INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

4 de Novembro, 2020 **PRIMEIRO TESTE 2020-21** **Uma resolução possível**

I

1. Pretende-se um modelo de regressão linear simples para prever o `Peso` (y).

   (a) A melhor variável preditora (x) é a variável `NBagos`, por ser a variável mais fortemente correlacionada com a variável resposta, tendo-se a correlação $r_{xy}=0.9530$. A proporção de variabilidade observada de `Peso` que pode ser explicada por essa regressão é $R^2=(r_{xy})^2=0.9530^2=0.908209$. Assim, quase 91% da variabilidade observada nos pesos dos cachos será explicada pela recta de regressão sobre `NBagos`. É previsível que um valor tão elevado de R^2 seja significativamente diferente de zero, facto que se comprova efectuando o teste F de ajustamento global do modelo:

   Hipóteses: $H_0 : \mathcal{R}^2 = 0 \quad vs. \quad H_1 : \mathcal{R}^2 > 0.$

   Estatística do Teste: $F = \frac{QMR}{QMRE} = (n-2)\frac{R^2}{1-R^2} \underset{\cdot}{\sim} F_{(1,n-2)}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[1,98]}$. Vamos aproximar esta fronteira da Região Crítica pelo valor tabelado $f_{0.05[1,120]}=3.92$.

   Conclusões: É preciso calcular o valor da estatística na nossa amostra, que é $F_{calc} = 98 \times \frac{0.908209}{1-0.908209}=969.6428$. A rejeição de H_0 é claríssima, pelo que o modelo ajustado é muito significativamente diferente do Modelo Nulo, como era de esperar.

   (b) O declive da recta ajustada é dado por $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x} = 0.9530\,\frac{98.24383}{18.74355} = 4.995125$. A ordenada na origem é dada por $b_0=\overline{y}-b_1\,\overline{x}=175.84-4.995125\times 46.15=-54.68502$. Logo, a equação da recta de regressão ajustada é $y=-54.68502+4.995125\,x$. O declive ajustado $b_1=4.995125$ tem unidades de medida g por bago. Em média, por cada bago a mais visível numa imagem, prevê-se um aumento de quase 5 g no peso do cacho.

2. Os gráficos de resíduos (no eixo vertical) contra valores ajustados $\hat{y}_i$ (no eixo horizontal) servem para validar os pressupostos do modelo de regressão linear. Em caso de validade desses pressupostos deve ver-se uma nuvem de pontos sem qualquer padrão especial, e disposta numa banda horizontal (contendo o valor médio dos resíduos, zero). Essa é a situação do gráfico mais à direita (correspondente à logaritmização de ambas as variáveis). Os dois outros gráficos revelam padrões que põem em dúvida pressupostos do modelo. No gráfico da esquerda (modelo das variáveis não transformadas) é visível uma forma em funil, que indica que a variabilidade dos resíduos aumenta, à medida que aumenta também o valor de $\hat{y}_i$. Um tal padrão sugere heterogeneidade na variabilidade dos resíduos que, sendo acentuada, põe em causa o pressuposto de homogeneidade das variâncias dos erros aleatórios. No gráfico a meio (correspondente ao modelo com a logaritmização dos pesos, mas não do número de bagos), existe uma evidente curvatura na nuvem de pontos, o que indicia que a relação linear da equação do modelo não acompanha a tendênca de fundo da nuvem de pontos na relação entre log-pesos e número de bagos. Assim, apenas a dupla log-transformação parece linearizar a relação entre as variáveis, ao mesmo tempo que torna plausível a hipótese de homogeneidade das variâncias.

3. Foi ajustado o modelo com equação de base $y^* = b_0 + b_1\,x^*$, onde y^* indica log-pesos do cacho e x^* indica log-número de bagos visíveis na imagem.

(a) A variância σ^2 dos erros aleatórios é estimada pelo Quadrado Médio Residual ($QMRE$). A raíz quadrada desse valor é dada no enunciado (com a designação erro padrão residual, em inglês). Logo, $QMRE = 0.163^2 = 0.026569$. As unidades de medida deste valor são o quadrado das unidades de medida dos resíduos (iguais às unidades de medida da variável resposta), ou seja, log-gramas ao quadrado.

(b) O intervalo de predição para y^* (log-peso do cacho), quando o número de bagos é 50, logo o log-número de bagos é $x^* = \ln(50) = 3.912023$, é dado no enunciado: $]\,4.9051\,,\,5.5556\,[$.

i. A estimativa da variabilidade associada à previsão, $\hat{\sigma}_{indiv}$ *não* pode ser calculada a partir da sua definição, $\hat{\sigma}_{indiv} = \sqrt{QMRE \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x^*})^2}{(n-1)\, s^2_{x^*}}\right]}$, uma vez que o preditor foi log-transformado e não se conhecem, nem a média, nem a variância, dos valores de log-número de bagos. No entanto, é possível calcular o valor de $\hat{\sigma}_{indiv}$ sabendo que a amplitude do intervalo de precisão é $2 \times t_{0.025(n-2)} \times \hat{\sigma}_{indiv}$. A amplitude do intervalo é $5.5556 - 4.9051 = 0.6505$. Por outro lado, $t_{0.025(98)} \approx t_{0.025(100)} = 1.98397$. Logo, tem-se (aproximadamente) $\hat{\sigma}_{indiv} = \frac{0.6505}{2 \times 1.98397} = 0.163939$.

ii. Exponenciando os extremos do intervalo do enunciado obtém-se um intervalo para o peso dos cachos (em g) entre $\mathrm{e}^{4.9051} = 134.9764$ e $\mathrm{e}^{5.5556} = 258.6821$. Logo, 95% dos cachos com 50 bagos visíveis nas imagens devem ter peso entre 134.9764 gramas e 258.6821 gramas.

(c) Pede-se se é admissível considerar que $y^2 = c\,x^3$, onde y indica o peso do cacho e x o número de bagos visíveis (sendo c a constante de proporcionalidade). Essa equação corresponde a admitir que $y = d\,x^{\frac{3}{2}}$ (com $d = c^{\frac{1}{2}}$), ou seja, que existe uma relação potência entre peso dos cachos e número de bagos visível, com a potência igual a $\frac{3}{2} = 1.5$. Sabemos que uma relação linear entre a log-transformação de y e x corresponde a uma relação potência entre y e x:

$$\ln(y) = b_0 + b_1\, \ln(x) \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0 + b_1\, \ln(x)} \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0}\mathrm{e}^{\ln(x^{b_1})} \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0}\, x^{b_1}\ .$$

Nesta relação potência, o expoente é o declive da recta entre log-y e log-x.

Alternativamente, pode aplicar-se logaritmos aos dois membros da igualdade $y^2 = c\,x^3$ e obter

$$\ln(y^2) = \ln(c) + \ln(x^3) \quad \Leftrightarrow \quad 2\ln\, y = \ln\, c + 3\ln\, x \quad \Leftrightarrow \quad \ln\, y = \frac{\ln\, c}{2} + \frac{3}{2}\ln\, x\ ,$$

ou seja, a recta populacional relacionando as variáveis log-transformadas, teria declive $\frac{3}{2}$.

Assim, o enunciado pergunta se um intervalo de confiança para o declive populacional β_1 contém o valor 1.5. A estimativa amostral de β_1 é $b_1 = 1.40715$. O intervalo de confiança é centrado nesse valor, e dado por: $]\ b_1 - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\ ,\ b_1 + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\ [$. No enunciado pode ver-se que $\hat{\sigma}_{\hat{\beta}_1} = 0.03633$. Como já se viu que $t_{0.025(98)} \approx t_{0.025(100)} = 1.98397$, tem-se o intervalo de confiança a 95% (aproximadamente) dado por: $]\ 1.335\ ,\ 1.479\ [$. Este intervalo não contém o valor 1.5, pelo que esse valor não é (a 95% de confiança) um valor admissível para β_1. Não é justificável descrever a relação entre peso dos cachos e número de bagos visíveis nas imagens da forma indicada no enunciado.

4. Tem-se um modelo de regressão linear múltipla envolvendo as três variáveis, log-transformadas.

(a) O valor exacto de R^2 não é conhecido com base na informação disponível, mas não pode ser inferior ao valor $R^2 = 0.9387$ do modelo do ponto 3, uma vez que esse modelo é um *submodelo* do actual. De facto, toda a variabilidade da variável resposta que é explicada por um

submodelo pode também ser explicada por um modelo completo (a equação do submodelo também se pode escrever com o modelo completo, associando zeros às restantes variáveis), embora o modelo completo explique, em geral, mais variabilidade (tem combinações lineares que nunca poderiam ser criadas no submodelo). Assim, apenas se pode afirmar que, para o modelo deste ponto, se verifica $R^2 \geq 0.9387$.

(b) Pede-se o resíduo para uma observação em que o peso observado do cacho foi $y = 19.20$ (o menor peso observado), e correspondente a uma imagem onde eram visíveis 11 bagos e uma área 24.22. Não esquecendo as log-transformações em todas as variáveis, e usando os valores de b_0, b_1 e b_2 do enunciado, tem-se:

$$e_i = y_i^* - \hat{y^*}_i = \ln(19.20) - (-0.6237 + 1.0366 \cdot \ln(11) + 0.3835 \cdot \ln(24.22)) = -0.129331 \ .$$

(c) Para obter a relação não linear ajustada nas escalas originais, tem de exponenciar-se a relação ajustada:

$$\begin{array}{rrcl} & \ln(y) & = & -0.6237 + 1.0366 \cdot \ln(x_1) + 0.3835 \cdot \ln(x_2) \\ \Leftrightarrow & y & = & \mathrm{e}^{-0.6237+1.0366\cdot\ln(x_1)+0.3835\cdot\ln(x_2)} = \mathrm{e}^{-0.6237} \cdot \mathrm{e}^{1.0366\cdot\ln(x_1)} \cdot \mathrm{e}^{0.3835\cdot\ln(x_2)} \\ \Leftrightarrow & y & = & 0.5359577 \cdot \mathrm{e}^{\ln(x_1^{1.0366})} \cdot \mathrm{e}^{\ln(x_2^{0.3835})} \\ \Leftrightarrow & y & = & 0.5359577 \cdot x_1^{1.0366} \cdot x_2^{0.3835} \end{array}$$

Nesta relação, o peso correspondente a um cacho com, visíveis na imagem, 11 bagos e área 24.22, é $0.5359577 \cdot 11^{1.0366} \cdot 24.22^{0.3835} = 21.85088$. Este valor ajustado é maior (sobre-estima) o valor realmente observado, $y = 19.20$. A diferença entre o valor observado e este valor previsto é $19.20 - 21.85088 = -2.65088$ gramas. Esta diferença *não* resulta de exponenciar o resíduo obtido na alínea anterior.

II

1. A taxa de variação relativa de $y(x)$ define-se como $\frac{y'(x)}{y(x)}$. Afirmar que esta taxa é proporcional a $y(x)$ significa que existe uma constante α tal que $\frac{y'(x)}{y(x)} = \alpha\, y(x)$. Assim, tem-se $\frac{y'(x)}{y^2(x)} = \alpha$. Recordando a regra de primitivação $P f^\alpha f' = \frac{f^{\alpha+1}}{\alpha+1}$, tem-se, primitivando os dois lados da equação:

$$y^{-2}(x) \cdot y'(x) = \alpha \quad \Leftrightarrow \quad \frac{y^{-1}(x)}{-1} = \alpha\, x + K \quad \Leftrightarrow \quad \frac{-1}{y(x)} = \alpha\, x + K \quad \Leftrightarrow \quad y(x) = \frac{1}{c + d\, x} \ ,$$

com $c = -K$ e $d = -\alpha$, obtendo-se assim a relação indicada no enunciado. Trata-se duma relação linear entre o *recíproco de* y e x, como mostra a penúltima equação, ou seja, a relação do enunciado será válida se um gráfico relacionando $\frac{1}{y}$ e x revelar uma relação linear.

2. (a) O modelo de regressão linear simples é ajustado com base em n pares de observações $\{(x_i, Y_i)\}_{i=1}^n$, sendo Y_i variáveis aleatórias que indicam as observações da variável resposta e x_i os correspondentes valores da variável preditora, que se admite serem fixados pelo experimentador (não aleatórios). Admite-se ainda que:

   i. $Y_i = \beta_0 + \beta_1\, x_i + \epsilon_i$ é a equação do modelo indicando a relação linear de fundo, dada pela recta populacional $y = \beta_0 + \beta_1\, x$, bem como os desvios a essa relação nas observações individuais, associada aos erros aleatórios ϵ_i;

ii. $\epsilon_i \frown \mathcal{N}(0, \sigma^2)$, para todo o i, indicando que os erros aleatórios que descrevem a oscilação das observações em torno da recta populacional são identicamente distribuidos, com distribuição Normal de média zero e variância comum σ^2;

iii. $\{\epsilon_i\}_{i=1}^n$ são variáveis aleatórias independentes.

(b) As variáveis aleatórias Y_i são a soma de parcelas não aleatórias $(\beta_0 + \beta_1\, x_i)$ e duma variável aleatória com distribuição Normal (ϵ_i). As propriedades da distribuição Normal garantem assim que cada Y_i também terá distribuição Normal. O valor médio e variância de cada Y_i podem calcular-se a partir das propriedades operatórias desses indicadores. Tem-se

$$E[Y_i] = E[\beta_0 + \beta_1\, x_i + \epsilon_i] = \beta_0 + \beta_1\, x_i + \underbrace{E[\epsilon_i]}_{=0} = \beta_0 + \beta_1\, x_i$$

De forma análoga,

$$V[Y_i] = V[\beta_0 + \beta_1\, x_i + \epsilon_i] = V[\epsilon_i] = \sigma^2\ .$$

Logo, $Y_i \frown \mathcal{N}(\ \beta_0 + \beta_1\, x_i\ ,\ \sigma^2\ )$. Embora as observações tenham todas distribuição Normal e variâncias homogéneas, não são identicamente distribuídas, pois o seu valor médio varia consoante o correspondente valor do preditor, x_i.

(c) Sabe-se que, numa distribuição χ^2, o valor esperado é igual ao parâmetro da distribuição. Assim (e usando as propriedades operatórias do valor esperado), tem-se:

$$\begin{aligned} E\left[\frac{SQRE}{\sigma^2}\right] = n-2 \quad &\Leftrightarrow \quad \frac{1}{\sigma^2}\, E[SQRE] = n-2 \quad \Leftrightarrow \quad E[SQRE] = \sigma^2 \cdot (n-2) \\ &\Leftrightarrow \quad \frac{1}{n-2}\, E[SQRE] = \sigma^2 \quad \Leftrightarrow \quad E\left[\frac{SQRE}{n-2}\right] = \sigma^2\ . \end{aligned}$$

Logo, o quociente $\frac{SQRE}{n-2}$ é um estimador centrado da variância σ^2 dos erros aleatórios. Esse quociente chama-se o Quadrado Médio Residual.

3. Considera-se uma regressão linear múltipla com p preditores e n observações.

(a) A matriz do modelo $\mathbf{X}$ é uma matriz de dimensão $n \times (p+1)$, cuja primeira coluna é constituída por n uns (vector $\vec{\mathbf{1}}_n$) e cada uma das p colunas seguintes, que designamos por $\vec{\mathbf{x}}_j$ (com $j = 1, ..., p$), é dada pelos n valores observados correspondentes à j-ésima variável preditora. Trata-se duma matriz não aleatória, uma vez que no modelo se admitem fixados pelo experimentador os valores das variáveis preditoras. O subespaço das suas colunas, $\mathcal{C}(\mathbf{X})$, é por definição o conjunto de todas as possíveis combinações lineares das colunas de $\mathbf{X}$, ou seja, das combinações lineares da forma $a_0\, \vec{\mathbf{1}}_n + a_1\, \vec{\mathbf{x}}_1 + a_2\, \vec{\mathbf{x}}_2 + ... + a_p\, \vec{\mathbf{x}}_p$.

(b) Sabemos que a matriz de projecção ortogonal referida é $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, onde $\mathbf{X}$ é a matriz do modelo. Ora,

i. A idempotência de $\mathbf{H}$ significa que $\mathbf{HH} = \mathbf{H}$. Tendo em conta que $(\mathbf{X}^t\mathbf{X})^{-1}$ é a matriz inversa de $\mathbf{X}^t\mathbf{X}$, vem:

$$\mathbf{H}\,\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\cancel{\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\,.$$

A simetria de $\mathbf{H}$ significa que $\mathbf{H}^t = \mathbf{H}$. Recordem-se três propriedades conhecidas de matrizes, duas das quais constam do formulário: a transposta duma matriz transposta é a matriz original $((\mathbf{A}^t)^t = \mathbf{A})$; a transposta dum produto de matrizes é o produto das correspondentes transpostas, pela ordem inversa $((\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t)$; e a transposta duma matriz inversa é a inversa da transposta $((\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1})$. Tem-se então:

$$\mathbf{H}^t = [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t = [\mathbf{X}^t]^t[(\mathbf{X}^t\mathbf{X})^{-1}]^t\mathbf{X}^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\,.$$