Exercícios - Estatística e Delineamento - 2017-18

2 Regressão Linear Simples

1. Com base nos dados do Instituto Nacional de Estatística (INE), foi criado um ficheiro em formato CSV (*Comma separated values*) chamado `cereais.csv` e contendo a evolução da superfície agrícola utilizada anualmente na produção de cereais para grão (variável `area`, em km^2) em Portugal, no período de 1986 a 2011 (variável `ano`). O ficheiro encontra-se disponível na página *web* da disciplina (secção *Materiais de apoio*, subsecção *Dados*). O ficheiro `cereais.csv` deve ser descarregado e guardado na directoria onde se localiza a sessão de trabalho do `R`, que convém ser a pasta (de preferência com o nome `AulasED`) onde tem estado a guardar o seu trabalho. Seguidamente o seu conteúdo deve ser lido para a sessão do `R` através do comando:

```
> Cereais <- read.csv("cereais.csv")
```

Atenção: Ao guardar o ficheiro na directoria da sua sessão de trabalho, não deve alterar o tipo (CSV) de ficheiro. A melhor forma de garantir isso será o de apenas *guardar* o ficheiro, sem o abrir. Se preferir abrir o ficheiro antes de o guardar, certifique-se de que está a utilizar um editor de texto que não altera o conteúdo do ficheiro. Os programas do tipo *Office* podem não oferecer essa garantia, pelo que se recomenda evitar a sua utilização.

(a) Construa uma nuvem de pontos de superfície agrícola *vs.* ano e comente.

(b) A partir do gráfico obtido na alínea anterior, sugira um valor para o coeficiente de correlação entre superfície agrícola e ano. Depois, utilize os comandos do `R` para calcular esse mesmo coeficiente de correlação. Comente o seu significado.

(c) Ajuste uma recta de regressão de superfície agrícola utilizada sobre anos. Discuta o significado dos parâmetros da recta ajustada, no contexto do problema sob estudo.

(d) Comente a qualidade da recta obtida, calculando o respectivo coeficiente de determinação e interpretando o valor obtido.

(e) Trace a recta de regressão ajustada em cima da nuvem de pontos e comente.

(f) Calcule a Soma de Quadrados Total (SQT), a partir do cálculo da variância amostral de y.

(g) Calcule o valor da Soma de Quadrados da Regressão (SQR).

(h) Calcule a Soma de Quadrados dos Resíduos (SQRE), directamente a partir dos resíduos, e verifique numericamente a relação fundamental da Regressão Linear: SQT=SQR+SQRE.

(i) Altere as unidades de medida da variável `area`, de km^2 para hectares (`area` $\rightarrow$ `area`$\times 100$). Ajuste novamente a regressão, após efectuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação R^2? Comente.

(j) De novo a partir dos dados originais, transforme a variável `ano` num contador dos anos do estudo (`ano` $\rightarrow$ `ano`-1985). Ajuste novamente a regressão, após efectuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação R^2? Comente.

2. O ficheiro `azeite.xls`, disponível na página *web* da disciplina (secção *Materiais de apoio*, subsecção *Dados*), é um ficheiro de tipo folha de cálculo, comum a aplicações de escritório como o LibreOffice, OpenOffice ou MicrosoftOffice. A folha de cálculo contém dados relativos à produção de azeite em Portugal no período 1995-2010, disponibilizados pelo Instituto Nacional de Estatística (www.ine.pt). As colunas "Azeitona" e "Azeite" correspondem à produção de azeitona oleificada (em t) e azeite (em hl), respectivamente.

(a) Abra o ficheiro `azeite.xls` com um programa do tipo *Office* e guarde a folha de cálculo num ficheiro de texto de nome `azeite.txt`, utilizando o *Save as* com a opção *Ficheiro de Texto*. Coloque esse ficheiro na pasta de trabalho do R.

(b) Numa sessão do R, guarde os dados do ficheiro `azeite.txt` (criado na alínea anterior) numa *data frame* de nome `azeite`, através do comando:

```
> azeite <- read.table("azeite.txt", header=TRUE)
```

(c) Crie a nuvem de pontos relacionando as produções de Azeite (eixo vertical, variável y) e Azeitona (eixo horizontal, variável x).

(d) Com base na nuvem de pontos, sugira um valor para o coeficiente de correlação entre as duas variáveis. Avalie a sua sugestão calculando o valor de r_{xy}. Comente o valor obtido.

(e) Calcule as estimativas de mínimos quadrados para os parâmetros da recta de regressão, e comente o seu significado.

(f) Calcule a precisão da recta de regressão estimada de y sobre x e comente o valor obtido.

3. Demonstre as seguintes relações algébricas:

(a) $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, para qualquer conjunto de n valores, $\{x_i\}_{i=1}^{n}$, de média $\overline{x}$.

(b) $(n-1)\text{cov}_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n}(y_i - \overline{y})x_i$, para quaisquer conjuntos de n valores, $\{x_i\}_{i=1}^{n}$, e $\{y_i\}_{i=1}^{n}$ de médias $\overline{x}$ e $\overline{y}$, respectivamente.

4. Deduza as expressões para o declive e ordenada na origem da recta de regressão, resultantes de minimizar a soma dos quadrados dos resíduos:

$$\begin{aligned} b_1 &= \frac{cov_{xy}}{s_x^2} \\ b_0 &= \overline{y} - b_1\overline{x} \end{aligned}$$

5. Mostre que, numa Regressão Linear Simples, baseada em n pares de observações $\{(x_i, y_i)\}_{i=1}^{n}$, se verificam:

(a) a igualdade da média dos valores observados e da média dos valores ajustados de y.

(b) a média dos resíduos ($e_i = y_i - \hat{y}_i$) é nula.

(c) as três Somas de Quadrados da regressão são múltiplos de variâncias:

$$\begin{aligned} SQT &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)\cdot s_y^2 \\ SQR &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = (n-1)\cdot s_{\hat{y}}^2 \\ SQRE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = (n-1)\cdot s_e^2 \, , \end{aligned}$$

onde $s_\star^2$ indica a variância amostral das quantidades representadas por $\star$.

(d) $SQR = b_1^2 \cdot (n-1)\cdot s_x^2$, onde $(n-1)\cdot s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

(e) $SQT = SQR + SQRE$.

6. Considere uma regressão linear simples, ajustada com n pares de observações $\{(x_i, y_i)\}_{i=1}^n$. Mostre que:

   (a) O declive da recta de regressão de y sobre x pode-se escrever em termos do desvio padrão de cada variável e do coeficiente de correlação entre as duas variáveis, sendo dado por:

   $$b_1 \;=\; r_{xy} \cdot \frac{s_y}{s_x} \;.$$

   (b) O coeficiente de determinação R^2 é igual ao quadrado do coeficiente de correlação entre as observações da variável preditora x e da variável resposta y.

   (c) O quadrado do coeficiente de correlação entre os n valores observados y_i e os n correspondentes valores ajustados, $\hat{y}_i$, é também igual ao coeficiente de determinação: $(r_{y\hat{y}})^2 = R^2$.

7. O programa `R` tem vários conjuntos de dados disponíveis. Um desses conjuntos de dados designa-se `anscombe` e pode ser visto apenas escrevendo o nome do objecto. Utilizando estes dados, determine, e comente os valores obtidos para:

   (a) As médias de cada variável x_i e y_i $(i = 1:4)$.

   (b) As variâncias de cada variável x_i e y_i $(i = 1:4)$.

   (c) O valor dos parâmetros b_0 e b_1 nas quatro rectas de regressão de y_i sobre x_i $(i = 1, 2, 3, 4)$.

   (d) Os Coeficientes de Determinação associados às quatro rectas indicadas na alínea anterior.

   Após comentar os resultados obtidos, construa as quatro nuvens de pontos $\{(x_i^{(j)}, y_i^{(j)})\}_{i=1}^{11}$, para $j = 1:4$. Comente esses gráficos, à luz dos valores anteriormente obtidos.

8. Utilizando os dados das medições morfométricas sobre 150 lírios, contidos no objecto `iris` do `R`, responda às seguintes questões:

   (a) Construa a nuvem de pontos de comprimento das pétalas (eixo horizontal, variável x) e largura das pétalas (eixo vertical, variável y).

   (b) Ajuste a recta de regressão de largura (y) sobre comprimento (x) das pétalas, e desenhe-a sobre a nuvem de pontos.

   (c) Ajuste a recta de regressão de comprimento sobre largura, mantendo os nomes de x (comprimento) e y (largura), ou seja, calcule a "recta de x sobre y", de equação $x = b_0^* + b_1^* \, y$.

   (d) Sobre a nuvem de pontos original, trace agora a recta de regressão de comprimento sobre largura - a "recta de x sobre y". (NOTA: Tenha em atenção que uma equação $x = b_0^* + b_1^* \, y$ tem, na forma canónica, equação $y = -\frac{b_0^*}{b_1^*} + \frac{1}{b_1^*} \, x$). Verifique que a recta de regressão de y sobre x é diferente da recta de regressão de x sobre y.

   (e) Explique o facto de as rectas obtidas nas alíneas anteriores serem diferentes.

9. O programa `R` tem um grande número de pacotes adicionais disponíveis. Um desses pacotes adicionais designa-se `MASS` e pode ser carregado mediante o comando `library(MASS)`.

   Considere o conjunto de dados `Animals`, disponível no referido módulo `MASS`, onde se listam pesos médios dos cérebros (em g) e dos corpos (em kg) para 28 espécies animais. Pretende-se estudar uma relação entre pesos do cérebro (variável resposta, y) e pesos do corpo (variável preditora, x).

   (a) Construa uma nuvem de pontos de pesos do corpo (eixo horizontal) e pesos do cérebro (eixo vertical). Calcule o coeficiente de correlação correspondente e comente.

(b) Construa nuvens de pontos com as seguintes transformações de uma ou ambas as variáveis:
   i. $\ln(y)$ vs. x;
   ii. y vs. $\ln(x)$;
   iii. $\ln(y)$ vs. $\ln(x)$.

(c) Considere uma relação linear entre $\ln(y)$ e $\ln(x)$. Explicite a relação de base correspondente entre as variáveis originais (não logaritmizadas). Comente.

Nas alíneas seguintes considere sempre os *dados logaritmizados*.

(d) Calcule os coeficientes de correlação e de determinação associados à relação entre $\ln(x)$ e $\ln(y)$. Interprete os valores obtidos. Como se explica que o Coeficiente de Determinação não seja particularmente elevado, sendo evidente a partir da nuvem de pontos que existe uma boa relação linear entre log-peso do corpo e log-peso do cérebro para a generalidade das espécies?

(e) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo (utilizando a totalidade das observações). Trace essa recta sobre a nuvem de pontos e comente.

(f) Considere agora a estimativa para o declive da recta, $b_1 = 0.49599$. Qual o significado biológico deste valor, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

(g) Considere a nuvem de pontos das variáveis logaritmizadas. Identifique os três pontos que se destacam na parte inferior direita da nuvem. (NOTA: explore o comando `identify` do R). Comente.

Nas restantes alíneas, considere apenas os dados (logaritmizados) respeitantes a espécies que *não sejam de dinossáurios*.

(h) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo. Trace essa recta sobre a nuvem de pontos e comente. (NOTA: Aproveite a nuvem de pontos anterior, com a totalidade das espécies, para melhor compreender o efeito da exclusão das três espécies de dinossáurios sobre a recta ajustada).

(i) Analise os principais resultados associados à regressão ajustada na alínea 9h). Compare com os resultados obtidos na alínea 9e) e comente. Em particular, como se explica a elevação considerável no valor do coeficiente de determinação?

(j) Considere agora a estimativa para o declive da nova recta, $b_1 = 0.75226$. Qual o significado biológico deste valor, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

10. Num estudo sobre poluição numa grande cidade, foram efectuadas medições, em 116 dias, da quantidade de ozono no ar (em partes por mil milhões) às 14h00 e da temperatura máxima (em ^{o}C) no respectivo dia. Essas observações encontram-se num ficheiro em formato `csv` de nome `ozono.csv`, que se encontra disponível na página *web* da disciplina e pode ser descarregado para a directoria de trabalho da sessão do `R`, como indicado no Exercício 1. Seguidamente, o conteúdo desse ficheiro deve ser lido para dentro da sessão do `R` e armazenado num objecto de nome `ozono`, através do comando `read.csv`:

```
> ozono <- read.csv("ozono.csv")
```

(a) Construa a nuvem de pontos de ozono (eixo vertical) *vs.* temperatura máxima (eixo horizontal).

(b) Tendo em conta a curvatura observada no gráfico, foi sugerido o ajustamento dum modelo exponencial, da forma $y = a\,e^{b\,x}$.

i. Construa a nuvem de pontos com as transformações adequadas para verificar se o modelo exponencial é, efectivamente, uma boa opção.
ii. Ajuste o modelo *linearizado* recorrendo ao comando `lm` do `R`. Determine o respectivo coeficiente de determinação e comente.
iii. Interprete os parâmetros da recta que ajustou, directamente em termos do modelo exponencial.
iv. Indique, justificando, qual o teor médio de ozono (em partes por mil milhões) estimado pelo modelo ajustado, para um dia em que a temperatura máxima seja de 25ºC.

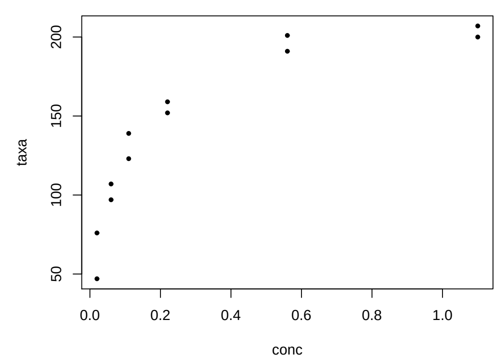
(c) Considere novamente a nuvem de pontos original. Trace a curva exponencial correspondente ao ajustamento efectuado na alínea anterior.

11. Num estudo sobre reacções enzimáticas, procura-se analisar a "velocidade" da reacção em células tratadas com puromicina. Para diferentes concentrações do substrato (variável *conc*), medidas em partes por milhão (ppm), registou-se o número de emissões radioactivas por minuto, e a partir destas calculou-se a taxa inicial ou "velocidade" da reacção, em contagens/minuto/minuto (variável *taxa*). Os resultados obtidos são dados na tabela seguinte e encontram-se *nas duas primeiras colunas* e nas *doze primeiras linhas* da *data frame* `Puromycin` do `R`, com as designações `conc` e `rate`, respectivamente (são as linhas a que corresponde o nível `treated` no factor `state`):

conc	0.02	0.02	0.06	0.06	0.11	0.11	0.22	0.22	0.56	0.56	1.10	1.10
taxa	76	47	97	107	123	139	159	152	191	201	207	200

A relação entre taxas da reacção e concentrações do substrato é representada no gráfico à direita. Admite-se que o modelo de Michaelis-Menten é adequado à descrição da relação referida, e decide-se usar este modelo com a seguinte parametrização (onde y representa a *taxa* e x a concentração *conc*),

$$y = \frac{ax}{b+x} \qquad (a > 0, b > 0 \quad \text{e} \quad x > 0) \ .$$



(a) Mostre que o modelo referido pode ser linearizado, indicando a relação linearizada e as transformações de variáveis necessárias.

(b) Ajuste o modelo linearizado que escolheu na alínea anterior, através do comando `lm` do `R`.

(c) Estime os parâmetros a e b na relação original no modelo de Michaelis-Menten. Como interpreta o valor estimado do parâmetro a?

Na resolução dos Exercícios seguintes, de natureza inferencial, admita válido o Modelo da Regressão Linear Simples.

12. Considere os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros duma recta de regressão.

(a) Mostre que a média ($\overline{Y}$) das observações de Y, é uma variável aleatória não correlacionada com o estimador do declive da recta, $\hat{\beta}_1$, ou seja, mostre que:

$$Cov[\overline{Y}, \hat{\beta}_1] \;=\; 0 \;.$$

(b) Mostre que a covariância entre os dois estimadores dos parâmetros da recta é dada por:

$$Cov[\hat{\beta}_0, \hat{\beta}_1] \;=\; -\,\frac{\overline{x}\sigma^2}{(n-1)\cdot s_x^2} \;.$$

(c) Deduza da alínea anterior que *os estimadores de β_0 e de β_1 não são, em geral, independentes.* Indique uma condição necessária para que o possam ser.

13. Mostre que o estimador da ordenada na origem da recta tem a seguinte distribuição:

$$\hat{\beta}_0 \;\cap\; \mathcal{N}\left(\beta_0\,,\,\sigma^2\left[\frac{1}{n}+\frac{\overline{x}^2}{(n-1)\cdot s_x^2}\right]\right) \;,$$

onde $(n-1)\cdot s_x^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2$.

14. Considere de novo os dados do Exercício 8 (medições sobre lírios), admitindo agora que se trata da concretização duma amostra aleatória extraída duma população mais vasta. Considere, em particular, a relação entre largura da pétala (`Petal.Width`, variável y) e comprimento da pétala (`Petal.Length`, variável x), ambas em *cm*. Responda às seguintes alíneas.

(a) Obtenha estimativas das variâncias e dos desvios padrão dos estimadores dos parâmetros da recta, β_0 e β_1.

(b) Obtenha um intervalo a 95% de confiança para o declive β_1 da correspondente recta populacional.

(c) Obtenha um intervalo a 95% de confiança para a ordenada na origem β_0 da recta populacional.

(d) Utilize um teste de hipóteses para validar a seguinte afirmação: "por cada centimetro a mais no comprimento da pétala, a largura da pétala cresce, em média, $0.5cm$".

(e) Utilize um teste de hipóteses para validar a seguinte afirmação: "por cada centimetro a mais no comprimento da pétala, a largura da pétala cresce, em média, menos de $0.5cm$".

(f) Utilize um teste de hipóteses sobre o declive da recta populacional β_1 para validar a seguinte afirmação: "não existe uma relação linear significativa entre comprimentos e larguras das pétalas, nos lírios".

(g) Valide de novo a afirmação anterior, mas agora utilizando um teste de ajustamento global do Modelo (teste F).

(h) Preveja o valor esperado da largura da pétala para lírios cuja pétala tenha comprimento 4.5cm. Construa um intervalo de confiança para esse valor esperado.

(i) Construa um intervalo de predição (95%) associado à largura duma pétala cujo comprimento seja 4.5cm. Compare com o intervalo de confiança obtido na alínea anterior e comente.

(j) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões.

(k) Para cada uma das seguintes transformações dos dados, verifique os efeitos sobre os parâmetros ajustados e sobre o coeficiente de determinação. Comente.

i. os comprimentos das pétalas são dados em milímetros ($x \to 10 \times x$), mantendo-se as larguras (y) em centimetros.
ii. as larguras das pétalas são dadas em milímetros ($y \to 10 \times y$), mantendo-se os comprimentos (x) em centimetros.
iii. em simultâneo, larguras e comprimentos das pétalas são expressas em milimetros ($x \to 10 \times x$ e $y \to 10 \times y$).

15. A estatística do teste de ajustamento global do modelo (teste F) é dada por $F = \frac{QMR}{QMRE}$. O Coeficiente de Determinação define-se como $R^2 = \frac{SQR}{SQT}$. Com base nestas definições, e tendo em conta as propriedades das somas de quadrados,

    (a) Mostre que a estatística F se pode escrever também como:

    $$F = (n-2) \cdot \frac{R^2}{1-R^2}$$

    (b) Verifique, a partir da expressão anterior, que a estatística F é (para n fixo) uma *função crescente do Coeficiente de Determinação*. Interprete esse facto, em termos do significado de R^2 e a natureza do teste de ajustamento global.

16. Mostre que, numa Regressão Linear Simples, a estatística $F = \frac{QMR}{QMRE}$ do teste de ajustamento global é o quadrado da estatística $T = \frac{\hat{\beta}_1}{\sqrt{\frac{QMRE}{(n-1)\cdot s_x^2}}}$ do teste t para a hipótese $H_0: \ \beta_1 = 0$. Tendo em conta os resultados dados na disciplina de Estatística (dos 1^{os} ciclos do ISA), relacionando as distribuições t e F, conclua que, numa Regressão Linear Simples, estes dois testes são equivalentes.

17. Considere os dados do Exercício 9 (`Animals`). Trabalhe sempre com os *dados logaritmizados*, para a totalidade das espécies.

    (a) Considere a presença de erros aleatórios na relação linear entre as variáveis logaritmizadas: $\log(Y) = \beta_0 + \beta_1 \log(x) + \epsilon$. Qual a consequência para a relação entre as variáveis originais (não logaritmizadas) associada à presença dos erros aleatórios? E como se traduzem os restantes pressupostos do Modelo de Regressão Linear em termos dessa relação entre as variáveis originais (não logaritmizadas)?

    (b) Efectue um teste de ajustamento global da regressão ajustada na alínea 9e). Como se explica que o valor do coeficiente de determinação não seja particularmente bom, quando o teste F sugere que a rejeição da hipótese nula do teste de ajustamento é muito enfática?

    (c) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. É admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro (isto é, y é proporcional a x: $y \propto x$)?

    (d) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Em particular, veja como a presença das três espécies de natureza diferente das restantes está a afectar estes gráficos.

    Nas restantes alíneas, considere apenas os dados (logaritmizados) respeitantes a espécies que *não sejam de dinossáurios*.

    (e) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. Perante o novo valor de b_1, será agora admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?

(f) Preveja o valor esperado do log-peso do cérebro para espécies com peso de corpo igual a $250kg$. Construa um intervalo de confiança para esse valor esperado.

(g) Construa um intervalo de predição associado ao log-peso do cérebro duma espécie cujo peso do corpo seja $250kg$. Como obter um intervalo de predição associado *ao peso* do cérebro?

(h) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões, tendo presente os gráficos análogos obtidos com a presença das 3 espécies de dinossáurios.

18. Considere agora o conjunto de dados relativos a 62 espécies de mamíferos, que é dado no objecto *mammals* do pacote *MASS*, e cuja natureza é semelhante aos dados do Exercício 9.

(a) Construa uma nuvem de pontos de pesos do corpo (eixo horizontal) e pesos do cérebro (eixo vertical).

(b) Tendo em vista uma relação alométrica entre as duas variáveis, construa agora uma segunda nuvem de pontos, desta vez entre os *logaritmos* de cada variável. Comente os dois gráficos.

(c) Explicite a relação de base entre as variáveis originais (não logaritmizadas) associada a uma relação linear entre as variáveis logaritmizadas. Comente.

Nas alíneas seguintes considerar os *dados logaritmizados*, para a totalidade das espécies.

(d) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo. Trace essa recta sobre a nuvem de pontos e comente.

(e) Analise os principais resultados associados à regressão ajustada. Considere em particular os valores do Coeficiente de Determinação, e os resultados do teste F de ajustamento global.

(f) Qual o significado biológico da estimativa do declive da recta, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

(g) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. Será agora admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?

(h) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo.

19. Dado o Modelo de Regressão Linear Simples, considere o estimador do valor esperado de Y, associado a $X = x$, ou seja, o estimador $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$. Mostre que a sua variância é $V[\hat{\mu}_{Y|x}] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right]$. NOTA: Tenha em atenção o Exercício 12.

20. No contexto do Modelo de Regressão Linear Simples,

(a) Mostre que a covariância entre o i-ésimo valor observado e o correspondente valor ajustado da variável resposta é dada por $Cov[Y_i, \hat{Y}_i] = \sigma^2 h_{ii}$, onde $h_{ii} = \left[\frac{1}{n} + \frac{(x_i-\overline{x})^2}{(n-1)\cdot s_x^2}\right]$ (que é também o efeito alavanca da i-ésima observação). SUGESTÃO: Recorde que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

(b) Calcule a covariância entre cada observação de Y e o respectivo resíduo, ou seja, $Cov[Y_i, E_i]$.

(c) Mostre que a covariância entre cada valor ajustado de Y e o respectivo resíduo é nula, ou seja, mostre que $Cov[\hat{Y}_i, E_i] = 0, \forall\, i = 1, ..., n$. Com base neste resultado, justifique a utilização do gráfico de resíduos vs. valores ajustados de Y para estudar o comportamento dos resíduos (em vez de, por exemplo, o gráfico de resíduos vs. valores observados de Y).

(d) Com o auxílio dos resultados anteriores, mostre que os resíduos têm a distribuição indicada nas aulas teóricas, ou seja, mostre que $E_i \quad \cap \quad \mathcal{N}\left(0\,,\,\sigma^2\,(1-h_{ii})\right)$.

21. No contexto da Regressão Linear Simples,

    (a) Determine o valor esperado da soma de quadrados associada à Regressão (SQR).
    SUGESTÃO: Utilize a fórmula para SQR obtida na alínea d) do Exercício 5.

    (b) Compare os valores esperados dos quadrados médios QMR e $QMRE$. Com base nessa comparação, justifique a natureza unilateral direita da região de rejeição associada ao teste de ajustamento global, cuja estatística de teste é $F = \frac{QMR}{QMRE}$.