \tag{4.44}$$

e sendo **Q** a matriz definida em (4.38), verifica-se que

$$\mathbf{G} = \mathbf{MQM} \tag{4.45}$$

Se for utilizado o método de Gauss-Newton, as relações (4.41) e (4.45) mostram como a matriz das estimativas das variâncias e covariância assintóticas de *b* e *r* pode ser obtida a partir da matriz **Q**.

4.5. A distribuição assintótica de uma função de estimadores

As propriedades estatísticas de uma combinação *linear* de variáveis aleatórias são bem conhecidas. Nesta seção vamos examinar algumas propriedades que são válidas para funções não lineares de variáveis aleatórias (ou, especificamente, de estimadores de parâmetros). Trata-se, obviamente, de propriedades muito úteis ao trabalhar com modelos de regressão não linear.

Seja *a* um estimador não-tendencioso do parâmetro α de uma população, isto é,

$$E(a) = \alpha \tag{4.46}$$

Seja $V(a)$ a variância desse estimador. Admitamos que

$$\lim_{n\to\infty} V(a) = 0 \tag{4.47}$$

De (4.46) e (4.47) segue-se que o estimador *a* converge em média quadrática para α. Então

$$\operatorname{plim} a = \alpha, \tag{4.48}$$

isto é, *a* é um estimador consistente de α.

Seja $\phi(a)$ uma função com derivadas de primeira e segunda ordem contínuas numa vizinhança de $a = \alpha$. Consideremos que a função $\phi(a)$ não depende do tamanho (*n*) da amostra utilizada para obter o valor da estimativa de α. Nestas condições pode-se demonstrar que $b = \phi(a)$ é um estimador consistente de $\beta = \phi(a)$, com distribuição assintoticamente normal com variância

$$V(b) = [\phi_a(\alpha)]^2 V(a), \tag{4.49}$$

onde $\phi_a(\alpha)$ representa o valor de $\phi_a(a) = \dfrac{d\phi(a)}{da}$ para $a = \alpha$.

A normalidade da distribuição de $\phi(a)$, quando $n \to \infty$, se deve ao teorema do limite central. A seguir, vamos mostrar que $\operatorname{plim} b = \beta$ e que a variância assintótica de *b* é dada por (4.49).

De acordo com a série de Taylor, temos

$$b = \phi(a) = \phi(\alpha) + [\phi_a(\alpha)](a - \alpha) + Q\,, \tag{4.50}$$

onde *Q* é o resto.

Desde que $\operatorname{plim} a = \alpha$, segue-se que $\operatorname{plim} b = \phi(\alpha) = \beta$.

Para *n* suficientemente grande podemos desprezar o resto em (4.50), obtendo

$$(b - \beta)^2 = [\phi_a(\alpha)]^2 (a - \alpha)^2$$

Então, a variância assintótica de *b* é $V(b) = [\phi_a(\alpha)]^2 V(a)$, c.q.d.

Conhecida a estimativa (*a*) de α e a estimativa da respectiva variância, $\hat{V}(a)$, a estimativa da variância assintótica de $b = \phi(a)$ é dada por

$$\hat{V}(b) = [\phi_a(a)]^2 \hat{V}(a) \tag{4.51}$$

Pode-se demonstrar que, se *a* é uma variável aleatória com distribuição assintoticamente normal com média α e variância *V*(*a*), e se $b = \phi(a)$ é uma função com derivada de segunda ordem contínua numa vizinhança de $a = \alpha$, então $b = \phi(a)$ tem distribuição assintoticamente normal com média $\beta = \phi(\alpha)$ e variância dada por (4.49). Note que não é necessário que *a* seja um estimador não-tendencioso; basta que seja consistente.

Podemos adotar, como regra prática para obter o estimador (4.51), o seguinte procedimento:

a) Diferenciamos $b = \phi(a)$, obtendo

$$db = \phi_a(a)\, da \tag{4.52}$$

b) Elevamos os dois membros de (4.52) ao quadrado e substituímos os quadrados das diferenças das variáveis pelas respectivas estimativas de variância. O resultado é a expressão (4.51).

Esse procedimento pode ser estendido para o caso de funções de duas ou mais variáveis. Assim, sejam $\hat{V}(a_1)$, $\hat{V}(a_2)$ e $\hat{\text{cov}}(a_1, a_2)$ as estimativas das variâncias e da covariância dos estimadores consistentes (a_1, a_2) dos parâmetros α_1 e α_2. Então, $b = \phi(a_1, a_2)$ é um estimador consistente de $\beta = \phi(\alpha_1, \alpha_2)$ e a estimativa de sua variância assintótica pode ser obtida como segue:

a) De $b = \phi(a_1, a_2)$, diferenciando, obtemos

$$db = \phi_1 \, da_1 + \phi_2 \, da_2, \tag{4.53}$$

onde

$$\phi_h = \frac{\partial \phi(a_1, a_2)}{\partial a_h} \qquad h = 1, 2$$

b) Elevando ao quadrado os dois membros de (4.53), substituindo os quadrados dos diferenciais das variáveis pelas estimativas das respectivas variâncias e o produto dos diferenciais de duas variáveis pela estimativa da respectiva covariância, obtemos

$$\hat{V}(b) = \phi_1^2 \hat{V}(a_1) + \phi_2^2 \hat{V}(a_2) + 2\phi_1\phi_2 \, \hat{\text{cov}}(a_1, a_2) \tag{4.54}$$

Generalizando, seja $\boldsymbol{\alpha}$ um vetor-coluna com k parâmetros $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ cujas estimativas $(a_1, a_2, \ldots, a_k)$ constituem o vetor-coluna **a**. Se **a** tem distribuição k-dimensional assintoticamente normal com vetor de médias $\boldsymbol{\alpha}$ e matriz de variâncias e covariâncias **W**, e se $b = \phi(a_1, a_2, \ldots, a_k)$ é uma função com derivados de segunda ordem contínuas numa vizinhança de $\mathbf{a} = \boldsymbol{\alpha}$, então $b = \phi(a_1, a_2, \ldots, a_k)$ é um estimador consistente de $\beta = \phi(\alpha_1, \alpha_2, \ldots, \alpha_k)$ com distribuição assintoticamente normal de média β e variância

$$V(b) = \boldsymbol{\phi}' \mathbf{W} \boldsymbol{\phi} \tag{4.55}$$

onde $\boldsymbol{\phi}$ é o vetor-coluna cujos elementos são os valores de

$$\phi_i = \frac{\partial \phi(a_1, a_2, \ldots, a_k)}{\partial a_i}, \text{ com } i = 1, \ldots, k, \text{ para } \mathbf{a} = \boldsymbol{\alpha}.$$

Se $c = \psi(a_1, a_2, \ldots, a_k)$ é outra função com derivadas de segunda ordem contínuas numa vizinhança de $\mathbf{a} = \boldsymbol{\alpha}$, então a covariância assintótica de b e c é

$$\text{cov}(b, c) = \boldsymbol{\phi}' \mathbf{W} \boldsymbol{\psi}, \tag{4.56}$$

onde $\boldsymbol{\psi}$ é o vetor-coluna com os valores de $\psi_i = \dfrac{\partial \psi(a_1, a_2, \ldots, a_k)}{\partial a_i}$, com $i = 1, \ldots, k$, para $\mathbf{a} = \boldsymbol{\alpha}$.

Como exemplo de aplicação da técnica de determinação da variância de uma transformação não linear de uma variável aleatória, vamos examinar um método de obter boas estimativas preliminares dos parâmetros β e ρ do modelo (4.25).

Ignorando o erro u_j e aplicando logaritmos à relação funcional entre *Y* e *X*, obtemos

$$Z = \ln Y = \ln \beta + (\ln \rho) X$$

Isso indica que estimativas preliminares de β e ρ podem ser obtidas fazendo uma regressão linear simples de *Z* contra *X*. No modelo (4.25), a variância de *Y*, dado *X*, é constante. Como se comporta a variância de $Z = \ln Y$?

Temos

$$dZ = \frac{dY}{Y},$$

$$(dZ)^2 = \frac{(dY)^2}{Y^2}$$

e

$$V(Z) = \frac{V(Y)}{Y^2}$$

Com *V*(*Y*) constante, *V*(*Z*) é inversamente proporcional a Y^2. Portanto, para obter boas estimativas preliminares de β e ρ, ao fazer a regressão de *Z* contra *X*, devemos usar mínimos quadrados ponderados, usando Y^2 como fator de ponderação.

Exercícios

1. Considere o modelo

$$Y_i = \beta^{X_i} + u_i$$

onde os u_i são erros aleatórios independentes de média zero e variância σ^2.

a) Dado um conjunto de *n* pares de valores (X_i, Y_i), mostre (deduzindo) como se pode obter uma estimativa de mínimos quadrados de β.

b) Pressupondo que os u_i têm distribuição normal, quais são as propriedades desse estimador?

c) Obtenha as estimativas de β e do respectivo desvio padrão com base na seguinte amostra

X_i	Y_i
0	1,000
1	0,480
2	0,255
3	0,145

Experimente $b_o = 0{,}5$ como estimativa preliminar de β.

2. Considere o seguinte modelo de regressão não linear

$$Y_i = X_i^{\beta} + u_i \qquad (i = 1, ..., n)$$

onde β é um parâmetro e os u_i são erros aleatórios independentes com distribuição normal, média zero e variância σ^2.

a) Dada uma estimativa preliminar (b_o) de β, mostre como será obtida a correção Δb.

b) Descreva uma maneira de obter uma estimativa preliminar de β.

c) Com base nos dados da tabela ao lado (3 observações), calcule Δb para $b_o = 0{,}5$.

d) Idem, para $b_o = \frac{1}{3}$.

X	Y
8	6
64	6
512	7

e) Determine a estimativa do desvio padrão assintótico da estimativa de β.

3. Considere o modelo

$$Y_i = \beta\rho^{X_i} + u_i$$

onde os u_i são erros aleatórios independentes de média zero e variância σ^2.

a) Pressupondo que os u_i têm distribuição normal, quais são as propriedades das estimativas de β e ρ obtidas pelo método de mínimos quadrados?

b) Obtenha as estimativas de β e ρ e dos respectivos desvios padrões com base na seguinte amostra:

X_i	Y_i
0	8,00
1	3,95
2	2,20
3	0,80

Experimente $r_o = 0{,}5$ como estimativa preliminar de ρ.

4. Considere o seguinte modelo de regressão não linear

$$Y_i = \frac{\alpha X_i}{X_i + \beta} + u_i \qquad (i = 1, 2, ..., n)$$

onde α e β são parâmetros e u_i são erros aleatórios independentes com distribuição normal, média zero e variância σ^2.

a) Dada uma amostra de valores de *X* e *Y* , e admitindo que se tenham estimativas preliminares de α e β (ou só de β), mostre como são obtidas as estimativas dos parâmetros (*a* e *b*) pelo método de Gauss-Newton.

b) São dados os valores de *X* e *Y* em uma amostra com 4 observações. Dada as estimativas preliminares $a_o = 120$ e $b_o = 3$, mostre que a correção Δb é igual a zero.

X	*Y*
1	35
3	60
5	67
9	95

c) Determine, com base nessa amostra, as estimativas dos desvios padrões assintóticos de *a* e *b*.

d) Descreva uma maneira de obter boas estimativas preliminares de α e β.

5. Colocando um termômetro, inicialmente a 32,9 ^{o}C, em um ambiente com temperatura inferior, foram lidas, de minuto em minuto, as seguintes temperaturas:

Tempo Em minutos (X)	Temperatura em graus centígrados (Y)
0	32,9
1	29,1
2	26,9
3	25,4
4	24,5
5	23,8

Pode-se demonstrar, com base em princípios da termologia, que a temperatura do termômetro (Y) decresce no tempo de acordo com a função $f(X) = \alpha + \beta\rho^{X}$, onde α é a temperatura do ambiente, que se admite constante.

Admitindo que os valores de Y_i incluam erros aditivos independentes, com distribuição normal de média zero e variância constante, obtenha estimativas dos parâmetros α, β e ρ e dos respectivos desvios padrões assintóticos.

Observação: para fazer esse exercício e os dois seguintes, é recomendável o uso de computador.

6. Os dados a seguir se referem ao crescimento da altura (Y), em decímetros, de certa espécie vegetal, em função do número (X) de anos após o plantio da muda no campo.

X	Y
0	2
1	8
2	27
3	43
4	60
5	76

Admite-se que Y seja relacionado com X de acordo com o modelo

$$\ln Y_i = \alpha + \beta\rho^{X_i} + u_i$$

onde os u_i são erros independentes com distribuição normal de média zero e variância σ^2.

a) Obtenha as estimativas dos parâmetros α, β e ρ de acordo com o método de mínimos quadrados.

b) Qual é a estimativa da ordenada da assíntota superior de *Y*?

c) Calcule a soma dos quadrados dos desvios.

d) Obtenha estimativas dos desvios padrões das estimativas dos parâmetros α, β e ρ.

7. Refaça o exercício anterior considerando o modelo

$$\ln Y_i = \ln \alpha - \ln[1 + e^{-(\beta+\gamma X_i)}] + u_i,$$

onde os u_i são erros independentes com distribuição normal de média zero e variância σ^2.

Considere $a_o = 70$ como estimativa preliminar de α.

Qual dos dois modelos (Gompertz ou logística) se ajusta melhor?

8. Considere o seguinte modelo de regressão não linear:

$$Y_i = (\alpha + X_i)^2 + u_i$$

Admite-se que os u_i são aleatórios independentes, com $u_i \sim N(0, \sigma^2)$.

a) Descreva uma maneira de obter uma estimativa preliminar de α.

b) Deduza a fórmula da correção a ser feita em dada estimativa preliminar (α_0) de α.

c) Deduza a fórmula para $V(a)$, sendo *a* a estimativa de máxima verossimilhança de α.

d) Calcule a correção (Δa) de *a* para $a_0 = 2{,}9$, com base na seguinte amostra:

X	Y
1	14
2	23
3	39

e) Calcule a correção (Δa) de *a* para $a_0 = 3$.

f) Obtenha a estimativa da variância de *a*.

g) Teste, ao nível de significância de 1%, a hipótese de que $\alpha = 0$.

9. Supõe-se que as variáveis *X* e *Y* estão relacionadas de acordo com o modelo

$$Y_i = \frac{\alpha}{X_j + \beta} + u_j$$

Os u_j são aleatórios independentes e admite-se que $u_j \sim N(0, \sigma^2)$.

Sejam *a* e *b* as estimativas de máxima verossimilhança de α e β, respectivamente.

a) Dada uma amostra de valores de X_j e Y_j (*j* = 1,..., *n*) , obtenha o sistema de duas equações cuja solução fornece os valores de *a* e *b*.

b) A tabela ao lado mostra os valores de X_j e Y_j em uma amostra com 6 observações. Mostre que, para essa amostra, os valores *a* = 60 e *b* = 3 satisfazem o sistema de equações obtido no item (*a*).

X_j	Y_j
0	20
1	16
2	12
3	7
7	6
9	8

c) Determine a estimativa de σ^2, associada a 4 graus de liberdade, por analogia com modelos lineares.

d) Determine a estimativa da matriz de variâncias e covariância de *a* e *b*.

Respostas

1. a) Dada uma estimativa preliminar (b_o) de β, a correção, de acordo com o método de Newton, é

$$\Delta b = \frac{\sum (Y_i - b_o^{X_i}) X_i b_o^{X_i - 1}}{\sum X_i^2 b_o^{2X_i - 2} - \sum (Y_i - b_o^{X_i}) X_i (X_i - 1) b_o^{X_i - 2}}$$

e, de acordo com o método de Gauss-Newton, é

$$\Delta b = \frac{\sum (Y_i - b_o^{X_i}) X_i b_o^{X_i - 1}}{\sum X_i^2 b_o^{2X_i - 2}}$$

A estimativa preliminar de β pode ser obtida pela fórmula

$$b_o = \left(\frac{Y_2}{Y_1} \right)^{\frac{1}{X_2 - X_1}},$$

Onde (X_1, Y_1) e (X_2, Y_2) são duas observações quaisquer da amostra ou as coordenadas de dois pontos da curva, traçada "a olho" em um gráfico.

b) A estimativa (*b*) obtida de acordo com o método de mínimos quadrados coincide com a estimativa de máxima verossimilhança. Portanto, é uma estimativa consistente e assintoticamente eficiente. A variância assintótica de *b*, igual ao limite inferior de Cramér-Rao, é

$$V(b) = \frac{\sigma^2}{\sum X_i^2 \beta^{2X_i - 2}}$$

c) $\Delta b = 0$, $b = 0{,}5$ e $s(b) = 0{,}01036$.

2. a) $\hat{Y}_i = X_i^b$. Então, aproximadamente, $\hat{Y}_i = X_i^{b_o} + X_i^{b_o} (\ln X_i) \Delta b$

Fazendo uma regressão de $Y_i - X_i^{b_o}$ contra $X_i^{b_o} \ln X_i$, sem termo constante, obtemos

$$\Delta b = \frac{\sum X_i^{b_o} (\ln X_i)(Y_i - X_i^{b_o})}{\sum X_i^{2b_o} (\ln X_i)^2}$$

b) Fazer uma regressão linear de $Z_i = \ln Y_i$ contra $W_i = \ln X_i$, sem termo constante, com fator de ponderação Y_i^2. Então,

$$b_o = \frac{\sum (\ln X_i)(\ln Y_i) Y_i^2}{\sum (\ln X_i)^2 Y_i^2}$$

c) $\Delta b = -0{,}107$

d) $\Delta b = 0$

e) $s(b) = 0{,}0614$

Outra alternativa é obter a correção pelo método de Newton, que é

$$\Delta b = \frac{\sum X_i^{b_o} (\ln X_i)(Y_i - X_i^{b_o})}{\sum X_i^{2b_o} (\ln X_i)^2 - (Y_i - X_i^{b_o}) X_i^{b_o} (\ln X_i)^2}$$

ou

$$\Delta b = \frac{\sum X_i^{b_o} (\ln X_i)(Y_i - X_i^{b_o})}{\sum (\ln X_i)^2 X_i^{b_o} (2X_i^{b_o} - Y_i)}$$

Nesse caso a resposta ao item (c) é $\Delta b = -0{,}064$.

3. a) As estimativas de mínimos quadrados para β e ρ coincidem com as estimativas de máxima verossimilhança. Portanto, são estimativas consistentes e assintoticamente eficientes. A matriz de variâncias e covariâncias assintóticas é

$$\begin{bmatrix} \sum \rho^{2X_i} & \beta \sum X_i \rho^{2X_i - 1} \\ \beta \sum X_i \rho^{2X_i - 1} & \beta^2 \sum X_i^{\,2} \rho^{2X_i - 2} \end{bmatrix}^{-1} \sigma^2$$

b) $b = 8$, $\Delta r = 0$, $r = 0{,}5$, $s(b) = 0{,}1982$ e $s(r) = 0{,}0178$.

4. a) Temos $\hat{Y} = \dfrac{aX}{X + b}$

Segue-se que, aproximadamente, $\hat{Y} = \dfrac{a_o X}{X + b_o} + \dfrac{X}{X + b_o} \Delta a - \dfrac{a_o X}{(X - b_o)^2} \Delta b$

Então, Δa e $-\Delta b$ são obtidos através de uma regressão linear múltipla de

$$Z = Y - \frac{a_o X}{X + b_o} \text{ contra } W_1 = \frac{X}{X + b_o} \text{ e } W_2 = \frac{a_o X}{(X + b_o)^2}$$

As estimativas são corrigidas ($a = a_o + \Delta a$, $b = b_o + \Delta b$) e a regressão é refeita até que Δa e Δb sejam desprezíveis.

b) $\mathbf{W'W} = \dfrac{1}{128} \begin{bmatrix} 162 & 2350 \\ 2350 & 38450 \end{bmatrix}$, $(\mathbf{W'W})^{-1} = \dfrac{8}{22075} \begin{bmatrix} 19225 & -1175 \\ -1175 & 81 \end{bmatrix}$ e $\mathbf{W'z} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Conclui-se que $\Delta a = \Delta b = 0$

c) $s(a) = 19{,}928$ e $s(b) = 1{,}2935$

Outra alternativa é considerar $\hat{Y}$ como função apenas de *b* ao desenvolver pela série de Taylor, obtendo, aproximadamente,

$$\hat{Y} = \frac{aX}{X + b_o} - \frac{aX}{(X - b_o)^2}\Delta b$$

Então *a* e $c = -a\Delta b$ são obtidos fazendo uma regressão linear múltipla de

Y contra $W_1 = \dfrac{X}{X + b_o}$ e $W_2 = \dfrac{X}{(X + b_o)^2}$

Obtém-se $\mathbf{W'W} = \dfrac{1}{36864}\begin{bmatrix} 46656 & 5640 \\ 5640 & 769 \end{bmatrix}$, $(\mathbf{W'W})^{-1} = \dfrac{8}{883}\begin{bmatrix} 769 & -5640 \\ -5640 & 46656 \end{bmatrix}$

$\mathbf{W'y} = \begin{bmatrix} 29160 \\ 3525 \end{bmatrix}\dfrac{1}{192}$, $a = 120$ e $c = 0$

e) Fazer uma regressão linear simples ponderada de $Z = 1/Y$ contra $W = 1/X$, com Y^4 como fator de ponderação. Sejam *c* e *d* os coeficientes da equação estimada $(\hat{Z} = c + dW)$. Então as estimativas preliminares de α e β são $a_o = 1/c$ e $b_o = d/c$.

5. $a = 22{,}847$, $b = 10{,}025$, $r = 0{,}6323$, $s(a) = 0{,}141$, $s(b) = 0{,}139$ e $s(r) = 0{,}0102$.

6. a) $a = 4{,}7020$, $b = -4{,}0445$ e $r = 0{,}6160$.

   b) $\exp(a) = 110{,}2$.

   c) A S.Q.Res. para $Z = \ln(Y)$ é igual a 0,0358.

   d) $s(a) = 0{,}1895$, $s(b) = 0{,}1897$ e $s(r) = 0{,}0363$.

7. a) $g = \ln(a) = 4{,}2452$, $b = -3{,}4987$, $c = 1{,}4479$.

   b) $a = \exp(g) = 69{,}8$.

   c) A S.Q.Res. para $Z = \ln(Y)$ é igual a 0,0396.

   d) $s(g) = 0{,}0916$, $s(b) = 0{,}1288$, $s(c) = 0{,}1026$

8. a) Calcular a média dos valores de $\sqrt{y_i} - X_i$

   b) $\Delta a = \dfrac{\sum\left[Y_i - (a_0 + X_i)^2\right](a_0 + X_i)}{\sum\left[3(a + X_i)^2 - Y_i\right]}$

   c) $V(a) = \dfrac{\sigma^2}{308}$

   d) $\Delta a = 0{,}103$

   e) $\Delta a = 0$

f) $\hat{V}(a) = \frac{8,5}{308} = 0,0276$

g) t = 18,059 , significativo $(t_0 = 9,925)$

9. a) $\sum_j \left(Y_j - \frac{a}{X_j + b} \right) \frac{1}{X_j + b} = 0$ $\sum_j \left(Y_j - \frac{a}{X_j + b} \right) \frac{a}{(X_j + b)^2} = 0$

b) As duas equações são satisfeitas para a = 60 e b = 3.

c) $s^2 = \frac{19}{4} = 4,75$

d) $\begin{bmatrix} 236,06 & 14,015 \\ 14,015 & 0,90239 \end{bmatrix}$

5. VARIÁVEL DEPENDENTE BINÁRIA: LÓGITE E PRÓBITE

5.1. Introdução

No modelo usual de análise de regressão a variável dependente deve ser contínua, pois para fazer os testes *t* e *F* é necessário pressupor que o erro (que é um componente da variável dependente) tem distribuição normal.
Mas há muitos problemas em que queremos avaliar como diversas variáveis estão associadas com a ocorrência ou não de algum fato (variável binária). Exemplo: analisar como a participação (sim ou não) de uma mulher no mercado de trabalho depende de sua idade, escolaridade, estado conjugal, número e idade dos filhos, etc. Outro exemplo: analisar como o fato de um domicílio ter ou não um carro depende da renda, da idade dos membros da família, da localização do domicílio, etc. Nestes casos não é apropriado aplicar as técnicas usuais de análise de regressão, existindo métodos específicos, como os modelos de lógite e próbite. Cabe ressaltar que o problema ocorre apenas quando a variável *dependente* é binária. A existência de variáveis binárias como variáveis *explanatórias* não afeta as fórmulas básicas da análise de regressão usual (incluindo estimadores de mínimos quadrados generalizados e os diversos métodos de estimação usados em equações simultâneas).

Consideremos a relação entre o fato de uma família ter ou não um carro e a renda familiar. Por simplicidade, vamos considerar apenas a renda familiar como variável explanatória. Se a análise for feita utilizando os dados individuais (cada família constituindo uma observação), a variável dependente é binária. Mas se as famílias forem agrupadas por faixas de renda (como ocorre nas publicações das Pesquisas de Orçamentos Familiares — POF — do IBGE), então a variável dependente passa a ser a *proporção* de famílias, em cada faixa de renda, que têm carro. Mesmo nesse caso não seria apropriado aplicar os métodos usuais de análise de regressão. Note-se que se isso fosse feito poderíamos obter valores estimados da variável dependente negativos ou maiores do que um, incompatíveis com a natureza da variável.
A tabela 1 mostra dados sobre a porcentagem de famílias com automóvel em uma amostra de 16.013 famílias distribuídas em 10 estratos de recebimento mensal. Trata-se de dados que foram calculados com base em resultados da POF (Pesquisa de Orçamentos Familiares) de 1995/96, arredondando os números e desconsiderando os problemas de ponderação decorrentes do processo de amostragem dessa pesquisa, que abrange o município de Goiânia, o Distrito Federal e as regiões metropolitanas de Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre. Por simplicidade, vamos admitir que se trata de uma amostra aleatória simples.
Note-se, na tabela 5.1, como a proporção de famílias com carro cresce com o seu recebimento.

Tabela 5.1. Dados (artificiais) sobre o número de famílias que possuem automóvel em 10 estratos de recebimento familiar, com base em resultados obtidos para o total das áreas pesquisadas na POF de 1995/96.

Estratos de recebimento mensal familiar (salário mínimo[(1)])	Recebimento Médio (R$)	Nº de famílias na amostra		Porcentagem com automóvel
		Total	com automóvel	
até 2	148	1.666	64	3,8
mais de 2 a 3	282	1.340	105	7,8
mais de 3 a 5	445	2.440	286	11,7
mais de 5 a 6	617	1.139	219	19,2
mais de 6 a 8	786	1.770	421	23,8
mais de 8 a 10	1.017	1.241	403	32,5
mais de 10 a 15	1.381	2.121	978	46,1
mais de 15 a 20	1.969	1.231	724	58,8
mais de 20 a 30	2.761	1.207	889	73,7
mais de 30	6.700	1.858	1.592	85,7
Total	1.636	16.013	5.681	35,5

(1) Salário mínimo em setembro de 1996: R$112,00.

Nos trabalhos originais dos criadores do próbite (Bliss, 1935) e do lógite (Berkson, 1944) o problema analisado era a reação de um animal submetido a determinada dose de uma droga (ou algum outro tipo de "estímulo"). Em entomologia, para analisar a susceptibilidade de determinado inseto a um inseticida, diversos grupos de insetos (cada um com $n_i = 20$ insetos, por exemplo) são submetidos a doses crescentes do produto, verificando-se, em cada grupo, qual a proporção de insetos mortos depois de certo tempo. A análise estatística dos dados permite estimar a dose do inseticida que mata 50% dos insetos, denominada *dose letal mediana* (ou DL_{50}). Os dados obtidos de experimentos desse tipo, onde o resultado é uma variável binária (sobrevivência ou morte do inseto), são denominados *dados de resposta quântica* (ou *reação quântica*, conforme o "Dicionário Brasileiro de Estatística", de Rodrigues, 1970).

5.2. O Lógite

Vamos admitir que haja *k* variáveis explanatórias para a resposta quântica. O vetor-linha com os valores dessas variáveis explanatórias na *j*-ésima observação é

$$\mathbf{x}'_j = \begin{bmatrix} 1 & x_{1j} & \dots & x_{kj} \end{bmatrix}$$

com $j = 1, ..., L$.

O correspondente vetor de parâmetros é

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

No modelo do lógite admite-se que, dado $\mathbf{x}_j$, a probabilidade de obter uma resposta favorável é

$$P_j = \left[1 + \exp(-\mathbf{x}'_j\boldsymbol{\beta})\right]^{-1} = \frac{1}{1 + \exp(-\mathbf{x}'_j\boldsymbol{\beta})} \tag{5.1}$$

Note-se que, se houver apenas uma variável explanatória, trata-se de uma curva logística com assíntota superior com ordenada 1.
Obtemos

$$Q_j = 1 - P_j = \frac{\exp(-\mathbf{x}'_j\boldsymbol{\beta})}{1 + \exp(-\mathbf{x}'_j\boldsymbol{\beta})}$$

e

$$\frac{P_j}{Q_j} = \exp(\mathbf{x}'_j\boldsymbol{\beta})$$

Então

$$Y_j = \ln\frac{P_j}{Q_j} = \mathbf{x}'_j\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_k x_{kj} \tag{5.2}$$

que é o *lógite* correspondente a P_j. Note-se que, partindo do modelo não linear (5.1), o lógite é, por construção, uma função linear das variáveis explanatórias. Note-se, também, que quando P_j varia de zero a 1, o lógite varia de –∞ a +∞.

Considerando dados como os apresentados na tabela 5.1, seja n_j o número de elementos (famílias) submetidos ao "estímulo" x_j (o recebimento familiar[13]), e seja m_j o correspondente número de respostas "favoráveis". Então a proporção de respostas favoráveis observada é

$$p_j = \frac{m_j}{n_j} \tag{5.3}$$

e o lógite observado é

$$y_j = \ln\frac{p_j}{q_j},$$

[13] Usualmente é melhor usar como variável explanatória (x_j) o logaritmo do recebimento (ou rendimento).

onde $q_j = 1 - p_j$

Quando os dados são individuais, $n_j = 1$ para todo j. Se $p_j = m_j = 1$, temos $q_j = 0$; se $p_j = m_j = 0$, temos $q_j = 1$. Em qualquer dos dois casos não se define o lógite observado. Quando são usados dados agrupados $(n_j > 1)$ e há algumas observações com p_j igual a zero ou 1, para determinados cálculos preliminares em que é necessário o lógite é usual substituir o zero por $1/(2n_j)$ e substituir o 1 por $1 - 1/(2n_j)$.

5.3. Estimação dos parâmetros por meio de uma regressão linear ponderada

De (5.3), diferenciando, obtemos

$$dy_j = \frac{dp_j}{p_j q_j}$$

Então, conforme o que foi visto na seção 4.5,

$$V(y_j) = \frac{V(p_j)}{P_j^2 Q_j^2} \tag{5.4}$$

Como $V(p_j) = \frac{P_j Q_j}{n_j}$,

segue-se que $V(y_j) = \frac{1}{n_j P_j Q_j}$ (5.5)

Assim, as estimativas dos parâmetros podem ser obtidas fazendo uma regressão linear de

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix} \quad \text{contra} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_L' \end{bmatrix}$$

com fatores de ponderação $n_j p_j (1 - p_j)$, isto é,

$$[b_i] = [a_{hi}]^{-1} [c_i] \tag{5.6}$$

onde

b_i é a estimativa de β_i ,

$$a_{hi} = \sum_j x_{ij} x_{hj} n_j p_j (1 - p_j) \ , \tag{5.7}$$

$$c_i = \sum_j x_{ij} y_j n_j p_j (1 - p_j), \tag{5.8}$$

$x_{0j} = 1$ para todo *j*, [b_i] representa o vetor-coluna dos b_i, [c_i] representa o vetor-coluna dos c_i e [a_{hi}] representa a matriz quadrada com elementos a_{hi}, com *h* e *i* variando de zero a *k*. A matriz das estimativas das variâncias e covariâncias das estimativas dos parâmetros é

$$[a_{hi}]^{-1} s^2 \tag{5.9}$$

onde

$$s^2 = \text{Q.M.Res.} = \frac{1}{L-k-1}\left(\sum_{j=1}^{L} y_j^2 n_j p_j (1 - p_j) - \sum_{i=0}^{k} b_i c_i \right) \tag{5.10}$$

Esse procedimento não tem fundamentação estatística rigorosa. Note-se que a variância de y_i, dada pela expressão (5.5), depende da probabilidade P_j, mas a ponderação da regressão é feita com base na proporção observada p_j. Note-se, também, que esse procedimento exige o cálculo do lógite observado (y_i), que não é definido quando p_j é igual a zero ou 1. Como os cálculos são relativamente simples, não envolvendo um processo iterativo, esse procedimento era muito utilizado antes da grande expansão das facilidades computacionais. Para os dados da tabela 5.1 obtêm-se os seguintes resultados (estimativas dos desvios padrões entre parênteses):

$$\hat{Y} = \underset{(0,4269)}{-10,6320} + \underset{(0,0598)}{1,4374x} \ ,$$

onde *x* é o logaritmo neperiano do recebimento familiar médio do estrato.

A proporção estimada será 1/2 quando $\hat{Y}$ for igual a zero, isto é, quando *x* for igual a $-b_0/b_1$. Então a estimativa da renda familiar para a qual se espera que metade das famílias tenha carro é

$$\exp\left(-\frac{b_0}{b_1} \right) = \exp\left(\frac{10,6320}{1,4374} \right) = \text{R\$1.631}$$

5.4. Estimativas dos parâmetros pelo método de mínimos quadrados, com processo iterativo

Como se sabe que os estimadores de máxima verossimilhança são, em geral, consistentes e assintoticamente eficientes, a tendência é usar esse método, que será apresentado na próxima seção. O leitor menos interessado em variantes metodológicas pode pular a presente seção.

Uma vez que $V(p_j) = \frac{P_j Q_j}{n_j}$, no método de mínimos quadrados devemos minimizar a soma de quadrados *ponderados*

$$Z = \sum_j \frac{n_j}{\hat{P}_j \hat{Q}_j} (p_j - \hat{P}_j)^2 \tag{5.11}$$

onde $\hat{Q}_j = 1 - \hat{P}_j$

e $$\hat{P}_j = \frac{1}{1 + \exp\{-\mathbf{x}'_j \mathbf{b}\}} \tag{5.12}$$

Obtemos

$$\frac{\partial Z}{\partial b_i} = \sum_j \left[-2(p_j - \hat{P}_j) \frac{n_j}{\hat{P}_j \hat{Q}_j} - \frac{n_j (p_j - \hat{P}_j)^2}{\hat{P}_j^2 \hat{Q}_j^2} (1 - 2\hat{P}_j) \right] \frac{\partial \hat{P}_j}{\partial b_i} = 0$$

Mas $$\frac{\partial \hat{P}_j}{\partial b_i} = \hat{P}_j \hat{Q}_j x_{ij}$$

Então

$$\sum_j \frac{(p_j - \hat{P}_j) n_j x_{ij}}{\hat{P}_j \hat{Q}_j} \left[-2\hat{P}_j \hat{Q}_j - (p_j - \hat{P}_j)(1 - 2\hat{P}_j) \right] = 0$$

ou

$$F_i = \sum_j n_j x_{ij} \frac{p_j - \hat{P}_j}{\hat{P}_j \hat{Q}_j} (2 p_j \hat{P}_j - p_i - \hat{P}_j) = 0$$

ou, ainda, $$F_i = \sum_j n_j x_{ij} \frac{\hat{P}_j - p_j}{\hat{P}_j \hat{Q}_j} (p_j + \hat{P}_j - 2 p_i \hat{P}_j) = 0 \tag{5.13}$$

com $i = 0, 1, ..., k$
Esse é o sistema de equações normais

A seguir obtemos

$$\frac{\partial F_i}{\partial b_h} = \sum_j n_j x_{ij} \left[\frac{\hat{P}_j \hat{Q}_j - (\hat{P}_j - p_j)(\hat{Q}_j - \hat{P}_j)}{\hat{P}_j^2 \hat{Q}_j^2} (p_i + \hat{P}_j - 2 p_j \hat{P}_j) + \frac{\hat{P}_j - p_j}{\hat{P}_j \hat{Q}_j} (1 - 2 p_j) \right] \frac{\partial \hat{P}_j}{\partial b_h}$$

Lembrando que $\frac{\partial \hat{P}_j}{\partial b_h} = \hat{P}_j \hat{Q}_j x_{hj}$, obtemos, após algumas simplificações,

$$\frac{\partial F_i}{\partial b_h} = \sum_j n_j x_{ij} x_{hj} \frac{p_j^2 + \hat{P}_j^2 + 2 p_j \hat{P}_j (p_j \hat{P}_j - p_j - \hat{P}_j)}{\hat{P}_j \hat{Q}_j} \tag{5.14}$$

De acordo com (5.13) e (5.14), as correções Δb_i nas estimativas preliminares dos parâmetros são dadas por

$$[\Delta b_i] = [a_{hi}]^{-1}[c_i] \tag{5.15}$$

onde

$$c_i = -F_i = \sum_j n_j x_{ij} \frac{\hat{P}_j - p_j}{\hat{P}_j \hat{Q}_j}(2p_j\hat{P}_j - p_j - \hat{P}_j) \tag{5.16}$$

e

$$a_{hi} = \sum_j n_j x_{ij} x_{hj} \frac{p_j^2 + \hat{P}_j^2 + 2p_j\hat{P}_j(p_j\hat{P}_j - p_j - \hat{P}_j)}{\hat{P}_j\hat{Q}_i} \tag{5.17}$$

Como estimativas preliminares para iniciar o processo iterativo podem ser utilizadas as obtidas por meio da regressão linear ponderada descrita na seção anterior.
Cabe ressaltar que os cálculos indicados em (5.15), (5.16) e (5.17) não envolvem o lógite observado. Essas expressões podem ser usadas mesmo no caso de observações individuais, quando $n_j = 1$ para todo j e p_j é igual a zero ou 1.

5.5. Estimativas dos parâmetros pelo método da máxima verossimilhança

A função de verossimilhança é

$$£ = \prod_j \binom{n_j}{m_j} P_j^{m_j}(1-P_j)^{n_j - m_j} \tag{5.18}$$

onde n_j é o número de observações (indivíduos) no *j*-ésimo ensaio e m_j é o número de indivíduos afetados pelo tratamento.

Segue-se que

$$\ln £ = \sum_j \left[\ln \binom{n_j}{m_j} + m_j \ln P_j + (n_j - m_j)\ln(1-P_j) \right]$$

Então

$$\frac{\partial \ln £}{\partial \beta_i} = \sum_j \left(\frac{m_j}{P_j} - \frac{n_j - m_j}{1-P_j} \right) \frac{\partial P_j}{\partial \beta_i} \tag{5.19}$$

De (5.1) obtemos

$$\frac{\partial P_j}{\partial \beta_i} = x_{ij} P_j Q_j = x_{ij} P_j (1-P_j) \tag{5.20}$$

Substituindo (5.20) em (5.19) e simplificando, obtemos

$$\frac{\partial \ln £}{\partial \beta_i} = \sum_j x_{ij}(m_j - n_j P_j)$$

Como $p_j = \frac{m_j}{n_j}$, segue-se que

$$\frac{\partial \ln £}{\partial \beta_i} = \sum_j n_j x_{ij} (p_j - P_j) \tag{5.21}$$

O sistema de equações cuja solução consiste nos estimadores de máxima verossimilhança é

$$\sum_j n_j x_{ij} (p_j - P_j) = 0 \quad (i = 0, 1, ..., k) \tag{5.22}$$

De (5.21) segue-se que

$$\frac{\partial^2 \ln £}{\partial \beta_i \partial \beta_h} = -\sum_j n_j x_{ij} \frac{\partial P_j}{\partial \beta_h}$$

De acordo com (5.20) temos

$$\frac{\partial P_j}{\partial \beta_h} = x_{hj} P_j Q_j$$

Então

$$\frac{\partial^2 \ln £}{\partial \beta_i \partial \beta_h} = -\sum_j n_j x_{ij} x_{hj} P_j Q_j \tag{5.23}$$

De acordo com (5.21) e (5.23), as correções Δb_i nas estimativas preliminares dos parâmetros são dadas por

$$[\Delta b_i] = [a_{hi}]^{-1} [c_i]$$

onde (trocando o sinal de $\frac{\partial^2 \ln £}{\partial \beta_i \partial \beta_h}$ e mantendo o sinal de $\frac{\partial \ln £}{\partial \beta_i}$)

$$c_i = \sum_j n_j x_{ij} (p_j - \hat{P}_j)$$

e

$$a_{hi} = \sum_j n_j x_{ij} x_{hj} \hat{P}_j \hat{Q}_j \tag{5.24}$$

A matriz de informação é $[w_{hi}]$, onde

$$w_{hi} = -E\left(\frac{\partial^2 \ln £}{\partial \beta_i \partial \beta_h} \right)$$

De (5.23) segue-se que

$$-E\left(\frac{\partial^2 \ln £}{\partial \beta_i \partial \beta_h} \right) = \sum_j n_j x_{ij} x_{hj} P_j Q_j$$

Uma vez que os estimadores de máxima verossimilhança são consistentes e assintoticamente eficientes, a matriz das variâncias e covariâncias assintóticas das estimativas dos parâmetros é $[w_{hi}]^{-1}$.

Lembrando (5.24), conclui-se que a matriz das *estimativas* das variâncias e covariâncias assintóticas das estimativas dos parâmetros é

$$[a_{hi}]^{-1} \tag{5.25}$$

Para verificar se o ajustamento é bom, calcula-se a soma de quadrados dos desvios ponderados das proporções (o fator de ponderação é o inverso da estimativa da variância de p_j, que é $\hat{V}(p_j) = \hat{P}_j\hat{Q}_j / n_j$):

$$S = \sum_j \frac{n_j}{\hat{P}_j\hat{Q}_j}(p_j - \hat{P}_j)^2 \tag{5.26}$$

Se o modelo adotado for o verdadeiro, essa soma de quadrados tem, assintoticamente, distribuição de qui-quadrado com $L - k - 1$ graus de liberdade.[14]

Se essa soma de quadrados tiver valor elevado (lembrar que o valor esperado de uma variável com distribuição de qui-quadrado é igual ao seu número de graus de liberdade), as estimativas das variâncias e covariâncias assintóticas das estimativas dos parâmetros devem ser corrigidas, multiplicando-as por $S/(L-k-1)$. Nesse caso a matriz das estimativas das variâncias e covariâncias assintóticas das estimativas dos parâmetros passa a ser

$$[a_{hi}]^{-1}\frac{S}{L-k-1} \tag{5.27}$$

Aplicando o método da máxima verossimilhança aos dados da tabela 1 obtemos

$$\hat{Y} = -10{,}7043 + 1{,}4478x,$$

onde *x* é o logaritmo neperiano do recebimento familiar.

A estimativa do recebimento para o qual se espera que metade das famílias tenha automóvel é

$$\exp\left(-\frac{b_0}{b_1}\right) = \exp\left(\frac{10{,}7043}{1{,}4478}\right) = \text{R\$1.625}$$

Sem correção, as estimativas dos desvios padrões das estimativas dos parâmetros são

$$s(b_0) = 0{,}1752 \qquad \text{e} \qquad s(b_1) = 0{,}0246.$$

Mas a soma de quadrados dos desvios ponderados é bastante elevada: 51,50, com 8 graus de liberdade. Fazendo a correção, as estimativas dos desvios padrões das estimativas dos parâmetros passam a ser

$$s(b_0) = 0{,}4444 \qquad \text{e} \qquad s(b_1) = 0{,}0625.$$

Na figura 5.1 podemos observar a curva ajustada aos dados da tabela 5.1, mostrando como a proporção de famílias com carro cresce em função de *x*.

[14] Lembrar que uma distribuição binomial pode ser considerada aproximadamente normal se $np>5$ e $nq>5$.

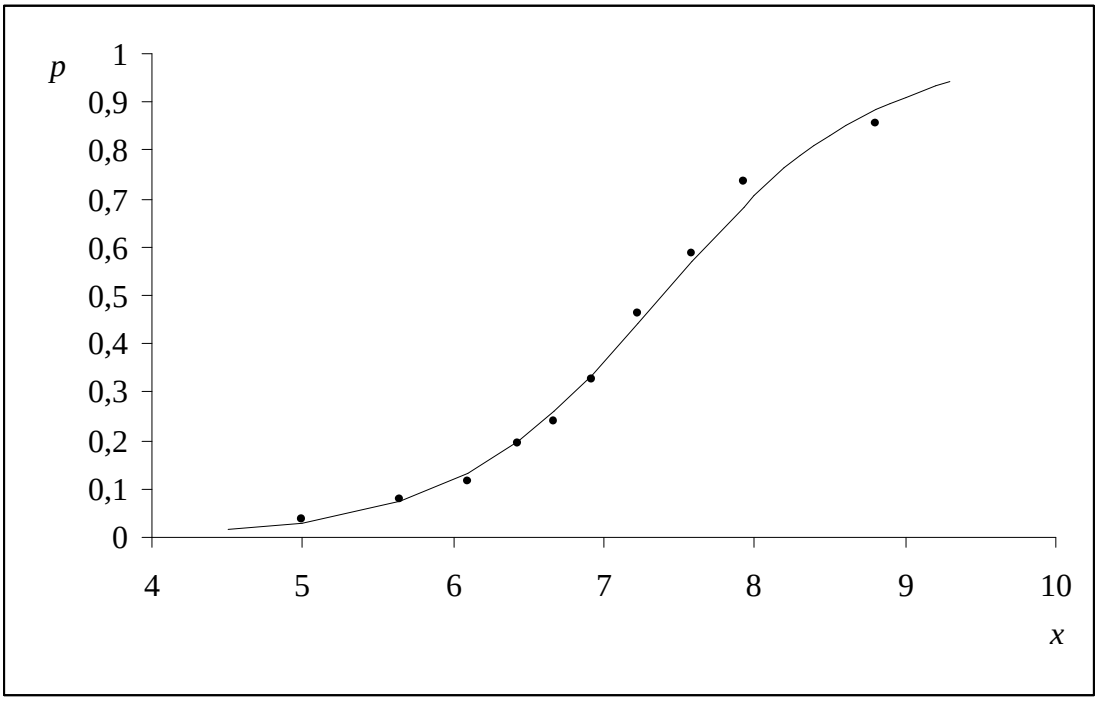


Figura 5.1. Proporção de famílias com carro em função do logaritmo do recebimento: valores observados (*p*) e curva ajustada conforme modelo de lógite.

5.6. O caso particular de uma única variável explanatória binária

É interessante considerar o caso particular de um modelo de lógite com uma única variável explanatória que é binária. Poderíamos estar analisando, por exemplo, como o fato de uma pessoa ser ou não fumante afeta a probabilidade de ela ter câncer no pulmão, ou como o fato de um empregado ter ou não curso superior afeta a probabilidade de ele ser sindicalizado. Os dados poderiam ser organizados como na tabela a seguir:

Tabela 5.2. Frequências para uma variável explanatória (*x*) binária.

Resultado	$x = 0$	$x = 1$
Favorável	m_0	m_1
Desfavorável	$n_0 - m_0$	$n_1 - m_1$
Total	n_0	n_1

Para $x = 0$, a proporção de resultados favoráveis é $p_0 = m_0/n_0$ e para $x = 1$, a proporção de resultados favoráveis é $p_1 = m_1/n_1$. De acordo com o modelo de lógite, as respectivas probabilidades esperadas são

$$P_0 = \frac{1}{1+\exp(-\alpha)}$$

e

$$P_1 = \frac{1}{1+\exp(-\alpha-\beta)}$$

Sendo a e b as estimativas de α e β, as probabilidades estimadas são

$$\hat{P}_0 = \frac{1}{1+\exp(-a)} \quad \text{ou} \quad a = \ln\frac{\hat{P}_0}{\hat{Q}_0} \tag{5.28}$$

e

$$\hat{P}_1 = \frac{1}{1+\exp(-a-b)} \quad \text{ou} \quad a+b = \ln\frac{\hat{P}_1}{\hat{Q}_1} \tag{5.29}$$

Veremos que, neste caso particular, as estimativas de máxima verossimilhança dos dois parâmetros podem ser obtidas sem necessidade de processo iterativo.
De acordo com (5.22), o sistema de equações que deve ser resolvido para obter as estimativas de máxima verossimilhança dos dois parâmetros é

$$\begin{cases} n_0(p_0 - \hat{P}_0) + n_1(p_1 - \hat{P}_1) = 0 & (5.30) \\ n_1(p_1 - \hat{P}_1) = 0 & (5.31) \end{cases}$$

Segue-se que

$$\begin{cases} \hat{P}_0 = p_0 & (5.32) \\ \hat{P}_1 = p_1 & (5.33) \end{cases}$$

De (5.28) e (5.32), com $q_0 = 1 - p_0$, obtemos

$$a = \ln\frac{p_0}{q_0} \qquad (5.34)$$

Em seguida, utilizando (5.29) e (5.33) e fazendo $q_1 = 1 - p_1$, deduz-se que

$$b = \ln\frac{p_1}{q_1} - \ln\frac{p_0}{q_0} = \ln\frac{\frac{p_1}{q_1}}{\frac{p_0}{q_0}} \tag{5.35}$$

A razão

$$\exp(b) = \frac{\frac{p_1}{q_1}}{\frac{p_0}{q_0}} \tag{5.36}$$

é denominada *odds ratio*. Ela é uma medida da intensidade do efeito de x (mudança de $x = 0$ para $x = 1$) sobre a probabilidade de obter resultado “favorável”. É óbvio que a *odds ratio* não

pode ser negativa. Valores de *b* positivo, igual a zero ou negativo correspondem a valores da *odds ratio* maior do que 1, igual a 1 ou menor do que 1, respectivamente.

De acordo com (5.24) e (5.25), as estimativas das variâncias e covariância assintóticas de *a* e *b* são dadas por

$$\begin{bmatrix} n_0\hat{P}_0\hat{Q}_0 + n_1\hat{P}_1\hat{Q}_1 & n_1\hat{P}_1\hat{Q}_1 \\ n_1\hat{P}_1\hat{Q}_1 & n_1\hat{P}_1\hat{Q}_1 \end{bmatrix}^{-1}$$

Lembrando (5.32) e (5.33), essa matriz de estimativas de variâncias e covariância fica

$$\begin{bmatrix} n_0 p_0 q_0 + n_1 p_1 q_1 & n_1 p_1 q_1 \\ n_1 p_1 q_1 & n_1 p_1 q_1 \end{bmatrix}^{-1} \quad (5.37)$$

Verifica-se que

$$\hat{V}(b) = \frac{n_0 p_0 q_0 + n_1 p_1 q_1}{n_0 p_0 q_0 n_1 p_1 q_1} \qquad (5.38)$$

5.7. Variâncias dos lógites estimados e das probabilidades estimadas

Temos

$$\hat{Y}_j = \mathbf{x}'_j \mathbf{b} \qquad (5.39)$$

onde **b** é o vetor-coluna das estimativas dos parâmetros de acordo com o método da máxima verossimilhança.

Como $\text{plim}\mathbf{b} = \boldsymbol{\beta}$, temos

$$\text{plim}\hat{Y}_j = \mathbf{x}'_j\boldsymbol{\beta} = Y_j$$

De acordo com a relação (4.55), na seção 4.5, a estimativa da variância assintótica de $\hat{Y}_j$ é dada por

$$\hat{V}(\hat{Y}_j) = \mathbf{x}'_j [a_{hi}]^{-1} \mathbf{x}_j \qquad (5.40)$$

ou

$$\hat{V}(\hat{Y}_j) = \mathbf{x}'_j [a_{hi}]^{-1} \mathbf{x}_j \frac{S}{L-k-1}, \qquad (5.41)$$

conforme se considere (5.25) ou (5.27) como sendo a matriz das estimativas das variâncias e covariâncias assintóticas das estimativas dos parâmetros.

Analogamente a (5.4), temos

$$V(\hat{Y}_j) = \frac{V(\hat{P}_j)}{P_j^2 Q_j^2}$$

ou

$$V(\hat{P}_j) = P_j^2 Q_j^2 V(\hat{Y}_j)$$

Então a estimativa da variância assintótica da proporção estimada é dada por

$$\hat{V}(\hat{P}_j) = \hat{P}_j^2 \hat{Q}_j^2 \hat{V}(\hat{Y}_j) \tag{5.42}$$

5.8. Efeitos marginais

Em uma regressão linear múltipla o coeficiente β_i é o efeito marginal de x_i sobre a variável dependente *Y*. Analogamente, no modelo de lógite o coeficiente β_i é o efeito marginal de x_i sobre o lógite *Y*. Mas o lógite é uma variável artificial, e o pesquisador está efetivamente interessado no efeito marginal de x_i sobre a probabilidade *P*.

De $$Y = \ln \frac{P}{1-P}$$

obtemos

$$\frac{\partial Y}{\partial x_i} = \frac{1}{P(1-P)} \frac{\partial P}{\partial x_i} \tag{5.43}$$

Mas $$\frac{\partial Y}{\partial x_i} = \beta_i \tag{5.44}$$

De (5.43) e (5.44) segue-se que

$$\frac{\partial P}{\partial x_i} = \beta_i P(1-P) \tag{5.45}$$

Note-se que o efeito marginal de x_i sobre *P* depende do ponto da curva (ou da superfície) que for considerado. Dado um vetor de valores das variáveis explanatórias $\mathbf{x}'_h$, podemos calcular $\hat{Y}_h = \mathbf{x}'_h \mathbf{b}$ e

$$\hat{P}_h = \frac{1}{1+\exp(\hat{Y}_h)} \tag{5.46}$$

De acordo com (5.45), a estimativa do efeito marginal de x_i sobre *P* no ponto $\mathbf{x}'_h$ é $b_i \hat{P}_h (1-\hat{P}_h)$.

5.9. Pares concordantes e discordantes

Quando são analisados dados individuais (*p* é uma variável binária), uma maneira de avaliar se o modelo de lógite se ajustou bem aos dados consiste em verificar, para cada par de observações com valores diferentes de *p*, se o sentido de variação de *p* coincide ou não com o sentido de variação de $\hat{P}$ (a probabilidade estimada).

Consideremos o par de observações $p_h = 0$ e $p_i = 1$. Se $\hat{P}_i > \hat{P}_h$ o par é *concordante*. Se $\hat{P}_i < \hat{P}_h$ o par é *discordante*. Se $\hat{P}_i = \hat{P}_h$ diz-se que ocorre *empate*.

Seja N o número total de observações, havendo N_0 observações com $p = 0$ e N_1 observações com $p = 1$. Então o número de pares com valores distintos de *p* é $N_0 N_1$. Vamos indicar por n_c e n_d o número de pares concordantes e discordantes, respectivamente. Obviamente o número de empates é

$$N_0 N_1 - n_c - n_d$$

Há quatro índices de correlação de ordem que podem ser calculados para avaliar a qualidade do ajustamento do modelo de lógite:

(I) *D* de Somer: $D = \dfrac{n_c - n_d}{N_0 N_1}$

(II) $\gamma = \dfrac{n_c - n_d}{n_c + n_d}$

(III) $\tau_a = \dfrac{n_c - n_d}{0{,}5N(N-1)}$

(IV) $c = \dfrac{n_c + 0{,}5(N_0 N_1 - n_c - n_d)}{N_0 N_1}$

Para todos esses índices um valor maior indica um melhor ajustamento, ou seja, uma equação estimada que leva a previsões mais corretas.

Pode-se verificar que

$$D = 2(c - 0{,}5)$$

Para ilustrar o cálculo dessas medidas, vamos considerar os dados artificiais da tabela 5.2.

Tabela 5.2. Dados artificiais para 20 valores de *x* e *p*, onde *p* é uma variável dependente binária.

x	P			
1	0	0	0	0
2	1	0	0	0
3	0	1	1	0
4	1	0	1	1
5	1	1	1	1

Pode-se verificar que há $10 \cdot 10 = 100$ pares de valores com valores distintos de *p*.
A equação estimada é

$$\hat{P} = -4{,}762 + 1{,}587x$$

Mesmo sem calcular as probabilidades estimadas, neste exemplo é possível verificar que há $n_c = 85$ pares concordantes, $n_d = 5$ pares discordantes e 10 empates, obtendo-se $D = 0{,}8$, $\gamma = 0{,}889$, $\tau_a = 0{,}421$ e $c = 0{,}9$.

5.10. O Próbite

No caso do próbite, em lugar da curva logística dada em (5.1) usa-se a função de distribuição de uma variável normal reduzida (*u*). Seja *f*(*u*) a função de densidade de probabilidade de uma variável normal reduzida e seja $\Phi(u)$ a correspondente função de distribuição, isto é,

$$\Phi(u) = \int_{-\infty}^{u} f(t)dt$$

Então o modelo de próbite é

$$P_j = \Phi(Z_j) \tag{5.47}$$

onde $$Z_j = \mathbf{x}_j'\boldsymbol{\beta}\ , \tag{5.48}$$

$$\mathbf{x}_j' = [1 \quad x_{1j} \quad x_{2j} \quad \cdots \quad x_{kj}] \quad \text{e} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

com $j = 1, 2, ..., L$
Z_j é o próbite correspondente a P_j.

Seja $p_j = \dfrac{m_j}{n_j}$ a proporção observada. Então

$$p_j = \Phi(z_i) \tag{5.49}$$

onde z_i é o próbite correspondente a p_j.

Antes de descrever o procedimento para obter as estimativas dos parâmetros, consideremos a aplicação clássica em que P_j é a probabilidade de morrer para insetos submetidos à dose de veneno x_j. Neste caso

$$Z_j = \mathbf{x}'_j\boldsymbol{\beta} = \beta_0 + \beta_1 x_j$$

A dose máxima que não chega a causar a morte do inseto é denominada *tolerância* do inseto ao veneno. Se a tolerância for T_j, o inseto não morre se a dose de veneno não ultrapassar

$$x_j = T_j$$

Se subtrairmos dos dois membros $\mu_T = E(T_j)$ e dividirmos por σ_T (o desvio padrão de T_j), obtemos

$$\frac{x_j - \mu_T}{\sigma_T} = \frac{T_j - \mu_T}{\sigma_T}$$

Se a tolerância T_j tiver distribuição normal, verifica-se que o segundo membro é uma variável normal reduzida que podemos igualar a $Z_j = \beta_0 + \beta_1 x_j$. Então a condição fica

$$-\frac{\mu_T}{\sigma_T} + \frac{1}{\sigma_T} x_j = \beta_0 + \beta_1 x_j \quad ,$$

havendo as seguintes correspondências entre coeficientes das duas expressões lineares em x_j:

$$\frac{1}{\sigma_T} = \beta_1 \qquad \text{e} \qquad -\frac{\mu_T}{\sigma_T} = \beta_0$$

Dessas relações obtemos

$$\sigma_T = \frac{1}{\beta_1} \qquad \text{e} \qquad \mu_T = -\frac{\beta_0}{\beta_1}$$

Depois de obtidas estimativas dos parâmetros β_0 e β_1, essas expressões permitem obter estimativas da média e do desvio padrão da tolerância.
Vejamos, em seguida, como obter as estimativas dos parâmetros β_i (com $i = 0,...,k$).

De (5.49) segue-se que

$$dp_j = f(z_i)dz_i$$

Então $V(p_i) = f(Z_j)V(z_i)$

Como $V(p_i) = \dfrac{P_j Q_j}{n_j}$, segue-se que

$$V(z_i) = \frac{P_j Q_j}{n_j [f(Z_j)]^2} \tag{5.50}$$

De acordo com (5.48) e (5.50) as estimativas dos parâmetros β_0, β_1, ..., β_k podem ser obtidas fazendo uma regressão de

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_L \end{bmatrix} \quad \text{contra} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \cdots \\ \mathbf{x}_L' \end{bmatrix}$$

com fatores de ponderação $\dfrac{n_j[f(z_j)]^2}{p_j(1-p_j)}$, isto é,

$$[b_i] = [a_{hi}]^{-1}[c_i] \tag{5.51}$$

onde

b_i é a estimativa de β_i,

$$a_{hi} = \sum_j x_{ij} x_{hj} \frac{n_j[f(z_i)]^2}{p_j(1-p_j)} \tag{5.52}$$

e

$$c_i = \sum_j x_{ij} z_j \frac{n_j[f(z_i)]^2}{p_j(1-p_j)} \tag{5.53}$$

Para obter as estimativas de máxima verossimilhança dos parâmetros do modelo de próbite, note-se, inicialmente, que as expressões (5.18) e (5.19) permanecem válidas. Em lugar de (5.20) temos

$$\frac{\partial P_j}{\partial \beta_i} = \frac{dP_j}{dZ_j} \cdot \frac{\partial Z_j}{\partial \beta_i} = x_{ij} f(Z_j) \tag{5.54}$$

De (5.19) obtemos

$$\frac{\partial \ln £}{\partial \beta_i} = \sum_j \frac{n_j}{P_j Q_j}(p_j - P_j)\frac{\partial P_j}{\partial \beta_i} \tag{5.55}$$

Lembrando (5.54), segue-se que

$$\frac{\partial \ln £}{\partial \beta_i} = \sum_j \frac{n_j}{P_j Q_j}(p_j - P_j) x_{ij} f\left(Z_j\right)$$

Igualando a zero essas expressões para i = 0, 1, ..., k obtemos o sistema de equações cuja solução, por processo iterativo, consiste nos estimadores de máxima verossimilhança. Ao desenvolver as fórmulas para efetuar esse processo iterativo deve-se notar que para a função de densidade de uma distribuição normal reduzida temos que $f'(Z) = -Zf(Z)$. Então, de (5.54) segue-se que

$$\frac{\partial^2 P_j}{\partial \beta_h \partial \beta_j} = -x_{ij} x_{hj} Z_j f\left(Z_j\right) \tag{5.56}$$

Sabemos que a matriz de informação é $[w_{hi}]$, onde

$$w_{hi} = -E\left(\frac{\partial^2 \ln £}{\partial\beta_i \partial\beta_h}\right)$$

De (5.55) segue-se que

$$\frac{\partial^2 \ln £}{\partial\beta_i \partial\beta_h} = -\sum_j \frac{n_j}{P_j Q_j} \cdot \frac{\partial P_j}{\partial\beta_i} \cdot \frac{\partial P_j}{\partial\beta_h} + (3 \text{ termos envolvendo } p_j - P_j) =$$

$$= -\sum_j \frac{n_j}{P_j Q_j} x_{ij} x_{hj} \left[f\left(Z_j\right)\right]^2 + (3 \text{ termos envolvendo } p_j - P_j)$$

Então

$$-E\left(\frac{\partial^2 \ln £}{\partial\beta_i \partial\beta_h}\right) = \sum_j \frac{n_j}{P_j Q_j} x_{ij} x_{hj} \left[f\left(Z_j\right)\right]^2 \tag{5.57}$$

Comparando (5.52) e (5.57) verifica-se que a matriz $[a_{hi}]$ no método de regressão ponderada corresponde a uma estimativa da matriz de informação associada ao método de máxima verossimilhança.

Vejamos, em seguida, quais são os efeitos marginais de x_i sobre *P* no modelo de próbite. De acordo com (5.47) e (5.48), temos

$$\frac{\partial P}{\partial x_i} = f(Z)\frac{\partial Z}{\partial x_i} = \beta_i f(Z) \tag{5.58}$$

Dado um conjunto de valores das variáveis explanatórias $\mathbf{x}'_h$, a estimativa do efeito marginal de x_i sobre *P* nesse ponto é

$$b_i f(\mathbf{x}'_h \mathbf{b}) = b_i f(\hat{Z}_h) \tag{5.59}$$

Vamos comparar os efeitos marginais nos modelos lógite e próbite em um ponto da superfície em que $\hat{P}_h = 1/2$. Nesse ponto temos $\hat{Y}_h = \hat{Z}_h = 0$ e $f(\hat{Z}_h) = 1/\sqrt{2\pi}$. Lembrando a expressão obtida no final da seção 5.8, os efeitos marginais dos dois modelos serão os mesmos nesse ponto se

$$0{,}25 b_{i(\text{lógite})} = \frac{1}{\sqrt{2\pi}} b_{i(\text{próbite})}$$

ou, aproximadamente,

$$b_{i(\text{lógite})} = 1{,}6 b_{i(\text{próbite})} \tag{5.60}$$

Aplicando o método de regressão ponderada descrito pelas equações (5.51), (5.52) e (5.53) aos dados da tabela 5.1, obtemos (estimativas dos desvios padrões entre parênteses)

$$\hat{Z} = -6{,}146 + 0{,}830x$$
$$\quad (0{,}283) \quad (0{,}040)$$

A estimativa do recebimento para o qual se espera que metade das famílias tenham automóvel é R$1.642.
Pelo método da máxima verossimilhança obtemos

$$\hat{Z} = -6{,}1813 + 0{,}8350x$$

A estimativa do recebimento para o qual se espera que metade das famílias tenham automóvel é R$1.641.
Sem correção, as estimativas dos desvios padrões das estimativas dos parâmetros são

$$s(b_0) = 0{,}0911 \quad \text{e} \quad s(b_1) = 0{,}0129$$

Mas a soma de quadrados dos desvios ponderados é elevada: 84,25, com 8 graus de liberdade. Fazendo a correção, as estimativas dos desvios padrões das estimativas dos parâmetros passam a ser

$$s(b_0) = 0{,}2955 \quad \text{e} \quad s(b_1) = 0{,}0417$$

É interessante comparar as estimativas dos parâmetros do próbite com as estimativas dos parâmetros do lógite, verificando que a relação (5.49) é aproximadamente válida.
É comum que os modelos de lógite e próbite produzam resultados muito semelhantes. Para o caso do exemplo analisado poderíamos optar pelo lógite, que levou a uma soma de quadrados de desvios ponderados menor.

5.11. Lógite multinomial

Há situações em que é necessário considerar a classificação em mais do que duas categorias. Para facilitar a exposição, vamos admitir que cada observação corresponde a uma pessoa. Vamos indicar as *k* diferentes categorias por E_h , com *h* = 1, 2,...,*k*. A probabilidade de *i*-ésima pessoa pertencer à categoria *h* é indicada por P_{ih}. Deseja-se analisar como a probabilidade P_{ih} depende das características da pessoa. Para isso pode ser usado o modelo de lógite multinomial descrito a seguir. É evidente que o modelo de lógite binomial, analisado anteriormente, é o caso particular em que $k = 2$.

Como se trata de *k* categorias exaustivas e mutualmente exclusivas, temos

$$\sum_{h=1}^{k} P_{ih} = 1 \qquad \text{para todo } i. \tag{5.61}$$

Seja $\mathbf{x}_i'$ o vetor-linha com os valores das variáveis explanatórias para a *i*-ésima pessoa e sejam $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,..., $\boldsymbol{\beta}_k$ os vetores-coluna de parâmetros para as categorias E_1 , E_2 ,..., E_k , respectivamente.

Devido à restrição (5.61), só podemos determinar $k-1$ vetores de parâmetros, sendo que uma categoria deve ser adotada como base. Isso corresponde ao fato de que no lógite

binomial é estimado um único vetor de parâmetros. Sem perda de generalidade, vamos adotar E_1 como categoria base, fazendo $\boldsymbol{\beta}_1 = 0$. Então o modelo de lógite multinomial é expresso pelas seguintes equações:

$$P_{i1} = \frac{1}{1 + \sum_{j=2}^{k} \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_j\right)} \tag{5.62}$$

e

$$P_{ih} = \frac{\exp\left(\mathbf{x}_i' \boldsymbol{\beta}_h\right)}{1 + \sum_{j=2}^{k} \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_j\right)} \qquad \text{para } h = 2, 3, \ldots k \tag{5.63}$$

Note-se, nas expressões (5.62) e (5.63), que as probabilidades P_{ih} (com $h = 1, \ldots, k$) dependem de todos os $k-1$ vetores de parâmetros.

É fácil verificar, a partir de (5.62) e (5.63), que

$$\frac{P_{ih}}{P_{i1}} = \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_h\right) \quad \text{para } h = 2, 3, \ldots, k \tag{5.64}$$

De maneira mais geral, indicando por h e m duas categorias quaisquer, tem-se

$$\frac{P_{ih}}{P_{im}} = \exp\left[\mathbf{x}_i' \left(\boldsymbol{\beta}_h - \boldsymbol{\beta}_m\right)\right] \quad \text{para } h, m = 1, 2, \ldots, k \tag{5.65}$$

A relação (5.65) mostra que no modelo de lógite multinomial a *relação* entre as probabilidades associadas às categorias h e m depende apenas dos vetores $\boldsymbol{\beta}_h$ e $\boldsymbol{\beta}_m$, não sendo afetada por uma redifinição das outras categorias.

Seja X_{mi} uma determinada variável do vetor $\mathbf{x}_i'$ e sejam β_{2m}, β_{3m}, ..., β_{km} os respectivos parâmetros. Se, por exemplo, $\beta_{2m} > 0$, pode-se concluir, de acordo com (5.64), que a relação P_{i2} / P_{i1} cresce em função de X_{mi}. Mas, havendo mais de duas categorias, isso não permite concluir que P_{i2} cresce em função de X_{mi}. É perfeitamente possível que a *relação* P_{i2} / P_{i1} cresça com os valores de P_{i2} e P_{i1} se reduzindo, desde que a redução de P_{i1} seja proporcionalmente maior do que a redução de P_{i2}.[15]

No caso do lógite binomial, com apenas duas categorias, o crescimento da relação P_{i2} / P_{i1} significa, obviamente, que P_{i2} cresce e P_{i1} diminui, já que $P_{i1} + P_{i2} = 1$.

[15] Como exemplo de análise em que isso ocorreu, Ver Hoffmann (2010).

Exercícios

1. Seja P_i a proporção esperada de resultados "favoráveis" em n_i tentativas, quando o valor da variável explanatória é X_i. Considere o modelo

$$P_i = 1 - \frac{1}{e^{\alpha+\beta X_i}} \qquad \text{com} \qquad \beta > 0 \quad \text{e} \quad X_i \geq -\frac{\alpha}{\beta}$$

A proporção observada de resultados "favoráveis" em *k* ensaios, com *k* diferentes valores de X_i, é

$$p_i = \frac{m_i}{n_i} \qquad \text{com} \qquad i = 1, ..., k.$$

a) Mostre que $Y = \ln\frac{1}{1-P}$ é uma função linear de *X*.

b) Determine o fator de ponderação que deve ser utilizado para estimar α e β através de uma regressão ponderada de $y_i = \ln\frac{1}{1-p_i}$ contra X_i.

2. Admite-se que a probabilidade de uma família possuir determinado bem cresce com a renda familiar (X_i) de acordo com a seguinte função:

$$P_i = \frac{X_i}{X_i + \alpha} \quad , \quad \text{com } \alpha > 0$$

Seja n_i o número de famílias com renda familiar X_i em uma amostra de *N* famílias, com

$$N = \sum_{i=1}^{k} n_i$$

Seja m_i (com $m_i \leq n_i$) o número de famílias com renda X_i, na amostra, que possuem aquele bem. Então a proporção de famílias com renda X_i que possuem o bem, na amostra, é

$$p_i = \frac{m_i}{n_i} \qquad \text{É claro que, em geral,} \quad p_i \neq P_i$$

Mostre como o estimador de máxima verossimilhança (*a*) de α pode ser obtido a partir dos valores de X_i, n_i, m_i (com $i = 1, ..., k$). Obtenha, inicialmente, a equação cuja solução é o estimador de máxima verossimilhança. Em seguida mostre como será calculada a correção $\Delta a = a - a_0$, dada uma estimativa preliminar a_0.

3. Admite-se que a probabilidade de uma família possuir determinado bem cresce com a renda familiar (X_i) de acordo com a seguinte função:

$$P_i = \exp\left(-\frac{\beta}{X_i}\right) \quad , \quad \text{com } \beta > 0$$

Seja n_i o número de famílias com renda familiar X_i em uma amostra de *N* famílias, com

$$N = \sum_{i=1}^{k} n_i$$

Seja m_i (com $m_i \leq n_i$) o número de famílias com renda X_i, na amostra, que possuem aquele bem. Então a proporção de famílias com renda X_i que possuem o bem, na amostra, é

$$p_i = \frac{m_i}{n_i}$$ É claro que, em geral, $p_i \neq P_i$

a) Mostre como o estimador de máxima verossimilhança (b) de β pode ser obtido a partir dos valores de X_i, n_i, m_i (com i = 1, ..., k). Obtenha, inicialmente, a equação cuja solução é o estimador de máxima verossimilhança. Em seguida mostre como será calculada a correção $\Delta b = b - b_0$, dada uma estimativa preliminar b_0.

b) Para os dados na tabela abaixo, verifique que $\Delta b = 0$ para b_0 = 6 ln 4.

X_i	n_i	m_i
2	3	0
4	4	0
6	4	1
12	2	2

c) Obtenha a estimativa da renda para a qual se espera que metade das famílias tenham o bem.

d) Determine a estimativa do desvio padrão de b (sem correção).

4. Vamos admitir que a proporção de famílias (P) que possui determinado bem de consumo cresce com o logaritmo da renda familiar (x) de acordo com a função logística. Diferentemente do que acontece no modelo tradicional de lógite, vamos admitir que algumas famílias não adquirem o bem, por maior que seja sua renda, de maneira que a assíntota superior da função tem ordenada $\theta < 1$, isto é

$$P_i = \frac{\theta}{1 + \exp[-(\alpha + \beta x_i)]}$$

Descreva, resumidamente, como você pode estimar os 3 parâmetros desse modelo com base em um conjunto de valores de x_i e p_i (com p_i indicando o valor observado de P_i).

5. Admite-se que a probabilidade de uma família possuir determinado bem cresce com a renda familiar *per capita* (x_i) de acordo com a seguinte função

$$P_i = \beta^{\left(\frac{1}{2}\right)^{x_i}} \text{ , com } 0 < \beta < 1$$

Seja n_i o número de famílias com renda *per capita* igual a x_i em uma amostra de N famílias. Admitindo que haja k diferentes níveis de renda *per capita*, temos

$$N = \sum_{i=1}^{k} n_i$$

Seja m_i (com $m_i \leq n_i$) o número de famílias com renda x_i, na amostra, que possuem aquele bem. Então, a proporção de famílias com renda x_i que possuem o bem, na amostra, é

$$p_i = \frac{m_i}{n_i}$$

É claro que, em geral, $p_i \neq P_i$.

a) Mostre como o estimador de máxima verossimilhança (b) de β pode ser obtido a partir dos valores de x_i, n_i, m_i (com $i = 1, ..., k$).
Obtenha, inicialmente, a equação cuja solução é o estimador de máxima verossimilhança. Em seguida mostre como será calculada a correção $\Delta b = b - b_0$, dada uma estimativa preliminar b_0.

b) Para os dados na tabela abaixo, verifique que $\Delta b = 0$ para $b_0 = \frac{1}{256}$.

x_i	n_i	m_i
1	1000	60
2	1000	260
3	1000	492

c) Obtenha a estimativa da renda para a qual se espera que 80% das famílias tenham o bem.
d) Determine a estimativa do desvio padrão de b (sem correção).
e) Mostre como pode ser obtida a estimativa preliminar (b_o) de b.

6. Admite-se que a probabilidade de uma família possuir determinado bem cresce com o valor da renda familiar *per capita* que excede a linha de pobreza (X_i), de acordo com a seguinte função:

$$P_i = 1 - \frac{\alpha}{X_i + 1} \quad , \qquad \text{com } 0 < \alpha < 1.$$

Seja n_i o número de famílias com renda *per capita* "excedente" X_i em uma amostra de N famílias, com

$$N = \sum_{i=1}^{k} n_i$$

Seja m_i (com $m_i \leq n_i$) o número de famílias com renda *per capita* "excedente" X_i, na amostra, que possuem aquele bem. Então a proporção de famílias com renda *per capita* "excedente" X_i que possuem o bem, na amostra, é

$$p_i = \frac{m_i}{n_i}$$

a) Mostre como o estimador de máxima verossimilhança (a) de α pode ser obtido a partir dos valores de X_i, n_i e m_i (com $i = 1, ..., k$). Obtenha, inicialmente, a equação cuja

solução é o estimador de máxima verossimilhança. Em seguida mostre como será calculada a correção $\Delta a = a - a_0$, dada uma estimativa preliminar a_0.

b) Considere os seguintes dados:

x_i	n_i	m_i
0	50	11
1	50	28
3	50	36
7	50	48

Calcule a correção Δa para $a_0 = 0{,}8$.

c) Escreva a equação da probabilidade estimada de acordo com o método da máxima verossimilhança.

d) Determine a renda *per capita* "excedente" para a qual se estima que metade das famílias tenha o bem.

e) Determine a estimativa do desvio padrão de *a* (sem correção).

f) Descreva uma maneira de obter a estimativa preliminar de *a*.

7. Admite-se que a probabilidade de uma pessoa possuir determinado bem depende de sua renda (X_i) de acordo com a seguinte função:

$$P_i = 1 - \theta^{X_i} \text{ , com } 0 < \theta < 1.$$

Seja n_i o número de pessoas com renda X_i em uma amostra de *N* pessoas, com $N = \sum_{i=1}^{k} n_i$.

Seja m_i (com $m_i \le n_i$) o número de pessoas com renda X_i, na amostra, que possuem aquele bem. Então a proporção de pessoas com renda X_i que possuem o bem, na amostra, é

$$p_i = \frac{m_i}{n_i}$$

a) Mostre como o estimador de máxima verossimilhança ($\hat{\theta}$) de θ pode ser obtido a partir dos valores de X_i, n_i e m_i (com $i = 1, ..., k$). Obtenha, inicialmente, a equação cuja solução é o estimador de máxima verossimilhança. Em seguida mostre como será calculada a correção $\Delta\hat{\theta}$, dada uma estimativa preliminar $\hat{\theta}_0$.

Observação: $\frac{dP_i}{d\theta} = -X_i \theta^{X_i - 1} = -\frac{X_i Q_i}{\theta}$, com $Q_i = 1 - P_i$ e $\frac{dP_i}{dX_i} = -Q_i \ln \theta$.

b) Considere os seguintes dados:

x_i	n_i	m_i
1	1000	220
2	1000	342
3	1000	488

Calcule a correção $\Delta\hat{\theta}$ para $\hat{\theta}_0 = 0{,}8$.

c) Calcule a correção $\Delta\hat{\theta}$ para $\hat{\theta}_0 = 0{,}7$.

d) Determine o valor de *X* para o qual $\hat{P} = 0{,}5$.

e) Determine a estimativa do desvio padrão de $\hat{\theta}$ (sem correção).

f) A estimativa preliminar de θ pode ser obtida fazendo uma regressão de $Z_i = \ln(1 - p_i)$ contra X_i. Qual é o fator de ponderação que deve ser usado nessa regressão? Apresente a fórmula de cálculo de $\hat{\theta}_0$ em função de Z_i, X_i, n_i, p_i e $q_i = 1 - p_i$.

8. Admite-se que a probabilidade de uma família possuir determinado bem cresce com a renda familiar *per capita* (X_i), de acordo com a seguinte função:

$$P_i = \frac{X_i^2}{X_i^2 + \alpha} \quad , \qquad \text{com } \alpha > 0 .$$

Seja n_i o número de famílias com renda *per capita* X_i em uma amostra de *N* famílias, com

$$N = \sum_{i=1}^{k} n_i$$

Seja m_i (com $m_i \le n_i$) o número de famílias com renda *per capita* X_i, na amostra, que possuem aquele bem. Então a proporção de famílias com renda *per capita* X_i, que possuem o bem, na amostra, é

$$p_i = \frac{m_i}{n_i}$$

a) Mostre como o estimador de máxima verossimilhança (*a*) de α pode ser obtido a partir dos valores de X_i, n_i e m_i (com $i = 1, ..., k$). Obtenha, inicialmente, a equação cuja solução é o estimador de máxima verossimilhança. Em seguida mostre como será calculada a correção $\Delta a = a - a_0$, dada uma estimativa preliminar a_0.

b) Considere os seguintes dados:

x_i	n_i	m_i
2	4	0
3	8	2
6	8	4
8	10	6
12	7	6

Calcule a correção Δa para $a_0 = 30$.

c) Calcule a correção Δa para $a_0 = 36$.

d) Obtenha a renda *per capita* para a qual se estima que metade das famílias tenha o bem.

e) Determine a estimativa do desvio padrão de *a* (sem correção).

9. Admite-se que a probabilidade (*P*) de um domicílio ter um microcomputador depende da sua renda *per capita* (*X*), de acordo com o seguinte modelo:

$$P = \frac{1}{1+\exp(-\alpha - \beta X)}$$

São fornecidos os seguintes dados referentes a 500 domicílios, sendo n_i o número de domicílios com renda *per capita* X_i, e p_i a proporção desses domicílios que têm microcomputador:

X_i	n_i	p_i
4	100	0,1
8	100	0,2
15	100	0,5
22	100	0,8
26	100	0,9

a) Obtenha as estimativas de α e β pelo método de Berkson (regressão ponderada).

b) Verifique se há "falta de ajustamento", adotando um nível de significância de 10%.

c) Teste, ao nível de significância de 1%, a hipótese de que $\beta = 0$.

10. Considere os seguintes dados para duas variáveis binárias:

Número de observações conforme valores de *x* e *y*

y	*X*		Total
	0	1	
0	40	60	100
1	60	40	100
Total	100	100	200

a) Determine as estimativas dos parâmetros α e β do modelo linear

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

ou seja, faça a regressão linear simples de y_i contra x_i.

b) Obtenha as estimativas de máxima verossimilhança de α e β do modelo de lógite, no qual a probabilidade de $y_i = 1$ é dada por

$$P_i = \frac{1}{1 + \exp[-(\alpha + \beta x_i)]}$$

ou

$$\ln \frac{P_i}{1 - P_i} = \alpha + \beta x_i$$

Observação: nesse caso especial, a determinação das estimativas de máxima verossimilhança dos parâmetros de um modelo de lógite **não** exige processo iterativo.

11. Temos 30 observações para uma variável binária Y_i, classificadas em 3 categorias. Para distinguir as 3 categorias são usadas duas variáveis binárias (Z_{1i} e Z_{2i}), como mostra a tabela a seguir:

Categoria	Z_{1i}	Z_{2i}	Nº de observações		
			com $Y_i = 0$	com $Y_i = 1$	Total
A	0	0	8	2	10
B	1	0	4	6	10
C	0	1	2	8	10

Obtenha as estimativas de máxima verossimilhança de α, β_1 e β_2 do modelo de lógite no qual a probabilidade de $Y_i = 1$ é dada por

$$P_i = \frac{1}{1+\exp[-(\alpha+\beta_1 Z_{1i}+\beta_2 Z_{2i})]}$$

ou

$$\ln\frac{P_i}{1-P_i} = \alpha+\beta_1 Z_{1i}+\beta_2 Z_{2i}$$

Observação: nesse caso especial, a determinação das estimativas de máxima verossimilhança dos parâmetros de um modelo de lógite **não** exige processo iterativo.

12. Admite-se que a probabilidade de uma família possuir determinado bem cresce com o logaritmo da renda familiar *per capita* (X_i), de acordo com a seguinte função:

$$P_i = \alpha + \frac{\beta}{X_i} \text{ com } 0 < \alpha < 1, \ \beta < 0 \text{ e } X_i \geq -\frac{\beta}{\alpha}$$

Seja n_i o número de famílias cujo logaritmo da renda *per capita* é igual a X_i, em uma amostra de *N* famílias, com

$$N = \sum_{i=1}^{k} n_i$$

Seja m_i o número de famílias que possuem o bem entre aquelas cujo logaritmo da renda *per capita* é igual a X_i. Então a proporção de famílias que possui o bem, na amostra, para esse nível de renda, é

$$p_i = \frac{m_i}{n_i}$$

a) Obtenha o sistema de duas equações cuja solução fornece as estimativas de α e β (*a* e *b*) pelo método da máxima verossimilhança, com base em uma amostra com *k* valores de X_i, n_i e m_i.

b) Verifique se as equações são satisfeitas para os valores da tabela abaixo, com $a = 0{,}8$ e $b = -2$

X_i	n_i	m_i
5	200	84
10	200	108
20	200	147

c) Faça um teste de qui-quadrado para verificar se o ajustamento do modelo aos dados é bom, adotando um nível de significância de 5%.

d) Calcule as estimativas dos desvios padrões de *a* e de *b*.

13. Considere os seguintes dados para duas variáveis binárias:

Número de observações			
y	*X*		Total
	0	1	
0	4	27	31
1	16	9	25
Total	20	36	56

a) Obtenha as estimativas de máxima verossimilhança de α e β do modelo de lógite, no qual a probabilidade de $y_i = 1$ é dada por

$$P_i = \frac{1}{1+\exp[-(\alpha+\beta x_i)]} \qquad \text{ou} \qquad \ln\frac{P_i}{1-P_i} = \alpha + \beta x_i$$

b) Obtenha a estimativa da variância assintótica da estimativa de β e teste, ao nível de significância de 1%, a hipótese de que $\beta = 0$.

Observação: nesse caso especial, a determinação das estimativas de máxima verossimilhança dos parâmetros de um modelo de lógite **não** exige processo iterativo.

14. Admite-se que a probabilidade de um empregado ser sindicalizado depende de sua escolaridade (x_i), de acordo com a seguinte função:

$$P_i = \frac{1}{1+\alpha 0{,}5^{x}}$$

Seja n_i o número de empregados com escolaridade x_i em uma amostra de *N* empregados, com $N = \sum n_i$. Seja m_i (com $m_i \le n_i$) o número de empregados com escolaridade x_i, na amostra, que são sindicalizados.

a) Mostre como o estimador de máxima verossimilhança (*a*) de α pode ser obtido a partir dos valores de x_i, n_i e m_i (com *i* = 1, ..., *k*). Obtenha, inicialmente, a equação cuja solução é o estimador de máxima verossimilhança. Em seguida mostre como pode ser calculada a correção Δa, dada a estimativa preliminar a_0.

Considere os seguintes dados:

x_i	n_i	m_i
1	90	13
2	90	15
3	90	27
4	90	48

b) Calcule a correção Δa para $a_0 = 16$.

c) Calcule a soma de quadrados dos desvios ponderados e verifique se o modelo é apropriado, considerando um nível de significância de 5%.

d) Determine a estimativa do desvio padrão de *a* (sem correção).

15. Uma variável binária Y_i é observada em uma amostra com 3 categorias de pessoas, como mostra a tabela abaixo. Para distinguir as 3 categorias são utilizadas duas variáveis binárias, Z_{1i} e Z_{2i}.

Categoria	Z_1	Z_2	Número de observações (pessoas)		
			Com $Y = 1$	Com $Y = 0$	Total
A	0	0	m_0	$n_0 - m_0$	n_0
B	1	0	m_1	$n_1 - m_1$	n_1
C	0	1	m_2	$n_2 - m_2$	n_2

Define-se $p_0 = \frac{m_0}{n_0}$, $p_1 = \frac{m_1}{n_1}$ e $p_2 = \frac{m_2}{n_2}$

Considere o modelo de lógite no qual a probabilidade de $Y = 1$ é dada por:

$$P = \frac{1}{1 + \exp[-(\alpha + \beta_1 Z_1 + \beta_2 Z_2)]} \qquad \text{ou} \qquad \ln = \frac{P}{1 - P} = \alpha + \beta_1 Z_1 + \beta_2 Z_2$$

a) Obtenha as expressões que fornecem as estimativas de máxima verossimilhança de α, β_1 e β_2 (a, b_1 e b_2, respectivamente) em função de p_0, p_1 e p_2.

A seguir, admita que $n_0 = 72$, $n_1 = 40$, $n_2 = 72$, $m_0 = 12$, $m_1 = 20$ e $m_2 = 60$.

b) Determine os valores de a, b_1 e b_2.

c) Determine as estimativas das variâncias assintóticas de a, b_1 e b_2.

16. Considere um modelo de lógite no qual se inclui x e seu quadrado como variáveis explanatórias:

$$P = \frac{1}{1+\exp\left[-(\alpha+\beta x)+\gamma x^2\right]}$$

a) Deduza a expressão para $\frac{dP}{dx}$.

b) De acordo com a expressão (5.45) da apostila, temos

$$\frac{\partial P}{\partial x_1} = \beta_i P(1-P)$$

e o efeito marginal de x_i sobre P tem sempre o mesmo sinal que β_1. Pode-se dizer, então, que o sinal do efeito marginal não depende do valor das variáveis explanatórias. Isso vale para o efeito marginal obtido no item (a)? (comentar sumariamente).

17. Um modelo de lógite multinomial para 3 categorias pode ser apresentado por

$$P_{ih} = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_h)}{1+\sum_{j=1}^{2}\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)} \quad \text{para } h = 0, 1 \text{ ou } 2$$

com $\boldsymbol{\beta}_0 = \mathbf{0}$. Então $\exp(\mathbf{x}_i'\boldsymbol{\beta}_0) = 1$. Isso significa que a categoria $h = 0$ é adotada como base.

É considerada apenas uma variável explanatória (x) e seu quadrado, de maneira que

$$\mathbf{x}_i' = \begin{bmatrix} 1 & x_i & x_i^2 \end{bmatrix}$$

Admite-se que *x* pode variar de zero a 10 e que

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 3 \\ -0{,}2 \\ 0 \end{bmatrix} \text{ e } \boldsymbol{\beta}_2 = \begin{bmatrix} 0 \\ 1 \\ -0{,}1 \end{bmatrix}$$

Calcule os valores de P_0 , P_1 e P_2 para *x* = 0, *x* = 5 e *x* = 10.

18. Admite-se que a probabilidade de um agricultor utilizar determinada técnica depende de sua escolaridade (x_i), de acordo com a seguinte função:

$$P_i = \exp[-\exp(2 - \beta x_i)]$$

Para uma amostra de *N* agricultores é informada a sua escolaridade $(x_i, \text{com } i = 1,...,N)$ e o fato de usar $(Y_i = 1)$ ou não usar $(Y_i = 0)$ a técnica em questão.

a) Com base no método da máxima verossimilhança, obtenha a equação cuja solução é o estimador (*b*) de β.

b) Dada uma estimativa preliminar b_0 de β, obtenha a expressão para a correção (Δb) a ser feita nessa estimativa.

c) Com base na amostra os valores 4 observações apresenta na tabela ao lado, calcule a correção Δb para $b_0 = 1$

x_j	Y_j
1	0
2	0
4	1
5	1

Respostas

1. a) $\ln \dfrac{1}{1 - P_i} = \alpha + \beta X_i$

b) Com $y_i = \ln \frac{1}{1-p_i}$, verifica-se que $V(y_i) = \frac{P_i}{n_i Q_i}$

Então o fator de ponderação é $\frac{n_i q_i}{p_i} = \frac{n_i(1-p_i)}{p_i}$

2. A verossimilhança da amostra é

$$£ = \prod_{i=1}^{k} \binom{n_i}{m_i} P_i^{m_i} (1-P_i)^{n_i - m_i}$$

com $P_i = \frac{X_i}{X_i + \alpha}$ e $1 - P_i = \frac{\alpha}{X_i + \alpha} = Q_i$

Então $\ln £ = \sum_i \left[\ln \binom{n_i}{m_i} + m_i \ln P_i + (n_i - m_i) \ln Q_i \right]$

e $\frac{d \ln £}{d\alpha} = \sum_i \frac{n_i (p_i - P_i)}{P_i Q_i} \cdot \frac{dP_i}{d\alpha}$

Como $\frac{dP_i}{d\alpha} = -\frac{P_i}{X_i + \alpha} = -\frac{P_i Q_i}{\alpha}$, segue-se que

$$\frac{d \ln £}{d\alpha} = \frac{1}{\alpha} \sum_i n_i (P_i - p_i)$$

O estimador de máxima verossimilhança é a solução da equação

$$\sum_i n_i (\hat{P}_i - p_i) = 0 \quad , \quad \text{com} \quad \hat{P}_i = \frac{X_i}{X_i + a}$$

Dada uma estimativa preliminar a_0, a correção é dada por

$$\Delta a = a_0 \frac{\sum n_i (\hat{P}_{0i} - p_i)}{\sum n_i (\hat{P}_{0i} \hat{Q}_{0i} + \hat{P}_{0i} - p_i)}$$

onde $\hat{P}_{0i} = \frac{X_i}{X_i + a_0}$ e $\hat{Q}_{0i} = 1 - \hat{P}_{0i} = \frac{a_0}{X_i + a_0}$

3. a) $\sum_i \frac{n_i (\hat{P}_i - p_i)}{X_i (1 - \hat{P}_i)} = 0$, com $\hat{P}_i = \exp\left(-\frac{b}{X_i} \right)$

$$\Delta b = \frac{\sum_i \frac{n_i(\hat{P}_{0i} - p_i)}{X_i(1-\hat{P}_{0i})}}{\sum_i \frac{n_i \hat{P}_{0i}(1-p_i)}{X_i^{\,2}(1-\hat{P}_{0i})^2}}$$

b)

X_i	n_i	p_i	$\hat{P}_{0i}$
2	3	0	1/64
4	4	0	1/8
6	4	1/4	1/4
12	2	1	1/2

c) $X = 12$

d) Verifica-se que $-E\left(\frac{d^2 \ln £}{d\beta^2}\right) = \sum_i \frac{n_i P_i}{X_i^{\,2}(1-P_i)}$; $s(b) = 3{,}186$

4. a) Examinando um gráfico estabelece-se uma estimativa preliminar (θ_0) para θ.

b) Obtêm-se estimativas preliminares de α e β fazendo uma regressão linear simples ponderada de

$$y_i = \ln \frac{p_i}{\theta_0 - p_i} \quad \text{contra} \quad x_i$$

com fatores de ponderação $\frac{n_i p_i (\theta_0 - p_i)^2}{\theta_0^2 (1-p_i)}$ ou $\frac{n_i p_i (\theta_0 - p_i)^2}{1-p_i}$

c) As estimativas de máxima verossimilhança para θ, α e β são a solução do sistema

$$\begin{cases} \sum_i \frac{(p_i - P_i)n_i}{1-P_i} = 0 \\ \sum_i \frac{(p_i - P_i)n_i(\theta - P_i)}{1-P_i} = 0 \\ \sum_i \frac{(p_i - P_i)(\theta - P_i)n_i x_i}{1-P_i} = 0 \end{cases}$$

com $$P_i = \frac{\theta}{1 + \exp\{-(\alpha + \beta x_i)\}}$$

Como se trata de um sistema não linear, a solução deverá ser obtida através de um processo iterativo, como no método de Newton.

5. a) $\dfrac{d\ln £}{d\beta} = \dfrac{1}{\beta}\sum_i (p_i - P_i)\dfrac{n_i}{Q_i}\left(\dfrac{1}{2}\right)^{x_i}$

$$\sum_i \frac{n_i(p_i - \hat{P}_i)}{1-\hat{P}_i}\left(\frac{1}{2}\right)^{x_i} = 0 \quad , \quad \text{onde} \quad \hat{P}_i = b^{\left(\frac{1}{2}\right)^{x_i}}$$

$$\Delta b = \frac{b_0 \sum_i (p_i - \hat{P}_{0i})\dfrac{n_i}{1-\hat{P}_{0i}}\left(\dfrac{1}{2}\right)^{x_i}}{\sum_i \dfrac{(1-p_i)\hat{P}_{0i}}{(1-\hat{P}_{0i})^2} n_i \left(\dfrac{1}{2}\right)^{2x_i}}$$

b)

x_i	n_i	$p_i - \hat{P}_i$	$\left(\frac{1}{2}\right)^{x_i}$	$1-\hat{P}_i$
1	1000	–0,0025	½	15/16
2	1000	0,01	¼	3/4
3	1000	–0,008	1/8	1/2

$$\sum_{i=1}^{3} (p_i - \hat{P}_i)\frac{n_i}{1-\hat{P}_i}\left(\frac{1}{2}\right)^{x_i} = 0$$

c) 4,635 d) 0,0005359

e) $\ln b_0 = \dfrac{\sum_i \left(\frac{1}{2}\right)^{x_i} (\ln p_i)\dfrac{n_i p_i}{1-p_i}}{\sum_i \left(\frac{1}{2}\right)^{2x_i} \dfrac{n_i p_i}{1-p_i}}$

6. a) $\dfrac{d\ln £}{d\alpha} = \dfrac{1}{\alpha}\sum_i \left(n_i - \dfrac{m_i}{P_i}\right)$

$$\Delta a = \frac{a_0 \sum_i \left(n_i - \dfrac{m_i}{\hat{P}_{0i}}\right)}{\sum_i \dfrac{m_i \hat{Q}_{0i}}{\hat{P}_{0i}^2}}$$

b) $\Delta a = 0$ para $a_0 = 0{,}8$. Então, $a = 0{,}8$

c) $\hat{P}_i = 1 - \dfrac{0{,}8}{X_i + 1}$ d) $X^* = 0{,}6$

e) $\hat{V}(a) = \dfrac{a^2}{\sum_i \dfrac{n_i \hat{Q}_i}{\hat{P}_i}} = 0{,}002546$, $s(a) = 0{,}050456$

f) $a_0 = \dfrac{\sum_i \dfrac{n_i}{(X_i + 1) p_i}}{\sum_i \dfrac{n_i}{(X_i + 1)^2 p_i q_i}}$

7. a) $\dfrac{d \ln £}{d\theta} = -\dfrac{1}{\theta} \sum_i (m_i - n_i P_i) \dfrac{X_i}{P_i}$

$$\sum_i \left(\frac{m_i}{\hat{P}_i} - n_i \right) X_i = 0 \quad \text{com} \quad \hat{P}_i = 1 - \hat{\theta}^{X_i}$$

$$\Delta\hat{\theta} = -\hat{\theta}_0 \frac{\sum_i \left(\dfrac{m_i}{\hat{P}_{0i}} - n_i \right) X_i}{\sum_i \dfrac{m_i \hat{Q}_{0i} X_i^2}{\hat{P}_{0i}^2}} \quad \text{com} \quad \hat{Q}_{0i} = \hat{\theta}_0^{X_i} \text{ e } \hat{P}_{0i} = 1 - \hat{Q}_{0i}$$

b) Para $\hat{\theta}_0 = 0{,}8$, obtemos $\Delta\hat{\theta} = 0$

c) Para $\hat{\theta}_0 = 0{,}7$, obtemos $\Delta\hat{\theta} = 0{,}1527$ ou $\Delta\hat{\theta} = 0{,}1254$

d) $X = \dfrac{\ln 0{,}5}{\ln 0{,}8} = 3{,}106$

e) $\dfrac{d^2 \ln £}{d\theta^2} = -\dfrac{1}{\theta^2} \sum_i \dfrac{n_i Q_i X_i^2}{P_i} + [\text{termos com fator } (m_i - n_i P_i)]$

$$-E\left(\frac{d^2 \ln £}{d\theta^2} \right) = \frac{1}{\theta^2} \sum_i \frac{n_i Q_i X_i^2}{P_i} \quad \text{e} \quad V(\hat{\theta}) = \frac{\theta^2}{\sum_i \dfrac{n_i Q_i X_i^2}{P_i}}$$

$\hat{V}(\hat{\theta}) = 0{,}000031$, $s(\hat{\theta}) = 0{,}00558$

f) Como $\hat{Q}_i = \hat{\theta}^{X_i}$, temos $\ln \hat{Q}_i = (\ln \hat{\theta}) X_i$

De $Z_i = \ln q_i$ obtemos $V(Z_i) \approx \dfrac{P_i}{n_i Q_i}$

Então o fator de ponderação é $\dfrac{n_i q_i}{p_i}$ e $\hat{\theta}_0 = \exp\left(\dfrac{\sum_i X_i Z_i \dfrac{n_i q_i}{p_i}}{\sum_i X_i^2 \dfrac{n_i q_i}{p_i}} \right)$

8. $P_i = \dfrac{X_i^2}{X_i^2 + \alpha}$ e $Q_i = \dfrac{\alpha}{X_i^2 + \alpha}$ $\dfrac{dP_i}{d\alpha} = -\dfrac{P_i}{X_i^2 + \alpha} = -\dfrac{P_i Q_i}{\alpha}$

$\dfrac{d \ln £}{d\alpha} = -\dfrac{1}{\alpha} \sum n_i (p_i - P_i)$ $\qquad$ $\dfrac{d^2 \ln £}{d\alpha^2} = \dfrac{1}{\alpha^2} \sum n_i (p_i - P_i) - \dfrac{1}{\alpha^2} \sum n_i P_i Q_i$

a) O estimador de máxima verossimilhança (a) deve satisfazer a equação

$-\dfrac{1}{a} \sum n_i (p_i - \hat{P}_i) = 0$, com $p_i = \dfrac{m_i}{n_i}$ e $\hat{P}_i = \dfrac{X_i^2}{X_i^2 + a}$

Para uma estimativa preliminar a_0, a correção é $\Delta a = \dfrac{a_0 \sum (m_i - n_i \hat{P}_{0i})}{\sum [m_i - n_i \hat{P}_{0i} (1 + \hat{Q}_{0i})]}$

b) Para $a_0 = 30$ a correção é $\Delta a = \dfrac{30 \cdot 1{,}281993}{8{,}272547} = 4{,}649$

c) Para $a_0 = 36$ a correção é $\Delta a = 0$. Então a = 36.

d) $\hat{P} = 0{,}5$ para X = 6.

e) $\hat{V}(a) = \dfrac{a^2}{\sum n_i \hat{P}_i \hat{Q}_i} = \dfrac{36^2}{7{,}064} = 183{,}465$ e $s(a)$ = 13,545

9. a) a = –2,9855 e b = 0,1990

b) $\chi^2 = 0{,}00265$ (com 3 graus de liberdade), não-significativo.

c) t = 409,7, significativo.

10. a) $\hat{y} = 0{,}6 - 0{,}2x$

b) $\hat{\alpha} = \ln \dfrac{0{,}6}{1 - 0{,}6} = 0{,}405465$

$\hat{\beta} = \ln \dfrac{\dfrac{0{,}4}{1 - 0{,}4}}{\dfrac{0{,}6}{1 - 0{,}6}} = -0{,}810930$

11. $\hat{\alpha} = \ln \dfrac{0,2}{1-0,2} = -1,3863$,

$$\hat{\beta}_1 = \ln \frac{\dfrac{0,6}{1-0,6}}{\dfrac{0,2}{1-0,2}} = 1,7918 \quad \text{e} \quad \hat{\beta}_2 = \ln \frac{\dfrac{0,8}{1-0,8}}{\dfrac{0,2}{1-0,2}} = 2,7726$$

12.a) Com $\hat{P}_i = a + \dfrac{b}{X_i}$ e $\hat{Q}_i = 1 - \hat{P}_i$, o sistema é

$$\sum_i \frac{m_i - n_i \hat{P}_i}{\hat{P}_i \hat{Q}_i} = 0$$

$$\sum_i \frac{m_i - n_i \hat{P}_i}{\hat{P}_i \hat{Q}_i X_i} = 0$$

b) Sim, são satisfeitas.

c) $\sum_i \dfrac{n_i (p_i - \hat{P}_i)^2}{\hat{P}_i \hat{Q}_i} = 4,5$, significativo (o valor crítico é 3,841).

Conclui-se que o modelo não é apropriado.

d) Sem correção: $s(a) = 0,04055$ e $s(b) = 0,3127$.

Com correção: $s(a) = 0,08602$ e $s(b) = 0,6633$.

13.a) $\hat{\alpha} = 1,3863$ e $\hat{\beta} = -2,4849$.

b) $\hat{V}(\hat{\beta}) = 0,4606$, $Z = -3,66$, significativo ($Z_0 = 2,576$).

14.a) $\sum_{i=1}^{k} (m_i - n_i \hat{P}_i) = 0$, com $\hat{P}_i = \dfrac{1}{1 + a 0,5^x}$

$$\Delta a = \frac{-a_0 \sum (m_i - n_i \hat{P}_{0i})}{\sum n_i \hat{P}_{0i} \hat{Q}_{0i}}$$

b) $\Delta a = 0$

c) $\chi_3^2 = 2,4875$, não-significativo (o valor crítico ara um nível de significância de 5% é 7,815);

d) 1,9726.

15. a) $a = \ln \dfrac{p_0}{q_0}$ $b_1 = \ln \dfrac{p_1}{q_1} - a$ e $b_2 = \ln \dfrac{p_2}{q_2} - a$

b) $a = -1,6094$, $b_1 = 1,6094$ e $b_2 = 3,2189$

c) $\hat{V}(a) = 0{,}1$, $\hat{V}(b_1) = 0{,}2$ $\hat{V}(b_2) = 0{,}2$

16. a) $\dfrac{dP}{dX} = PQ(\beta + 2\gamma x)$

b) É claro que o sinal de $\beta + 2\gamma x$ pode mudar dependendo dos valores dos parâmetros β e γ e do intervalo de variação de *x*. A expressão (5.45) só é válida quando uma variável x_i entra apenas linearmente, em um único elemento de vetor $\mathbf{x}_i'$.

17.

Prob.	$x = 0$	$x = 5$	$x = 10$
P_0	0,0453	0,0486	0,2119
P_1	0,9094	0,3592	0,5761
P_2	0,0453	0,5922	0,2119

18. a) Com $\hat{P}_i = \exp[-\exp(2 - bx_i)]$, devemos ter $\sum (Y_i - \hat{P}_i)\dfrac{x_i}{1-\hat{P}_i}\exp(2 - bx_i) = 0$

b) Com $\hat{P}_{0i} = \exp[-\exp(2 - b_0 x_i)]$, $N = \sum (Y_i - \hat{P}_{0i})\dfrac{x_i}{1-\hat{P}_{0i}}\exp(2 - b_0 x_i)$,

$$D_1 = \sum (Y_i - \hat{P}_{0i})\frac{x_i}{1-\hat{P}_{0i}}\exp(2 - b_0 x_i)\left[1 - \frac{\hat{P}_{0i}}{1-\hat{P}_{0i}}\exp(2 - b_0 x_i)\right] \quad \text{e}$$

$$D_2 = \sum \frac{x_i^2 \hat{P}_{0i}}{1-\hat{P}_{0i}}\exp[2(2 - b_0 x_i)] \text{ , } \Delta b = \frac{N}{D_1 + D_2}$$

c) $\Delta b = -0{,}11024$

6. COMPONENTES PRINCIPAIS E ANÁLISE FATORIAL

6.1. Introdução

Análise fatorial (*factor analysis*) é um conjunto de métodos estatísticos que, em certas situações, permite "explicar" o comportamento de um número relativamente grande de variáveis observadas em termos de um número relativamente pequeno de variáveis latentes ou *fatores*. Admite-se que a relação entre variáveis observadas e fatores é *linear*. A análise fatorial será encarada aqui como uma técnica estatística exploratória destinada a "resumir" as informações contidas em um conjunto de variáveis em um conjunto de fatores, com o número de fatores sendo geralmente bem menor do que o número de variáveis observadas.
A análise fatorial foi desenvolvida inicialmente dentro da psicologia, como uma tentativa de estabelecer um modelo matemático para explicar habilidades e comportamentos humanos. Os resultados de um grande número de testes aplicados a um conjunto de pessoas eram submetidos à análise fatorial para identificar um número pequeno de características básicas ou fatores da mente.

A análise fatorial pode ser utilizada, por exemplo, para obter medidas do grau de modernização da agricultura nas Microrregiões Homogêneas (MRH) do país. Inicialmente são determinados, em cada MRH, os valores de um conjunto de variáveis indicadoras de modernização (número de tratores por hectare ou por unidade de mão-de-obra, uso de energia elétrica por hectare, uso de herbicidas por hectare etc.). A seguir essas variáveis são submetidas à análise fatorial visando obter uma medida sintética do grau de modernização. Em várias pesquisas, partindo-se de um conjunto com cerca de 30 variáveis observadas foi possível extrair dois fatores que se mostram associados com as duas dimensões básicas da modernização da agricultura: a produtividade da terra e a produtividade do trabalho.[16]

A análise de componentes principais é uma técnica estatística estreitamente associada com a análise fatorial. Dado um conjunto de variáveis, os componentes principais são combinações lineares dessas variáveis, construídas de maneira a "explicar" o máximo da *variância* das variáveis originais. Na seção 6.6 veremos que no modelo básico de análise fatorial a definição dos fatores é feita visando, precipuamente, explicar as *correlações* entre as variáveis originais. Vamos admitir que dispomos de L observações para n variáveis. No espaço L-dimensional das observações as n variáveis correspondem a n vetores. Um grupo de variáveis fortemente correlacionadas entre si corresponde a um feixe de vetores. A análise fatorial (ou a análise de componentes principais) permite detectar esses feixes. Se houver um número substancial de variáveis formando um desses feixes, deverá ser obtido um fator altamente correlacionado com as variáveis que formam o feixe.
A análise de componentes principais é formalmente apresentada e ilustrada nas seções 6.2 a 6.5. A análise fatorial, incluindo a rotação dos fatores, é exposta nas seções 6.6 a 6.10.

[16] Ver Hoffmann e Kageyama (1985), Hoffmann e Kassouf (1989) e Hoffmann (1992).

6.2. A matriz das correlações simples e a determinação do primeiro componente principal

Vamos admitir que dispomos de L observações para n variáveis. Seja X_{ij} (com $i = 1, ..., n$ e $j = 1, ..., L$) a j-ésima observação da i-ésima variável.

A média da i-ésima variável é $\overline{X}_i = \frac{1}{L}\sum_j X_{ij}$

Fazemos
$$x_{ij} = \frac{X_{ij} - \overline{X}_i}{\sqrt{\sum_j \left(X_{ij} - \overline{X}_i\right)^2}}. \qquad (6.1)$$

Com essa transformação temos $\sum x_{ij}^2 = 1$, isto é, no espaço L-dimensional das observações, o vetor $\mathbf{x}_i$, para cada variável, tem módulo igual a 1. Após essa transformação todas as variáveis têm a mesma variância e a participação de uma variável na determinação dos componentes principais irá depender apenas das suas correlações com as demais variáveis.

Definimos a matriz
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1L} \\ x_{21} & x_{22} & \dots & x_{2L} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nL} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \dots \\ \mathbf{x}_n' \end{bmatrix}$$

Verifica-se que a matriz $n \times n$ das correlações simples entre as variáveis é dada por
$$\mathbf{R} = \mathbf{X}\mathbf{X}' \qquad (6.2)$$

Vamos considerar uma combinação linear das n variáveis transformadas

$$g_{1j} = c_{11}x_{1j} + c_{12}x_{2j} + \dots + c_{1n}x_{nj} \qquad (j = 1, ..., L)$$

ou, em notação matricial,
$$\mathbf{g}_1' = \mathbf{c}_1'\mathbf{X}, \qquad (6.3)$$

onde $\mathbf{g}_1'$ é um vetor-linha com L elementos e $\mathbf{c}_1'$ é um vetor-linha com os n coeficientes da combinação linear considerada.

Por definição, o *primeiro componente principal* de $\mathbf{X}$ é a combinação linear $\mathbf{g}_1' = \mathbf{c}_1'\mathbf{X}$ com variância máxima, sujeita à restrição $\mathbf{c}_1'\mathbf{c}_1 = 1$.

Note-se que não teria sentido definir o primeiro componente como a combinação linear $\mathbf{g}_1' = \mathbf{c}_1'\mathbf{X}$ com variância máxima, sem impor uma restrição aos coeficientes em $\mathbf{c}_1'$, pois nesse caso a variância da combinação linear poderia crescer ilimitadamente.

Uma vez que g_{1j} é uma combinação linear de variáveis com média zero, sua média também é igual a zero. Então sua variância é dada por

$$V(g_1) = \frac{1}{L}\mathbf{g}_1'\mathbf{g}_1 .$$

Estamos usando a notação para variância da população, apesar de se tratar, geralmente, de uma amostra.

Lembrando (6.3) e (6.2), segue-se que

$$V(g_1) = \frac{1}{L}\mathbf{c}_1'\mathbf{X}\mathbf{X}'\mathbf{c}_1 = \frac{1}{L}\mathbf{c}_1'\mathbf{R}\mathbf{c}_1 . \tag{6.4}$$

Portanto, para obter o primeiro componente principal de $\mathbf{X}$ devemos determinar o vetor $\mathbf{c}_1$, com $\mathbf{c}_1'\mathbf{c}_1 = 1$, que maximiza $\mathbf{c}_1'\mathbf{R}\mathbf{c}_1$. De acordo com o método de Lagrange, formamos a função

$$\phi = \mathbf{c}_1'\mathbf{R}\mathbf{c}_1 - \lambda(\mathbf{c}_1'\mathbf{c}_1 - 1)$$

e obtemos

$$\frac{d\phi}{d\mathbf{c}_1} = 2\mathbf{R}\mathbf{c}_1 - 2\lambda\mathbf{c}_1 = \mathbf{0}$$

ou
$$\mathbf{R}\mathbf{c}_1 = \lambda\mathbf{c}_1 . \tag{6.5}$$

Substituindo esse resultado em (6.4) e lembrando que $\mathbf{c}_1'\mathbf{c}_1 = 1$, obtemos

$$V(g_1) = \frac{\lambda}{L} . \tag{6.6}$$

De (6.5) obtemos

$$(\mathbf{R} - \lambda\mathbf{I})\mathbf{c}_1 = \mathbf{0} . \tag{6.7}$$

Este é um sistema de *n* equações lineares homogêneas com incógnitas c_{1i} (i = 1, ..., n). Para que haja uma solução não-trivial devemos ter

$$|\mathbf{R} - \lambda\mathbf{I}| = 0 , \tag{6.8}$$

que é denominada *equação característica*. Admitindo que **R** seja uma matriz não-singular, essa equação tem *n* raízes reais positivas, denominadas *autovalores* ou *raízes características* de **R**, que passamos a indicar por λ_i.

Para obter o primeiro componente principal devemos, de acordo com (6.6), escolher o maior autovalor de **R**, que indicamos por λ_1.

A seguir, fazendo $\lambda = \lambda_1$ em (6.7) e lembrando que $\mathbf{c}_1'\mathbf{c}_1 = 1$, podemos obter $\mathbf{c}_1$, que é o *autovetor* ou *vetor característico* correspondente à maior raiz característica de **R**.

Com $\lambda = \lambda_1$ as relações (6.5) e (6.6) ficam:

$$\mathbf{R}\mathbf{c}_1 = \lambda_1\mathbf{c}_1 \tag{6.9}$$

e

$$V(g_1) = \frac{\lambda_1}{L} . \tag{6.10}$$

6.3. Os n componentes principais

O *segundo componente principal* de $\mathbf{X}$ é a combinação linear $\mathbf{g}_2' = \mathbf{c}_2'\mathbf{X}$ com variância máxima, sujeita às restrições $\mathbf{c}_2'\mathbf{c}_2 = 1$ e $\mathbf{g}_2'\mathbf{g}_1 = 0$. Esta última restrição significa que $\mathbf{g}_1$ e $\mathbf{g}_2$ são vetores ortogonais entre si, ou seja, que g_1 e g_2 são variáveis não-correlacionadas. Como $\mathbf{g}_1' = \mathbf{c}_1'\mathbf{X}$ e $\mathbf{g}_2' = \mathbf{c}_2'\mathbf{X}$, a condição de ortogonalidade fica

$$\mathbf{g}_2'\mathbf{g}_1 = \mathbf{c}_2'\mathbf{X}\mathbf{X}'\mathbf{c}_1 = 0$$

Lembrando (6.2) e (6.9) verifica-se que os dois primeiros componentes serão ortogonais entre si se

$$\mathbf{c}_2'\mathbf{c}_1 = 0$$

Então o *segundo componente principal* pode ser definido como a combinação linear $\mathbf{g}_2' = \mathbf{c}_2'\mathbf{X}$ com variância máxima, sujeita às restrições $\mathbf{c}_2'\mathbf{c}_2 = 1$ e $\mathbf{c}_2'\mathbf{c}_1 = 0$.

É fácil verificar que a variância de g_2 é dada por

$$V(g_2) = \frac{1}{L}\mathbf{g}_2'\mathbf{g}_2 = \frac{1}{L}\mathbf{c}_2'\mathbf{R}\mathbf{c}_2 \tag{6.11}$$

De acordo com o método de Lagrange, para maximizar $\mathbf{c}_2'\mathbf{R}\mathbf{c}_2$, com $\mathbf{c}_2'\mathbf{c}_2 = 1$ e $\mathbf{c}_2'\mathbf{c}_1 = 0$, formamos a função

$$\varphi = \mathbf{c}_2'\mathbf{R}\mathbf{c}_2 - \lambda_2(\mathbf{c}_2'\mathbf{c}_2 - 1) - \omega\,\mathbf{c}_2'\mathbf{c}_1$$

e obtemos

$$\frac{d\phi}{d\mathbf{c}_2} = 2\mathbf{R}\mathbf{c}_2 - 2\lambda_2\mathbf{c}_2 - \omega\,\mathbf{c}_1 = \mathbf{0} \tag{6.12}$$

Pré-multiplicando por $\mathbf{c}_1'$ e lembrando que $\mathbf{c}_2'\mathbf{c}_1 = 0$ e $\mathbf{c}_1'\mathbf{c}_1 = 1$, obtemos

$$2\mathbf{c}_1'\mathbf{R}\mathbf{c}_2 = \omega$$

Uma vez que, de acordo com (6.9), temos $\mathbf{c}_1'\mathbf{R} = \lambda_1\mathbf{c}_1'$, segue-se que

$$\omega = 2\lambda_1\mathbf{c}_1'\mathbf{c}_2 = 0$$

Substituindo esse resultado em (6.12) obtemos

$$\mathbf{R}\mathbf{c}_2 = \lambda_2\mathbf{c}_2 \tag{6.13}$$

ou

$$(\mathbf{R} - \lambda_2\mathbf{I})\mathbf{c}_2 = \mathbf{0}\ ,$$

mostrando que λ_2 é um autovalor de $\mathbf{R}$ e $\mathbf{c}_2$ é o correspondente autovetor.

Substituindo (6.13) em (6.11) e lembrando que $\mathbf{c}_2'\mathbf{c}_2 = 1$, obtemos

$$V(g_2) = \frac{\lambda_2}{L}$$

Portanto, para maximizar $V(g_2)$ com $\mathbf{c}_2'\mathbf{c}_2 = 1$ e $\mathbf{c}_2'\mathbf{c}_1 = 0$, devemos escolher o segundo maior autovalor de **R**.

Vamos admitir que os *n* autovalores de **R** sejam distintos[17] e estejam ordenados de maneira que

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n$$

Sabe-se que $\text{tr}(\mathbf{R}) = n = \sum_i \lambda_i$

Seja $\mathbf{c}_i$, com $\mathbf{c}_i'\mathbf{c}_i = 1$ e $\mathbf{c}_i'\mathbf{c}_k = 0$ para $k \neq i$, o autovetor correspondente a λ_i. Generalizando o raciocínio desenvolvido anteriormente, podemos concluir que os sucessivos componentes principais de $\mathbf{X}$ são dados por

$$\mathbf{g}_i' = \mathbf{c}_i'\mathbf{X} \tag{6.14}$$

com

$$V(g_i) = \frac{\lambda_i}{L} \tag{6.15}$$

Definindo as matrizes

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1L} \\ g_{21} & g_{22} & \cdots & g_{2L} \\ \cdots & \cdots & \cdots & \cdots \\ g_{n1} & g_{n2} & \cdots & g_{nL} \end{bmatrix}$$

e

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \ldots & \mathbf{c}_n \end{bmatrix},$$

onde cada coluna é um vetor característico de **R**, obtemos

$$\mathbf{G} = \mathbf{C}'\mathbf{X} \tag{6.16}$$

6.4. Decomposição da variância das variáveis e as correlações entre variáveis e componentes principais

De acordo com (6.9) e (6.13), generalizando, temos

$$\mathbf{R}\mathbf{c}_i = \lambda_i \mathbf{c}_i \tag{6.17}$$

ou

[17] Quando há raízes características iguais também é possível obter os componentes principais, mas a solução não é única.

$$\mathbf{RC} = \mathbf{C\Lambda}\,, \tag{6.18}$$

onde

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$$

Uma vez que $\mathbf{c}_i'\mathbf{c}_i = 1$ e $\mathbf{c}_i'\mathbf{c}_k = 0$ para $k \neq i$, a matriz **C** é ortogonal, isto é,

$$\mathbf{C'C} = \mathbf{CC'} = \mathbf{I} \tag{6.19}$$

Então, pré-multiplicando os dois membros de (6.18) por $\mathbf{C}'$, obtemos

$$\mathbf{C'RC} = \mathbf{\Lambda}\,. \tag{6.20}$$

A partir dessa relação é fácil verificar que

$$\sum \lambda_i = \text{tr}(\mathbf{R}) = n\,.$$

Com a transformação

$$\mathbf{f}_i' = \frac{1}{\sqrt{\lambda_i}}\mathbf{g}_i' \tag{6.21}$$

os componentes principais passam a ser vetores com módulo igual a 1. Se **F** é uma matriz $n \times L$ onde cada linha corresponde a um componente principal assim transformado, temos

$$\mathbf{F} = \mathbf{\Lambda}^{-1/2}\mathbf{G}\,. \tag{6.22}$$

De acordo com (6.16) segue-se que

$$\mathbf{F} = \mathbf{\Lambda}^{-1/2}\mathbf{C'X}\,. \tag{6.23}$$

Lembrando (6.2) e (6.20) verifica-se que

$$\mathbf{FF'} = \mathbf{I}_n \tag{6.24}$$

Pré-multiplicando os dois membros de (6.23) por

$$\mathbf{A} = \mathbf{C\Lambda}^{1/2} \tag{6.25}$$

e lembrando (6.19), obtemos

$$\mathbf{X} = \mathbf{AF} \tag{6.26}$$

ou $$x_{ij} = a_{i1}f_{1j} + a_{i2}f_{2j} + \ldots + a_{in}f_{nj}\quad, \tag{6.27}$$

com $i = 1, ..., n$ e $j = 1, ..., L$

Tendo definido os componentes principais como combinações lineares das variáveis X_i, verificamos agora que cada uma dessas variáveis pode ser considerada como uma combinação linear de *n* componentes principais ortogonais entre si.

De (6.26), lembrando (6.2) e (6.24), obtemos

$$\mathbf{R} = \mathbf{XX'} = \mathbf{AA'} \tag{6.28}$$

Então

$$\mathbf{x}_i'\mathbf{x}_i = \sum_j x_{ij}^2 = 1 = \sum_k a_{ik}^2 \tag{6.29}$$

Essa expressão mostra que a_{ik}^2 representa a fração da variância de x_i associada ao *k*-ésimo componente principal.

Multiplicando (6.27) por f_{kj} , somando em relação a *j*, e lembrando (6.24), verifica-se que o coeficiente de correlação entre x_i e f_k é

$$r(x_i, f_k) = a_{ik} \tag{6.30}$$

Analogamente, verifica-se que

$$r(x_i, x_k) = a_{i1}a_{k1} + a_{i2}a_{k2} + \ldots + a_{in}a_{kn} \tag{6.31}$$

Note que a expressão (6.31) é equivalente à expressão (6.28). Note, também, que pós-multiplicando (6.26) por $\mathbf{F}'$ e lembrando (6.24) obtemos

$$\mathbf{XF}' = \mathbf{A} \quad ,$$

que é a expressão matricial equivalente a (6.30), mostrando que o elemento a_{ik} da matriz **A** é a correlação entre a *i*-ésima variável e o *k*-ésimo componente principal.

De (6.25), lembrando (6.19), obtemos

$$\mathbf{A}'\mathbf{A} = \mathbf{\Lambda} \tag{6.32}$$

Então

$$\sum_i a_{ik}^2 = \lambda_k \quad , \tag{6.33}$$

isto é, a soma das "contribuições" $\left(a_{ik}^2\right)$ do *k*-ésimo componente principal para "explicar" as variâncias das n variáveis x_i é igual a λ_k, que é o correspondente autovalor de **R**. Uma vez que $\sum \lambda_k = n$, concluímos que a fração da variância das *n* variáveis x_i "explicada" pelo *k*-ésimo componente principal é λ_k / n.

Na análise de um problema é comum passar a utilizar apenas os primeiros componentes principais, aos quais corresponde, geralmente, grande parte da variância das *n* variáveis. É claro que alguma informação é perdida quando substituímos as *n* variáveis por um número menor de componentes principais. Por outro lado, há vantagens óbvias em substituir um número relativamente grande de variáveis, com problemas de multicolinearidade, por um número relativamente pequeno de variáveis (componentes principais) não-correlacionadas entre si.

Nesta exposição consideramos sempre as variáveis transformadas de acordo com (6.1), que são números puros (sem unidade de medida). Dessa maneira a base para extração dos componentes principais é a matriz (**R**) das correlações simples entre as variáveis. Também é possível considerar as variáveis apenas centradas $\left(X_{ij} - \overline{X}_i\right)$. Neste caso os componentes principais serão obtidos a partir da matriz de variâncias e covariâncias das variáveis. Note-se que esse procedimento só deve ser utilizado se a unidade de medida das variáveis for homogênea, pois os componentes principais serão combinações lineares das variáveis centradas (que mantêm a unidade de medida das variáveis originais). Quando os

componentes principais são extraídos da matriz de variâncias e covariâncias, a influência de uma variável nos resultados cresce com a sua variância.

6.5. Um exemplo numérico

A seguir vamos desenvolver um exemplo numérico, tendo em vista deixar mais claro o método de determinação dos componentes principais. Para isso consideremos os dados artificiais apresentados na tabela 6.1, referentes a 10 observações de 3 variáveis.

Tabela 6.1. Valores das variáveis X_1, X_2 e X_3 em 10 observações

X_1	X_2	X_3
5	7	4
10	12	6
4	7	7
5	3	6
5	7	7
6	11	5
5	9	7
2	2	6
5	7	7
3	5	5

Pode-se verificar que $\overline{X}_1 = 5$, $\overline{X}_2 = 7$ e $\overline{X}_3 = 6$. De acordo com (6.1), temos

$$x_{1j} = \frac{X_{1j} - 5}{\sqrt{40}}$$

$$x_{2j} = \frac{X_{2j} - 7}{\sqrt{90}}$$

e

$$x_{3j} = \frac{X_{3j} - 6}{\sqrt{10}}$$

A matriz de correlações simples entre as 3 variáveis é

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}8 & 0 \\ 0{,}8 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

De acordo com (6.8), a correspondente equação característica é

$$\begin{vmatrix} 1-\lambda & 0{,}8 & 0 \\ 0{,}8 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{vmatrix} = 0$$

ou

$$(1-\lambda)^3 - 0{,}8^2(1-\lambda) = 0$$

ou ainda,

$$(1-\lambda)\left[(1-\lambda)^2 - 0{,}8^2\right] = 0$$

Verifica-se, portanto, que os autovalores de **R**, já colocadas em ordem decrescente, são:

$$\lambda_1 = 1{,}8$$
$$\lambda_2 = 1$$

e $\lambda_3 = 0{,}2$

Então o primeiro componente principal "explica" (ou "capta")

$$\frac{1{,}8}{3} = 0{,}6 \quad \text{ou} \quad 60\%$$

da variância das variáveis x_1, x_2 e x_3, e os dois primeiros componentes principais "explicam" (ou "captam")

$$\frac{1{,}8+1}{3} = 0{,}933 \quad \text{ou} \quad 93{,}3\%$$

dessa variância.

De acordo com (6.7), o vetor característico correspondente a $\lambda = 1{,}8$ deve satisfazer as equações

$$\begin{bmatrix} -0{,}8 & 0{,}8 & 0 \\ 0{,}8 & -0{,}8 & 0 \\ 0 & 0 & -0{,}8 \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{12} \\ c_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

com $\mathbf{c}_1'\mathbf{c}_1 = 1$, ou seja,

$$c_{11}^2 + c_{12}^2 + c_{13}^2 = 1$$

Resolvendo, obtemos

$$\mathbf{c}_1 = \pm \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$$

Por conveniência vamos optar pelo sinal positivo, fazendo com que o primeiro componente principal seja positivamente correlacionado com as variáveis X_1 e X_2.

Os autovetores correspondentes a $\lambda_2 = 1$ e $\lambda_3 = 0{,}2$ podem ser obtidos de maneira semelhante, verificando-se que a matriz com os três vetores característicos é

$$\mathbf{C} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix}$$

Cabe ressaltar que o sinal de qualquer coluna dessa matriz é arbitrário. Trocar o sinal de uma coluna corresponde a trocar o sinal do respectivo componente principal.

De acordo com (6.25), multiplicando os elementos de cada coluna de **C** pela raiz quadrada do correspondente autovalor, obtemos

$$\mathbf{A} = \begin{bmatrix} \sqrt{0{,}9} & 0 & \sqrt{0{,}1} \\ \sqrt{0{,}9} & 0 & -\sqrt{0{,}1} \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0{,}9487 & 0 & 0{,}3162 \\ 0{,}9487 & 0 & -0{,}3162 \\ 0 & 1 & 0 \end{bmatrix}$$

Essa matriz mostra que o primeiro componente principal tem correlação positiva e forte com X_1 e X_2, e não tem correlação com X_3. Mostra, também, que o segundo componente principal é ortogonal a X_1 e X_2 e colinear com X_3, e que o terceiro componente principal tem correlações com X_1 e X_2 que são relativamente fracas e de sinais opostos.

Verifica-se que a soma dos quadrados dos elementos de qualquer coluna de **A** reproduz a correspondente raiz característica, e que a soma dos quadrados dos elementos de qualquer linha de **A** é igual a 1 (já que os 3 componentes principais "explicam" ou "captam" 100% da variância de cada uma das variáveis).

Para ilustrar o cálculo dos valores dos componentes principais, vamos determinar o valor de g_1 e f_1 para a sexta observação. Temos

$$x_{16} = \frac{1}{\sqrt{40}} \quad , \quad x_{26} = \frac{4}{\sqrt{90}} \quad \text{e} \quad x_{36} = -\frac{1}{\sqrt{10}} .$$

De acordo com (6.3) ou (6.16), e tendo em vista o vetor $\mathbf{c}_1$, obtemos

$$g_{16} = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{40}} + \frac{1}{\sqrt{2}} \cdot \frac{4}{\sqrt{90}} = 0{,}409946 .$$

Finalmente, de acordo com (6.21), obtemos

$$f_{16} = \frac{1}{\sqrt{1{,}8}}\, g_{16} = 0{,}3056 .$$

Os valores do primeiro componente principal para as 10 observações, calculados dessa maneira, são apresentados na segunda coluna da tabela 6.2. Cabe ressaltar que, por construção, obtemos um vetor $\mathbf{f}_1$ com módulo igual a 1, ou seja,

$$\sum f_{1j}^2 = 1 .$$

Então a estimativa da variância de f_1 será igual a $1/(L-1)$. Geralmente, na prática, é preferível definir os componentes principais de maneira que a estimativa de sua variância seja igual a 1. Neste caso os componentes principais serão variáveis reduzidas e fica mais fácil identificar os valores relativamente baixos ou relativamente altos (abaixo de –2 ou acima de 2, por exemplo). Para obter componentes principais com estimativa de variância igual a 1, em lugar de (6.22) e (6.23) devemos utilizar as relações

$$\mathbf{F}^* = \sqrt{L-1}\ \mathbf{\Lambda}^{-1/2}\mathbf{G}$$

$$\mathbf{F}^* = \sqrt{L-1}\ \mathbf{\Lambda}^{-1/2}\mathbf{C}'\mathbf{X}$$

Para o exemplo numérico apresentado temos $\sqrt{L-1} = 3$ e os valores do primeiro componente principal de maneira que a estimativa de sua variância seja 1 estão na última coluna da tabela 6.2.

Tabela 6.2. Valores do primeiro componente principal para as 10 observações de maneira que o vetor tenha módulo igual a 1 $\left(f_1\right)$ ou de maneira que a estimativa de sua variância seja igual a 1 $\left(f_1^*\right)$.

Observação	f_1	f_1^*
1	0	0
2	0,6944	2,0833
3	–0,0833	–0,2500
4	–0,2222	–0,6667
5	0	0
6	0,3056	0,9167
7	0,1111	0,3333
8	–0,5278	–1,5833
9	0	0
10	–0,2778	–0,8333

É interessante analisar a representação geométrica desse exemplo numérico no espaço das observações. A matriz **R** mostra que o vetor $\mathbf{x}_3$ é ortogonal a $\mathbf{x}_1$ e $\mathbf{x}_2$, que formam entre si um ângulo θ tal que $\cos\theta = 0{,}8$. Então $\theta = 36{,}87°$. A figura 6.1 mostra a posição dos três vetores. Note-se que $\mathbf{x}_3$ é perpendicular ao plano onde estão os vetores $\mathbf{x}_1$ e $\mathbf{x}_2$.

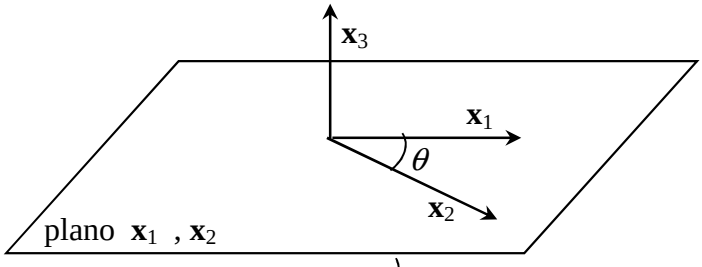


Figura 6.1. Representação geométrica dos vetores $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$.

Nesse caso particular é possível perceber que o primeiro componente principal será colinear com a soma dos vetores $\mathbf{x}_1$ e $\mathbf{x}_2$, "captando" assim boa parte da variância de X_1 e X_2. O ângulo de $\mathbf{f}_1$ com $\mathbf{x}_1$ ou $\mathbf{x}_2$ será

$$\frac{\theta}{2} = 18{,}435^{o}.$$

Então a correlação de f_1 com X_1 ou X_2 será

$$a_{11} = a_{21} = \cos 18{,}435° = 0{,}9487.$$

Uma vez que $\mathbf{x}_3$ é ortogonal a $\mathbf{f}_1$, a correlação entre essas variáveis é igual a zero, isto é, $a_{31} = 0$.

A visualização dos vetores $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ no espaço também permite perceber que o segundo componente principal é colinear com $\mathbf{x}_3$ e, portanto, ortogonal a $\mathbf{x}_1$ e $\mathbf{x}_2$. Então $a_{32} = 1$ e $a_{12} = a_{22} = 0$.

Como o terceiro componente principal tem de ser ortogonal aos dois primeiros, verifica-se que ele estará no plano de $\mathbf{x}_1$ e $\mathbf{x}_2$, sendo colinear com uma diferença entre esses dois vetores. Se considerarmos que $\mathbf{f}_3$ é colinear com $\mathbf{x}_1 - \mathbf{x}_2$, o ângulo entre $\mathbf{f}_3$ e $\mathbf{x}_1$ é

$$90 - \frac{\theta}{2} = 71{,}565^{o}$$

e o ângulo entre $\mathbf{f}_3$ e $\mathbf{x}_2$ é

$$90 + \frac{\theta}{2} = 108{,}435^{\circ}.$$

Então

$$a_{13} = \cos 71{,}565° = 0{,}3162$$

$$a_{23} = \cos 108{,}435° = -0{,}3162$$

e $$a_{33} = \cos 90° = 0.$$

Verifica-se, portanto, que nesse caso particular é possível obter a matriz **A** com base na representação geométrica no espaço das observações. É óbvio que isso é quase sempre impossível quando estão sendo analisados dados observados.
Para se familiarizar com a metodologia de análise dos componentes principais é aconselhável o estudante resolver os 3 primeiros exercícios no final do capítulo. Na próxima seção é apresentado o modelo de análise fatorial.

6.6. O Modelo da análise fatorial

O modelo de análise fatorial tem grande semelhança com a relação (6.26) ou (6.27), onde cada variável x_i é representada como uma combinação linear dos n componentes principais.

Nos modelos de análise fatorial cada uma das n variáveis é uma combinação linear de m (com $m < n$) fatores comuns e de um fator específico (*unique factor*). Para a i-ésima variável temos:

$$x_{ij} = a_{i1} f_{1j} + a_{i2} f_{2j} + \ldots + a_{im} f_{mj} + u_i y_{ij} \tag{6.34}$$

ou

$$x_{ij} = \sum_{p=1}^{m} a_{ip} f_{pj} + u_i y_{ij} \quad ,$$

onde f_{pj} representa o valor do p-ésimo fator comum para a j-ésima observação, a_{ip} (com $p = 1, ..., m$) e u_i são coeficientes e y_{ij} representa o valor do i-ésimo fator específico para a j-ésima observação.

Embora estejamos usando as letras f e a para representar os fatores comuns e os respectivos coeficientes, da mesma maneira que fizemos com os componentes principais, os m fatores comuns não se confundem, necessariamente, com os m primeiros componentes principais. A diferença ficará clara no final dessa exposição. Entretanto, é óbvio que há grande semelhança entre a relação (6.27) e o modelo (6.34).

No modelo de análise fatorial pressupõe-se que os fatores específicos (y_i, com $i = 1, ..., n$) são ortogonais entre si. Pressupõe-se, também, que cada um dos fatores específicos é ortogonal com todos os m fatores comuns (f_p, com $p = 1, ..., m$).

Vamos pressupor que os fatores comuns são ortogonais (não-correlacionados) entre si, não considerando, no que se segue, o caso de fatores oblíquos.

Vamos admitir, ainda, que todos os fatores são variáveis com média zero e que os respectivos vetores, no espaço L-dimensional das observações, tem módulo igual a 1, isto é,

$$\sum_j f_{pj} = \sum_j y_{ij} = 0$$

e

$$\sum_j f_{pj}^2 = \sum_j y_{ij}^2 = 1 \tag{6.35}$$

para $p = 1, ..., m$ e $i = 1, ..., n$

Em notação matricial o modelo (6.34) fica

$$\mathbf{X} = \mathbf{AF} + \mathbf{UY} \ , \tag{6.36}$$

onde $\mathbf{X}$ é a matriz $n \times L$ definida no início da seção anterior,

$$\mathbf{F} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1L} \\ f_{21} & f_{22} & \cdots & f_{2L} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mL} \end{bmatrix},$$

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1L} \\ y_{21} & y_{22} & \cdots & y_{2L} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n1} & \cdots & y_{nL} \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix} \quad \text{e} \quad \mathbf{U} = \begin{bmatrix} u_1 & 0 & \ldots & 0 \\ 0 & u_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & u_n \end{bmatrix}$$

As condições (6.35), juntamente com a ortogonalidade entre os *n* fatores específicos e entre os *m* fatores comuns, fazem com que tenhamos

$$\mathbf{YY}' = \mathbf{I}_n \qquad \text{e} \qquad \mathbf{FF}' = \mathbf{I}_m . \tag{6.37}$$

A ortogonalidade dos *m* fatores comuns com os *n* fatores específicos permite escrever que

$$\mathbf{FY}' = \mathbf{0} \quad , \tag{6.38}$$

onde o segundo membro é uma matriz $m \times n$ de zeros.

Como $\mathbf{R} = \mathbf{XX}'$, de (6.36), (6.37) e (6.38) obtemos

$$\mathbf{R} = \mathbf{XX}' = \mathbf{AA}' + \mathbf{UU}'$$

ou

$$\mathbf{R} = \mathbf{AA}' + \mathbf{U}^2 . \tag{6.39}$$

De acordo com essa relação, considerando um elemento da diagonal de **R**, temos

$$1 = \sum_{j=1}^{L} x_{ij}^2 = \sum_{p=1}^{m} a_{ip}^2 + u_i^2 \quad \text{para } i = 1, ..., n. \tag{6.40}$$

Os termos do último membro dessa expressão nos dão as proporções da variância de x_i devidas a cada um dos fatores. A parte associada aos m fatores comuns é denominada *comunalidade* da variável e será indicada por

$$h_i^2 = \sum_{p=1}^{m} a_{ip}^2 . \tag{6.41}$$

A proporção da variância da *i*-ésima variável devida ao fator específico é igual a u_i^2 e é denominada *especificidade* (*uniqueness*) da variável.

De acordo com (6.40) temos

$$h_i^2 + u_i^2 = 1. \tag{6.42}$$

De (6.39), considerando um elemento fora da diagonal de **R**, obtemos

$$r(x_i, x_k) = \sum_{p=1}^{m} a_{ip} a_{kp}. \tag{6.43}$$

Essa relação mostra que, de acordo com o modelo de análise fatorial, as correlações entre as variáveis x_i podem ser obtidas a partir da matriz **A**.

Multiplicando os dois membros de (6.34) por f_{pj}, somando em relação a j e lembrando que os fatores comuns (f_n) e os fatores específicos (y_i) são vetores ortogonais entre si e com módulo igual a 1, obtemos

$$r(x_i, f_p) = a_{ip}.$$

Em notação matricial temos

$$\mathbf{XF}' = \mathbf{A} \tag{6.44}$$

Verifica-se, portanto, que a *i*-ésima linha da matriz **A** é constituída pelos coeficientes de correlação da *i*-ésima variável com cada um dos *m* fatores comuns. Essa matriz é denominada de estrutura dos fatores ou, simplesmente, *estrutura*.

Os coeficientes a_{ip}, que no caso de fatores ortogonais coincidem com os elementos da estrutura, são denominados *cargas fatoriais* (*factor loadings*).[18]

6.7. Existência de solução

Dada uma matriz de correlações **R**, sempre é possível obter os correspondentes componentes principais, de acordo com o que foi apresentado na seção 2. Entretanto, dada uma matriz **R**, nem sempre é possível obter matrizes **A** e **U** de acordo com o modelo de análise fatorial ou, mais especificamente, que satisfaçam a relação (6.39). Para ilustrar essa questão vamos apresentar alguns exemplos encontrados no livro de Lawley e Maxwell (1971, p. 10-11). Nesses exemplos admite-se que há 3 variáveis (a matriz **R** é 3×3) e que se deseja fazer uma análise fatorial com apenas um fator comum ($m = 1$ e a matriz **A** é 3×1).

a) Neste primeiro exemplo a matriz das correlações entre as 3 variáveis é

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}56 & 0{,}40 \\ 0{,}56 & 1 & 0{,}35 \\ 0{,}40 & 0{,}35 & 1 \end{bmatrix}$$

[18] No caso de fatores oblíquos os coeficientes dos fatores comuns (a_{ip}) não coincidem com os elementos da estrutura, isto é, $r(x_i, f_p) \neq a_{ip}$.

Temos

$$\mathbf{AA'} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & a_{31} \end{bmatrix} = \begin{bmatrix} a_{11}^2 & a_{11}a_{21} & a_{11}a_{31} \\ a_{11}a_{21} & a_{21}^2 & a_{21}a_{31} \\ a_{11}a_{31} & a_{21}a_{31} & a_{31}^2 \end{bmatrix}$$

Uma vez que a matriz **U** é diagonal, de acordo com (6.39) devemos ter

$$\begin{cases} 0{,}56 = a_{11}a_{21} \\ 0{,}40 = a_{11}a_{31} \\ 0{,}35 = a_{21}a_{31} \end{cases}$$

Resolvendo esse sistema de 3 equações com 3 incógnitas obtemos (o sinal do vetor é arbitrário)

$$a_{11} = 0{,}8 \quad , \qquad a_{21} = 0{,}7 \qquad \text{e} \qquad a_{31} = 0{,}5.$$

Lembrando (6.39) verifica-se que as especificidades são $u_1^2 = 0{,}36$, $u_2^2 = 0{,}51$ e $u_3^2 = 0{,}75$. Considerando que o sinal das colunas de **A** é arbitrário, observa-se que nesse exemplo há uma solução única para a análise fatorial.

b) Vamos admitir, agora, que

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}84 & 0{,}60 \\ 0{,}84 & 1 & 0{,}35 \\ 0{,}60 & 0{,}35 & 1 \end{bmatrix}$$

Considerando um único fator comum, devemos ter

$$\begin{cases} 0{,}84 = a_{11}a_{21} \\ 0{,}60 = a_{11}a_{31} \\ 0{,}35 = a_{21}a_{31} \end{cases}$$

Resolvendo esse sistema obtemos

$$a_{11} = 1{,}2 \quad , \qquad a_{21} = 0{,}7 \qquad \text{e} \qquad a_{31} = 0{,}5.$$

Segue-se que

$u_1^2 = 1 - a_{11}^2 = -0{,}44$, o que é um absurdo.

Verifica-se, portanto, que para essa matriz **R** não há solução para a análise fatorial com um único fator comum. Se forem considerados dois fatores comuns há infinitas soluções.

c) Como outro exemplo, consideremos a matriz

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & 0 \\ r_{12} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , \quad \text{com} \quad 0 < r_{12}^2 < 1$$

Para uma análise fatorial com um único fator comum devemos ter

$$\begin{cases} r_{12} = a_{11}a_{21} \\ 0 = a_{11}a_{31} \\ 0 = a_{21}a_{31} \end{cases}$$

Neste caso há infinitas soluções, com $a_{31} = 0$ e valores de a_{11} e a_{21} satisfazendo as condições

$$a_{11}a_{21} = r_{12} \quad , \quad a_{11}^2 \leq 1 \quad \text{e} \quad a_{21}^2 \leq 1$$

d) Finalmente, vamos admitir que os elementos fora da diagonal principal de **R** são negativos. Resolvendo o sistema de equações em a_{11} , a_{21} e a_{31} verifica-se que as raízes são imaginárias (não há solução no campo real).

6.8. Métodos de Análise Fatorial

Há vários métodos para efetuar uma análise fatorial. Uma exposição bastante completa do assunto pode ser encontrada em Harman (1976) ou Johnson e Wichern (1982).

O método da *máxima verossimilhança* é o que tem melhor fundamentação estatística, se for feita a pressuposição de que os fatores $(f_p$ e $y_i)$ têm distribuição normal.

Um método bastante usado é o dos *fatores principais*. De acordo com esse método, os m fatores comuns correspondem às m maiores raízes características da matriz $\mathbf{R}^*$, que é obtida a partir de **R** substituindo os elementos da diagonal por estimativas das comunalidades das n variáveis (ver Harman, 1976, p. 135-141). Pode-se provar que o limite inferior para o valor da comunalidade de x_i é dado pelo coeficiente de determinação da regressão de x_i contra as $n-1$ variáveis restantes (ver Harman, 1976, p. 87). Por outro lado, de acordo com (6.42) a comunalidade não pode ser superior a 1. O método dos fatores principais pode ser usado iterativamente: obtida uma solução, calculam-se as comunalidades, cujos valores são inseridos na diagonal da matriz **R** para obter uma nova solução, e assim por diante, até que as alterações nos valores das comunalidades possam ser consideradas desprezíveis.

A análise fatorial pelo método dos componentes principais é a mais simples. Partindo-se da matriz **R**, adotam-se, como fatores comuns, os m primeiros componentes principais dessa matriz.

Na expressão (6.28) (referente a componentes principais), se a matriz $\mathbf{A}(n \times n)$ for decomposta em $\mathbf{A}_1(n \times m)$ e $\mathbf{A}_2(n \times l)$, com $l = n - m$, obtemos

$$\mathbf{R} = \mathbf{A}_1\mathbf{A}_1' + \mathbf{A}_2\mathbf{A}_2'$$

Em geral $\mathbf{A}_2\mathbf{A}_2'$ não será uma matriz diagonal e, consequentemente, a análise fatorial pelo método dos componentes principais só satisfaz a expressão (6.39) de maneira aproximada.

6.9. Rotação dos fatores

É comum fazer uma rotação dos fatores, mantendo a ortogonalidade entre eles. O objetivo dessa rotação ortogonal é obter uma estrutura simples, isto é, obter uma nova matriz $n \times m$ de coeficientes dos fatores de maneira que os valores absolutos dos elementos de cada coluna dessa matriz se aproximem, na medida do possível, de zero ou 1. Isso facilita a interpretação

dos fatores pois cada um dos novos fatores, após a rotação, deverá apresentar correlação relativamente forte com uma ou mais variáveis e correlação relativamente fraca com as demais variáveis.
Um dos critérios mais usados para obter a matriz (**T**) de transformação ortogonal é o VARIMAX (ver Harman, 1976, p. 290-299). Tratando-se de uma transformação ortogonal, temos

$$\mathbf{TT}' = \mathbf{I}_m \tag{6.45}$$

A partir de (6.36), desprezando o termo referente aos fatores singulares, temos

$$\hat{\mathbf{X}} = \mathbf{AF} \tag{6.46}$$

onde **A** é uma matriz $n \times m$ e **F** é uma matriz $m \times L$.
De acordo com (6.45) podemos escrever

$$\hat{\mathbf{X}} = \mathbf{ATT}'\mathbf{F} \tag{6.47}$$

O produto

$$\mathbf{T}'\mathbf{F} = \mathbf{Q} \tag{6.48}$$

nos dá a matriz dos fatores após a rotação, e o produto

$$\mathbf{AT} = \mathbf{B} \tag{6.49}$$

fornece a nova matriz $n \times m$ de cargas fatoriais.

É interessante assinalar que a rotação ortogonal não altera a comunalidade das variáveis. De (6.49), lembrando (6.45), obtemos

$$\mathbf{BB}' = \mathbf{ATT}'\mathbf{A}' = \mathbf{AA}' \tag{6.50}$$

Finalmente, considerando um elemento da diagonal desses produtos matriciais e lembrando a definição de comunalidade dada em (6.41), conclui-se que

$$\sum_{p=1}^{m} b_{ip}^2 = \sum_{p=1}^{m} a_{ip}^2 = h_i^2 \tag{6.51}$$

De (6.50) e (6.39) segue-se que

$$\mathbf{R} = \mathbf{BB}' + \mathbf{U}^2 \tag{6.52}$$

Pós-multiplicando os dois membros de (6.44) por **T** e lembrando (6.48) e (6.49), verifica-se que

$$\mathbf{XQ}' = \mathbf{B} \quad , \tag{6.53}$$

comprovando que os elementos de **B** são os coeficientes de correlação entre as variáveis x_i e os fatores, após a rotação.

6.10. Medida de adequação da amostra à análise fatorial

A medida de adequação de Kaiser (também denominada medida de Kaiser-Meyer-Olkin) é dada por

$$K = \frac{\sum_{i=1}^{n}\sum_{k\neq i} r^2(x_i, x_k)}{\sum_{i=1}^{n}\sum_{k\neq i} r^2(x_i, x_k) + \sum_{i=1}^{n}\sum_{k\neq i} \pi^2(x_i, x_k)}$$

onde $r(x_i, x_k)$ é o coeficiente de correlação entre x_i e x_k e $\pi(x_i, x_k)$ é o coeficiente de correlação *parcial* entre x_i e x_k, dadas as demais variáveis do conjunto analisado.

Fixando *i*, podemos obter uma medida de adequação para cada variável:

$$K_i = \frac{\sum_{k\neq i} r^2(x_i, x_k)}{\sum_{k\neq i} r^2(x_i, x_k) + \sum_{k\neq i} \pi^2(x_i, x_k)}$$

Para o exemplo numérico apresentado na seção 6.5, as três correlações parciais são iguais às correspondentes correlações simples e obtemos K = 0,5, $K_1 = 0{,}5$ e $K_2 = 0{,}5$. Para x_3 as correlações são nulas e considera-se $K_3 = 0$ por definição especial.

É usual afirmar-se que uma medida geral de adequação da amostra igual ou maior do que 0,8 é boa e que um valor abaixo de 0,5 é inaceitável. Em minha opinião esses limites são muito arbitrários e é melhor basear a avaliação de uma análise fatorial pelo método dos componentes principais na proporção da variância das variáveis originais que é "explicada" pelos fatores extraídos, ou seja, nos valores de λ_k / n (como discutido na seção 6.4) e nas comunalidades, considerando, também, a interpretação dos fatores com base na matriz de cargas fatoriais após a rotação. Os exercícios 25 e 27 mostram exemplos artificiais de análises fatoriais com resultados que poderiam ser considerados de interesse, apesar de o valor da medida de adequação da amostra ser relativamente baixo.

Exercícios

1. Para cada um dos conjuntos de dados a seguir, determine a matriz **A** e os valores dos componentes principais transformados de maneira que as estimativas das suas variâncias sejam iguais a 1.

a)

X_1	X_2
1	0
1	1
0	1
−1	0
0	−1
−1	−1

b)

X_1	X_2
2	1
1	2
−2	−1
−1	−2

c)

X_1	X_2
2	1
2	–4
–2	–1
–2	4

d)

X_1	X_2	X_3
2	0	0
2	2	0
0	2	0
–2	0	0
0	–2	0
–2	–2	0
1	1	1
1	1	–1
–1	–1	1
–1	–1	–1

e)

X_1	X_2	X_3
1	1	2
1	2	1
2	1	1
–1	–1	–2
–1	–2	–1
–2	–1	–1

f)

X_1	X_2	X_3
1	1	2
1	2	1
2	1	1
1	1	1
2	2	2
–1	–1	–2
–1	–2	–1
–2	–1	–1
–1	–1	–1
–2	–2	–2

2. No caso do item (d) do exercício 1, sejam $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ os vetores correspondentes às variáveis X_1, X_2 e X_3, no espaço 10-dimensional das observações.
   a) Determine os ângulos entre $\mathbf{x}_1$ e $\mathbf{x}_2$, entre $\mathbf{x}_1$ e $\mathbf{x}_3$ e entre $\mathbf{x}_2$ e $\mathbf{x}_3$.
   b) Mostre que o primeiro componente principal é proporcional a $\mathbf{x}_1 + \mathbf{x}_2$.
   c) Mostre que o segundo componente principal é proporcional a $\mathbf{x}_3$.
   d) Mostre que o terceiro componente principal é proporcional a $\mathbf{x}_1 - \mathbf{x}_2$.
   e) Determine os ângulos entre $\mathbf{x}_1$ e cada um dos três componentes principais.

3. É dada a seguinte amostra de valores de X_1 e X_2

X_1	X_2
2	24
4	36
6	30
6	36
7	45
9	39
9	45
9	51
11	45
12	54
12	60
14	54
16	66

   a) Qual é o coeficiente de correlação entre X_1 e X_2?
   b) Sejam $\mathbf{x}_1$ e $\mathbf{x}_2$ os vetores-coluna com as variáveis X_1 e X_2 centradas e transformadas de maneira que $\mathbf{x}_1'\mathbf{x}_2$ seja o coeficiente de correlação entre X_1 e X_2. Qual é o ângulo entre os vetores $\mathbf{x}_1$ e $\mathbf{x}_2$ no espaço das observações (com 13 dimensões)?
   c) Qual é a correlação entre X_1 e o primeiro componente principal? E entre X_2 e o primeiro componente principal?
   d) Qual é o ângulo entre $\mathbf{x}_1$ e o primeiro componente principal?

4. A tabela ao lado mostra os valores de X_1, X_2 e X_3 em uma amostra com 10 observações.

X_1	X_2	X_3
1	2	2
2	1	2
4	4	0
–1	–2	2
–2	–1	2
–4	–4	0
1	2	–2
2	1	–2
–1	–2	–2
–2	–1	–2

   a) Determine a matriz (**R**) das correlações simples entre essas 3 variáveis.
   b) Qual é o ângulo entre os vetores $\mathbf{x}_1$ e $\mathbf{x}_2$ no espaço das observações (com 10 dimensões)?
   c) A partir da matriz **R**, determine a matriz (**A**) das correlações entre os dois primeiros componentes principais e as 3 variáveis.
   d) Qual é a proporção da variância total das 3 variáveis (após "normalização") que pode ser "explicada" pelos dois primeiros componentes principais?
   e) Calcule os valores do primeiro componente principal para as 10 observações, isto é, determine o vetor $\mathbf{f}_1$ (com $\mathbf{f}_1\mathbf{f}_1' = 1$).

5. Considerando os dados apresentados no exercício anterior, obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com um único fator comum e admitindo que as cargas fatoriais de X_1 e X_2 sejam iguais.
   Compare o resultado obtido com uma análise de componentes principais em que fosse considerado apenas o 1º componente principal. Os resultados são iguais? Tendo em vista "explicar" o comportamento das variáveis X_1, X_2 e X_3, em que sentido a análise fatorial (com um fator comum) é superior? Em que sentido a análise de componentes principais (extraindo apenas o 1º) é superior?

6. A tabela ao lado mostra os valores de X_1, X_2 e X_3 em uma amostra com 5 observações.

X_1	X_2	X_3
7	3	7
1	9	7
6	8	9
3	5	6
3	5	6

   a) Determine a matriz **R** das correlações simples entre essas 3 variáveis.

b) Seja $\mathbf{x}_i$, com $i = 1, 2, 3$, o vetor-coluna com a variável X_i centrada (ou centrada e "normalizada" de maneira que o vetor tenha módulo igual a 1). Qual é o ângulo entre os vetores $\mathbf{x}_1$ e $\mathbf{x}_2$ no espaço das observações? E entre $\mathbf{x}_1$ e $\mathbf{x}_3$?

c) Determine a matriz (**A**) das correlações entre as 3 variáveis e os dois primeiros componentes principais (Uma das raízes características de **R** é $\lambda = 1{,}5$).

d) Qual é a proporção da variância total das 3 variáveis (após "normalização") que é "explicada" pelos dois primeiros componentes principais?

e) Qual é o ângulo entre $\mathbf{x}_1$ e o primeiro componente principal?

f) Calcule os valores do primeiro componente principal para as 5 observações, isto é, determine o vetor $\mathbf{f}_1$ (com $\mathbf{f}_1'\mathbf{f}_1 = 1$).

g) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com um único fator comum.

7. A tabela a seguir mostra os valores de X_1, X_2 e X_3 em uma amostra com 6 observações.

X_1	X_2	X_3
9	5	8
5	4	7
4	3	4
3	2	6
2	3	6
1	1	5

a) Determine a matriz (**R**) das correlações simples entre essas três variáveis.

b) Qual é o ângulo (em graus) entre os vetores $\mathbf{x}_1$ e $\mathbf{x}_2$ (com as variáveis centradas ou "normalizadas") no espaço das observações?

c) Sabendo que uma das raízes características da matriz **R** é igual a 0,1, determine a matriz **A** das correlações entre o primeiro componente principal e cada uma das 3 variáveis.

d) Qual é a proporção da variância total das 3 variáveis (após "normalização") que é "explicada" pelo primeiro componente principal?

e) Calcule os valores do primeiro componente principal para as 6 observações, isto é, determine o vetor $\mathbf{f}_1$, com módulo igual a 1.

f) Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{f}_1$?

g) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com um único fator comum.

h) Compare as duas alternativas: extrair o primeiro componente principal ou fazer uma análise fatorial com um único fator comum.

8. A tabela a seguir mostra os valores X_1, X_2 e X_3 em uma amostra com 7 observações.

X_1	X_2	X_3
9	7	10
5	5	8
4	6	11
3	5	4
2	4	6
1	3	2
4	5	1

a) Determine a matriz (**R**) das correlações simples entre essas 3 variáveis.

b) Qual é o ângulo (em graus) entre os vetores $\mathbf{x}_1$ e $\mathbf{x}_2$ (com as variáveis centradas ou "normalizadas") no espaço das observações?

c) Sabendo que duas das raízes características da matriz **R** são 0,437178 e 0,088757, determine a matriz **A** das correlações entre o primeiro componente principal e cada uma das 3 variáveis.

d) Qual é a proporção da variância total das 3 variáveis (após "normalização") que é "explicada" pelo primeiro componente principal?

e) Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{f}_1$?

f) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com um único fator comum.

9. A tabela a seguir mostra valores de X_1, X_2 e X_3 em uma amostra com 6 observações.

X_1	X_2	X_3
9	1	4
7	5	2
5	3	6
9	1	4
7	5	2
5	3	6

a) Determine a matriz das correlações simples entre essas 3 variáveis.

b) Sejam $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ os vetores com os valores centrados das variáveis X_1, X_2 e X_3, respectivamente, no espaço hexadimensional das observações. Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{x}_2$? E entre $\mathbf{x}_1$ e $\mathbf{x}_3$?

c) Determine a matriz (**A**) das correlações entre os dois primeiros componentes principais e as 3 variáveis.

d) Qual é a proporção da variância total das 3 variáveis (após "normalização") que é "explicada" pelo primeiro componente principal? E pelos dois primeiros componentes principais?

e) Quais são os ângulos (em graus) entre $\mathbf{x}_1$ e dois primeiros componentes principais?

f) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com um único fator comum.

10. É dada a seguinte amostra com 12 observações de quatro variáveis:

X_1	X_2	X_3	X_4
4	7	9	8
5	9	4	4
3	5	8	6
3	5	4	4
4	7	5	4
2	6	5	3
5	9	8	6
6	8	5	3
4	7	7	6
6	8	7	7
2	6	7	7
4	7	3	2

a) Determine a matriz das correlações entre as 4 variáveis.
b) Qual é o ângulo entre $\mathbf{x}_1$ e $\mathbf{x}_2$?
c) Qual é o ângulo entre $\mathbf{x}_1$ e $\mathbf{x}_4$?
d) Determine a matriz das cargas fatoriais para uma análise fatorial pelo método dos componentes principais, considerando apenas os dois primeiros componentes principais.
e) Considerando as variáveis transformadas de maneira que todas fiquem com a mesma variância, qual a proporção da variância total "explicada" pelos dois fatores?
f) Qual é o ângulo entre $\mathbf{x}_1$ e o primeiro componente principal?
g) Qual é o ângulo entre $\mathbf{x}_1$ e o segundo componente principal?
h) Qual é a comunalidade de X_1 nesta análise fatorial pelo método de componentes principais?

11. Se uma matriz quadrada **R** for bloco-diagonal, isto é, se

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 \end{bmatrix},$$

onde $\mathbf{R}_1$ e $\mathbf{R}_2$ são matrizes quadradas, então

$$|\mathbf{R}| = |\mathbf{R}_1| \cdot |\mathbf{R}_2|$$

Admite-se que há uma amostra com 5 observações para as variáveis X_1, X_2, X_3 e X_4. Os valores de X_1 e X_2 são apresentados na tabela a seguir

X_1	X_2
4	3
6	9
4	9
7	9
4	5

A matriz de correlações entre as 4 variáveis é

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & 0 & 0 \\ r_{12} & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

a) Determine r_{12}.
b) Determine os autovalores de **R**.
c) Determine o ângulo entre os vetores $\mathbf{x}_1$ e $\mathbf{x}_3$ no espaço das observações.
d) Determine o ângulo entre os vetores $\mathbf{x}_3$ e $\mathbf{x}_4$ no espaço das observações.
e) Vamos considerar uma análise fatorial pelo método dos componentes principais, com dois fatores, utilizando as variáveis transformadas de maneira que todas fiquem com a mesma variância. Qual é a proporção da variância total "explicada" pelos dois fatores (dois primeiros componentes principais)?
f) A matriz das cargas fatoriais para essa análise é

$$\mathbf{A} = \begin{bmatrix} 0 & a_{21} \\ 0 & a_{22} \\ 1 & 0 \\ -1 & 0 \end{bmatrix}$$

Determine a_{21} e a_{22}.
g) Determine o ângulo entre $\mathbf{x}_3$ e o primeiro componente principal.
h) Determine as comunalidades de X_2 e de X_4.

12. Se uma matriz quadrada **R** for bloco-diagonal, isto é, se

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 \end{bmatrix}$$

onde $\mathbf{R}_1$ e $\mathbf{R}_2$ são matrizes quadradas, então

$$|\mathbf{R}| = |\mathbf{R}_1| \cdot |\mathbf{R}_2|$$

É dada a seguinte amostra com 12 observações de cinco variáveis:

X_1	X_2	X_3	X_4	X_5
5	7	16	2	6
6	3	18	4	8
4	11	14	4	8
3	9	27	6	7
7	5	5	6	7
5	7	16	6	8
7	5	5	10	11
3	9	27	10	11
5	7	16	10	10
4	11	14	12	10
6	3	18	12	10
5	7	16	14	12

A matriz de correlações entre essas variáveis é

$$\mathbf{R} = \begin{bmatrix} 1 & -0{,}8 & -0{,}8 & 0 & 0 \\ -0{,}8 & 1 & 0{,}28 & 0 & 0 \\ -0{,}8 & 0{,}28 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & r \\ 0 & 0 & 0 & r & 1 \end{bmatrix}$$

São dados dois autovalores de **R**: 2,28 e 0,72.

a) Determine *r*.
b) Determine o ângulo entre $\mathbf{x}_1$ e $\mathbf{x}_2$ no espaço das observações.
c) Determine os outros 3 autovalores de **R**.
d) Vamos considerar uma análise fatorial pelo método dos componentes principais, com dois fatores, utilizando as variáveis transformadas de maneira que todas fiquem com a mesma variância. Qual é a proporção da variância total "explicada" pelos dois fatores?
e) A matriz das cargas fatoriais, pelo método dos componentes principais, é

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & 0 \\ a_{31} & 0 \\ 0 & \sqrt{0{,}95} \\ 0 & \sqrt{0{,}95} \end{bmatrix}$$

Determine a_{11} , a_{21} e a_{31}.

f) Determine o ângulo entre $\mathbf{x}_2$ e o primeiro componente principal
g) Determine as comunalidades de X_2 e de X_4.

13. Determine o valor das medidas de adequação da amostra de Kaiser para os dados do exercício 1 (f).

14. É dada a seguinte amostra com 8 observações de quatro variáveis:

X_1	X_2	X_3	X_4
4	3	8	8
8	5	6	2
7	3	9	8
7	3	5	2
5	5	9	8
8	5	8	8
5	5	5	2
4	3	6	2

a) Determine a matriz de correlações entre as quatro variáveis.
b) Sendo $\mathbf{x}_i$ o vetor com os valores centrados da variável X_i, com i = 1, 2, 3, 4. Qual é o ângulo entre $\mathbf{x}_1$ e $\mathbf{x}_2$? E entre $\mathbf{x}_1$ e $\mathbf{x}_4$?
c) Determine a matriz das cargas fatoriais para uma análise fatorial pelo método dos componentes principais, considerando apenas os componentes principais com raiz característica (autovalor) maior do que 1.
d) Considerando as variáveis transformadas de maneira que todos fiquem com a mesma variância, qual a proporção da variância total "explicada" pelos dois fatores?
e) Qual é o ângulo entre $\mathbf{x}_1$ e o primeiro fator? E entre $\mathbf{x}_1$ e o segundo fator?
f) Qual é a comunalidade de X_1 ?
g) Determine o valor dos escores fatoriais (valor dos dois componentes principais) para a primeira observação (de maneira que cada fator seja uma variável com média zero e estimativa de variância igual a 1).

15. Vamos admitir que em uma amostra com n observações de X_1, X_2 e X_3 as correlações entre X_1 e X_2, X_1 e X_3 e entre X_2 e X_3 sejam todas iguais a r, com $r > 0$.
a) Vamos admitir que seja feita uma análise fatorial pelo método dos componentes principais, a partir da matriz de correlações entre as três variáveis, com apenas 1 fator

(o primeiro componente principal). Deduza as expressões para as cargas fatoriais, as comunalidades e a proporção da variância total das 3 variáveis (considerada igual a 3) "explicada" por esse fator, sempre em função de *r*.

b) Considere, em seguida, uma análise fatorial propriamente dita, com apenas 1 fator comum. Obtenha, novamente, as expressões para as cargas fatoriais, as comunalidades e a proporção "explicada" da variância total das 3 variáveis, sempre em função de *r*.

c) Compare os resultados em (a) e (b).

16. A tabela a seguir mostra os valores de X_1, X_2 e X_3 em uma amostra com 4 observações:

X_1	X_2	X_3
16	29	10
10	13	4
8	17	4
2	1	10

a) Determine a matriz das correlações simples entre essas três variáveis.

b) Sendo $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ os vetores com os valores centrados das variáveis X_1, X_2 e X_3, respectivamente, no espaço tetradimensional das observações. Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{x}_2$? E entre $\mathbf{x}_1$ e $\mathbf{x}_3$?

c) Determine a matriz (**A**) das correlações entre os dois primeiros componentes principais e as 3 variáveis.

d) Qual é a proporção da variância total das 3 variáveis (após "normalização") que é "explicada" pelo primeiro componente principal? E pelos dois primeiros componentes principais?

e) Quais são os ângulos (em graus) entre $\mathbf{x}_1$ e os dois primeiros componentes principais?

f) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com um único fator comum.

17. Vamos admitir que em uma amostra com *n* observações de X_1, X_2, X_3 e X_4, as correlações entre duas variáveis sejam todas iguais a *r*, com $r > 0$.

a) Vamos admitir que seja feita uma análise fatorial pelo método dos componentes principais, a partir da matriz de correlações entre as quatro variáveis, com apenas 1 fator (o primeiro componente principal). Deduza as expressões para as cargas fatoriais, as comunalidades e a proporção da variância total das 4 variáveis (considerada igual a 4) "explicada" por esse fator, sempre em função de *r*.

b) Considere, em seguida, uma análise fatorial propriamente dita, com apenas 1 fator comum. Obtenha, novamente, as expressões para as cargas fatoriais, as comunalidades e a proporção "explicada" da variância total das 4 variáveis, sempre em função de *r*.

c) Compare os resultados em (a) e (b).

Observação: Para obter as raízes características da matriz **R** de dimensões 4×4, com equicorrelação, lembre quais são essas raízes características no caso de 2 e de 3 variáveis, e procure "extrapolar" para o caso de 4 variáveis. A seguir verifique se as expressões assim obtidas efetivamente satisfazem a equação característica $|\mathbf{R} - \lambda\mathbf{I}| = 0$.

18. A partir de uma amostra de valores das variáveis X_1, X_2, X_3, X_4 e X_5 foi obtida a matriz de correlações

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}5 & 1 & 0{,}5 & 0 & 0 \\ 0{,}5 & 0{,}5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0{,}8 \\ 0 & 0 & 0 & 0{,}8 & 1 \end{bmatrix}$$

a) Sejam $\mathbf{x}_i$ (com $i = 1, ..., 5$) os vetores com os valores centrados das variáveis, no espaço das observações. Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{x}_2$? Entre $\mathbf{x}_1$ e $\mathbf{x}_5$? Entre $\mathbf{x}_4$ e $\mathbf{x}_5$?

b) Determine os autovalores de **R**.

c) Determine a matriz $\mathbf{A}(5 \times 2)$ das cargas fatoriais dos dois primeiros componentes principais.

d) Qual é a proporção da variância total das 5 variáveis (após "normalização") que é "explicada" pelos dois primeiros componentes principais?

e) Qual é o ângulo (em graus) entre $\mathbf{x}_3$ e o primeiro componente principal?

f) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com dois fatores comuns.

19. A partir de uma amostra de valores das variáveis X_1, X_2, X_3, X_4 e X_5 foi obtida a matriz de correlações

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}94 & 0{,}94 & 0 & 0 \\ 0{,}94 & 1 & 0{,}94 & 0 & 0 \\ 0{,}94 & 0{,}94 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0{,}96 \\ 0 & 0 & 0 & 0{,}96 & 1 \end{bmatrix}$$

a) Sejam $\mathbf{x}_i$ (com $i = 1, ..., 5$) os vetores com os valores centrados das variáveis, no espaço das observações. Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{x}_2$? Entre $\mathbf{x}_1$ e $\mathbf{x}_5$? Entre $\mathbf{x}_4$ e $\mathbf{x}_5$?

b) Determine os autovalores de **R**.

c) Determine a matriz **A** (com 5 linhas e 2 colunas) das cargas fatoriais dos dois primeiros componentes principais.

d) Qual é a proporção da variância total das 5 variáveis (após "normalização") que é "explicada" pelos dois primeiros componentes principais?

e) Qual é o ângulo (em graus) entre $\mathbf{x}_3$ e o primeiro componente principal?

f) Determine a comunalidade de cada variável nessa análise fatorial pelo método dos componentes principais.

20. Admite-se que a matriz das correlações entre 3 variáveis (X_1, X_2 e X_3) é aquela apresentada no início da seção 6.7:

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}56 & 0{,}40 \\ 0{,}56 & 1 & 0{,}35 \\ 0{,}40 & 0{,}35 & 1 \end{bmatrix}$$

a) Determine a comunalidade de cada variável em uma análise fatorial (propriamente dita) com um único fator comum.

b) Determine a porcentagem da variância das 3 variáveis (após a padronização que as deixa com a mesma variância) que é "explicada" pelo fator comum.

c) Sabendo que $|\mathbf{R} - \lambda\mathbf{I}| = 0$ para $\lambda = 0{,}436$, determine a matriz de correlações entre as 3 variáveis e o *primeiro componente principal*.

d) Determine a porcentagem da variância das 3 variáveis que é captada pelo 1º componente principal e compare com o resultado obtido no item (b).

21. Admite-se que a matriz das correlações entre 3 variáveis (X_1, X_2 e X_3) é

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}64 & 0{,}48 \\ 0{,}64 & 1 & 0{,}48 \\ 0{,}48 & 0{,}48 & 1 \end{bmatrix}$$

a) Em uma análise fatorial (propriamente dita) com um único fator comum, determine o vetor das cargas fatoriais e a comunalidade de cada variável.

b) Determine a porcentagem da variância das 3 variáveis (após a padronização que as deixa com a mesma variância) que é "explicada" pelo fator comum.

c) Sabendo que $|\mathbf{R} - \lambda\mathbf{I}| = 0$ para $\lambda = 0{,}36$, determine a matriz de correlações entre as 3 variáveis e o *primeiro componente principal*.

d) Determine a porcentagem da variância das 3 variáveis que é captada pelo 1º componente principal e compare com o resultado obtido no item (b).

22. A partir de uma amostra de valores das variáveis X_1, X_2, X_3 e X_4 foi obtida a matriz de correlações

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}6 & 0 & 0 \\ 0{,}6 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0{,}4 \\ 0 & 0 & 0{,}4 & 1 \end{bmatrix}$$

a) Sejam $\mathbf{x}_i$ (com i = 1, 2, 3 ou 4) os vetores com os valores centrados das variáveis, no espaço das observações. Qual é o ângulo (em graus) entre $\mathbf{x}_1$ e $\mathbf{x}_3$? Entre $\mathbf{x}_3$ e $\mathbf{x}_4$?

b) Determine os autovalores de **R**.

c) Determine a matriz **A** (com 4 linhas e 2 colunas) das cargas fatoriais dos dois primeiros componentes principais.

d) Qual é a proporção da variância total das 4 variáveis (após "normalização") que é "explicada" pelos dois primeiros componentes principais?

e) Qual é o ângulo (em graus) entre $\mathbf{x}_3$ e o primeiro componente principal? E entre $\mathbf{x}_3$ e o segundo componente principal?

f) Determine a comunalidade de cada variável nessa análise fatorial pelo método dos componentes principais.

g) Obtenha as cargas fatoriais de um modelo como o descrito pelas relações (6.36) a (6.39), com dois fatores comuns, e determine a comunalidade de cada variável para esse modelo.

23. O coeficiente de correlação entre X_1 e X_2 em uma amostra de *n* observações é igual a *r*, com $r < 0$. São definidas as variáveis

$$x_{ij} = \frac{X_{ij} - \overline{X}_i}{\sqrt{\sum_j \left(X_{ij} - \overline{X}_i\right)^2}} \quad \text{, para } i = 1, 2$$

Seja $\mathbf{x}_1$ o vetor cujos elementos são x_{ij}. Determine, sempre em função de *r*:

a) O maior autovalor da matriz de correlações de X_1 e X_2.

b) O vetor **A** das correlações de X_1 e X_2 com o primeiro componente principal.

c) A proporção da variância de cada variável "explicada" pelo 1º componente principal.

d) O vetor das cargas fatoriais para uma análise fatorial propriamente dita com 1 único fator comum, de maneira que, no espaço das observações, esse fator comum seja um vetor com a mesma direção e o mesmo sentido que o vetor $\mathbf{x}_1$.

e) A proporção da variância total das variáveis x_{1i} e x_{2i} que é "explicada" por essa análise fatorial, comparando-a com o valor correspondente para a análise que utiliza o 1º componente principal.

24. É dada a amostra de 9 valores de X_1 e X_2, apresentada na tabela ao lado.

Forneça, agora, os valores numéricos do que foi pedido nos itens da questão anterior.

Além disso, determine:

f) o ângulo, em graus, entre $\mathbf{x}_1$ e $\mathbf{x}_2$.

g) o ângulo, em graus, entre $\mathbf{x}_1$ e o 1º componente principal.

X_1	X_2
3	10
4	10
5	9
8	9
8	6
8	3
11	3
12	2
13	2

h) o ângulo, em graus, entre $\mathbf{x}_2$ e o 1º componente principal.

i) o ângulo, em graus, entre $\mathbf{x}_2$ e o fator comum obtido no item (d).

25. A tabela ao lado apresenta 20 valores das variáveis U_h, com $h = 1, 2, ..., 6$. Note-se que, por construção, essas 6 variáveis são não-correlacionadas entre si. Usando um computador, gere os 20 valores de $X_{1i} = 7 + 5U_{1i}$, $X_{2i} = 9 + 5U_{1i} + U_{2i}$, $X_{3i} = 10 + 5U_{3i}$ e $X_{4i} = 6 + 5U_{3i} + U_{4i}$, com $i = 1, 2, ..., 20$.

Note que, por construção, X_1 e X_2 são correlacionadas, pois têm um termo em comum. O mesmo acontece com X_3 e X_4. Por outro lado, não há correlação de X_1 ou X_2 com X_3 ou X_4.

U_1	U_2	U_3	U_4	U_5	U_6
1	1	1	1	0	0
1	1	1	–1	0	0
1	1	–1	1	0	0
1	1	–1	–1	0	0
1	–1	1	1	0	0
1	–1	1	–1	0	0
1	–1	–1	1	0	0
1	–1	–1	–1	0	0
0	0	0	0	2	2
0	0	0	0	2	–2
0	0	0	0	–2	2
0	0	0	0	–2	–2
–1	1	1	1	0	0
–1	1	1	–1	0	0
–1	1	–1	1	0	0
–1	1	–1	–1	0	0
–1	–1	1	1	0	0
–1	–1	1	–1	0	0
–1	–1	–1	1	0	0
–1	–1	–1	–1	0	0

A seguir faça a análise fatorial da matriz 20 x 4
Fazendo a análise fatorial pelo método dos com fatores extraem 99,03% da variância total e que 0,9903. Verifique, também, que a medida de ad 0,5.

26. Usando os valores de U_{hi} apresentados no exercício anterior, calcule, por meio de um computador, os valores de

$$X_{1i} = 7 + 5U_{1i}$$
$$X_{2i} = 9 + 5U_{1i} + U_{2i}$$
$$X_{3i} = 8 + 5U_{1i} + U_{5i}$$
$$X_{4i} = 15 + 5U_{3i}$$
$$X_{5i} = 19 + 5U_{3i} + U_{4i}$$
$$X_{6i} = 17 + 5U_{3i} + U_{6i}$$

Note que, por construção, há correlação entre as três primeiras variáveis (que têm em comum o termo $5U_{1i}$) e também entre as três últimas (que têm em comum o termo $5U_{3i}$), mas não há correlação entre qualquer variável do primeiro grupo com qualquer variável do segundo grupo.

A seguir faça a análise fatorial pelo método dos componentes principais, extraindo 2 fatores, da matriz 20 x 6 de valores de X_{hi} (com *h* = 1, 2, ..., 6 e *i* = 1, 2, ..., 20). Mostre que os 2 fatores "explicam" 98,28% da variância total, as comunalidades são iguais a 0,9913 ou 0,9786 e a medida de adequação da amostra de Kaiser é igual a 0,7439.

27. Utilizando, novamente, os valores de U_{hi} apresentados no exercício 25, calcule os valores das variáveis

$$X_{1i} = 11 + 5U_{5i}$$
$$X_{2i} = 9 + 4U_{5i} + 3U_{1i}$$
$$X_{3i} = 10 + 4U_{5i} + 3U_{2i}$$
$$X_{4i} = 15 + 5U_{6i}$$
$$X_{5i} = 19 + 4U_{6i} + 3U_{3i}$$
$$X_{6i} = 17 + 4U_{6i} + 3U_{4i}$$

Faça a análise fatorial pelo método dos componentes principais, extraindo 2 fatores, da matriz 20 x 6 de valores de X_{hi}. Mostre que cada fator é associado a apenas 3 das 6 variáveis, com cargos fatoriais iguais a 0,9530 ou 0,8909. Verifique, também, que os dois fatores "explicam" 83,19% da variância total, as comunalidades são iguais a 0,9082 ou 0,7938 e a medida de adequação da amostra é igual a 0,6840.

Respostas

1. a) $\mathbf{A} = \begin{bmatrix} 0{,}866 & 0{,}500 \\ 0{,}866 & -0{,}500 \end{bmatrix}$

Componentes principais	
1º	2º
0,646	1,118
1,291	0
0,646	–1,118
–0,646	–1,118
–0,646	1,118
–1,291	0

b) $\mathbf{A} = \begin{bmatrix} 0{,}9487 & 0{,}3162 \\ 0{,}9487 & -0{,}3162 \end{bmatrix}$

Componentes principais	
1º	2º
0,866	0,866
0,866	–0,866
–0,866	–0,866
–0,866	0,866

c) $\mathbf{A} = \begin{bmatrix} 0{,}8702 & 0{,}4927 \\ -0{,}8702 & 0{,}4927 \end{bmatrix}$

Componentes principais	
1º	2º
0,327	1,180
1,180	–0,327
–0,327	–1,180
–1,180	0,327

d) $\mathbf{A} = \begin{bmatrix} 0{,}894 & 0 & 0{,}447 \\ 0{,}894 & 0 & -0{,}447 \\ 0 & 1 & 0 \end{bmatrix}$

Componentes principais		
1º	2º	3º
0,75	0	1,5
1,50	0	0
0,75	0	–1,5
–0,75	0	–1,5
–0,75	0	1,5
–1,50	0	0
0,75	1,5	0
0,75	–1,5	0
–0,75	1,5	0
–0,75	–1,5	0

e) Uma das soluções é

$\mathbf{A} = \begin{bmatrix} 0{,}943 & 0{,}289 & 0{,}167 \\ 0{,}943 & -0{,}289 & 0{,}167 \\ 0{,}943 & 0 & -0{,}333 \end{bmatrix}$

Componentes principais		
1º	2º	3º
0,913	0	–1,291
0,913	–1,118	0,646
0,913	1,118	0,646
–0,913	0	1,291
–0,913	1,118	–0,646
–0,913	–1,118	–0,646

f) Uma das soluções é

$\mathbf{A} = \begin{bmatrix} 0{,}969 & 0{,}213 & -0{,}123 \\ 0{,}969 & -0{,}213 & -0{,}123 \\ 0{,}969 & 0 & 0{,}246 \end{bmatrix}$

Componentes principais		
1º	2º	3º
0,880	0	1,732
0,880	–1,5	–0,866
0,880	1,5	–0,866
0,660	0	0
1,320	0	0
–0,880	0	–1,732
–0,880	1,5	0,866
–0,880	–1,5	0,866
–0,660	0	0
–1,320	0	0

2. a) 53,13°, 90° e 90°
   e) 26,57°, 90° e 63,43°

3. a) $r = \frac{11}{12} = 0{,}9167$
   b) $\varphi = 23{,}556°$
   c) 0,9789 e 0,9789
   d) $\frac{\varphi}{2} = 11{,}778°$

4. a) $\mathbf{R} = \begin{bmatrix} 1 & \frac{12}{13} & 0 \\ \frac{12}{13} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ b) $\cos\varphi_{12} = \frac{12}{13}$

$\varphi_{12} = 22{,}62^\circ$

c) $\mathbf{A} = \begin{bmatrix} 0{,}9806 & 0 \\ 0{,}9806 & 0 \\ 0 & 1 \end{bmatrix}$ d) $\frac{\lambda_1 + \lambda_2}{3} = \frac{1}{3}\left(\frac{25}{13} + 1\right)$

$= 0{,}974$ ou 97,4%

e) $\mathbf{f}_1' = [\theta \quad \theta \quad \omega \quad -\theta \quad -\theta \quad -\omega \quad \theta \quad \theta \quad -\theta \quad -\theta]$

onde $\theta = 0{,}2121$ e $\omega = 0{,}5657$

5. $\mathbf{A} = \begin{bmatrix} 0{,}9608 \\ 0{,}9608 \\ 0 \end{bmatrix}$

O 1º componente principal "explica" uma parte maior da variância (64,1% contra 61,5% na análise fatorial). Por outro lado, a análise fatorial "explica" perfeitamente as correlações, o que não acontece com o 1º componente principal.

6. a) $\mathbf{R} = \begin{bmatrix} 1 & -0{,}5 & 0{,}5 \\ -0{,}5 & 1 & 0{,}5 \\ 0{,}5 & 0{,}5 & 1 \end{bmatrix}$

b) Ângulo entre $\mathbf{x}_1$ e $\mathbf{x}_2$ = 120º

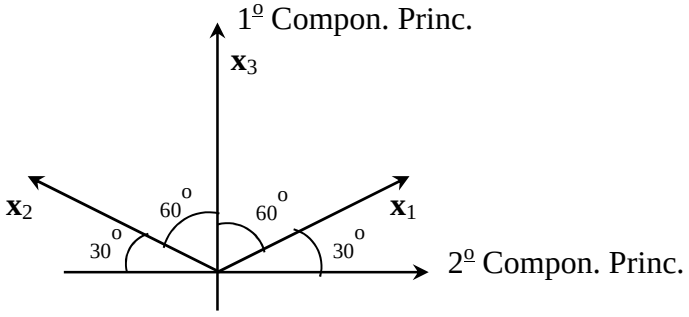
Ângulo entre $\mathbf{x}_1$ e $\mathbf{x}_3$ = 60º

c) $\lambda_1 = \lambda_2 = 1{,}5$ e $\lambda_3 = 0$

Uma solução possível para as matrizes **C** e **A** é

$$\mathbf{C} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 1 & 0 \end{bmatrix}$$

d) 100%



e) 60^o

f)

f_1	f_2
0	$\frac{1}{\sqrt{2}}$
0	$-\frac{1}{\sqrt{2}}$
$\frac{2}{\sqrt{6}}$	0
$-\frac{1}{\sqrt{6}}$	0
$-\frac{1}{\sqrt{6}}$	0

g) Impossível.

7. a) $\mathbf{R} = \begin{bmatrix} 1 & 0,9 & 0,7 \\ 0,9 & 1 & 0,7 \\ 0,7 & 0,7 & 1 \end{bmatrix}$ b) $\widehat{\mathbf{x}_1\mathbf{x}_2} = 25,8^\circ$

c) $\lambda_1 = 2,537428$ $\mathbf{A} = \begin{bmatrix} 0,94703 \\ 0,94703 \\ 0,86238 \end{bmatrix}$ d) 84,6%

e) $\mathbf{f}_1' = [\ 0,746 \quad 0,285 \quad -0,215 \quad -0,177 \quad -0,118 \quad -0,521]$

f) $\widehat{\mathbf{x}_1\mathbf{f}_1} = 18,7^\circ$ g) $\mathbf{A} = \begin{bmatrix} 0,94868 \\ 0,94868 \\ 0,73786 \end{bmatrix}$

h) A análise fatorial "explica" as correlações e 78,1% da variância total. O primeiro componente principal não "explica" tão bem as correlações mas "explica" uma proporção maior da variância (84,6%).

8. a) $\mathbf{R} = \begin{bmatrix} 1 & 0,9 & 0,6 \\ 0,9 & 1 & 0,7 \\ 0,6 & 0,7 & 1 \end{bmatrix}$ b) $25,84^\circ$

c) $\lambda_1 = 2,474065$ $\mathbf{A} = \begin{bmatrix} 0,9257 \\ 0,9608 \\ 0,8331 \end{bmatrix}$

d) 82,47% e) 22,22°

f) impossível

9. a) $\mathbf{R} = \begin{bmatrix} 1 & -0{,}5 & -0{,}5 \\ -0{,}5 & 1 & -0{,}5 \\ -0{,}5 & -0{,}5 & 1 \end{bmatrix}$ b) 120° e 120° (ou 240°)

c) $\lambda_1 = \lambda_2 = 1{,}5$ e $\lambda_3 = 0$

$\mathbf{A} = \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \\ 0 & -1 \end{bmatrix}$ ou $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix}$

Há outras alternativas.

d) 50% ; 100% e) 30° e 60° (Para a 2ª alternativa esses ângulos são 0° e 90°)

f) Impossível

Observação: É interessante notar que os vetores $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ estão em um mesmo plano.

10. a) $\mathbf{R} = \begin{bmatrix} 1 & 0{,}8 & 0 & 0 \\ 0{,}8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0{,}9 \\ 0 & 0 & 0{,}9 & 1 \end{bmatrix}$ b) 36,870° c) 90°

d) $\lambda_1 = 1{,}9$ e $\lambda_2 = 1{,}8$

$\mathbf{A} = \begin{bmatrix} 0 & \sqrt{0{,}9} \\ 0 & \sqrt{0{,}9} \\ \sqrt{0{,}95} & 0 \\ \sqrt{0{,}95} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0{,}9487 \\ 0 & 0{,}9487 \\ 0{,}9747 & 0 \\ 0{,}9747 & 0 \end{bmatrix}$ e) 92,5%

f) 90° g) 18,435° h) 0,9

11. a) $r_{12} = 0{,}625$ b) 2, 1,625, 0,375 e 0
c) 90^o d) 180^o
e) 0,9063 f) $a_{21} = a_{22} = 0{,}90139$
g) 0^o h) 0,8125 e 1

12. a) $r = 0{,}9$ b) 143,13° ou 2,498 radianos
c) $\lambda_1 = 2{,}28$, $\lambda_2 = 1{,}9$, $\lambda_3 = 0{,}72$, $\lambda_4 = 0{,}1$, $\lambda_5 = 0$
d) 83,6% e) $a_{11} = -1$, $a_{21} = a_{31} = 0{,}8$
f) 36,87° ou 0,6435 radianos g) 64% e 95%

13. $K = K_1 = K_2 = K_3 = 0{,}7847$

14. a) $r_{12} = \frac{1}{\sqrt{10}} = 0{,}31623$, $r_{34} = \frac{3}{\sqrt{10}} = 0{,}94868$

As outras 4 correlações são iguais a zero.

b) $71{,}565^o$ e 90^o

c) $\mathbf{A} = \begin{bmatrix} 0 & 0{,}8112 \\ 0 & 0{,}8112 \\ 0{,}9871 & 0 \\ 0{,}9871 & 0 \end{bmatrix}$

d) 0,8162 e) 90^o e $35{,}78^o$

f) comunalidade = 0,6581 g) 0,7735 e –1,3058

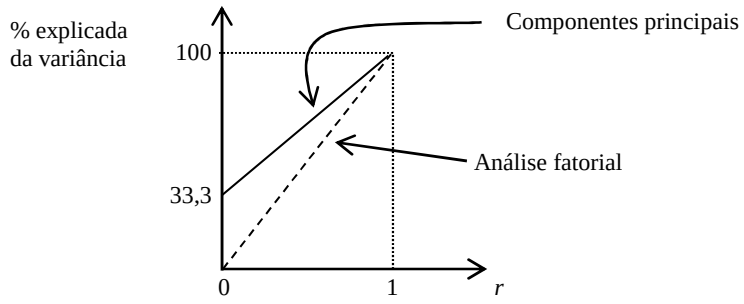
15. a) $\lambda_1 = 1 + 2r \qquad a_{11} = a_{21} = a_{31} = \sqrt{\dfrac{1+2r}{3}}$

Comunalidades iguais a $\dfrac{1+2r}{3}$ e parte da variância "explicada" = $\dfrac{1+2r}{3}$

b) $a_{11} = a_{21} = a_{31} = \sqrt{r}$

Comunalidade = Parte da variância "explicada" = r

c)



16. a) $\mathbf{R} = \begin{bmatrix} 1 & 0{,}96 & 0 \\ 0{,}96 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ b) $\widehat{\mathbf{x}_1\mathbf{x}}_2 = 16{,}26^\circ \quad \widehat{\mathbf{x}_1\mathbf{x}}_3 = 90^\circ$

c) $\mathbf{A} = \begin{bmatrix} 0{,}98995 & 0 \\ 0{,}98995 & 0 \\ 0 & 1 \end{bmatrix}$ d) 65,33% , 98,67%

e) $\widehat{\mathbf{x}_1\mathbf{f}}_1 = 8{,}13^\circ \quad \widehat{\mathbf{x}_1\mathbf{f}}_2 = 90^\circ$ f) Uma solução é $\mathbf{A} = \begin{bmatrix} 0{,}9798 \\ 0{,}9798 \\ 0 \end{bmatrix}$

17. a) $a_{11} = a_{12} = a_{13} = a_{14} = \sqrt{\dfrac{1+3r}{4}}$

Comunalidade e fração da variância explicada = $\dfrac{1+3r}{4}$

b) $a_{11} = a_{12} = a_{13} = a_{14} = \sqrt{r}$

Comunalidade e fração da variância explicada = r

18. a) $\cos\theta_{12} = 0{,}5$, $\theta_{12} = 60^\circ$

$\cos\theta_{15} = 0$, $\theta_{15} = 90^\circ$

$\cos\theta_{45} = 0{,}8$, $\theta_{45} = 36{,}87^\circ$

b) $\lambda_1 = 2$, $\lambda_2 = 1{,}8$, $\lambda_3 = \lambda_4 = 0{,}5$ e $\lambda_5 = 0{,}2$

c) $\mathbf{A} = \begin{bmatrix} 0{,}8165 & 0 \\ 0{,}8165 & 0 \\ 0{,}8165 & 0 \\ 0 & 0{,}9487 \\ 0 & 0{,}9487 \end{bmatrix}$

d) 76% e) $35{,}26^o$

f) $\mathbf{A} = \begin{bmatrix} 0{,}7071 & 0 \\ 0{,}7071 & 0 \\ 0{,}7071 & 0 \\ 0 & 0{,}8944 \\ 0 & 0{,}8944 \end{bmatrix}$

19. a) $\cos\theta_{12} = 0{,}94$, $\theta_{12} = 19{,}95^\circ$

$\cos\theta_{15} = 0$, $\theta_{15} = 90^\circ$

$\cos\theta_{45} = 0{,}96$, $\theta_{45} = 16{,}26^\circ$

b) $\lambda_1 = 2{,}88$ $\lambda_2 = 1{,}96$, $\lambda_3 = \lambda_4 = 0{,}06$ e $\lambda_5 = 0{,}04$

c) $\mathbf{A} = \begin{bmatrix} 0{,}97980 & 0 \\ 0{,}97980 & 0 \\ 0{,}97980 & 0 \\ 0 & 0{,}98995 \\ 0 & 0{,}98995 \end{bmatrix}$

d) 96,8% e) $11{,}54^o$

f) $h_1^2 = h_2^2 = h_3^2 = 0{,}96$ e $h_4^2 = h_5^2 = 0{,}98$

20. a) 64% para X_1, 49% para X_2 e 25% para X_3.

b) $100 \cdot \dfrac{1{,}38}{3} = 46\%$

c) A maior raiz característica de **R** é $\lambda_1 = 1{,}88$

$$\mathbf{A} = \begin{bmatrix} 0{,}8424 \\ 0{,}8178 \\ 0{,}7083 \end{bmatrix}$$

d) $100 \cdot \dfrac{1{,}88}{3} = 62{,}7\%$ (substancialmente maior do que 46%).

21. a) $\mathbf{A} = \begin{bmatrix} 0{,}8 \\ 0{,}8 \\ 0{,}6 \end{bmatrix}$ Comunalidades: 0,64, 0,64 e 0,36

b) $\frac{1{,}64}{3} \cdot 100 = 54{,}7\%$.

c) A maior raiz característica de **R** é 2,0704665.

$$\mathbf{A} = \begin{bmatrix} 0{,}8593 \\ 0{,}8593 \\ 0{,}7706 \end{bmatrix}$$

e) 69,0% (substancialmente maior do que no item b)

22. a) 90^o e $66{,}42^o$

b) $\lambda_1 = 1{,}6$, $\lambda_2 = 1{,}4$, $\lambda_3 = 0{,}6$ e $\lambda_4 = 0{,}4$

c) $\mathbf{A} = \begin{bmatrix} 0{,}8944 & 0 \\ 0{,}8944 & 0 \\ 0 & 0{,}8367 \\ 0 & 0{,}8367 \end{bmatrix}$ d) 75%

e) 90^o e $33{,}21^o$ f) 0,8, 0,8, 0,7 e 0,7

g) $\mathbf{A} = \begin{bmatrix} 0{,}7746 & 0 \\ 0{,}7746 & 0 \\ 0 & 0{,}6325 \\ 0 & 0{,}6325 \end{bmatrix}$

Comunalidades: 0,6, 0,6, 0,4 e 0,4.

23. a) $\lambda_1 = 1 - r$ b) $a_{11} = -a_{12} = \sqrt{\frac{1-r}{2}}$ c) $\varphi = \frac{1-r}{2}$

d) $\begin{bmatrix} 1 \\ r \end{bmatrix}$ e) $\psi = \frac{1+r^2}{2}$

Com $r < 0$, temos $r^2 \leq -r$. Segue-se que $1 + r^2 \leq 1 - r$ e $\psi \leq \varphi$, com a igualdade sendo válida apenas se $r = -1$.

24. a) 1,9

b) $a_{11} = \sqrt{0,95} = 0,97468$ e $a_{12} = -a_{11}$

c) 0,95

d) $\begin{bmatrix} 1 \\ -0,9 \end{bmatrix}$

e) 0,905

f) 154,16º

g) 12,92º

h) 167,08º

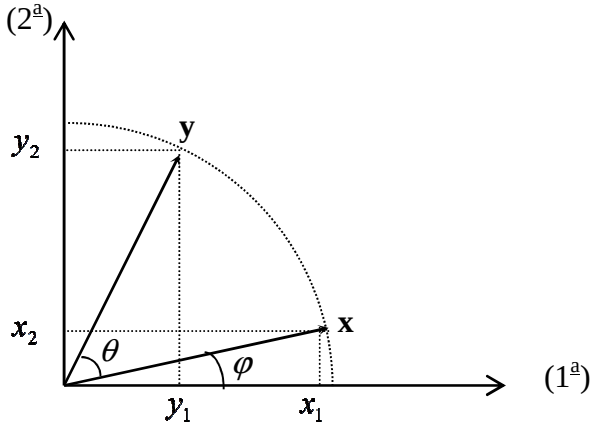
i) 154,16º

APÊNDICE: ROTAÇÃO DE VETORES

A1. Rotação de um vetor em um plano (espaço bidimensional)

A figura A1 mostra os eixos do espaço bidimensional, o vetor na posição inicial (**x**) e o vetor após a rotação (**y**). O ângulo de rotação (em sentido anti-horário) é igual a θ.

Figura A1.



Temos as seguintes relações:

$$x_1 = R\cos\varphi$$
$$x_2 = R\,\mathrm{sen}\,\varphi$$
$$y_1 = R\cos(\varphi+\theta) = R(\cos\varphi\cos\theta - \mathrm{sen}\,\varphi\,\mathrm{sen}\,\theta)$$
$$y_2 = R\,\mathrm{sen}(\varphi+\theta) = R(\mathrm{sen}\,\varphi\cos\theta + \mathrm{sen}\,\theta\cos\varphi)$$
$$y_1 = x_1\cos\theta - x_2\,\mathrm{sen}\,\theta$$
$$y_2 = x_2\cos\theta + x_1\,\mathrm{sen}\,\theta = x_1\,\mathrm{sen}\,\theta + x_2\cos\theta$$
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\mathrm{sen}\,\theta \\ \mathrm{sen}\,\theta & \cos\theta \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

ou

$$\begin{bmatrix} y_1 & y_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} \cos\theta & \mathrm{sen}\,\theta \\ -\mathrm{sen}\,\theta & \cos\theta \end{bmatrix}$$

Essa última expressão pode ser escrita como

$$\mathbf{y}' = \mathbf{x}'\mathbf{T}'$$

Notar que $\mathbf{T}\mathbf{T}' = \mathbf{I}$

Verifica-se que a matriz de transformação ortogonal (**T**), quando multiplica **x**, faz a rotação desse vetor para a posição **y**.

A2. Rotação de dois vetores no plano definido pelos dois vetores, em um espaço tridimensional

a) Primeiro exemplo

Vamos considerar 2 vetores com módulo igual a 1 e ortogonais entre si, como mostra a figura A2.

$$\mathbf{f}_1' = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

$$\mathbf{f}_2' = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Consideremos a matriz de transformação ortogonal

$$\mathbf{T}' = \begin{bmatrix} \cos\theta & \operatorname{sen}\theta \\ -\operatorname{sen}\theta & \cos\theta \end{bmatrix}$$

Para um ângulo de rotação $\theta = 45^o$ temos

$$\mathbf{T}' = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

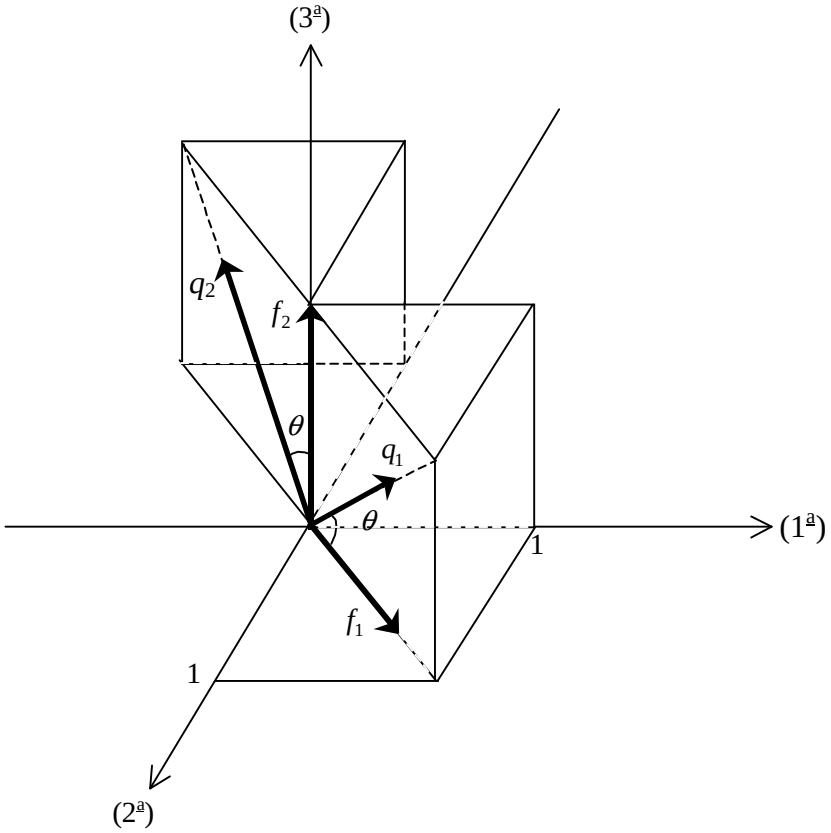
$\mathbf{Q} = \mathbf{T}'\mathbf{F}$ [expressão (6.48)].

Verifica-se que

$$\mathbf{Q} = \mathbf{T}'\mathbf{F} = \begin{bmatrix} \mathbf{q}_1' \\ \mathbf{q}_2' \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{\sqrt{2}}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

Note-se que após a rotação os dois vetores continuam com módulo igual a 1 e ortogonais entre si.

Figura A2



b) Segundo exemplo

Vamos considerar, novamente, dois vetores ($\mathbf{f}_1$ e $\mathbf{f}_2$) com módulo igual a 1 e ortogonais entre si.

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}$$

A figura A3 ilustra a rotação no espaço tridimensional. O vetor $\mathbf{f}_1$ está sobre a linha OC e o vetor $\mathbf{f}_2$ está sobre a linha OA. Para facilitar a compreensão, o leitor deve assinalar os vetores $\mathbf{f}_1$ e $\mathbf{f}_2$ na Figura 3, lembrando que são vetores com módulo igual a 1 e notando que OA, como diagonal de um quadrado com aresta igual a 1, tem comprimento $\sqrt{2}$ e que OC, como diagonal de um cubo com aresta igual a 1, tem comprimento $\sqrt{3}$.

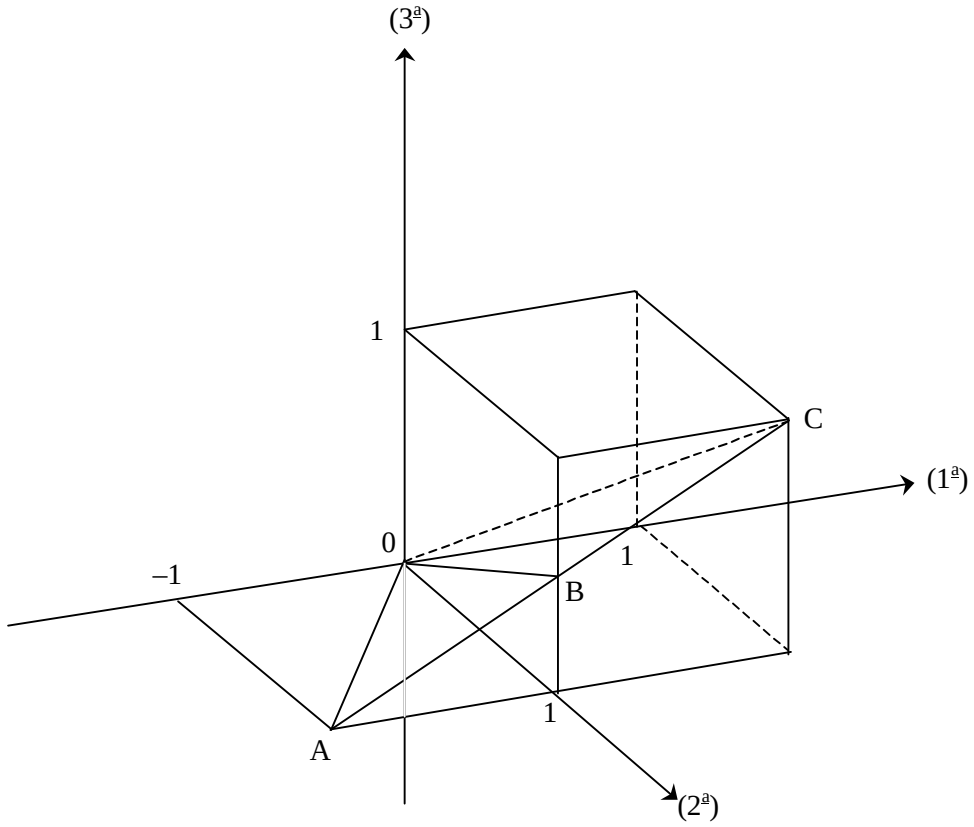
É realizada uma rotação com ângulo $\theta = \beta$, tal que $\cos\theta = \sqrt{\frac{3}{5}}$.

Segue-se que $\text{sen}\,\theta = \sqrt{\frac{2}{5}}$ e $\theta = \beta = 39{,}23^{o}$

Após a rotação, os novos vetores são $\mathbf{q}_1$ e $\mathbf{q}_2$, de maneira que

$$\begin{bmatrix} q_1' \\ q_2' \end{bmatrix} = \mathbf{Q} = \mathbf{T'F} = \begin{bmatrix} \sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} \\ -\sqrt{\frac{2}{5}} & \sqrt{\frac{3}{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\sqrt{\frac{5}{6}} & \frac{1}{\sqrt{30}} & -\sqrt{\frac{2}{15}} \end{bmatrix}$$

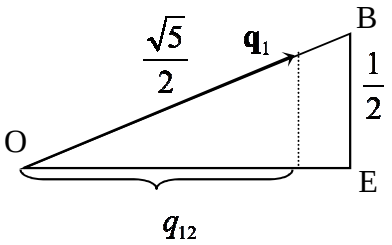
Figura A3



Verifica-se que o primeiro fator "gira" da direção $\overline{\mathrm{OC}}$ $(\mathbf{f}_1)$ para a direção $\overline{\mathrm{OB}}$ $(\mathbf{q}_1)$

Tendo em vista determinar os elementos de $\mathbf{q}_1$ com base na geometria, destaca-se, na figura A4, o triângulo OBE da figura A3, no plano dos eixos da 2ª e da 3ª dimensão.

Figura A4



$$\overline{|\mathrm{OB}|} = \frac{\sqrt{5}}{2}$$

$$\overline{|\mathrm{OE}|} = 1$$

Por semelhança de triângulos, e lembrando que o módulo de $\mathbf{q}_1$ é igual a 1, temos

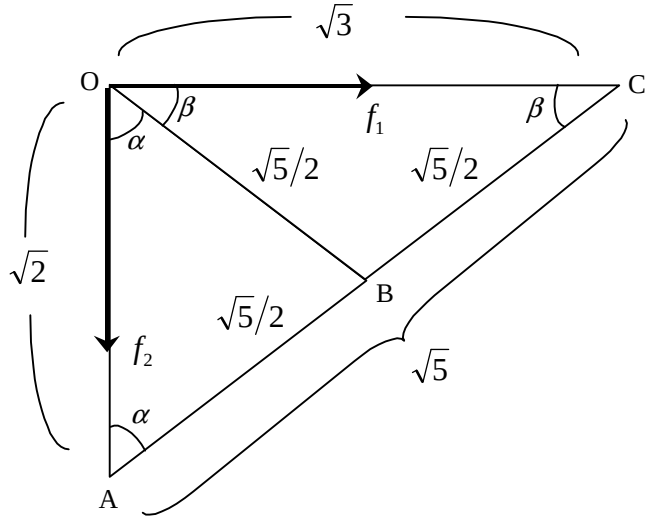
$$\frac{q_{12}}{1} = \frac{1}{\sqrt{5}/2}$$

Segue-se que $q_{12} = \frac{2}{\sqrt{5}}$

Analogamente, de $\frac{q_{13}}{1/2} = \frac{1}{\sqrt{5}/2}$, obtemos $q_{13} = \frac{1}{\sqrt{5}}$

A figura A5 destaca o triângulo OAC da figura A3, facilitando a visualização das dimensões e das rotação de $\mathbf{f}_1$ (na linha OC) para $\mathbf{q}_1$ (na linha OB).

Figura A5



$$\frac{1}{\sqrt{3}}\overline{\mathrm{OC}} = f_1$$

$$\frac{1}{\sqrt{2}}\overline{\mathrm{OA}} = f_2$$

A seguir vamos considerar, alternativamente, um ângulo de rotação $\theta = -\alpha$. Então

$$\cos\alpha = \sqrt{\frac{2}{5}} \quad , \quad \operatorname{sen}\alpha = \sqrt{\frac{3}{5}} \quad , \; \theta = -\alpha = -50{,}77^{o}$$

$$\cos\theta = \sqrt{\frac{2}{5}} \quad , \quad \operatorname{sen}\theta = -\sqrt{\frac{3}{5}}$$

$$\mathbf{Q} = \mathbf{T}'\mathbf{F} = \begin{bmatrix} \sqrt{\frac{2}{5}} & -\sqrt{\frac{3}{5}} \\ \sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{5}{\sqrt{30}} & -\frac{1}{\sqrt{30}} & \sqrt{\frac{2}{15}} \\ 0 & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{5}{6}} & -\frac{1}{\sqrt{30}} & \sqrt{\frac{2}{15}} \\ 0 & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

Verifica-se que nesse caso é o segundo fator que “gira” da direção $\overline{\mathrm{OA}}$ $(\mathbf{f}_2)$ para a direção $\overline{\mathrm{OB}}$ $(\mathbf{q}_2)$

A3. Rotação de dois vetores no plano definido pelos dois vetores, em um espaço com duas ou mais dimensões

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_1' \\ \mathbf{q}_2' \end{bmatrix} = \mathbf{T}'\mathbf{F} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \end{bmatrix}$$

Se os vetores iniciais tem comprimento igual a 1 e são ortogonais entre si, temos

$$\mathbf{f}_1'\mathbf{f}_1 = 1 \qquad (1)$$

$$\mathbf{f}_2'\mathbf{f}_2 = 1 \qquad (2)$$

$$\mathbf{f}_1'\mathbf{f}_2 = \mathbf{f}_2'\mathbf{f}_1 = 0 \qquad (3)$$

Se θ é o ângulo de rotação, temos

$$\cos\theta = \cos(\mathbf{f}_1 \widehat{\ } \mathbf{q}_1) = \cos(\mathbf{f}_2 \widehat{\ } \mathbf{q}_2)$$

ou

$$\cos\theta = \mathbf{q}_1'\mathbf{f}_1 = \mathbf{q}_2'\mathbf{f}_2$$

$$\cos\theta = (t_{11}\mathbf{f}_1' + t_{12}\mathbf{f}_2')\mathbf{f}_1 = (t_{21}\mathbf{f}_1' + t_{22}\mathbf{f}_2')\mathbf{f}_2$$

Lembrando (1), (2) e (3), obtemos

$$\cos\theta = t_{11} = t_{22} \qquad (4)$$

A rotação, obviamente, não altera o comprimento do vetor. Então

$$\mathbf{q}_1'\mathbf{q}_1 = 1, \text{ ou seja,}$$

$$(t_{11}\mathbf{f}_1' + t_{12}\mathbf{f}_2')(t_{11}\mathbf{f}_1 + t_{12}\mathbf{f}_2) = 1$$

Lembrando (1), (2) e (3), obtemos

$$t_{11}^2 + t_{12}^2 = 1 \qquad (5)$$

Como a rotação mantém a ortogonalidade entre os vetores, temos

$$\mathbf{q}_1'\mathbf{q}_2 = 0$$

$$(t_{11}\mathbf{f}_1' + t_{12}\mathbf{f}_2')(t_{21}\mathbf{f}_1 + t_{22}\mathbf{f}_2) = 0$$

Lembrando (1), (2) e (3), obtemos

$$t_{11}t_{21} + t_{12}t_{22} = 0 \qquad (6)$$

Lembrando (4), segue-se que

$$t_{21} = -t_{12} \qquad (7)$$

De (4), (5) e (7) conclui-se que a matriz $\mathbf{T}'$ sempre pode ser representada por

$$\mathbf{T}' = \begin{bmatrix} \cos\theta & \pm \operatorname{sen}\theta \\ \mp \operatorname{sen}\theta & \cos\theta \end{bmatrix}$$

BIBLIOGRAFIA

ANGRIST, Joshua D. e PISCHKE, Jörn - Steffen (2009). *Mostly Harmless Econometrics*. Princeton University Press.

BELSLEY, D.A.; KUH, E. e WELSCH, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley.

BERKSON, J. (1944) Application of the logistic function to bio-assay. *J. Am. Statist. Ass.*, Boston, *39*:357-65.

BLISS, C. I. (1935) The calculation of the dosage mortaly curve. *Ann. Appl. Biol.*, Cambridge, *22*:134-67.

CHATFIELD, C. e COLLINS, A.J. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall.

COOPER, J. C. B. (1983). Factor Analysis: An Overview. *The American Statistician 37*(2): 141-147, maio/1983.

DACHS, J.N.W. e CARVALHO, J.F. de (1984) *Diagnóstico em Regressão.* 6º SINAPE, Instituto de Matemática, UFRJ.

FINNEY, D.J. (1952) *Probit Analysis*. Cambridge, University Press.

GREENE, William H. (2000) *Econometric Analysis*. 4ª ed. Prentice-Hall.

HARMAN, H. H. (1976). *Modern Factor Analysis*. 3ª ed. The University of Chicago Press.

HOFFMANN, R. (1992). A dinâmica da modernização da agricultura em 157 microregiões homogêneas do Brasil. *Revista de Economia e Sociologia Rural. 30*(4):271-290.

HOFFMANN, R. e KAGEYAMA, A.A. (1985). Modernização da agricultura e distribuição da renda no Brasil. *Pesquisa e Planejamento Econômico. 15*(1): 171-208.

HOFFMANN, R. e KASSOUF, A.L. (1989). Modernização e desigualdade na agricultura brasileira. *Revista Brasileira de Economia. 43*(2):273-303.

HOFFMANN, R. (2016). Análise de regressão: uma introdução à econometria.

IBGE (1999) *Pesquisa de orçamentos familiares 1995-1996.* Vol. 1: Despesas, recebimentos e características das famílias, domicílios, pessoas e locais de compra. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística.

JOHNSON, R.A. e WICHERN, D.W. (1982). *Applied Multivariate Statistical Analysis*. Prentice Hall.

JUDGE, G.G.; HILL, R.C.; GRIFFITHS, W.E.; LÜTKEPOHL, H. e LEE, T.C. (1988) *Introduction to the Theory and Practice of Econometrics*. 2ª ed. John Wiley.

KASSOUF, Ana Lúcia (1988) *Previsão de preços na pecuária de corte do Estado de São Paulo*. Piracicaba, ESALQ-USP (Dissertação de Mestrado).

LANGE, O. (1967). *Introdução à econometria*. 2ª Ed. Rio de Janeiro, Fundo de Cultura.

LAWLEY, D.N. e MAXWELL, A.E. (1971). *Factor Analysis as a Statistical Method*. 2ª ed. American Elsevier.

MACHADO, Amauri de A. (1989) *Diagnóstico em Regressão Linear*. Depto. de Ciências Exatas, ESAL, Lavras.

MADDALA, G. 1983 *Limited Dependent and Qualitative Variables in Econometrics*. New York, Cambridge University Press.

MARINHO, Emerson e ARAUJO, Jair (2010). Pobreza e o sistema de seguridade social rural no Brasil. Rio de Janeiro, *Revista Brasileira de Economia* 64(2): 161-174.

MARINHO, Emerson; LINHARES, Fabrício e CAMPELO, Guaracyane (2011). Os programas de transferências de renda do governo impactam a pobreza no Brasil? Rio de Janeiro, *Revista Brasileira de Economia* 65(3):267-288.

OKAWA, Hiroshigue (1985) *Análise harmônica das variações dos preços e das quantidades de sardinha fresca no mercado atacadista de São Paulo - 1981/82*. Piracicaba, ESALQ-USP (Dissertação de Mestrado).

RODRIGUES, Milton da Silva. (1970) Dicionário Brasileiro de Estatística. 2ª ed. Rio de Janeiro, Fundação IBGE.

ROSSI, José W. (1982) Elasticidade de Engel para Dispêndios Familiares na Cidade do Rio de Janeiro. *Pesq. Plan. Econ. 12*(2):579-606.

ROSSI, José W. (1984) *As Variáveis Binárias em Análise de Regressão: Teoria e Aplicação*. IBGE.

SAS System for Regression, 1986 Edition.

THEIL, H. (1971). *Principles of econometrics*. New York, John Wiley.

ÍNDICE ANALÍTICO

T

V

W