INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

19 de Janeiro, 2016 **SEGUNDO EXAME 2015-16** **Uma resolução possível**

I

As distribuições Binomiais surgem associadas a variáveis aleatórias X que contam o número de êxitos em provas de Bernoulli. Recorde-se que provas de Bernoulli são experiências aleatórias com apenas dois possíveis resultados, que convencionamos chamar *êxitos* e *fracassos*, que se repetem de forma independente e em condições equivalentes, de tal forma que a probabilidade de um *êxito* seja igual em todas as provas. No nosso contexto, surge de forma natural a sugestão de uma distribuição Binomial para a variável aleatória X que conta o número de não enraizamentos (que, paradoxalmente, podemos chamar *êxitos*) em cada contentor, e da qual temos $N=164$ observações. A validade da distribuição Binomial pressupõe a independência das contagens e a equivalência das condições em cada contentor.

1. Uma distribuição Binomial tem dois parâmetros: m, que indica o número total de provas de Bernoulli em relação à qual se faz a contagem; e p, a probabilidade de êxito em cada prova individual de Bernoulli. No nosso contexto, o valor $m=10$ do primeiro parâmetro é determinado pela própria natureza do problema, já que X conta o número de êxitos em cada contentor, tendo cada contentor 10 plantas. Já o valor do segundo parâmetro, p, não é previamente especificado, tornando-se necessário estimá-lo a partir dos dados. A forma usual de fazer essa estimação baseia-se no facto de que se $X \cap B(m,p)$, então o seu valor esperado será $E[X]=mp$, pelo que $p=\frac{E[X]}{m}$. O valor esperado é desconhecido, mas pode ser estimado a partir da média amostral $\overline{x}$, ou seja, do número médio de estacas não enrazaidas por contentor, observadas nos $N=164$ contentores estudados. Tendo por base a tabela do enunciado, tem-se:

$$\overline{x}=\frac{0\times14+1\times54+2\times44+3\times22+4\times16+5\times8+6\times2+7\times3+8\times1}{164}=\frac{353}{164}=2.15244\ .$$

   Assim, o valor estimado de p será $\hat{p}=\frac{\overline{x}}{m}=\frac{2.15244}{10}=0.215244$.

2. A validade da distribuição assintótica da estatística de Pearson, nos teste χ^2, está associada aos critérios de Cochran, que exige que em nenhuma classe de valores haja menos de 1 contagem esperada, e que em não mais de 20% das classes haja menos de 5 contagens esperadas. Ora, se fossem usadas as classes de valores individuais constantes da tabela, o número esperado de observações correspondente a cada possível resultado seria dado por $\hat{E}_i=N\times P[X=i]=N\times\binom{m}{i}\hat{p}^i(1-\hat{p})^{m-i}$ (para $i=0,1,2,...,10$). Para simplificar as contas, nesta alínea será usado o valor aproximado da estimativa de p indicado no enunciado ($\hat{p}=0.2$), em vez do valor mais preciso calculado na alínea anterior. Assim, por exemplo, ao resultado $X=10$ corresponderia o valor esperado estimado $\hat{E}_{10}=164\times1\times(0.2)^{10}\times0.8^0$, ou seja, $\hat{E}_{10}=0.000016794\ll1$. Será, pois, necessário agrupar classes de resultados, a fim de respeitar os critérios de Cochran. O agrupamento sugerido no enunciado consiste em criar uma nova classe para resultados $X\geq5$. Nessa nova classe agrupada, o valor esperado estimado será $\hat{E}_{5+}=N\times P[X\geq5]$. Para calcular a probabilidade, recorre-se à tabela da distribuição cumulativa duma Binomial de parâmetros $m=10$ e $p=0.2$, obtendo-se: $P[X\geq5]=1-P[X\leq4]=1-0.9672=0.0378$. Logo, $\hat{E}_{5+}=164\times0.0378=6.1992$. Este valor, superior a 5, é o mais pequeno de todos os valores esperados estimados, uma vez que está associado à mais baixa probabilidade de ocorrência,

como se comprova olhando rapidamente para a tabela da Binomial. De facto, as probabilidades associadas aos valores individuais $X = i$ $(i = 0, 1, 2, 3, 4)$ são todas maiores que 0.0378, sendo a mais baixa $P[X = 4] = P[X \leq 4] - P[X \leq 3] = 0.9672 - 0.8791 = 0.0881$. Logo, é respeitado o critério de Cochran.

3. **Hipóteses:** $H_0 : X \cap B(10, 0.215244)$ vs. $H_1 : X \not\cap B(10, 0.215244)$.

   Estatística do Teste: A estatística de Pearson é dada por $X^2 = \sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$, sendo $k = 6$ (após o agrupamento de classes), O_i o número de observações na classe i e $\hat{E}_i$ o correspondente valor esperado estimado ao abrigo de H_0. A distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_{k-r-1} sendo $r = 1$ o número de parâmetros da distribuição que foi necessário estimar (o parâmetro p). Logo, a distribuição assintótica é χ^2_4.

   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 $\chi^2_{\text{calc}} > \chi^2_{0.05(4)} = 9.48773$.

   Conclusões: Como $\chi^2_{calc} = 17.5129 > 9.48778$, rejeita-se a hipótese nula de X seguir uma distribuição Binomial, com parâmetros $m = 10$ e $\hat{p} = 0.215244$. Tendo sido usados os valores de parâmetros duma Binomial mais adequados a este conjunto de dados (sendo p estimado a partir dos dados), esta conclusão aponta para a rejeição duma distribuição Binomial como a distribuição subjacente à variável de contagem X.

4. Pede-se o valor, na estatística do teste, da parcela correspondente a 3 estacas não enraizadas (que corresponde à classe $i = 4$). Essa parcela é dada por $\frac{(O_4 - \hat{E}_4)^2}{\hat{E}_4}$, sendo $\hat{E}_4 = N \times P[X = 3]$. Pode aproximar-se o valor de $P[X = 3]$ com recurso às tabelas da Binomial (usando o valor $\hat{p} = 0.2$ sugerido no enunciado), obtendo-se $\hat{E}_4 = N \times (P[X \leq 3] - P[X \leq 2]) = 164 \times (0.8791 - 0.6778) = 33.0132$. Assim, a parcela pedida terá valor (aproximado) $\frac{(O_4 - \hat{E}_4)^2}{\hat{E}_4} = \frac{(22 - 33.0132)^2}{33.0132} = 3.674$. Trata-se de cerca de um quinto do valor final da estatística de teste, mas a contribuição desta parcela não seria, por si só, suficiente para rejeitar a hipótese de distribuição Binomial.

II

1. A qualidade de ajustamento do modelo pode ser avaliada pelo coeficiente de determinação, que no nosso caso, é bastante modesto: $R^2 = 0.5841$. Assim, este modelo apenas explica pouco mais de 58% da variabilidade observada nos teores de antocianas. No entanto, este valor de R^2 é significativamente diferente de zero, como se comprova através dum teste de ajustamento global do modelo:

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

   Estatística do Teste: $F = \frac{n-(p+1)}{p} \frac{R^2}{1-R^2} \cap F_{[p, n-(p+1)]}$, sendo H_0 verdade.

   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 $F_{calc} > f_{\alpha(5,34)}$ que, pelas tabelas, corresppnde a um valor entre 2.53 e 2.45.

   Conclusões: O valor $F_{calc} = 9.551$, dado no enunciado, pertence à região crítica, pelo que se rejeita H_0, ou seja, tratando-se dum modelo que deixa a desejar, é no entanto um modelo significativamente diferente do modelo nulo, pelo que os preditores do modelo explicam alguma da variabilidade do teor de antocianas.

2. O coeficiente de determinação modificado tem neste exemplo o valor $R^2_{mod} = 0.523$. A definição deste coeficiente modificado é dada no formulário: $R^2_{mod} = 1 - \frac{QMRE}{QMT}$. Como se viu nas aulas, e se pode concluir a partir das definições dos dois Quadrados Médios referidos, tem-se $R^2_{mod} = 1 - \frac{n-1}{n-(p+1)} \frac{SQRE}{SQT} = 1 - \frac{n-1}{n-(p+1)}(1 - R^2)$ (onde R^2 indica o coeficiente de determinação usual). Assim, R^2_{mod} multiplica a proporção da variabilidade da variável resposta que *não* é explicada pelo modelo, $1 - R^2$, pelo factor $\frac{n-1}{n-(p+1)}$ que, sendo sempre maior do que 1, será tanto maior quanto menor fôr a diferença entre a informação usada para ajustar o modelo (reflectida na dimensão n da amostra) e a complexidade do modelo (expressa pelo número de parâmetros do modelo, $p+1$). No nosso caso, tem-se $n = 40$ e $p+1 = 6$, pelo que a referida diferença não é nem muito grande, nem muito pequena. No entanto, a modesta qualidade do modelo ($R^2 = 0.5841$) implica que o factor $\frac{n-1}{n-(p+1)} = \frac{39}{34} = 1.147$ vai incidir sobre uma proporção relativamente grande de variabilidade não explicada ($1 - R^2 = 0.4159$), penalizando esta proporção com um aumento de cerca de 15%. Assim se explica a diferença visível no valor de R^2_{mod}, quando comparado com o coeficiente de determinação usual.

3. A contribuição da variável `pH` para a previsão do teor de `antocianas` é traduzida pela parcela que lhe corresponde no modelo de regressão linear múltipla, ou seja, pela parcela β_4 `pH`. Se $\beta_4 = 0$, esta parcela em nada contribui para a estimação do teor de antocianas (independentemente do valor de `pH`). Logo, a veracidade da afirmação está associada ao facto de ser, ou não, admissível esse valor de β_4. Não é possível dar uma resposta apenas olhando para a magnitude da estimativa de β_4, ou seja, para o valor $b_4 = -54.928$ (esse valor depende também das unidades de medida das observações). É necessário ter também em conta a variabilidade associada a essa estimação, que é transmitida pelo erro padrão associado. O enunciado pede para responder através dum intervalo de confiança para β_4, que usa essa informação. A expressão genérica para esse IC a $(1-\alpha)\times 100\%$ é: $]\, b_4 - t_{\alpha/2(n-(p+1))}\, \hat{\sigma}_{\hat{\beta}_4}\, ,\ b_4 - t_{\alpha/2(n-(p+1))}\, \hat{\sigma}_{\hat{\beta}_4}\, [$. Tem-se $b_4 = -54.928$, $t_{0.025(34)} \approx 2.03$ e $\hat{\sigma}_{\hat{\beta}_4} = 51.261$, pelo que o IC a 95% de confiança é $]\, -158.9878\, ,\ 49.13183\, [$. Trata-se dum intervalo de grande amplitude, que inclui o valor zero. Assim, não é possível sustentar a afirmação do enunciado.

4. (a) A variável excluída no primeiro passo do algoritmo de exclusão sequencial foi a variável `rend`, uma vez que o valor de prova ($p = 0.75059$) no teste a que o respectivo coeficiente β_1 seja nulo é claramente superior aos níveis de significância usuais e é o mais elevado entre todos os preditores. Por outras palavras, entre todas as variáveis preditoras, `rend` é aquela em que se está mais longe de rejeitar a hipótese nula $\beta_j = 0$. A variável excluída no segundo passo do algoritmo tem de ser a variável `pH`, uma vez que é a outra variável do modelo original que não figura no submodelo, mas não era possível garantir esse facto apenas com base no ajustamento do modelo completo.

   (b) O valor de $QMRE$ no submodelo final (que tem $k = 3$ preditores, a saber, `brix`, `acidez` e `pesobago`), pode ser calculado a partir do valor do Critério de Informação de Akaike (AIC) desse submodelo, que pelo enunciado é $AIC = 213.02$. Ora, pelo formulário tem-se que, num submodelo com k preditores, $AIC = n \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1)$. Assim, $SQRE_k = n\, e^{\frac{AIC - 2(k+1)}{n}} = 6730.33$, pelo que $QMRE_k = 6730.33/36 = 186.9536$. O valor estimado da variância dos erros aleatórios no modelo original é o $QMRE$ desse modelo completo, que pelo enunciado do ajustamento do modelo completo é $QMRE_p = (13.83)^2 = 191.2689$. Assim, e apesar de no submodelo o valor de $SQRE$ ser (necessariamente) maior, a redução simultânea do número de preditores permitiu que o respectivo $QMRE$ seja menor do que no modelo completo.

5. Considerando agora apenas regressões lineares simples.

(a) A melhor variável preditora será a variável mais fortemente correlacionada com a variável resposta `antocianas`. Pela matriz de correlações (dada no enunciado) verifica-se que esse melhor preditor é a variável `brix`, cuja correlação com as `antocianas` é $r = 0.561$. Sabemos que, numa regressão linear simples, o coeficiente de determinação R^2 é o quadrado do coeficiente de correlação entre preditor e variável resposta, pelo que $R^2 = (0.561)^2 = 0.314721$. Assim, esta melhor regressão linear simples apenas explica cerca de $31,5\%$ da variabilidade observada no teor das antocianas, um valor que é muito baixo. Este valor nunca poderia ser maior que o R^2 do modelo completo (uma vez que é um submodelo desse modelo com cinco preditores), mas ainda se perde bastante capacidade explicativa em relação ao já baixo valor de R^2 do modelo completo.

(b) Independentemente da fraca capacidade explicativa do modelo de regressão linear simples, é possível ajustar a respectiva recta de regressão de teor de antocianas (y) sobre teor brix (x), que sabemos ter equação $y = b_0 + b_1\, x$, com $b_1 = r_{xy}\frac{s_y}{s_x} = (0.561) \times \frac{20.02151}{0.94994} = 11.82398$, e $b_0 = \overline{y} - b_1\overline{x} = 98.52 - 11.82398 \times 21.53 = -156.0502$ (sendo dadas no enunciado as médias e desvios padrão das duas variáveis do modelo). Ou seja, a equação da recta ajustada é $y = -156.0502 + 11.82398\, x$.

(c) Sabemos (ver formulário) que o estimador $\hat{\beta}_1$ do declive β_1 da recta de regressão considerada tem variância $V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)s_x^2}$, e o respectivo erro padrão é a raíz quadrada deste valor. Tem-se $n = 40$ e $s_x^2 = (0.94994)^2 = 0.902386$. O valor da variância σ^2 dos erros aleatórios é desconhecido, mas é estimado pelo Quadrado Médio Residual. Logo, podemos calcular a estimativa do erro padrão após calcularmos esse valor de $QMRE$. Pela definição de R^2 tem-se $SQR = R^2 \times SQT$, e como $R^2 = 0.314721$ e $SQT = (n-1)s_y^2 = 39 \times (20.02151)^2 = 15633.57$, vem $SQR = 4920.214$. Por outro lado, $SQRE = SQRT - SQR = 15633.57 - 4920.214 = 10713.36$. Assim, $QMRE = \frac{SQRE}{n-2} = 281.9304$. Logo, o erro padrão pedido é $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{281.9304}{39 \times 0.902386}} = 2.830365$.

III

1. Trata-se dum delineamento a um único factor (as variedades de tomate), sendo a variável resposta Y a resistência da película (em gf). Em cada um dos $k = 6$ níveis do factor há $n_c = 3$ repetições (as parcelas). O número igual de repetições nas 6 situações experimentais significa que o delineamento é equilibrado. **NOTA:** Não faz sentido considerar parcelas como um segundo factor num delineamento factorial a dois factores, uma vez que as parcelas duma variedade nada têm em comum com as parcelas de outra variedade; considerar parcelas como um segundo factor, subordinado ao factor variedade, numa relação hierarquizada, é conceptualmente possível, mas nesse caso haveria uma única observação em cada uma das 18 situações experimentais (parcelas) e não seria possível ajustar um modelo ANOVA.

   O modelo ANOVA a um factor é:

   - A resistência Y_{ij}, na j-ésima parcela ($j = 1, 2, 3$) associada à variedade i ($i = 1, ..., 6$), é dada por $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall\, i, j$, sendo μ_1 o rendimento esperado da primeira variedade; α_i o efeito (acréscimo à resistência) associado à variedade i (com a restrição $\alpha_1 = 0$); e ϵ_{ij} o erro aleatório da observação Y_{ij}.
   - Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogéneas: $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2)$, para qualquer i, j.

- Admite-se que os erros aleatórios ϵ_{ij} são independentes.

2. A tabela-resumo terá apenas duas linhas (além da linha correspondente aos Totais), associadas respectivamente aos efeitos do Factor e à variabilidade Residual. Sabemos que os graus de liberdade dos efeitos do factor são $k–1=5$ e que os graus de liberdade residuais são $n–k=18–6=12$. Por outro lado, as fórmulas para as Somas de Quadrados neste mais simples de todos os modelos são dadas no enunciado. A Soma de Quadrados Residual é $SQRE=\sum_{i=1}^{k}(n_i-1)s_i^2$ e, tratando-se dum delineamento equilibrado com $n_c=3$ tem-se, usando as variâncias amostrais de nível dadas no enunciado, $SQRE=2\times(14713.08+367.9434+5881.921+33132.64+5.414433+47.11163)=108\,296.2$. Também será possível calcular SQF a partir da fórmula do enunciado, uma vez que são disponibilizadas as médias amostrais de nível e globais. Mas a forma mais simples de obter esse valor vem da constatação de que, numa ANOVA a um factor, se tem $SQF=SQT-SQRE$ e também $SQT=(n-1)s_y^2=17\times 34\,517.82=586\,802.9$. Logo, $SQF=478\,506.7$. Dividindo estas Somas de Quadrados pelos graus de liberdade antes referidos obtêm-se os Quadrados Médios, e dividindo QMF por $QMRE$ obtem-se o valor calculado da estatística do teste F aos efeitos do factor. Assim, a tabela-resumo é:

	g.l.	SQs	Quadrados Médios	F_{calc}
Factor	5	478 506.7	$\frac{478\,506.7}{5}=95\,701.35$	$F_{calc}=\frac{QMF}{QMRE}=\frac{95\,701.35}{9\,024.685}=10.6044$
Residual	12	108 296.2	$\frac{108\,296.2}{12}=9\,024.685$	–

3. Eis o teste aos efeitos do factor (variedade):

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

   Estatística do Teste: $F = \frac{QMF}{QMRE} \cap F_{[k–1,n–k]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(5,12)}=3.11$.

   Conclusões: Como $F_{calc}=10.6044 > 3.11$, rejeita-se H_0, sendo possível concluir pela existência de efeitos de variedade (ao nível $\alpha = 0.05$). No nosso contexto, tal corresponde a afirmar que existem variedades de tomate cujas películas têm resistência média diferentes de outras.

4. Para comparar médias de variedade iremos utilizar o teste de Tukey. Sabemos que podemos considerar diferentes duas médias populacionais de nível, μ_i e $\mu_{i'}$, caso as respectivas médias amostrais de nível difiram mais do que o termo de comparação do teste de Tukey, ou seja, se $|\overline{y}_{i.} - \overline{y}_{i'.}| > q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}}$, onde $q_{\alpha(k,n-k)}$ indica o valor que deixa à sua direita uma região de probabilidade α, na distribuição de Tukey com parâmetros k e $n-k$. No nosso caso, $k=6$ e $n-k=12$, sendo, pelas tabelas da distribuição Tukey, $q_{0.05(6,12)}=4.75$. Como $\sqrt{\frac{QMRE}{n_c}}=\sqrt{\frac{9\,024.685}{3}}=54.84732$, podemos decretar a diferença significativa entre a média amostral das resistências da variedade $40C$, $\overline{y}_{4.}=705.8$ (que é a maior de todas), e a de qualquer outra variedade cuja média difira desta em mais de $4.75\times 54.84732=260.5248$ *gf*. Ora, $705.8-260.5248=445.2752$, e apenas a variedade *18* não tem média amostral inferior a esse valor. Logo, podemos concluir que as resistências médias das variedades *28*, *29*, *Ace* e *Roma* são diferentes (inferiores) à resistência média da variedade *40C*.

5. O teste de Bartlett visa optar entre a hipótese nula da igualdade de variâncias populacionais nas diferentes situações experimentais do delineamento, que neste caso corresponde às diferentes

variedades de tomate (níveis do factor) e a hipótese alternativa de que existe pelo menos um par de níveis com variâncias populacionais da variável resposta diferentes. Mais concretamente, visa optar entre $H_0 : \sigma_i^2 = \sigma_{i'}^2$, para qualquer par de níveis i, i', e $H_1 : \exists\, i, i'$ tal que $\sigma_i^2 \neq \sigma_{i'}^2$. A estatística do teste K^2 (que consta do formulário, mas tem uma expressão complicada) segue uma distribuição assintótica χ^2_{k-1}, com uma região crítica unilateral direita. Assim, no nosso problema deve rejeitar-se a hipótese nula da igualdade das variâncias caso $K^2_{calc} > \chi^2_{0.05(5)} = 11.0705$. Ora, segundo a listagem no enunciado, $K^2_{calc} = 24.2998$, pelo que se conclui pela rejeição de H_0, em favor de H_1 (ao nível $\alpha = 0.05$). Note-se que esta última hipótese representaria uma violação do pressuposto dos modelos ANOVA de igualdade das variâncias populacionais de todas as observações. No entanto, cabe referir que a aplicação do teste de Bartlett a este problema é duvidosa, uma vez que, com apenas 3 repetições em cada nível do factor, a validade da distribuição assintótica da estatística do teste é questionável (recorde-se a nossa convenção de que se deve exigir pelo menos 5 repetições em cada nível do factor para admitir essa distribuição assintótica).

IV

1. As três alíneas correspondem a mostrar que o estimador $\hat{\beta}_1$ tem distribuição $\mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$.

   (a) Como se recorda no enunciado, o estimador $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$ é uma combinação linear dos Y_i, que de acordo com o modelo linear são variáveis aleatórias Normais e independentes. É sabido que qualquer combinação linear de Normais independentes tem distribuição Normal, pelo que o estimador $\hat{\beta}_1$ tem uma distribuição dessa família. Nas alíneas seguintes calculam-se os dois parâmetros dessa distribuição.

   (b) Pela linearidade do valor esperado tem-se $E[\hat{\beta}_1] = E\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i E[Y_i]$. Tendo em conta que $E[Y_i] = \beta_0 + \beta_1 x_i$ (ver enunciado), vem:

   $$E[\hat{\beta}_1] = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} c_i\beta_0 + \sum_{i=1}^{n} c_i\beta_1 x_i = \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n} c_i x_i}_{=1} = \beta_1 \ ,$$

   uma vez que $\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{(n-1)s_x^2} = \frac{1}{(n-1)s_x^2} \sum_{i=1}^{n} (x_i - \overline{x}) = 0$, como se provou no Exercício 3a) da Regressão Linear Simples (sendo a igualdade $\sum_{i=1}^{n} c_i x_i = 1$ dada no enunciado).

   (c) De forma análoga, mas tendo em conta as propriedades das variâncias, nomeadamente que constantes multiplicativas passam para fora, elevadas ao quadrado, e que a variância da soma de v.a.s independentes é a soma das respectivas variâncias, tem-se:

   $$\begin{aligned} V[\hat{\beta}_1] &= V\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i^2 \underbrace{V[Y_i]}_{=\sigma^2} \\ &= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{[(n-1)s_x^2]^2} = \frac{\sigma^2}{[(n-1)s_x^2]^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)s_x^2} \\ &= \frac{\sigma^2}{[(n-1)s_x^2]^{\cancel{2}}} \cancel{(n-1)s_x^2} = \frac{\sigma^2}{(n-1)s_x^2} \ . \end{aligned}$$

2. (a) A soma dos elementos dum qualquer vector $\vec{\mathbf{x}} = (x_1, x_2, ..., x_n)^t$ obtém-se tomando o produto interno desse mesmo vector com o vector dos n uns: $\mathbf{1}_n^t \vec{\mathbf{x}} = \sum_{i=1}^{n} x_i$. Assim, a soma dos valores observados Y_i é dada por $\mathbf{1}_n^t \vec{\mathbf{Y}}$ e a soma dos valores ajustados $\hat{Y}_i$ é dada por $\mathbf{1}_n^t \vec{\hat{\mathbf{Y}}} = \mathbf{1}_n^t \mathbf{H} \vec{\mathbf{Y}} = (\mathbf{H}^t \mathbf{1}_n)^t \vec{\mathbf{Y}} = (\mathbf{H}\mathbf{1}_n)^t \vec{\mathbf{Y}}$, já que a matriz de projecção ortogonal $\mathbf{H}$ é simétrica ($\mathbf{H}^t = \mathbf{H}$). Mas $\mathbf{H}\mathbf{1}_n = \mathbf{1}_n$, uma vez que o vector $\mathbf{1}_n$ pertence ao subespaço $\mathcal{C}(\mathbf{X})$ sobre o qual $\mathbf{H}$ projecta, logo permanece invariante sob o efeito dessa projecção. Assim, as somas de valores observados e ajustados de Y coincidem ($\mathbf{1}_n^t \vec{\hat{\mathbf{Y}}} = \mathbf{1}_n \vec{\mathbf{Y}}$), pelo que coincidem também as respectivas médias.

   (b) Tem-se $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$. Sabemos pelas propriedades da distribuição Multinormal que, sendo $\vec{\mathbf{Y}}$ Multinormal, o produto $\mathbf{H}\vec{\mathbf{Y}}$ também o será (acetato 238, propriedade 7), sendo o seu vector médio dado por:

$$E[\vec{\hat{\mathbf{Y}}}] = E[\mathbf{H}\vec{\mathbf{Y}}] = \mathbf{H}E[\vec{\mathbf{Y}}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} \ ,$$

   uma vez que, sendo $\mathbf{H}$ a matriz de projecção ortogonal sobre o espaço das colunas da matriz $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$, qualquer vector desse subespaço (como é $\mathbf{X}\boldsymbol{\beta}$) permanece invariante sob o efeito dessa projecção. Por outro lado, e tendo em conta as propriedades das matrizes de variâncias-covariâncias, a matriz de (co-)variâncias do vector $\vec{\hat{\mathbf{Y}}}$ é dado por:

$$V[\vec{\hat{\mathbf{Y}}}] = V[\mathbf{H}\vec{\mathbf{Y}}] = \mathbf{H} \underbrace{V[\vec{\mathbf{Y}}]}_{=\sigma^2\mathbf{I}_n} \underbrace{\mathbf{H}^t}_{=\mathbf{H}} = \sigma^2 \mathbf{H}\mathbf{H} = \sigma^2 \mathbf{H},$$

   dada a idempotência de $\mathbf{H}$ ($\mathbf{H}\mathbf{H} = \mathbf{H}$). Logo, $\vec{\hat{\mathbf{Y}}} \cap \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H})$.