Licenciatura em Economia

# Econometria I

ISCTE-IUL, Business School, 2023/2024

# EXERCÍCIOS PRÁTICOS

Docente: Joaquim J. S. Ramalho
E-mail: jjsro@iscte-iul.pt
Gabinete: D5.10

## 1. Introdução

**1.1** A tabela seguinte contém despesas mensais com a habitação ($X$) de 10 famílias.

| Família | $X$ (euros) |
|---|---|
| 1 | 300 |
| 2 | 440 |
| 3 | 350 |
| 4 | 1100 |
| 5 | 640 |
| 6 | 480 |
| 7 | 450 |
| 8 | 700 |
| 9 | 670 |
| 10 | 530 |

a) Calcule a despesa média mensal com a habitação.
b) Calcule a despesa mediana mensal com a habitação.
c) Se as despesas com a habitação fossem medidas em centenas de euros, em vez de euros, qual seria a despesa média e mediana?
d) Suponha que a família 8 aumenta as suas despesas mensais com a habitação em 900 euros, mas as despesas de todas as outras famílias se mantêm. Calcule a despesa média e mediana com a habitação.

**1.2** Considere que $Y$ e $X$ são duas variáveis aleatórias. Um investigador recolheu uma amostra com os seguintes oito pares de elementos:

| $Y$ | $X$ |
|---|---|
| 5.1 | 10.3 |
| 4.5 | 8.4 |
| 7.6 | 9.3 |
| 9.5 | 15.4 |
| 13.2 | 14.1 |
| 12.0 | 11.3 |
| 2.1 | 5.2 |
| 3.5 | 4.7 |

a) Estime as médias, as variâncias e os desvios padrão de $Y$ e $X$.
b) Estime a covariância e o coeficiente de correlação entre $Y$ e $X$.
c) Se $X$ for multiplicado por 10, qual o efeito nos cálculos da alínea anterior? Comente.

**1.3** Com base numa amostra de 10 observações, foram obtidos os seguintes resultados:

$$\Sigma Y = 1110 \qquad \Sigma Y^2 = 132100 \qquad \Sigma XY = 205500$$
$$\Sigma X = 1700 \qquad \Sigma X^2 = 322000 \qquad r = 0.98$$

Ao verificar os cálculos encontraram-se dois pares de observações que estavam incorretos, ou seja, haviam sido utilizados:

| *Y* | *X* | | *Y* | *X* |
|---|---|---|---|---|
| 90 | 120 | em vez de | 80 | 110 |
| 140 | 220 | | 150 | 210 |

Qual será o efeito deste erro no coeficiente de correlação? Obtenha o valor de *r* correto.

**1.4** Seja *X* o salário anual de um professor universitário nos EUA, em milhares de dólares. Suponha que o salário médio é 52.3 com um desvio padrão de 14.6. Obtenha a média e o desvio padrão quando o salário é medido em dólares.

**1.5** Suponha que duas variáveis *Y* e *X* estão relacionadas pela expectativa condicionada:

$$E(Y|X) = 0.7 + 0.002X$$

a) Obtenha o valor esperado de *Y* quando *X* = 800. Obtenha também $E(Y|X = 1400)$. Comente a diferença.
b) Se o valor esperado de *X* for a média dos seus dois valores referidos na alínea anterior, qual acha que será o valor esperado de *Y*? Confirme a sua resposta usando a lei das expectativas iteradas.

**1.6** Suponha que um professor da população referida em (1.4) ganha 62760 dólares por ano.

a) Qual a percentagem exacta em que este salário excede a média?
b) Calcule agora a percentagem aproximada usando a diferença nos logaritmos naturais. Comente.

## 2. O Modelo de Regressão Linear Simples

**2.1** Seja *fert* o número de filhos de uma mulher, e *educ* o número de anos de escolaridade da mesma mulher. Um modelo simples relacionando a fertilidade com a educação será:

$$fert = \beta_0 + \beta_1 educ + u$$

onde *u* é o erro não observável.

a) Que fatores poderão estar incluídos em *u*? Estarão esses fatores provavelmente relacionados com o nível de educação?
b) Será que uma regressão simples poderá mostrar o efeito *ceteris paribus* da educação na fertilidade? Explique.

**2.2** Um dos fatores que poderá influenciar o peso do bebé à nascença (*peso*) é o número médio de cigarros fumado por dia pela mãe (*cigs*). Usando dados de 1388 nascimentos, estimou-se a seguinte regressão linear simples:

$$\widehat{peso} = 3.4 - 0.015cigs$$

a) Qual o peso à nascença previsto para o bebé de uma mãe não fumadora? E se a mãe fumar um maço (20 cigarros) por dia? Comente a diferença.
b) Será que esta regressão simples capta necessariamente uma relação causal entre o peso do bebé à nascença e o hábito de fumar da mãe?

**2.3** Usando observações do rendimento e consumo anuais (ambos medidos em dólares) para 100 famílias, foi estimada a seguinte função consumo linear:

$$\widehat{consumo} = -124.84 + 0.85rend\,; \quad n = 100; \;\; R^2 = 0.162$$

a) Interprete o termo constante nesta equação e comente o seu sinal.
b) Qual o consumo previsto se o rendimento familiar é 30 000 dólares?
c) Com *rend* no eixo das abcissas, trace um gráfico da propensão marginal ao consumo (PmgC) e da propensão média ao consumo (PMeC) estimadas.

**2.4** Assuma que lhe são fornecidos os seguintes dados sobre a despesa em publicidade (*X*) e as vendas de detergente (*Y*) de 10 empresas, em euros, num determinado período:

| *X* | 2 | 4.8 | 3 | 5 | 2.5 | 4 | 3.4 | 4.6 | 3.5 | 4.2 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Y* | 5 | 9.5 | 6 | 10 | 5.5 | 7.7 | 7.2 | 9 | 7 | 8.6 |

O modelo proposto para a relação existente entre as variáveis é: $Y = \beta_0 + \beta_1 X + u$.

a) Estime e interprete $\beta_0$ e $\beta_1$. Quanto se espera que as vendas aumentem, caso se gastem mais €3 em publicidade?
b) Represente graficamente os valores observados, bem como a reta estimada.

c) Calcule o valor do resíduo para cada observação, bem como a média de todos os resíduos.
d) Que proporção da variação das vendas destas 10 empresas é explicada pela publicidade? Justifique.
e) Qual o valor das vendas previsto para uma empresa que gaste €5,20 em publicidade?

**2.5** Use os dados em SONO.DTA para estudar se existe um "trade-off" entre o tempo passado a dormir por semana e o tempo passado em trabalho pago. Qualquer das variáveis poderia ser usada como independente. Para concretizar, estime o modelo

$$sono = \beta_0 + \beta_1 trab + u$$

onde *sono* é o tempo de sono noturno semanal, em minutos, e *trab* é o tempo de trabalho semanal, em minutos.

a) Apresente a equação estimada, juntamente com o número de observações e o $R^2$. O que é que significa o termo constante nesta equação?
b) Represente graficamente os valores observados e os valores previstos.
c) Se trabalharmos mais 2 horas por semana, em quanto se estima que diminua o tempo semanal de sono? O que pensa da magnitude deste efeito?

**2.6** Certo investigador recolheu uma amostra de 5 pares de valores ($X_i$, $Y_i$) e estimou uma recta cuja equação é:

$$\hat{Y}_i = 10 - 2X_i$$

Mais tarde, querendo estimar outro modelo, procurou os dados originais e verificou que apenas tinha a seguinte informação:

$$X_1 = 2; \quad X_2 = 3; \quad X_3 = 5; \quad X_4 = 4; \quad X_5 = 1.$$
$$\hat{u}_1 = -1; \quad \hat{u}_2 = ?; \quad \hat{u}_3 = 0; \quad \hat{u}_4 = ?; \quad \hat{u}_5 = 0.$$

Recordando as propriedades algébricas dos estimadores dos mínimos quadrados, recupere os valores de $Y_i$.

**2.7** Considere os seguintes pares de observações, relativos às variáveis *X* e *Y*:

| *Y* | 8 | 6 | 1 | 2 | 4 |
|---|---|---|---|---|---|
| *X* | 2 | 1 | 0 | 3 | 4 |

O modelo proposto é da forma: $Y = \beta_0 + \beta_1 X + u$. Usando o método dos mínimos quadrados, sem recorrer a software, estime:

a) Os parâmetros $\beta_0$ e $\beta_1$.
b) A soma de quadrados residuais.
c) A variância do erro.
d) A variância de $\hat{\beta}_0$ e a variância de $\hat{\beta}_1$.

e) O coeficiente de determinação.

**2.8** Um estudo das colheitas de 7 agricultores forneceu os seguintes resultados**,** onde *Y é* a produção em toneladas e *X* é a quantidade de chuva medida em *mm* por período de colheita:

$$\text{Colheita A:} \quad \hat{Y} = 200 + 0.8X\ ;\ \ \hat{\sigma}^2 = 30;\ R^2 = 0.49$$

$$\text{Colheita B:} \quad \hat{Y} = 500 + 1.2X\ ;\ \ \hat{\sigma}^2 = 20;\ R^2 = 0.81$$

a) Se no período em causa a chuva fosse de 95*mm,* qual das colheitas lucraria mais com a existência de chuva? Indique sucintamente as bases da sua resposta.
b) Para qual das duas colheitas se poderá prever a produção com maior rigor? Porquê?
c) Estime $\mathrm{Var}(\hat{\beta}_1)$ para a colheita A.

## 3. O Modelo de Regressão Linear Múltipla

**3.1** Considere a seguinte regressão estimada, referente a trabalhadores do sexo masculino:

$$\widehat{educ} = 10.36 - 0.094\, irmaos + 0.131\, educm + 0.210\, educp\,; n = 722;\ R^2 = 0.214$$

onde *educ* são os anos de escolaridade, *irmaos* é o número de irmãos e *educm* e *educp* são o número de anos de escolaridade da mãe e do pai, respetivamente.

a) O efeito de *irmaos* é o esperado? Explique. Mantendo *educm* e *educp* constantes, quanto é que o número de irmãos deve aumentar para que a educação prevista se reduza um ano? (É aceitável um número não inteiro.)
b) Discuta a interpretação do coeficiente de *educm*.
c) Suponha que o André não tem irmãos, e que os seus pais têm ambos 12 anos de escolaridade. O Bento também não tem irmãos, e os seus pais têm ambos 16 anos de escolaridade. Qual a diferença prevista nos anos de educação do Bento e do André?

**3.2** O modelo seguinte é uma versão simplificada do modelo de regressão múltipla de Biddle e Hamermesh (1990) para estudar o "trade-off" entre o tempo passado a dormir e a trabalhar e para analisar outros fatores que afectam o sono:

$$sono = \beta_0 + \beta_1 trab + \beta_2 educ + \beta_3 idade + u\,,$$

onde *sono* e *trab* são medidos em minutos por semana e *educ* e *idade* em anos.

a) Se os adultos renunciam a sono em troca de trabalho, qual é o sinal de $\beta_1$?
b) Que sinais acha que $\beta_2$ e $\beta_3$ irão ter?
c) Com base nas estimativas abaixo apresentadas, preveja quantos minutos de sono irá perder alguém que decida trabalhar mais 5 horas por semana.

$$\widehat{sono} = 3638.25 - 0.148\, trab - 11.13\, educ + 2.20\, idade\,;\ \ n = 706;\ R^2 = 0.113$$

d) Discuta o sinal e a magnitude do coeficiente de *educ*.
e) Diria que o trabalho, a educação e a idade explicam muita da variação no sono? Que outros fatores poderiam afetar o tempo passado a dormir? Estarão esses fatores provavelmente correlacionados com o trabalho?

**3.3** Com base numa amostra de 55 observações e no modelo abaixo apresentado, obtiveram-se os seguintes resultados:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$$

$$\left(X'X\right)^{-1}=\begin{bmatrix} 2.1 & -0.15 & -0.05 \\ -0.15 & 0.1 & -0.04 \\ -0.05 & -0.04 & 0.01 \end{bmatrix} \quad X'Y=\begin{bmatrix} 74 \\ 282 \\ 375 \end{bmatrix} \quad Y'Y=4019.75$$

a) Estime os coeficientes de regressão.
b) Comente o ajustamento do modelo.

**3.4** Use os dados em PRECASA.DTA para estimar o modelo

$$preco = \beta_0 + \beta_1 area + \beta_2 quartos + u$$

onde o preço de uma casa, em $10^3$ dólares, depende da área ($m^2$) e do número de quartos.

a) Escreva os resultados na forma de equação, incluindo *n* e $R^2$.
b) Qual o aumento previsto no preço de uma casa com mais um quarto, mantendo a área constante?
c) Qual o aumento estimado no preço de uma casa com um quarto adicional de $13m^2$ de tamanho? Compare a resposta com b).
d) A primeira casa da amostra tem uma área de $226m^2$ e 4 quartos. Encontre o preço de venda previsto para esta casa a partir da reta de regressão dos mínimos quadrados.
e) O preço de venda verificado para a primeira casa da amostra foi de 300 mil dólares. Encontre o resíduo para esta casa. O valor deste sugere que o comprador fez um bom ou um mau negócio?

**3.5** Considere que as vendas (em quantidade) de um determinado tipo de cigarros ($Y$) são função do preço ($X_1$) e dos gastos em publicidade ($X_2$), podendo a relação ser expressa por um modelo de regressão linear. Foram obtidos os seguintes dados, em euros, referentes a 10 empresas:

| $Y$ | 80 | 100 | 110 | 120 | 90 | 85 | 70 | 105 | 95 | 115 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 2 | 1 | 1 | 2 | 1.2 | 1.5 | 1.5 | 2 | 2.5 | 2 |
| $X_2$ | 25 | 30 | 40 | 45 | 30 | 20 | 30 | 60 | 50 | 60 |

a) Estime os coeficientes da regressão de $Y$ sobre $X_1$, bem como o erro padrão da regressão e o $R^2$. O que pensa da estimativa de $\beta_1$?
b) Estime a regressão de $Y$ sobre $X_1$ e $X_2$, e compare os resultados com a).
c) Qual o montante previsto de vendas de cigarros para uma empresa que pratique um preço de venda e tenha gastos em publicidade iguais aos valores médios da amostra?

**3.6** Num estudo que relaciona a média das classificações de alunos universitários com o tempo passado em várias atividades, foi distribuído um inquérito a vários estudantes. Perguntou-se quantas horas eles passavam por semana em quatro atividades: estudo, sono, trabalho e lazer. Dado que estas categorias esgotam o tempo disponível, a soma das horas passadas nas quatro atividades, para cada estudante, deve ser 168. O modelo proposto foi:

$$média = \beta_0 + \beta_1 estudo + \beta_2 sono + \beta_3 trabalho + \beta_4 lazer + u$$

a) No modelo acima, faz sentido manter constantes *sono*, *trabalho*, e *lazer*, enquanto se altera o *estudo*?
b) Por que razão este modelo viola a premissa da ausência de colinearidade perfeita?
c) Como é que o modelo poderá ser reformulado de forma a que satisfaça a premissa indicada em b) e os parâmetros tenham uma interpretação útil?

**3.7** Suponha que as seguintes observações são o resultado da medição da produção agrícola (*Y*) obtida em várias parcelas de terra, para cada nível de adubo (*X*):

| *X* | *Y* | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 1 | |
| 2 | 4 | 1 | 5 | |
| 3 | 5 | 6 | | |
| 4 | 2 | 1 | 3 | 2 |

O modelo proposto é $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$.

a) Estime a função de produção, e comente a diferença entre este tipo de dados e os usualmente obtidos na análise económica empírica.
b) Encontre o efeito na produção prevista de se usar mais uma unidade de adubo e preveja a produção máxima que é possível obter.
c) Esboce um gráfico para a produtividade marginal e a produtividade média estimadas. (Inclua explicitamente o ponto de interseção entre as duas linhas.)

**3.8** Admita a seguinte função de Custo: $C_i = \alpha Q_i^{\beta}$.
a) Mostre que $\beta$ é a elasticidade do custo em ordem à produção.
b) Construa o modelo de regressão linear adequado para obter estimativas de $\alpha$ e $\beta$ e calcule-as, com base na amostra seguinte:

| $C_i$ | 8 | 28 | 70 | 52 | 32 |
|---|---|---|---|---|---|
| $Q_i$ | 2 | 4 | 6 | 5 | 4 |

**3.9** Use os dados de SAL.DTA para este exercício.

a) Estime a equação $sal = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 antig + u$, calcule os resíduos e represente graficamente um histograma.
b) Repita a), mas com ln(*sal*) como variável dependente.
c) Com qual dos modelos anteriores lhe parece que o pressuposto da normalidade do erro está mais próximo de ser satisfeito?

**3.10** Considere o modelo $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_{i1}) + \beta_2 X_{i2} + u_i$ e os dados do exercício 3.5.

a) Demostre teoricamente que se os valores de *Y* e $X_1$ forem multiplicados pelas constantes *c* e *d*, respectivamente, apenas o valor de $\beta_0$ se alterará.
b) Comprove a alínea anterior estimando o modelo quer com os dados originais quer após alteração das unidades de medida, com *c* = 2 e *d* = 10.

**3.11** Considere o modelo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$ e o conjunto de observações:

| | | | | | |
|---|---|---|---|---|---|
| $X_1$ | -1 | 2 | 0 | 2 | 2 |
| $X_2$ | 1 | -1 | -2 | 1 | 1 |
| $Y$ | 2 | 6 | 5 | 6 | 6 |

a) Estime a regressão de $Y$ sobre $X_1$ e $X_2$.

b) Estime o modelo estandardizado $z_{iY} = \beta_0 + \beta_1 z_{i1} + \beta_1 z_{i2} + z_{iu}$, onde $z_{iY} = \frac{Y_i - \bar{Y}}{S_Y}$ e $z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_{X_j}}$, $j = 1, 2$, sendo $S_Y$ e os $S_{X_j}$'s desvios padrão amostrais.

c) Mostre como a partir do modelo estimado em b) poderia obter as estimativas encontradas em a).

d) Multiplique os valores de $X_1$ por 10 e volte a resolver as alíneas a) e b). Compare os resultados obtidos ao nível dos coeficientes estimados e do $R^2$.

## 4. Inferência e Previsão

**4.1** A variável *ID* é a despesa em investigação e desenvolvimento (I&D) como percentagem das vendas. As vendas estão medidas em milhões de dólares. A variável *mglucro* representa os lucros como percentagem das vendas. Usando dados de 32 empresas da indústria química, foi estimada a seguinte equação (por baixo de cada coeficiente, entre parênteses, é dada a estimativa do respectivo desvio-padrão):

$$\widehat{ID} = \underset{(1.369)}{0.472} + \underset{(0.216)}{0.321}\, ln(vendas) + \underset{(0.046)}{0.050} mglucro; \quad n = 32; \quad R^2 = 0.099$$

a) Interprete o coeficiente de ln(*vendas*). Em particular, se as vendas aumentam 10%, qual a variação em pontos percentuais estimada para *ID*? Trata-se de um efeito elevado em termos económicos?
b) Teste a hipótese de que a intensidade de I&D não se altera com as vendas, contra a alternativa de que aumenta com as vendas. Faça o teste usando um nível de significância de 5%.
c) A variável *mglucro* tem um efeito estatisticamente significativo em *ID*?

**4.2** Reconsidere o Exercício 3.4, baseado nos dados do ficheiro PRECASA.DTA. Use agora o logaritmo natural do preço da habitação como variável dependente, ou seja, $\ln(preco) = \beta_0 + \beta_1 area + \beta_2 quartos + u$.

a) Estime o modelo.
b) Considerando um nível de significância de 5%, teste as seguintes hipóteses: $H_0\colon \beta_1 = 0.1$; $H_0\colon \beta_2 = 0$
c) Calcule intervalos de confiança a 95% para $\beta_1$ e $\beta_2$.
d) Estime a variação percentual que ocorre no preço da habitação quando um quarto de $14\text{m}^2$ é acrescentado a essa habitação.
e) Determine o erro padrão da variação calculada na alínea anterior.
f) Use o erro padrão calculado em e) para estimar um intervalo de confiança a 95% para a variação calculada em d).

**4.3** A análise de regressão pode ser usada para testar se o mercado usa de forma eficiente a informação na avaliação das ações. Seja *y* a rentabilidade da ação, ou rendimento total obtido por guardar uma ação de uma empresa no período de quatro anos desde o final de 1990 ao final de 1994. A *hipótese dos mercados eficientes* diz que essas rentabilidades não devem estar sistematicamente relacionados com informação conhecida em 1990. Se as características da empresa conhecidas no princípio do período ajudarem a prever a rentabilidade da ação, então essa informação pode ser usada na escolha dos ativos. Seja *rdk* o rácio da dívida pelo capital da empresa, *ga* os ganhos por ação, *rendliq* o rendimento líquido, *sal* a remuneração total do executivo. Para 1990 foi estimada a seguinte equação:

$$\hat{y} = -\underset{(6.89)}{14.37} + \underset{(0.201)}{0.321} rdk + \underset{(0.078)}{0.043} ga - \underset{(0.0047)}{0.0051} rendliq + \underset{(0.0022)}{0.0035} sal; n = 142; R^2 = 0.0395$$

a) Teste se as variáveis explicativas são conjuntamente significativas ao nível de 5%. Verifique também se alguma variável explicativa é individualmente significativa.

b) O modelo foi reestimado usando a forma logarítmica para *rendliq* e *sal*, com os resultados abaixo indicados. Alguma das conclusões obtidas em a) se altera?

$$\hat{y} = -\underset{(39.37)}{36.30} + \underset{(0.203)}{0.327}\, rdk + \underset{(0.080)}{0.069}\, ga - \underset{(3.39)}{4.74}\, \ln(rendliq) + \underset{(6.31)}{7.24}\, \ln(sal);\; n = 142; \quad R^2 = 0.033$$

c) Porque é que não se usaram logaritmos de *rdk* e *ga* em b)?

d) No conjunto, a evidência de previsibilidade da rentabilidade das ações é fraca ou forte?

**4.4** Numa análise de dados em que se usou o modelo abaixo indicado e a amostra de (3.3), alargada com a inclusão de observações de duas novas variáveis, obtiveram-se os seguintes resultados:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

$$(X'X)^{-1} = \begin{bmatrix} 1.9 & -0.1 & -0.03 & 0.05 & -0.08 \\ -0.1 & 0.05 & -0.01 & -0.02 & 0.03 \\ -0.03 & -0.01 & 0.07 & -0.04 & -0.03 \\ 0.05 & -0.02 & -0.04 & 0.01 & 0.02 \\ -0.08 & 0.03 & -0.03 & 0.02 & 0.01 \end{bmatrix} \qquad X'y = \begin{bmatrix} 74 \\ 282 \\ 375 \\ 450 \\ 250 \end{bmatrix}$$

$$N{=}55 \qquad y'y{=}4019.75$$

a) Estime os coeficientes do modelo e escreva a equação de regressão, colocando por baixo de cada coeficiente a estimativa do respetivo desvio padrão.

b) Teste a hipótese de que $0.5\beta_2 - 2\beta_3 = 12$, ao nível de significância de 5%.

c) Calcule e interprete o coeficiente de determinação, $R^2$.

d) Teste a hipótese de que o modelo não se ajusta aos dados, ao nível de 5%.

e) Use um teste que lhe permita inferir qual dos dois modelos, (4.4) ou (3.3), se ajusta melhor aos dados.

**4.5** Considere os seguintes resultados de uma regressão estimada:

$$\hat{Y}_i = 6 + X_{i1} - X_{i2} + 2X_{i3}$$

$$\bar{Y} = 6;\;\; y'y = 360;\;\; y'X = [54 \quad 6 \quad -4 \quad 4]$$

a) Qual a percentagem da variação em *Y* que é explicada pelo modelo?

b) Será que o valor obtido em a) é significativamente diferente de zero, a um nível de 5%? Ou seja, poder-se-á considerar que o modelo como um todo se ajusta aos dados?

**4.6** Numa análise de regressão linear simples foram obtidos, a partir de uma amostra de 6 pares de valores de *X* e *Y,* os seguintes resultados: $R^2$ = 0.64, $S_x$ = 3 e $S_y$ = 5, sendo estes últimos os desvios padrão das variáveis *X* e *Y* na amostra.

a) Determine o intervalo de confiança a 95% para $\beta_1$, sabendo que $Y$ é uma função crescente de $X$.
b) Teste ao nível de significância de 5% a hipótese H$_0$: $\beta_1 = 0$ contra H$_1$: $\beta_1 \neq 0$; compare com a).

**4.7** Considere a seguinte função de produção, em que $Q$ *é* a quantidade produzida, $K$ representa o factor produtivo capital e $L$ o factor trabalho:

$$Q_i = \alpha L_i^{\beta_1} K_i^{\beta_2} e^{u_i}$$

Os resultados da estimação do logaritmo natural dessa função de produção (incluindo os erros padrão das estimativas por baixo das mesmas) são:

$$\hat{Y}_i = ? + \underset{(0.257)}{0.632} X_{1i} + \underset{(0.219)}{0.452} X_{2i}, \quad n = 20, \quad R^2 = 0.98 \quad X'Y = \begin{bmatrix} 88 \\ 240 \\ 214.4 \end{bmatrix}$$

$$\mathrm{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) = 0.055, \quad \bar{X}_1 = 2.56, \quad \bar{X}_2 = 2.1$$

a) Obtenha uma estimativa para $\alpha$.
b) Teste a hipótese de que as elasticidades relativas ao capital e ao trabalho são idênticas.
c) Elabore um quadro de análise de variância para o modelo transformado.

**4.8** De um estudo em que se utilizou o modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ conhecem-se os seguintes resultados:

$n = 6$ $\quad\quad$ $\hat{\beta}_1 = 3.5716$ $\quad\quad$ $\hat{\beta}_2 = 4.204$

$$(X'X)^{-1} = \begin{bmatrix} 2.078 & -0.224 & -0.286 \\ ? & ? & ? \\ ? & ? & 0.051 \end{bmatrix} \quad X'Y = \begin{bmatrix} 231 \\ 760 \\ 1050 \end{bmatrix} \quad Y'Y = 11050.098$$

a) Calcule $\hat{\beta}_0$ e os valores assinalados com "?".
b) Apresente um quadro de análise de variância para o modelo.

**4.9** Para estudar a racionalidade da avaliação dos preços das casas, propõe-se o modelo

$$preco = \beta_0 + \beta_1 aval + u$$

sendo a avaliação racional se $\beta_1 = 1$ e $\beta_0 = 0$. A equação estimada é:

$$\hat{preco} = -\underset{(16.27)}{14.47} + \underset{(0.049)}{0.976} aval; \quad n = 88; \quad SQR = 165645; \quad R^2 = 0.82$$

a) Teste separadamente as hipóteses H$_0$: $\beta_0 = 0$ e H$_0$: $\beta_1 = 1$. Que conclui?

b) Considerando agora a hipótese conjunta H0: $\beta_0 = 0$ e $\beta_1 = 1$, verifique que a *SQR* do modelo restringido é $\sum(preco_i - aval_i)^2$ . Sendo este valor $SQR^*$ = 209449, teste a hipótese indicada.

c) Teste agora H0: $\beta_2 = 0, \beta_3 = 0$ e $\beta_4 = 0$ no modelo

$$preco = \beta_0 + \beta_1 aval + \beta_2 area + \beta_3 lote + \beta_4 quartos + u$$

o qual foi estimado com as mesmas 88 casas, obtendo-se $R^2$ = 0.829.

**4.10** Use os dados de PRECASA.DTA para estimar o modelo

$$preco = \beta_0 + \beta_1 lote + \beta_2 area + \beta_3 quartos + u$$

a) Estime os coeficientes do modelo e o erro padrão da regressão e preveja o preço de uma casa construída num lote de 1000m$^2$, com 230m$^2$ de área e 4 quartos.
b) Estime um intervalo de confiança a 95% para o valor previsto em a).

## 5. Notas sobre Teoria Assimptótica

**5.1** Considere os dados em PESO.DTA e o seguinte modelo para explicar o peso à nascença:

$$peso = \beta_0 + \beta_1 cigs + \beta_2 ordn + \beta_3 rend + \beta_4 educm + \beta_5 educp + u$$

onde *peso* é o peso à nascença, *cigs* é o número médio de cigarros fumados pela mãe durante a gravidez, *ordn* é a ordem de nascimento da criança, *rend* é o rendimento anual familiar, *educm* e *educp* são o número de anos de escolaridade da mãe e do pai, respetivamente. Usando o teste *F* e o teste do multiplicador de Lagrange (*LM*):

a) Teste a hipótese de que, após ter em conta os outros fatores, a educação dos pais não tem efeito no peso do bebé à nascença.
b) Teste a significância global do modelo.

## 6. Análise da Especificação do Modelo

**6.1** A equação seguinte descreve o preço mediano da habitação em função do montante de poluição (*nox*, óxido de nitrogénio) e do número médio de quartos nas casas de uma comunidade (*quartos*):

$$ln(preco) = \beta_0 + \beta_1\, ln(nox) + \beta_2 quartos + u$$

a) Quais os sinais prováveis de $\beta_1$ e $\beta_2$? Qual a interpretação de $\beta_1$? Explique.
b) Porque é que *nox* (mais precisamente, o seu logaritmo) e *quartos* poderão estar negativamente correlacionadas? Se assim for, a regressão simples de ln(*preco*) sobre ln(*nox*) produzirá um estimador de $\beta_1$ enviesado para cima ou para baixo?
c) Após recolher os dados necessários foram estimadas as seguintes equações:

$$\ln(\widehat{preco}) = 11.71 - 1.043\ln(nox)\,; \quad n = 506; \quad R^2 = 0.264;$$

$$\ln(\widehat{preco}) = 9.23 - 0.718\ln(nox) + 0.306 quartos; \quad n = 506; \quad R^2 = 0.514.$$

A relação entre as estimativas da elasticidade do preço em relação a *nox* na regressão simples e na regressão múltipla é aquela que teria previsto, dada a sua resposta em b)? Isto quer dizer que (–0.718) está certamente mais próxima da verdadeira elasticidade que (–1.043)?

**6.2** Para explicar o salário dos executivos (*sal*) usou-se o modelo

$$\ln(sal) = \beta_0 + \beta_1\ln(vendas) + \beta_2 valmerc + \beta_3 mglucro + \beta_4 ase + \beta_5 ast + u$$

onde *valmerc* é o valor de mercado da empresa, *mglucro* é o lucro como percentagem das vendas, *ase* os anos de serviço como executivo na empresa atual e *ast* os anos de serviço totais na empresa atual. A estimação deste modelo pelo método dos mínimos quadrados, com $n$ = 177, forneceu um $R^2$ de 0.353. Quando $ase^2$ e $ast^2$ são adicionadas, obtém-se $R^2$ = 0.375. Há evidência de má especificação da forma funcional do modelo apresentado?

**6.3** Considere os dados de PRECASA.DTA e os modelos

(1) $preco = \beta_0 + \beta_1 lote + \beta_2 area + \beta_3 quartos + u$
(2) $preco = \beta_0 + \beta_1 lote + v$
(3) $preco = \beta_0 + \beta_1\, ln(lote) + \beta_2\, ln(area) + \beta_3 quartos + w$

a) Através de testes e critérios apropriados, verifique qual dos modelos é mais adequado para explicar a variação no preço da habitação.
b) Verifique se a forma funcional do modelo que selecionou na alínea anterior foi especificada corretamente, usando o teste RESET (versão F).
c) Repita a alínea anterior mas usando a versão LM do teste RESET.

## 7. Modelo de Regressão Linear com Variáveis Explicativas Qualitativas

**7.1** Pretende-se estudar as diferenças ao nível do salário (*salario*) mensal entre quatro grupos de indivíduos: homens licenciados (*HL*), mulheres licenciadas (*ML*), homens não licenciados (*HNL*) e mulheres não licenciadas (*MNL*). A partir de uma mostra de 498 observações, obteve-se o seguinte modelo estimado:

$$\ln(\widehat{salario}) = \underset{(0.100)}{0.321} + \underset{(0.055)}{0.213}HL + \underset{(0.058)}{0.148}ML - \underset{(0.054)}{0.110}MNL - \underset{(0.070)}{0.079}casado + \underset{(0.006)}{0.027}exper$$

onde *HL*, *ML* e *MNL* são variáveis *dummy* que indicam se o indivíduo pertence ao grupo referido (=1) ou não (=0), *casado* é outra variável *dummy* que indica se o indivíduo é casado (=1) ou não (=0) e *exper* representa o número de anos de experiência profissional.

a) Interprete os coeficientes estimados.
b) Indique como poderia testar se existe discriminação sexual no mercado dos trabalhadores licenciados.
c) Usando um nível de significância de 5%, verifique se existe discriminação sexual no mercado dos trabalhadores não licenciados.
d) Comente a seguinte afirmação: "Um trabalhador do sexo masculino, casado, com o 12º ano de escolaridade ganha significativamente menos que um trabalhador do mesmo sexo e com a mesma experiência, solteiro, com a 4ª classe".

**7.2** Considere o seguinte modelo para comparar os resultados de três métodos de ensinos diferentes:

$$Y = \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + u$$

onde *Y* é a classificação obtida pelo aluno num teste e

$$X_j = \begin{cases} 1 & \text{se foi usado o método } j \\ 0 & \text{caso contrário} \end{cases},$$

*j* = 1, 2, 3. Para cada método de ensino foram obtidas as classificações de $n_j$ alunos, *j* = 1, 2, 3, sendo $n_1 + n_2 + n_3 = n$.

a) De acordo com a informação dada escreva teoricamente as matrizes *Y*, *X*, $X'X$ e $X'Y$.
b) Mostre que o vetor das estimativas dos mínimos quadrados é $\hat{\delta} = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)'$.
c) Com base no modelo estimado abaixo apresentado (onde os erros padrão estão entre parênteses) teste a hipótese de que os alunos submetidos aos métodos 2 e 3 obtêm em média os mesmos resultados.

$$\hat{Y} = \underset{(2.18)}{10.55}X_1 + \underset{(3.21)}{12.86}X_2 + \underset{(4.11)}{13.67}X_3; \quad n = 63$$

**7.3** Os dados em SONO.DTA referem-se ao modelo

$$sono = \beta_0 + \beta_1 trab + \beta_2 educ + \beta_3 idade + \beta_4 idade^2 + \beta_5 filpeq + u$$

com *filpeq* = 1 se tem filhos pequenos, e *masc* = 1 se é do sexo masculino.

a) Estime a equação separadamente para homens e mulheres. Há diferenças substanciais?
b) Estime o modelo

$$\begin{aligned} sono = \beta_0 + \beta_1 trab + \beta_2 educ + \beta_3 idade + \beta_4 idade^2 + \beta_5 filpeq + \gamma_0 masc \\ + \gamma_1 masc \cdot trab + \gamma_2 masc \cdot educ + \gamma_3 masc \cdot idade + \gamma_4 masc \cdot idade^2 \\ + \gamma_5 masc \cdot filpeq + v \end{aligned}$$

e compare os resultados obtidos com os da alínea anterior.
c) Faça o teste de Chow para a igualdade dos parâmetros nas equações para ambos os sexos de duas maneiras diferentes (nível de significância: 10%).
d) Permita agora diferenças no termo independente para homens e mulheres, e teste se os termos de interação que envolvem *masc* são conjuntamente significativos.
e) Dados os resultados de c) e d), qual seria o seu modelo final?
f) Use o modelo escolhido em e) para prever quanto dorme em média por noite uma mulher de 30 anos, com o 12.º ano de escolaridade, que trabalha 8 horas por dia (de 2.ª a 6.ª) e tem 2 filhos pequenos. Comente.

## 8. Heterocedasticidade

**8.1** Considere um modelo linear para explicar o consumo mensal de cerveja:

$$cerveja = \beta_0 + \beta_1 rend + \beta_2 preco + \beta_3 educ + \beta_4 fem + u$$
$$E(u|rend, preco, educ, fem) = 0$$
$$\mathrm{Var}(u|rend, preco, educ, fem) = \sigma^2 rend^2$$

Escreva a equação transformada que tenha um erro homoscedástico.

**8.2** Use os dados em PRECASA.DTA para este exercício.

a) Estime o modelo $preco = \beta_0 + \beta_1 lote + \beta_2 area + \beta_3 quartos + u$, calculando os erros padrão usuais e os robustos à heteroscedasticidade. Discuta quaisquer diferenças importantes entre eles.
b) Repita a) para o modelo $\ln(preco) = \beta_0 + \beta_1 \ln(lote) + \beta_2 \ln(area) + \beta_3 quartos + u$.
c) O que é que este exemplo sugere sobre a heteroscedasticidade e a transformação usada na variável dependente?

**8.3** Use os dados de SONO.DTA para estimar a seguinte equação do sono:

$$sono = \beta_0 + \beta_1 trab + \beta_2 educ + \beta_3 idade + \beta_4 idade^2 + \beta_5 filpeq + \beta_6 masc + u$$

a) Escreva uma equação para a heteroscedasticidade que permita que a variância de *u* seja diferente para homens e mulheres, sem depender de outros fatores.
b) Estime os parâmetros da equação anterior. [Terá que estimar a equação do *sono* primeiro para obter os resíduos dos mínimos quadrados.] A variância estimada de *u* é maior para homens ou para mulheres?
c) A variância de *u* é estatisticamente diferente para homens e mulheres?

**8.4** Sabendo que o investimento de uma empresa (*Y*) depende do seu volume de vendas (*X*) pretende-se ajustar o modelo:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, \ldots, 9 \text{ empresas.}$$

Após estimar o modelo pelo método dos mínimos quadrados, e admitindo poder haver problemas com a utilização deste método, estimou-se posteriormente a seguinte regressão auxiliar:

$$\hat{u}_i^2 = 0.0118 + 0.0603 X_i; \quad n = 9; \quad R^2 = 0.0658$$

a) Qual lhe parece ter sido a razão que levou o investigador a realizar a regressão auxiliar? De acordo com os resultados obtidos pensa ser o método dos mínimos quadrados o melhor para a estimação do modelo?

b) Admitindo que a variância do erro é proporcional ao volume de vendas, ou seja $\mathrm{Var}(u_i | X_i) = \sigma^2 X_i$, use um método de estimação alternativo, considerando os seguintes valores para *X* e *Y*:

| *Y* | 1 | 3 | 2 | 4 | 3 | 5 | 4 | 4 | 4 |
|---|---|---|---|---|---|---|---|---|---|
| *X* | 1 | 5 | 3 | 7 | 8 | 9 | 6 | 5 | 7 |

**8.5** Para o modelo usado em 8.2.b) calcule sucessivamente as versões *F* e *LM*, bem como os valores-*p*, dos testes:

a) Breusch-Pagan;
b) White;
c) White, caso especial.

## 9. Problemas com os Dados

**9.1** Considere os seguintes dados:

| $Y$ | 80 | 100 | 110 | 120 | 90 | 85 | 70 | 105 | 95 | 115 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 2 | 1 | 1 | 2 | 1.2 | 1.5 | 1.5 | 2 | 2.5 | 2 |
| $X_2$ | 25 | 30 | 40 | 45 | 30 | 20 | 30 | 60 | 50 | 60 |
| $X_3$ | 2.5 | 2 | 5 | 4 | 3 | 1.5 | 3.5 | 5 | 5.5 | 6 |

a) Efectue a regressão de $Y$ nas restantes variáveis.
b) Suspeita-se que exista multicolineariedade neste modelo. Usando o critério VIF, verifique se há fundamento para esta suspeita.

**9.2** Considere os dados em ANSCOMBE.DTA para este exercício.

a) Confirme que as variáveis $Y_j, j = 1, 2, 3, 4$, por um lado, e as variáveis $X_j, j = 1, 2, 3, 4$, por outro lado, têm a mesma média e desvio-padrão amostrais.
b) Estime os seguintes modelos, guardando os valores estimados de $Y_j$ e confirmando que em todos os casos $\hat{\beta}_0 = 3$, $\hat{\beta}_1 = 0.5$ e $R^2 = 0.67$:

$$Y_1 = \beta_0 + \beta_1 X_1 + u_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + u_2$$
$$Y_3 = \beta_0 + \beta_1 X_3 + u_3$$
$$Y_4 = \beta_0 + \beta_1 X_4 + u_4$$

c) Represente graficamente os valores observados e os valores estimados e comente os resultados obtidos.
d) Reestime o segundo modelo adicionando a variável $X_2^2$ ao modelo.
e) Reestime o terceiro modelo omitindo a observação outlier.