INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

28 de Janeiro, 2015 **2^a CHAMADA de EXAME 2014-15 Uma resolução possível**

I

1. O número médio de colónias por quadrado é dado por uma média ponderada dos valores das contagens de colónias ($i = 0, 1, 2, ..., 9$), sendo as ponderações dadas pelas frequências observadas (O_i) para cada contagem:

$$\overline{x} \;=\; \frac{\sum_{i=0}^{9} i \times O_i}{\sum_{i=0}^{9} O_i} \;=\; \frac{0\times 27 + 1\times 50 + 2\times 11 + 3\times 6 + 4\times 3 + 5\times 1 + 6\times 1 + 7\times 1}{100} \;=\; 1.2\ .$$

2. Pede-se um teste χ^2 baseado na estatística de Pearson, para aferir a hipótese da variável aleatória X que representa as contagens de colónias por quadrado seguir uma distribuição de Poisson. Não é previamente especificado qualquer valor para o parâmetro λ dessa distribuição, pelo que será necessário estimá-lo. Sabe-se que, numa distribuição de Poisson, o parâmetro λ corresponde ao valor esperado da variável, pelo que é natural estimá-lo no nosso caso pelo número médio de contagens por colónia, que foi calculado na alínea anterior. Assim, toma-se $\hat{\lambda}=1.2$, e a hipótese nula corresponderá a que a variável de contagens X tenha distribuição Poisson (com esse valor do parâmetro), sendo a hipótese alternativa que X tenha outra qualquer distribuição. Ao abrigo desta hipótese nula, a contagem esperada (estimada) associada ao valor i ($i=0, 1, 2, ..., 9$) é dada por $\hat{E}_i = 100 \times p_i$, com $p_i = P[X = i] = \mathrm{e}^{-1.2}\,\frac{1.2^i}{i!}$. Estes valores são dados na tabela seguinte.

i	0	1	2	3	4	5	6	7	8	9
p_i	0.3012	0.3614	0.2169	0.0867	0.0260	0.0062	0.0012	0.0002	0.0000	0.0000
$\hat{E}_i$	30.1194	36.1433	21.6860	8.6744	2.6023	0.6246	0.1249	0.0214	0.0032	0.0004

Esta tabela merece dois comentários: (i) o valor da última classe deveria estar associado ao acontecimento "9 ou mais colónias", a fim de garantir que todos os possíveis resultados de X são contemplados; e (ii) sabemos que a validade da distribuição da estatística do teste χ^2 é apenas assintótica, considerando-se a aproximação aceitável quando se verificam as condições de Cochran. Assim, será necessário agrupar classes com $\hat{E}_i < 1$, e de forma a garantir que não mais de um quinto das classes remanescentes tenha valor esperado estimado inferior a 5. Esta dupla condição será garantida caso se agrupem as classes correspondentes a $X \geq 4$, resultando em cinco classes. Assim, a estatística de teste e respectiva distribuição assintótica sob H_0 é dada por:

$$\chi^2 \;=\; \sum_{i=0}^{4^+} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \;\sim\; \chi^2_{k-1-r}\ ,$$

onde $i = 4^+$ representa a classe de valores $X \geq 4$, $k = 5$ indica o número de classes após o agrupamento efectuado, e $r=1$ resulta da necessidade de subtrair um grau de liberdade por cada parâmetro estimado para a especificação completa da hipótese nula. Eis os valores esperados estimados, valores observados e parcelas da estatística do teste χ^2 resultantes:

i	0	1	2	3	4^+
$\hat{E}_i$	30.1194	36.1433	21.6860	8.6744	3.3769
O_i	27	50	11	6	6
$\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$	0.3231	5.3124	5.2656	0.8245	2.0376

O valor da estatística de teste é a soma das parcelas na última linha da tabela: $\chi^2_{calc}=13.7632$.

Este valor calculado da estatística tem de ser comparado com o valor fronteira duma região crítica unilateral direita, associada ao nível de significância pedido no enunciado, $\alpha=0.05$. Esse limiar é $\chi^2_{0.05(3)}=7.815$. Uma vez que $\chi^2_{calc}=13.7632>7.815=\chi^2_{0.05(3)}$, rejeita-se a hipótese nula e, uma vez que o parâmetro foi estimado a partir dos dados, sendo por isso o valor de λ mais favorável à hipótese de Poisson, podemos concluir que as contagens não seguem uma distribuição dessa natureza. Duas contagens ($i=1$ e $i=2$) contribuem com parcelas importantes para a rejeição desta hipótese nula, embora de formas diferentes: existem bastante mais quadrados observados na placa com uma única colónia, e bastante menos quadrados com duas colónias, do que seria de esperar ao abrigo da hipótese de Poisson. **Nota:** Este facto sugere a possibilidade de existir algum mecanismo de protecção das colónias de bacterias, visando dissuadir a presença na proximidade de colónias concorrentes.

Acrescente-se que se poderia objectar, em relação à experiência descrita no enunciado, que as contagens nas quais se baseia o teste não são, na realidade, contagens independentes (como deveriam ser para assegurar a validade do teste), uma vez que a contiguidade espacial dos quadrados permite supôr a existência de efeitos de interferência nas contagens. Da mesma forma, quadrados nas zonas periféricas da placa não estarão em igualdade de circunstâncias com quadrados nas zonas centrais da placa.

II

1. Neste ponto considera-se a regressão linear múltipla de `qres` sobre `zen`, `az` e `pos`.

   (a) O valor $R^2=0.7038$ significa que esta regressão linear múltipla explica mais de 70% da variabilidade das resoluções espaciais das imagens observadas, valor bastante aceitável.

   (b) Pede-se o valor ajustado da variável resposta `qres`, correspondente aos valores dos preditores indicados no enunciado. Ora a equação do modelo ajustado é $qres = -6.87563 + 6.78759\,zen - 0.12423\,az + 2.55470\,pos$. Substituindo os valores indicados, têm-se $\widehat{qres} = -6.87563 + 6.78759 \times 20 - 0.12423 \times (-100) + 2.55470 \times 60 = 294.5812\,m$. As unidades de medida (metros) são, naturalmente, as unidades de medida da variável resposta.

   (c) É pedido o valor de $s^2_y = \frac{SQT}{n-1}$. Sabe-se que $R^2_{mod} = 1 - \frac{QMRE}{QMT}$, sendo $QMT = s^2_y$. Pelo enunciado tem-se $\sqrt{QMRE} = 69.6$ e $R^2_{mod} = 0.7014$, pelo que $QMT = s^2_y = \frac{QMRE}{1-R^2_{mod}} = \frac{69.6^2}{1-0.7014} = 16222.9$.

   (d) Pede-se um teste F parcial para comparar este modelo completo de regressão múltipla com o submodelo de regressão linear simples, tendo por único preditor a variável explicativa mais correlacionada com a variável resposta `qres`. A partir da matriz de correlações dada no enunciado conclui-se que o melhor preditor é `zen`, e que o coeficiente de determinação na respectiva regressão linear simples será $R^2_s = 0.8173^2 = 0.6679793$. O teste F parcial pedido é assim:

Hipóteses: $H_0 : \mathcal{R}_c^2 = \mathcal{R}_s^2 \quad vs. \quad H_1 : \mathcal{R}_c^2 > \mathcal{R}_s^2$, onde $\mathcal{R}_c^2$ indica o coeficiente de determinação populacional do modelo completo e $\mathcal{R}_s^2$ quantidade idêntica para o submodelo.

Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \frac{R_c^2 - R_s^2}{1-R_c^2} \cap F_{(p-k,n-(p+1))}$, sob H_0, sendo $k=1$ o número de preditores do submodelo.

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 $F_{calc} > f_{0.05[2,368]} \approx 3.00$.

Conclusões: Tem-se $F_{calc} = \frac{368}{2} \frac{0.7038-0.6680}{1-0.7038} = 22.2390$. Logo rejeita-se tranquilamente H_0, i.e., considera-se que a qualidade de ajustamento do modelo completo é significativamente melhor (ao nível $\alpha = 0.05$) do que o da melhor regressão linear simples. Deve optar-se pelo modelo completo, apesar de ser menos parcimonioso.

(e) Designando a variável resposta `qres` por y e o preditor `zen` por x, tem-se $y = b_0 + b_1\, x$, com $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\, \frac{s_y}{s_x} = 0.8173 \times \sqrt{\frac{16223}{239.04292}} = 6.733$ e $b_0 = \overline{y} - b_1\, \overline{x} = 370.43 - 6.733 \times 30.379 = 165.88$.

(f) Pede-se um teste de hipóteses relativo à regressão linear simples, não se dispondo da listagem dos resultados do ajustamento de um tal modelo. No entanto, é possível responder, com base na informação disponível, já que sabemos que numa regressão linear simples, são equivalentes o teste t à hipótese de o declive β_1 da recta populacional ser nulo e o teste F de ajustamento global. Dispõe-se da informação necessária para testar as hipóteses indicadas no enunciado através dum teste de ajustamento global.

Hipóteses: $H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$.

Estatística do Teste: $F = (n-2) \frac{R^2}{1-R^2} \cap F_{(1,n-2)}$, 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 $F_{calc} > f_{0.05[1,370]} \approx 3.84$.

Conclusões: Como se viu acima, $R^2 = 0.6679793$. Logo, $F_{calc} = 370\, \frac{0.6679793}{1-0.6679793} = 744.3883$. Assim, o declive da recta de regressão é signficativamente diferente de zero, ao nível $\alpha = 0.05$ (e mesmo para níveis de significância muito inferiores, tendo em conta o valor muito elevado de F_{calc}).

2. Neste ponto considera-se o modelo cúbico para a variável resposta `qres`, usando como preditor a variável `zen`. A equação do modelo é $qres = \beta_0 + \beta_1\, zen + \beta_2\, zen^2 + \beta_3\, zen^3 + \epsilon$.

(a) Numa regressão linear, o coeficiente de determinação R^2 mede a proporção da variabilidade total da variável resposta que é explicada pelo modelo ajustado. Neste caso, existem dois modelos, *com a mesma variável resposta, ajustados com o mesmo conjunto de dados*. Assim, é perfeitamente legítimo comparar as proporções da variabilidade total das resoluções espaciais das imagens observadas que são explicadas por cada modelo. Concretamente, pode afirmar-se que modelo cúbico, isto é, polinomial de terceiro grau, na variável `zen`, explica uma proporção maior da variabilidade total (78.93%) do que o modelo de regressão linear múltipla inicial, com três preditores (que explica apenas 70.38% da variabilidade total das resoluções espaciais observadas). No entanto, não estamos perante um modelo e submodelo (uma vez que na equação de cada modelo há parcelas preditoras que não constam do outro modelo), razão pela qual não seria possível aplicar um teste F parcial para ver se as diferenças nos coeficientes de determinação se podem considerar significativas.

(b) i. Por definição, um resíduo (internamente) estandardizado é dado por $r_i = \frac{e_i}{\sqrt{QMRE\,(1-h_{ii})}}$ (ver formulário). No caso desta observação, são dados no enunciado os valores do resíduo

usual ($e_i = 230.9676$) e do efeito alavanca ($h_{ii} = 0.008574429$). Na listagem de resultados do ajustamento do modelo é ainda dado (sob a designação *Residual standard error*) o valor de $\sqrt{QMRE} = 58.71$. Substituindo na fórmula, tem-se $r_i = 3.951017 \approx 4$.

ii. Sabemos (ver formulário) que a distância de Cook duma observação pode ser calculada a partir do seu resíduo (internamente) estandardizado r_i e efeito alavanca h_{ii}, pela fórmula $D_i = r_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{1}{p+1}$. Tendo em conta os valores já referidos e o facto de existirem $p = 3$ parcelas de preditores, tem-se $D_i = 3.951017^2 \times \frac{0.008574429}{1-0.008574429} \frac{1}{4} = 0.03375227 \approx 0.035$.

(c) O valor do resíduo (internamente) estandardizado ($r_i \approx 4$) é bastante elevado, o que indica tratar-se duma observação relativamente afastada da curva polinomial de terceiro grau ajustada. Esse facto é visível na nuvem de pontos dada no enunciado onde (com o auxílio das coordenadas associadas ao ponto em cada eixo, dadas no enunciado) é possível identificar o ponto em questão como o ponto isolado mais acima, na quarta "coluna" a contar da direita. No entanto, não se trata duma observação influente no ajustamento do modelo, como se pode concluir pelo seu valor bastante baixo de distância de Cook ($D_i = 0.03375227$), muito abaixo do limite de guarda (0.5) e próximo de zero. Ou seja, a exclusão desta observação do conjunto de dados não conduziria a alterações sensíveis na curva ajustada. Também o valor do efeito alavanca associado a esta observação é modesto: $h_{ii} = 0.008574429$, um valor abaixo do valor alavanca médio, que sabemos ser $\overline{h} = \frac{p+1}{n} = \frac{4}{372} = 0.01075269$. Assim, esta observação não tem qualquer efeito importante de "atracção" da curva ajustada. Tratando-se duma observação atípica, não tem um grande impacto no ajustamento do modelo o que, em parte, reflecte também o elevado número de observações disponíveis. Esta alínea ilustra o facto de que os conceitos de resíduo elevado, efeito alavanca e influência, estando embora relacionados, não são equivalentes.

(d) Um modelo quadrático é um modelo polinomial de segundo grau, cuja equação é da forma $qres = \beta_0 + \beta_1\, zen + \beta_2\, zen^2 + \epsilon$. Trata-se dum submodelo do modelo cúbico originalmente ajustado, e os dois modelos são equivalentes caso se verifique a condição $\beta_3 = 0$ no modelo cúbico. Com base na informação disponibilizada, é possível testar esta hipótese.

Hipóteses: $H_0 : \beta_3 = 0 \quad vs. \quad H_1 : \beta_3 \neq 0.$

Estatística do Teste: $T = \frac{\hat{\beta}_3 - \beta_{3|H_0}}{\hat{\sigma}_{\hat{\beta}_3}} \cap t_{n-(p+1)}$, 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: (bilateral) Rejeitar H_0 se $|T_{calc}| > t_{0.025(368)} \approx 1.97$.

Conclusões: Tem-se, a partir do enunciado, $T_{calc} = 3.475$. Assim, rejeita-se a hipótese nula $\beta_3 = 0$, pelo que reduzir o grau do polinómio ajustado conduziria a resultados signficativamente piores. Esta conclusão poderia ser obtida de forma mais sintética, citando o valor de prova associado a estas hipóteses de teste ($p = 0.000572$), que é inferior a todos os níveis de significância usuais.

III

1. Trata-se dum delineamento factorial a dois factores, em que os $a = 6$ níveis do factor `ano` (factor A) são cruzados com os $b = 4$ níveis do factor `genotipo` (factor B), existindo observações em todas as $ab = 24$ situações experimentais resultantes.

   Inspeccionando a tabela-resumo do modelo ANOVA ajustado, rapidamente se conclui que foi utilizado um modelo ANOVA a dois factores, com efeitos de interacção. Na tabela existe uma

linha associada aos efeitos principais do factor A (`ano`), uma linha associada aos efeitos principais do factor B (`genotipo`), uma linha associada aos efeitos de interacção (`ano:genotipo`) e uma linha associada aos resíduos (`Residuals`). A partir dos graus de liberdade correspondentes a esta última linha, que sabemos serem, nos modelos lineares, sempre iguais ao número de observações (n) menos o número de parâmetros do modelo (ab, num modelo ANOVA a dois factores, com efeitos de interacção), tem-se $n-ab=192$, pelo que existem ao todo $n=192+24=216$ observações. Uma vez que o enunciado assegura tratar-se dum delineamento equilibrado, isto é, com um número comum n_c de observações nas $ab=24$ situações experimentais (células), conclui-se que em cada célula (combinação ano/genótipo) foram observadas $n_c=\frac{n}{ab}=\frac{216}{24}=9$ repetições.

2. Designando por Y_{ijk} o valor da variável resposta rendimento (`rend`), na k-ésima repetição ($k = 1, 2, ..., 9$) no ano i ($i = 1, ..., 6$), genótipo j ($j = 1, 2, 3, 4$), temos o seguinte modelo ANOVA para um delineamento factorial a dois factores, com efeitos de interacção:

   - $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall\, i, j, k$, sendo μ_{11} o rendimento esperado do primeiro genótipo (AI0122) no primeiro ano (1994); α_i o efeito principal (acréscimo ao rendimento) associado ao ano i (com a restrição $\alpha_1 = 0$); β_j o efeito principal (acréscimo ao rendimento) associado ao genótipo j (com a restrição $\beta_1 = 0$); $(\alpha\beta)_{ij}$ o efeito de interacção (acréscimo ao rendimento) associado à combinação do genótipo j com o ano i (e com as restrições $(\alpha\beta)_{ij}=0$ se $i=1$ ou $j=1$); e finalmente ϵ_{ijk} o erro aleatório da referida observação.
   - Admite-se que os erros aleatórios são todos 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.

3. Existem três tipos de efeitos no modelo: os efeitos principais do factor A (`ano`), α_i; os efeitos principais do factor B (`genotipo`), β_j; e os efeitos de interacção ano-genótipo, $(\alpha\beta)_{ij}$. Para cada tipo de efeito existe um teste F, cuja hipótese nula consiste na inexistência do referido tipo de efeito e a respectiva hipótese alternativa consiste na existência, em pelo menos um caso, do referido tipo de efeito. Eis em pormenor o teste F para os efeitos de interacção.

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

   Estatística do Teste: $F_{AB} = \frac{QMAB}{QMRE} \cap F_{[(a-1)(b-1),n-ab]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(15,192)} \approx 1.7$ (entre os valores tabelados 1.75 e 1.67).

   Conclusões: Como $F_{calc} = 0.759 < 1.7$, não se rejeita H_0, não sendo assim possível concluir pela existência de efeitos de interacção (ao nível $\alpha = 0.05$ e, pela análise do *p-value* $p = 0.72116$, também para todos os níveis usuais de α).

   NOTA: Esta conclusão não é surpreendente, uma vez que os genótipos estudados já tinham sido objecto duma análise prévia que conduziu à sua selecção precisamente por não terem revelado aquilo a que se chama interacção genótipo-ambiente, ou seja, revelaram rendimentos estáveis em diferentes meios ambiente.

   Uma inspecção rápida permite concluir que a situação será diferente no que concerne aos outros dois testes F. Assim, num teste à existência de efeitos principais do factor `ano`, em que as hipóteses em confronto são $H_0 : \alpha_i = 0, \forall\, i$ vs. $H_1 : \exists\, i$ tal que $\alpha_i \neq 0$, o valor de prova, resultante da estatística de teste $F = \frac{QMA}{QMRE} \cap F_{[a-1,n-ab]}$, sob H_0, é dado no enunciado com sendo

indistinguível de zero ($p < 21 \times 10^{-16}$), o que indica uma rejeição claríssima da hipótese nula. Analogamente, no teste aos efeitos principais do factor genótipo, em que as hipóteses em confronto são $H_0 : \beta_j = 0, \forall\, j$ vs. $H_1 : \exists\, j$ tal que $\beta_j \neq 0$, o valor de prova resultante da estatística de teste $F = \frac{QMB}{QMRE} \cap F_{[b-1,n-ab]}$, sob H_0, é $p = 0.00301$. Este *p-value* conduziria à rejeição de H_0 a níveis de significância como $\alpha = 0.05$ ou $\alpha = 0.01$, embora para um nível $\alpha = 0.001$ já a conclusão teria de ser não rejeição de H_0. Assim, é legítimo admitir a existência de efeitos principais de genótipo, embora estes não sejam tão claros como os efeitos do factor ano.

4. Trata-se dum gráfico de interacção, cuja natureza foi estudada nas aulas. A ausência de efeitos de interacção, acima discutida, é aqui reflectida no facto das curvas seccionalmente lineares associadas a cada nível do factor `ano` serem aproximadamente paralelos. Deste gráfico ressalta também que os efeitos principais de ano, cuja existência foi claramente indicada pelo respectivo teste F, resultam, em boa medida, dos rendimentos do ano 1999, que foram sistematicamente maiores que os restantes e, embora em menor medida, dos rendimentos do ano 1994, que foram sistematicamente menores. Esta realidade é também visível na tabela de médias do enunciado. Estes dois anos são os principais responsáveis pela conclusão de que nem todos os efeitos α_i serão nulos. A existência de efeitos de genótipo (que não era tão clara no teste F) é mais difícil de visualizar neste gráfico.

5. Caso se tivesse ignorado a existência do factor `ano` e estudado os dados através dum modelo ANOVA com um único factor (`genotipo`) a tabela-resumo teria de ser recalculada. Sabemos que numa tal tabela apenas existiram duas linhas: a do factor ano e a residual. Sabemos também que a Soma de Quadrados do Factor teria igual fórmula (logo igual valor) ao SQB do modelo para o delineamento factorial equilibrado a dois factores, e que também os graus de liberdade e o Quadrado Médio correspondente seriam iguais. Assim, $SQF = SQB = 36.2$, com $b-1 = 3$ graus de liberdade associados, pelo que $QMF = 12.1$. Por outro lado, as três Somas de Quadrados restantes, que constam da tabela-resumo no enunciado (isto é, SQA, SQAB e SQRE dos resíduos do modelo a dois factores, com interacção) terão de somar o valor de SQRE no modelo a um único factor. De facto, a soma de todas as SQs da tabela tem de ser igual a SQT, valor esse que não depende do modelo ajustado e, como tal, é igual nos dois casos. Indicando a Soma de Quadrados Residual do modelo apenas com o factor `genotipo` por $SQRE_g$, tem-se $SQRE_g = SQA + SQAB + SQRE = 1661.1 + 28.7 + 483.2 = 2173$. Os graus de liberdade associados a esta Soma de Quadrados são, como sempre, o número total de observações menos o número total de parâmetros do modelo que, neste caso do modelo apenas com o factor `genótipo` significa $n–b = 216–4 = 212$ (alternativamente, pode pensar-se que é a soma dos graus de liberdade correspondentes às somas de quadrado acima indicadas: $g.l.(SQRE_g) = 5+15+192 = 212$). Estes valores permitem construir o novo Quadrado Médio Residual, que será $QMRE_g = \frac{SQRE_g}{n-b} = \frac{2173}{212} = 10.25$. A estatística F do único teste F sobrante neste novo contexto (o teste aos efeitos de genótipo) será agora $F_{calc} = \frac{QMF}{QMRE} = \frac{12.1}{10.25} = 1.1805$. Assim, a tabela-resumo deste novo modelo a um único factor será:

Fonte de variação	g.l.	SQs	QMs	F_{calc}
Factor (genótipo)	3	36.2	12.1	1.1805
Residual	212	2173	10.25	–

O teste à existência de efeitos do factor terá agora como limiar (ao nível de significância $\alpha = 0.05$) o valor $f_{0.05(3,212)} \approx 2.65$. A conclusão é agora de não rejeição de H_0, ou seja uma conclusão diferente (para o mesmo α) do que no caso do modelo a dois factores com interacção. Esta diferença de conclusões resulta do facto de, para igual numerador na estatística F, o denominador

($QMRE$) ser agora consideravelmente maior (10.25 em vez de 2.5), o que por sua vez resulta do facto de toda a variabilidade associada aos efeitos agora retirados do modelo, e sobretudo a considerável variabilidade associada ao factor `ano`, ter ido parar à variabilidade residual (não explicada pelo modelo, que agora já não prevê efeitos de ano). Esta diminuição no valor de F_{calc} de 4.800 para 1.1805 significou (mesmo tendo em conta os diferentes graus de liberdade associados à estatística) que F_{calc} deixou de pertencer à região crítica do teste. Esta situação ilustra a ideia geral de que fontes de variabilidade nos valores da variável resposta que não estejam previstos no modelo tendem a inflacionar o Quadrado Médio Residual, logo a diminuir o valor de F_{calc}, podendo desse modo mascarar a existência de reais efeitos de outro tipo, previstos no modelo.

IV

1. O modelo de regressão linear múltipla relaciona uma variável resposta Y com p variáveis preditoras X_1, X_2, ..., X_p. Designando por $\mathbf{Y}$ o vector das n observações da variável resposta Y, $\boldsymbol{\varepsilon}$ o vector dos n erros aleatórios correspondentes, $\boldsymbol{\beta}$ o vector dos $p+1$ parâmetros do modelo, $\beta_0, \beta_1, ..., \beta_p$, e $\mathbf{X}$ a matriz $n \times (p+1)$, cuja primeira coluna é constituída por n uns e cada uma das restantes p colunas contém as n observações duma variável preditora, tem-se (ver também o acetato 222 das aulas teóricas):

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

   O modelo de regressão linear mútlipla é então dado por duas condições (acetato 232):

   (a) $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

   (b) $\boldsymbol{\varepsilon} \ \cap \ \mathcal{N}_n(\mathbf{0}, \ \sigma^2 \mathbf{I}_n)$,

   sendo $\mathbf{0}$ o vector de n zeros e $\mathbf{I}_n$ a matriz identidade $n \times n$. Na segunda condição, indica-se que o vector dos erros aleatórios segue uma distribuição Multinormal, com vector médio dado pelo vector de zeros (ou seja, cada erro aleatório individual tem valor esperado zero) e matriz de variâncias-covariâncias diagonal, com os elementos diagonais todos iguais a σ^2. Uma vez que, numa matriz de (co-)variâncias os elementos diagonais representam as variâncias de cada componente do vector, esta condição indica que $V[\epsilon_i] = \sigma^2$, $\forall\, i$. O facto de os elementos não diagonais da matriz $\sigma^2\, \mathbf{I}_n$ serem todos nulos equivale a dizer que a covariância entre elementos diferentes do vector aleatório dos erros é sempre nula (ou seja, $Cov[\epsilon_i, \epsilon_j] = 0$, sempre que $i \neq j$) e, como sabemos, numa distribuição Multinormal, tal facto implica a independência desses elementos.

2. A equação do modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ significa que $\mathbf{Y}$ é a soma dum vector aleatório ($\boldsymbol{\varepsilon}$) mais um vector não aleatório ($\mathbf{X}\boldsymbol{\beta}$). Ora, pela Propriedade 5 da Multinormal (acetato 231), a soma dum vector aleatório Multinormal (neste caso, $\boldsymbol{\varepsilon}$) com um vector não aleatório (neste caso, $\mathbf{X}\boldsymbol{\beta}$) preserva a Multinormalidade. Para obter os parâmetros da distribuição de $\mathbf{Y}$ basta recordar que esses parâmetros são, respectivamente, o vector esperado e a matriz de variâncias-covariâncias do vector e aplicar as propriedades operatórias dadas nos acetatos 226 e 227. Assim (e como para qualquer vector aleatório $\mathbf{W}$ e vector não aleatório $\mathbf{a}$ se verifica $E[\mathbf{W} + \mathbf{a}] = E[\mathbf{W}] + \mathbf{a}$),

tem-se: $E[\mathbf{Y}] = E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta} + \underbrace{E[\boldsymbol{\varepsilon}]}_{=\mathbf{0}} = \mathbf{X}\boldsymbol{\beta}$, e este é o primeiro parâmetro da distribuição Multinormal de $\mathbf{Y}$. Por outro lado, e tendo em conta que $V[\mathbf{W} + \mathbf{a}] = V[\mathbf{W}]$, tem-se $V[\mathbf{Y}] = V[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = V[\boldsymbol{\varepsilon}] = \sigma^2\,\mathbf{I}_n$ por hipótese do modelo. Logo, e como pedido no enunciado, tem-se $\mathbf{Y} \cap \mathcal{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\,\mathbf{I}_n\right)$. As propriedades acima referidas estão também sintetizadas no formulário disponível no exame.

3. Sabemos que o vector de estimadores dos parâmetros, $\hat{\boldsymbol{\beta}}$, é dado por: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\mathbf{Y}$, pelo que $\hat{\boldsymbol{\beta}}$ resulta de pré-multiplicar o vector aleatório $\mathbf{Y}$ pela matriz (não aleatória) $\mathbf{B} = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t$. Tendo em conta as propriedades de vectores esperanças e matrizes de (co-)variâncias deste tipo de produtos (ver Acetatos 226 e 227 das teóricas, e também o formulário: $E[\mathbf{BW}] = \mathbf{B}E[\mathbf{W}]$ e $V[\mathbf{BW}] = \mathbf{B}V[\mathbf{W}]\mathbf{B}^t$), tem-se:

   (a) $E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\mathbf{Y}] = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\,E[\mathbf{Y}] = \cancel{(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\mathbf{X}}\boldsymbol{\beta} = \boldsymbol{\beta}$.

   (b) Em relação à matriz de (co-)variâncias tem-se:

$$\begin{aligned} V[\hat{\boldsymbol{\beta}}] &= V[(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\mathbf{Y}] = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\,V[\mathbf{Y}]\,\left[(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\right]^t \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\,\sigma^2\mathbf{I}_n\,\mathbf{X}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]^t = \sigma^2\cancel{(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\mathbf{X}}\left[(\mathbf{X}^t\mathbf{X})^t\right]^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\,. \end{aligned}$$