INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

29 de Janeiro, 2018 **SEGUNDO EXAME 2017-18** **Uma resolução possível**

I

É dada uma tabela de contingências, com 16 contagens associadas às combinações de $a=4$ fases do ciclo vegetativo das oliveiras com $b=4$ classes de infestação.

1. É pedido para verificar se a distribuição das contagens nas 24 armadilhas de cada fase se distribui de forma idêntica pelas quatro classes de infestação. Efectuar-se-á um teste de homogeneidade χ^2, uma vez que o número de armadilhas em cada fase foi previamente fixado pelo experimentador.

   Hipóteses: Represente-se por $\pi_{j|i}$ a probabilidade de se recair na classe de infestação $j=1,2,3,4$ (correspondente às classes `A`, `B`, `C` e `D`, respectivamente), dada a fase $i=1,2,3,4$. Tem-se:
   Hipótese Nula (H_0): Para qualquer classe de infestação j verifica-se $\pi_{j|1}=\pi_{j|2}=\pi_{j|3}=\pi_{j|4}$ $[=\pi_{.j}$, sendo $\pi_{.j}$ a probabilidade de recair na classe de infestação $j]$.
   Hipótese Alternativa (H_1): pelo menos uma das igualdades em H_0 não se verifica.

   Estatística do Teste: É a estatística de Pearson, na forma de contagens bidimensionais: $X^2=\sum_{i=1}^{4}\sum_{j=1}^{4}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$, onde O_{ij} é a contagem correspondente à fase i e classe de infestação j, e $\hat{E}_{ij}$ é o correspondente valor esperado estimado, ao abrigo da hipótese nula de homogeneidade, dado por $\hat{E}_{ij}=\frac{N_{i.}\times N_{.j}}{N}$, onde $N=24\times 4=96$ é o número total de armadilhas observadas, $N_{i.}$ indica o número de armadilhas associadas à fase i e $N_{.j}$ indica o número total de armadilhas correspondentes à classe de infestação j. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$ (expressão geral num teste de homogeneidade).

   Nível de significância: $\alpha=$ P[Erro do tipo I] = P[Rej. H_0 | H_0 verdade] = 0.05

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $\chi^2_{calc}>\chi^2_{\alpha[(a-1)(b-1)]}=\chi^2_{0.05(9)}=16.919$.

2. É pedida a verificação das condições de Cochran, que permitem admitir a validade da distribuição assintótica $\chi^2_{(a-1)(b-1)}$. Essas condições exigem que em nenhuma célula da tabela de contingências haja um valor esperado estimado inferior a 1, e em não mais de 20% das células esse valor seja inferior a 5. Uma vez que o número de observações em cada linha é sempre igual, os valores esperados estimados serão constantes ao longo de cada coluna, e são dados por $\hat{E}_{ij}=\frac{24\times N_{.j}}{96}=\frac{N_{.j}}{4}$. Os totais de observações em cada coluna são dados por $N_{.1}=19$, $N_{.2}=21$, $N_{.3}=25$ e $N_{.4}=31$. Assim, o mais pequeno valor esperado estimado em qualquer célula corresponde às quatro célula da primeira coluna, sendo dado por $\hat{E}_{i1}=\frac{N_{.1}}{4}=4.75$. Trata-se de valores inferiores a 5, mas próximos desse limiar. Os valores esperados estimados seguintes (correspondentes às quatro células da segunda coluna) já são superiores a 5 (concretamente 5.25). Assim, têm-se em 25% das células um valor esperado estimado abaixo, mas bastante próximo de 5. Numa interpretação rigorosa dos critérios de Cochran, estes não se verificam. Mas uma interpretação flexível dos critérios permite admitir a validade da distribuição assintótica.

3. Tendo em conta a região crítica definida no ponto 1, o valor calculado da estatística, $X^2_{calc}=12.967$ não pertence à região crítica. Assim, não se rejeita H_0, ou seja, a informação disponível não permite rejeitar a hipótese de homogeneidade, ao longo das quatro fases do ciclo vegetativo das oliveiras, na distribuição das moscas capturadas.

4. As parcelas da estatística do teste são da forma $\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$ e medem o afastamento entre os valores esperados estimados, $\hat{E}_{ij}$, e os valores observados, O_{ij}. Sabendo que os valores esperados estimados são constantes ao longo de cada coluna, há um valor observado que salta à vista na tabela de contingências: o valor $O_{3,4}=13$. Esse valor corresponde a um valor esperado estimado $\hat{E}_{3,4}=\frac{24\times N_{.4}}{96}=\frac{31}{4}=7.75$. Assim, a parcela correspondente, na estatística do teste tem o valor 3.556452. Trata-se duma célula onde a contagem observada é quase o dobro da esperada. Mas, uma vez que não se rejeita H_0, mesmo esta discrepância não pode ser considerada significativa.

II

1. (a) Com excepção do Índice de Área Foliar, as restantes variáveis são medições da dimensão das oliveiras. As correlações entre qualquer par dessas restantes variáveis é positiva, o que é natural: uma árvore maior tende a ser maior nas diferentes variáveis medidas. Mas a correlação de `IAF` com qualquer das restantes variáveis é negativa. Assim, numa regressão linear simples de `IAF` com qualquer preditor individual (que representa sempre um aspecto da dimensão da oliveira) há uma relação decrescente: a maiores dimensões, menores valores do índice de área foliar. Logo, com base na informação disponível, pode afirmar-se que a oliveiras de maiores dimensões correspondem, em geral, menores valores de `IAF`, como afirmado no enunciado.

   (b) Estuda-se a regressão linear simples de `IAF` (Y) sobre o índice `D` (x).

      i. Tratando-se duma regressão linear simples, sabe-se que o Coeficiente de Determinação é o quadrado do coeficiente de correlação entre o preditor e a variável resposta, ou seja, $R^2=(-0.7746)^2=0.6000052$. Assim, aproximadamente 60% da variabilidade observada nos índices de área foliar é explicada pela regressão linear sobre o diâmetro da copa.

      ii. Na recta de regressão $y=b_0+b_1\,x$, o declive é dado por $b_1=r_{xy}\cdot\frac{s_y}{s_x}$, onde r_{xy} é o coeficiente de correlação referido na subalínea anterior, e s_x e s_y são os desvios padrões, respectivamente do preditor e da variável resposta. Tendo em conta os valores dados no enunciado, tem-se: $b_1=(-0.7746)\times\frac{\sqrt{3.8246}}{\sqrt{1.0770}}=-1.459697$. Por outro lado, a ordenada na origem é dada por $b_0=\overline{y}-b_1\,\overline{x}$, pelo que $b_0=4.928-(-1.459697)(3.323)=9.778573$. Logo, a equação da recta ajustada é $y=9.778573-1.459697\,x$.

      iii. O formulário indica a expressão dum intervalo de predição ($(1-\alpha)\times 100\%$) para uma observação individual de Y, dado $X=x$. Sabemos que o intervalo de confiança pedido no enunciado tem uma expressão análoga, mas sem a parcela "1+" debaixo da raíz quadrada, ou seja, é da forma:

$$\left]\,(b_0+b_1x)-t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\;,\;(b_0+b_1x)+t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\,\right[\,.$$

      Sabemos que $b_0=9.778573$, $b_1=-1.459697$, $x=4$, pelo que o valor estimado para o valor esperado de Y é $b_0+b_1\,x=3.939785$. Conhecemos ainda os valores $n=30$, $\overline{x}=3.323$, $s_x^2=1.0770$ e (a partir das tabelas) $t_{0.025(28)}=2.048$. Para obter $QMRE$ podemos começar por $s_y^2=3.8246$, donde $SQT=(n-1)\,s_y^2=29\times 3.8246=110.9134$. Assim, $SQR=R^2\times SQT=0.6000052\times 110.9134=66.54862$, pelo que $SQRE=SQT-SQR=110.9134-66.54862=44.36478$. Finalmente, $QMRE=\frac{SQRE}{n-2}=1.584456$. Substituindo estes valores na expressão do intervalo de confiança, obtemos o seguinte IC a 95% de confiança: $]\,3.3748\,,\,4.5047\,[$. Este intervalo a 95% de confiança é para o índice de área foliar *médio* para oliveiras com diâmetro da copa $4\,m$.

2. Neste ponto ajustamos uma regressão linear simples envolvendo os logaritmos (naturais) das mesmas variáveis consideradas no ponto anterior.

   (a) Uma relação linear entre os logaritmos de x e y corresponde a uma relação potência entre as variáveis originais. De facto,

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

   Assim, a curva ajustada entre `IAF` (y) e `D` (x) é $y = \mathrm{e}^{2.9442}\, x^{-1.2314} \Leftrightarrow y = \frac{18.99546}{x^{1.2314}}$.

   (b) A proporcionalidade inversa entre y e x corresponde a uma relação do tipo $y = \frac{c}{x}$, para alguma constante de proporcionalidade c. Tendo em conta a alínea anterior, verifica-se que o que é pedido é para testar se, na relação potência populacional, da forma $y = \mathrm{e}^{\beta_0}\, x^{\beta_1}$, é admissível considerar $\beta_1 = -1$. Como pedido no enunciado, a resposta será dada através dum teste de hipóteses.

   Hipóteses: $H_0 : \beta_1 = -1 \quad vs. \quad H_1 : \beta_1 \neq -1$.

   Estatística do Teste: $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}$, sob H_0.

   Nível de significância: $\alpha = \mathrm{P}[$ Erro do tipo I $] = \mathrm{P}[$ Rej. $H_0 \mid H_0$ verdade $] = 0.05$.

   Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{0.025(28)} = 2.048$.

   Conclusões: Tem-se $T_{calc} = \frac{b_1 - (-1)}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-1.2314+1}{0.1326} = -1.745098$. Logo, não se rejeita H_0, sendo de admitir (para o nível $\alpha = 0.05$) que o índice de área foliar `IAF` e o diâmetro da copa `D` sejam inversamente proporcionais.

   (c) O valor referido no enunciado (42.0678) é um valor comparável à Soma de Quadrados Residual do modelo linear, já que calcula a soma de quadrados da diferença entre valores observados e ajustados do índice de área foliar *na escala original do `IAF`*. Assim, é uma medida que permite comparar o desempenho dos dois modelos relacionando `IAF` e `D`: o modelo de regressão linear simples e o modelo potência ajustado através da transformação linearizante, após desfazer a transformação. A $SQRE$ do modelo linear fora já obtido na alínea 1(b)iii, e é $SQRE = 44.36478$. O valor correspondente no modelo potência (ajustado através da transformação linearizante) é menor (42.0678), pelo que se pode afirmar que corresponde a um melhor ajustamento ao abrigo do critério de minimizar a soma de quadrados dos resíduos de cada relação.

3. Tem-se um modelo de regressão linear múltipla entre `ln(IAF)` e os logaritmos (naturais) das restantes $p = 5$ variáveis.

   (a) O modelo agora ajustado tem um Coeficiente de Determinação $R^2 = 0.8493$, pelo que explica cerca de 85% da variabilidade observada nos logaritmos dos índices de área foliar, um valor bastante satisfatório. O valor do R^2 modificado é algo inferior, $R^2_{mod} = 0.8179$, verificando-se alguma penalização do valor de R^2 resultante do facto de um modelo com $p + 1 = 6$ parâmetros estar a ser ajustado com um número não muito elevado de observações, $n = 30$. De facto, e como $R^2_{mod} = 1 - (1 - R^2)\frac{n-1}{n-(p+1)}$ verifica-se que a parte da variabilidade de `ln(IAF)` não explicada pelo modelo ($1 - R^2 = 0.1507$) foi aumentada num factor $\frac{n-1}{n-(p+1)} = \frac{29}{24} = 1.208333$, ou seja, foi aumentada em cerca de 20%.

   (b) Pede-se para comparar este modelo de $p = 5$ preditores com um submodelo (do ponto II.2) de apenas $k = 1$ preditor. Usar-se-á um teste F parcial para comparar os dois modelos encaixados. Tem-se:

Hipóteses: $H_0 : \mathcal{R}^2_c = \mathcal{R}^2_s \quad vs. \quad H_1 : \mathcal{R}^2_c > \mathcal{R}^2_s$, onde $\mathcal{R}^2_c$ e $\mathcal{R}^2_s$ indicam os coeficientes de determinação populacional, respectivamente do modelo completo e do submodelo.

Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \frac{R^2_c - R^2_s}{1-R^2_c} \cap F_{(p-k,n-(p+1))}$, sob H_0.

Nível de significância: $\alpha = \text{P[Erro do tipo I]} = \text{P[Rej. } H_0 \mid H_0 \text{ verdade]} = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[4,24]} = 2.78$.

Conclusões: Tem-se $F_{calc} = \frac{24}{4} \frac{0.8493-0.7548}{1-0.8493} = 3.7624$. Logo, rejeita-se H_0, i.e., considerando-se que a qualidade de ajustamento do modelo completo difere significativamente (ao nível $\alpha = 0.05$) da do submodelo.

(c) O gráfico da esquerda tem no eixo vertical os resíduos usuais e_i de cada observação, e no eixo horizontal os correspondentes valores ajustados, $\hat{y}_i$. Sabemos que, a verificarem-se os pressupostos do modelo, este gráfico deveria ter os pontos dispostos numa banda horizontal, sem padrão aparente. No nosso caso, verifica-se alguma tendência (embora não muito pronunciada) para uma disposição numa banda oblíqua (assinalem-se os espaços em branco nos cantos superior esquerdo e inferior direito), que pode sugerir a presença duma observação com efeito alavanca importante, que atrai a superfície linear ajustada duma forma análoga às observações relativas a dinossáurios no exercício das aulas práticas. Em contrapartida, a dispersão dos resíduos parece ser constante ao longo de toda a gama de valores ajustados, pelo que o pressuposto de variâncias homogéneas parece admissível.

O gráfico da direita tem, no eixo vertical, os valores dos resíduos estandardizados (nenhum dos quais é, em módulo, maior do que aproximadamente 2, pelo que não há observações muito distantes da superfície ajustada) e, no eixo horizontal, os valores do efeito alavanca h_{ii} de cada observação. Sabemos que o valor alavanca tem de estar entre $\frac{1}{n} = \frac{1}{30} = 0.033333$ e 1, sendo o seu valor médio $\frac{p+1}{n} = \frac{6}{30} = 0.2$. O maior dos valores alavanca observados (para a observação 19) é cerca do dobro deste valor médio, e já é um valor respeitável (cerca de 0.4). Essa mesma observação tem uma distância de Cook muito elevada, já próximo do limiar 0.5, como se pode verificar pela sua proximidade à isolinha de Cook correspondente. É possível que a tendência para a obliquidade observada no gráfico da esquerda esteja associada a esta observação, que é bastante influente e tem um efeito alavanca razoável, embora não seja possível confirmar essa tese com a informação disponível no enunciado.

III

1. Tem-se a variável resposta rendimento das videiras (Y) e dois factores explicativos: o factor localidade e o factor ano. No entanto, o delineamento experimental não é factorial, uma vez que na grande maioria dos anos apenas houve observações numa única localidade. O modelo ajustado corresponde a considerar o delineamento como hierarquizado, com o factor dominante (A) a localidade (com $a = 3$ níveis) e o factor subordinado (B) o ano (que depende das localidades). O número de níveis do factor subordinado (ano) difere de localidade para localidade (isto é, difere consoante o nível do factor dominante), tendo-se $b_1 = 5$ anos para Mangualde (a localidade $i = 1$, pela ordem alfabética seguida pelo programa `R`); $b_2 = 4$ para Nelas e $b_3 = 2$ para a Vidigueira. O delineamento é equilibrado, uma vez que em cada uma das $\sum_{i=1}^{3} b_i = 11$ situações experimentais existem $n_c = 8$ repetições, perfazendo um total de 88 observações. Eis o modelo ANOVA ajustado:

   - $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, para qualquer $i = 1, 2, 3$, $j = 1, .., b_i$ e $k = 1, ..., 8$, sendo μ_{11} o rendimento esperado na primeira localidade (Mangualde), no primeiro ano observado (1994); α_i o efeito principal (aumento esperado no rendimento) associado à localidade i

(com a restrição $\alpha_1 = 0$); $\beta_{j(i)}$ o efeito (acréscimo no rendimento médio) associado ao ano j da localidade i (com a restrição $\beta_{1(i)} = 0$, para qualquer localidade i); e sendo ϵ_{ijk} o erro aleatório associado à observação Y_{ijk}.

- Admite-se que os erros aleatórios são Normais, de média zero e variâncias homogéneas: $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.
- Admite-se que os erros aleatórios ϵ_{ijk} são independentes.

2. Sabemos que os graus de liberdade associados ao factor dominante A (localidade) são $a - 1 = 2$, e que os graus de liberdade associados ao factor subordinado B são $\sum_{i=1}^{a} (b_i - 1) = 8$. Sabemos ainda que os graus de liberdade residuais são dados por $n - \sum_{i=1}^{3} b_i = 88 - 11 = 77$. A forma mais fácil de obter a Soma de Quadrados associada ao factor dominante A consiste em começar por obter a Soma de Quadrados Total $SQT = (n-1) \times s_y^2 = 87 \times 1.774264 = 154.361$, e depois usar a decomposição de SQT para obter $SQA = SQT - (SQB(A) + SQRE) = 154.361 - (49.11 + 44.16) = 61.091$. Também seria possível calcular SQA directamente pela sua definição, dada no formulário. Assim, o quadrado médio associado ao factor dominante é $QMA = \frac{SQA}{a-1} = \frac{61.091}{2} = 30.5455$. Finalmente, a estatística F_A para o teste aos efeitos do factor dominante é dada por $F_A = \frac{QMA}{QMRE} = \frac{30.5455}{0.573} = 53.30803$. **Nota:** Alguns destes valores podem sofrer erros de arredondamento nos cálculos.

   Logo, a tabela-resumo completa obtida é:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
local	2	61.09	30.545	53.31	3.01e-15
local:ano	8	49.11	6.139	10.71	5.44e-10
Residuals	77	44.16	0.573		

3. Neste caso há dois testes de interesse: aos efeitos do factor dominante A, e aos efeitos do factor subordinado B. Vejamos em pormenor o teste aos efeitos do factor subordinado B.

   Hipóteses: $H_0 : \beta_{j(i)} = 0 \ , \ \forall \ i, j$ vs. $H_1 : \exists \, i, j$ tal que $\beta_{j(i)} \neq 0$.

   Estatística do Teste: $F_B = \frac{QMB(A)}{QMRE} \cap F_{\left[\sum_{i=1}^{a}(b_i - 1) \, , \, n - \sum_{i=1}^{a} b_i\right]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(8,77)} \approx 2.05$.

   Conclusões: Como $F_{calc} = 10.71 > 2.05$, rejeita-se H_0, concluindo-se pela existência de efeitos significativos de ano (ao nível $\alpha = 0.05$). Esta conclusão seria semelhante para qualquer dos níveis de significância usuais, tendo em conta o *p-value* muito próximo de zero, disponível no enunciado ($p = 5.44 \times 10^{-10}$).

   No teste aos efeitos do factor dominante A (`local`) tem-se:

   Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i$ vs. $H_1 : \exists \, i$ tal que $\alpha_i \neq 0$.

   Estatística do Teste: $F_A = \frac{QMA}{QMRE} \cap F_{\left[a-1 \, , \, n - \sum_{i=1}^{a} b_i\right]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(2,77)} \approx 3.12$.

   Conclusões: Como $F_{calc} = 53.31 > 3.12$, rejeita-se H_0, concluindo-se pela existência de efeitos significativos de localidade (ao nível $\alpha = 0.05$). Esta conclusão seria semelhante para qualquer dos níveis de significância usuais, tendo em conta o *p-value* muito próximo de zero, disponível no enunciado ($p = 5.44 \times 10^{-10}$).

Os dois tipos de efeitos são claramente significativos.

4. Pretende-se comparar (com $\alpha = 0.05$) os pares de médias que se podem formar a partir das onze situações experimentais. Sabemos que duas médias populacionais, μ_{ij} e $\mu_{i'j'}$, são consideradas diferentes se as respectivas médias amostrais diferirem, em módulo, em mais do que a diferença significativa de Tukey, $q_{\alpha(\sum_{i=1}^{a} b_i\,,\,n-\sum_{i=1}^{a} b_i)}\sqrt{\frac{QMRE}{n_c}}$ (disponível no formulário), ou seja, se $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > q_{0.05(11,77)}\sqrt{\frac{0.573}{8}} \approx 4.69 \times \sqrt{0.071625} = 1.255178$. Ora, o rendimento médio amostral na Vidigueira, em 2002, foi o maior de todos, ou seja, $\overline{y}_{3,2} = 4.782$. As situações experimentais cujo rendimento seja inferior a $4.782 - 1.255 = 3.527$ diferem significativamente do rendimento na Vidigueira em 2002. Isso acontece em quase todas as restantes situações experimentais, sendo as excepções a Vidigueira em 2004 e Nelas em 2010.

IV

No contexto da regressão linear simples indicada no enunciado, tem-se:

1. Sabe-se que a recta de regressão tem equação $y = b_0 + b_1\,x$, sendo $b_0 = \overline{y} - b_1\overline{x}$. Esta última equação equivale a $\overline{y} = b_0 + b_1\,\overline{x}$. Mas isso significa que o ponto de coordenadas $(\overline{x}, \overline{y})$ satisfaz a equação da recta, ou seja, o centro de gravidade é um ponto da recta.

2. Exigir que para n pontos observados $\{(x_i, y_i)\}_{i=1}^{n}$ e uma qualquer recta $y = a + m\,x$, se verifique $\sum_{i=1}^{n}[y_i - (a + m\,x_i)] = 0$ equivale a exigir que $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(a + m\,x_i) = \underbrace{\sum_{i=1}^{n} a}_{=n\,a} + m\sum_{i=1}^{n} x_i$. Dividindo tudo por n resulta $\overline{y} = a + m\overline{x}$, ou seja, o ponto $(\overline{x}, \overline{y})$ será um ponto da recta. Mas a condição não impõe qualquer declive particular, pelo que não é uma condição suficiente para assegurar um bom ajustamento da recta à nuvem de pontos. **Comentário:** A exigência de, numa regressão linear, se minimizar a soma de quadrados dos resíduos contém em si uma dupla exigência, como se pode verificar considerando o sistema de equações que se obtém anulando as derivadas parciais de $SQRE$ em relação a b_0 e b_1. Uma dessa equações (de onde resulta a fórmula para b_0) corresponde a exigir que a soma dos resíduos seja nula. É a outra derivada parcial que irá determinar o declive óptimo da recta.

3. Sabemos que, num teste de ajustamento global duma regressão linear simples, ao nível de significância $\alpha = 0.05$, a rejeição de H_0 corresponde a ter-se $F_{calc} > f_{\alpha(1,n-2)}$, ou seja,

$$
\begin{aligned}
(n-2)\frac{R^2}{1-R^2} \;>\; f_{0.05(1,n-2)} \quad &\Leftrightarrow \quad \frac{n-2}{f_{0.05(1,n-2)}} \;>\; \frac{1-R^2}{R^2} \;=\; \frac{1}{R^2} - 1 \\
&\Leftrightarrow \quad 1 + \frac{n-2}{f_{0.05(1,n-2)}} \;>\; \frac{1}{R^2} \\
&\Leftrightarrow \quad R^2 \;>\; \frac{1}{1 + \frac{n-2}{f_{0.05(1,n-2)}}}
\end{aligned}
$$

O membro direito da desigualdade corresponde ao valor do coeficiente de determinação que define a fronteira de rejeição de H_0, e depende apenas da dimensão da amostra. Com base nas tabelas da distribuição F, verificamos que, para $n = 10$, tem-se $f_{0.05(1,8)} = 5.32$, logo o

limiar é $\frac{1}{1+\frac{8}{f_{0.05(1,8)}}} = 0.3993994$. Para $n = 62$, tem-se $f_{0.05(1,60)} = 4.00$, logo o limiar desce para $\frac{1}{1+\frac{60}{f_{0.05(1,60)}}} = 0.0625$. Considerando o limite quando n tende para ∞, a fracção no denominador do limiar tenderá para $+\infty$ (já que $\lim_{n\to\infty}(n-2) = +\infty$ e $\lim_{n\to+\infty} f_{0.05(1,n)} = 3.84$), pelo que a fracção tende para zero. Assim, para amostras muito grandes, mesmo valores muito pequenos de R^2 serão significativamente diferentes de zero. **Comentário:** Esta alínea chama a atenção para o facto de não ser possível decretar um dado valor de R^2 como sendo significativamente diferente de zero sem ter em conta a dimensão da amostra subjacente a esse valor. Numa regressão linear múltipla, além da dimensão da amostra, é também o número p de preditores que determina o limiar de significância de R^2.

4. Nesta pergunta, considera-se a notação das regressões lineares múltiplas no estudo duma regressão linear simples (no espírito do que foi feito no Exercício 3 de RLM nas aulas práticas).

   (a) Sabemos que qualquer matriz de (co-)variâncias do vector de estimadores, $\vec{\hat{\boldsymbol{\beta}}}$, tem na diagonal, as variâncias de cada estimador, e fora da diagonal, a covariância correspondente a cada linha/coluna da matriz. Tendo em conta que o nosso vector $\vec{\hat{\boldsymbol{\beta}}}$ apenas tem dois elementos, e tendo em conta as fórmulas para $V[\hat{\beta}_0]$, $V[\hat{\beta}_1]$ e $Cov[\hat{\beta}_0, \hat{\beta}_1]$ disponíveis no formulário, tem-se que a matriz pedida é a seguinte:

$$V[\vec{\hat{\boldsymbol{\beta}}}] \;=\; \begin{bmatrix} V[\hat{\beta}_0] & Cov[\hat{\beta}_0,\hat{\beta}_1] \\ Cov[\hat{\beta}_0,\hat{\beta}_1] & V[\hat{\beta}_1] \end{bmatrix} \;=\; \begin{bmatrix} \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\,s_x^2}\right] & \frac{-\overline{x}\,\sigma^2}{(n-1)\,s_x^2} \\ \frac{-\overline{x}\,\sigma^2}{(n-1)\,s_x^2} & \frac{\sigma^2}{(n-1)\,s_x^2} \end{bmatrix} .$$

   (b) Sabemos que o vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$ tem distribuição Multinormal (neste caso, de dimensão 2). O vector esperado corresponde ao vector dos verdadeiros valores dos parâmetros, ou seja, ao vector $\vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ (estimadores centrados). A matriz de (co-)variâncias foi indicada na alínea anterior. Assim, tem-se:

$$\vec{\hat{\boldsymbol{\beta}}} \;\cap\; \mathcal{N}_2\left(\vec{\boldsymbol{\beta}}, V[\vec{\hat{\boldsymbol{\beta}}}]\right) .$$

   Ora, o estimador $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1\, x$ corresponde a uma combinação linear dos estimadores no vector $\vec{\hat{\boldsymbol{\beta}}}$, com o vector de coeficientes dado por $\vec{\mathbf{a}} = \begin{bmatrix} 1 \\ x \end{bmatrix}$, ou seja, $\hat{\mu}_{Y|x} = \vec{\mathbf{a}}^t \vec{\hat{\boldsymbol{\beta}}}$. Pelas propriedades da Multinormal estudadas nas aulas, qualquer combinação linear dos elementos dum vector Multinormal tem distribuição Normal. Além disso, $E[\vec{\mathbf{a}}^t \vec{\hat{\boldsymbol{\beta}}}] = \vec{\mathbf{a}}^t E[\vec{\hat{\boldsymbol{\beta}}}] = \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}}$ e $V[\vec{\mathbf{a}}^t \vec{\hat{\boldsymbol{\beta}}}] = \vec{\mathbf{a}}^t\, V[\vec{\hat{\boldsymbol{\beta}}}]\, \vec{\mathbf{a}}$. Logo,

$$\hat{\mu}_{Y|x} = \vec{\mathbf{a}}^t \vec{\hat{\boldsymbol{\beta}}} \;\cap\; \mathcal{N}\left(\; \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} \;,\; \vec{\mathbf{a}}^t\, V[\vec{\hat{\boldsymbol{\beta}}}]\, \vec{\mathbf{a}} \;\right) .$$

Assim, o valor esperado de $\hat{\mu}_{Y|x}$ é $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_0 + \beta_1\, x$. A variância correspondente é dada por:

$$
\begin{aligned}
V[\hat{\mu}_{Y|x}] \;=\; \vec{\mathbf{a}}^t V[\vec{\hat{\boldsymbol{\beta}}}]\vec{\mathbf{a}} \;&=\; [1 \quad x] \begin{bmatrix} \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\,s_x^2}\right] & \frac{-\overline{x}\,\sigma^2}{(n-1)\,s_x^2} \\ \frac{-\overline{x}\,\sigma^2}{(n-1)\,s_x^2} & \frac{\sigma^2}{(n-1)\,s_x^2} \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} \\
&=\; \left[\; \frac{\sigma^2}{n} + \frac{\sigma^2}{(n-1)\,s_x^2}(\overline{x}^2 - x\overline{x}) \qquad \frac{\sigma^2}{(n-1)\,s_x^2}\,(-\overline{x} + x) \;\right] \begin{bmatrix} 1 \\ x \end{bmatrix} \\
&=\; \frac{\sigma^2}{n} + \frac{\sigma^2}{(n-1)\,s_x^2}\left[\overline{x}^2 - x\overline{x} - x\overline{x} + x^2\right] \\
&=\; \sigma^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right] .
\end{aligned}
$$

Logo,

$$\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1\, x \;\; \cap \;\; \mathcal{N}\left(\; \beta_0 + \beta_1\, x \;,\; \sigma^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right] \;\right)$$

que corresponde ao resultado visto nas aulas.