INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

13 Janeiro 2020 **Primeira Chamada de Exame 2019-20 Uma resolução possível**

I

Tem-se uma tabela de contagens com $k = 15$ categorias, com $N = 348$ observações duma variável aleatória X que conta o número de dentes por bolbo de alho.

1. É pedido para testar o ajustamento duma distribuição de Poisson com parâmetro $\lambda{=}11$. Usa-se o teste baseado na estatística de Pearson.

   Hipóteses: $H_0 : X \frown Po(\lambda{=}11)$ *vs.* $H_1 : X \not\frown Po(\lambda{=}11)$.

   Estatística do Teste: A estatística de Pearson é dada por $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. A distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_{k-1}, uma vez que não é necessário estimar o parâmetro da distribuição. Tem-se $k{-}1{=}14$ graus de liberdade..

   Nível de Significância Pode-se escolher $\alpha = \text{P[Erro de tipo I]} = \text{P[Rejeitar } H_0 | H_0 \text{ verdade]} = 0.05$.

   Região Crítica: Unilateral direita. Para um nível de significância $\alpha{=}0.05$, a regra de rejeição consiste em rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{0.05(14)} {=} 23.6848$.

   Conclusões O valor calculado da estatística de teste é dado no enunciado: $X^2_{calc} = 59.454$. Assim, rejeita-se H_0, concluindo-se (para $\alpha{=}0.05$) que a distribuição do número de dentes por bolbo não segue a distribuição de Poisson com parâmetro $\lambda = 11$. Esta conclusão *não* exclui a possibilidade que a distribuição de X seja uma Poisson com outro valor do parâmetro λ.

2. O Critério de Cochran permite considerar aceitável a aproximação assintótica da distribuição da estatística de Pearson quando nenhum dos valores esperados E_i fôr inferior a 1, e não mais de 20% forem inferiores a 5. Sendo os valores esperados dados por $E_i {=} N \times \pi_i$ (onde π_i indica a probabilidade associada à categoria i, caso seja verdade H_0), os menores valores esperados correspondem às categorias de menor probabilidade π_i, ou seja às categorias correspondentes a valores mais extremos de X. Uma vez que há duas categorias de valores agrupados ($X \leq 5$ e $X \geq 19$), pode dar-se o caso de as menores probabilidades estarem associadas aos valores individuais seguintes ($X = 6$ ou $X = 18$). Assim, é mais prático verificar quais são as probabilidades associadas aos valores esperados 1 e 5. Tem-se, como probabilidade associada ao valor esperado 1, $\pi = \frac{1}{N} = 0.002873563$. Uma consulta à tabela das distribuições de Poisson com parâmetro 11 permite verificar que $\pi_{\leq 5} = 0.038$; $\pi_{\geq 19} = 1 - \pi_{\leq 18} = 1 - 0.982 = 0.018$; $\pi_6 = \pi_{\leq 6} - \pi_{\leq 5} = 0.079 - 0.038 {=} 0.041$; $\pi_{18} {=} \pi_{\leq 18} - \pi_{\leq 17} {=} 0.982 - 0.968 {=} 0.014$. Assim, em nenhuma dessas (ou das restantes, com maiores valores de π_i) categorias o valor esperado será inferior a 1. Categorias com valor esperado inferior a 5 terão de ter probabilidade inferior a $\pi = \frac{5}{N} = 0.01436782$. As contas acima feitas permitem verificar que apenas uma das categorias ($X{=}18$) pode não verificar esta condição (depende do arredondamento que conduziu ao valor $\pi_{18} = 0.014$). Na categoria imediatamente a seguir ($X = 17$) tem-se $\pi_{17} {=} \pi_{\leq 17} - \pi_{\leq 16} {=} 0.968 - 0.944 {=} 0.024$, pelo que o valor esperado excederá seguramente 5. Havendo uma única categoria em 14 em que pode haver $E_i < 5$, estamos seguramente dentro das condiçoes exigidas pelo Critério de Cochran.

3. Por comodidade, vamos designar a categoria $X \leq 5$ como tendo valor observado $O_{\leq 5} = 7$ e valor esperado $E_{\leq 5} = N \times \pi_{\leq 5} = 348 \times 0.038 = 13.224$. A parcela da estatística correspondente a esta categoria é dada por $\frac{(E_{\leq 5} - O_{\leq 5})^2}{E_{\leq 5}} = 2.929$. Não se trata dum valor particularmente destacado, ou seja, contribui relativamente pouco para o valor observado da estatística, $X^2_{calc} = 59.454$.

II

1. Considera-se primeiro uma modelação de `Evap` sobre `Vvento`.

   (a) A nuvem de pontos revela tendência para uma relação curvilínea entre as variáveis e, ao mesmo tempo, alguma tendência para uma forma afunilada. A primeira destas características aconselha uma transformação de variáveis para *linearizar a relação* entre o preditor e a variável resposta, permitindo a utilização duma regressão linear simples. A segunda característica aponta para a violação do pressuposto de homogeneidade de variâncias dos erros aleatórios, que faz parte do modelo linear. Também aqui, uma transformação de variáveis pode contribuir para *estabilizar as variâncias dos erros aleatórios.*

   (b) A regressão linear foi ajustada apos uma logaritmização da variável resposta, mas não da variável preditora. Uma tal transformação é adequada no caso duma *relação exponencial* entre as variáveis originais (como se discutirá adiante).

      i. Neste gráfico de resíduos usuais E_i *vs.* valores ajustados $\hat{Y}_i$, o facto de os pontos estarem dispostos essencialmente em torno duma banda horizontal (que necessariamente contém o valor zero no eixo vertical) é coerente com a hipótese de linearidade entre variável resposta e preditor, bem como com a hipótese de variâncias homogéneas. Assim, a logaritmização de `Evap` parece ter simultaneamente linearizado a tendência de fundo e estabilizado as variâncias dos erros aleatórios.

      ii. Há dois aspectos a referir. Por um lado, o Coeficiente de Determinação $R^2 = 0.5716$ indica que a relação linear entre `log(Evap)` e `Vvento` explica $57,16\%$ da variância observada nos logaritmos das evaporações. Não é um valor muito elevado, correspondendo a um ajustamento relativamente modesto. No entanto, é um valor suficientemente elevado para ser considerado significativamente diferente de zero, como se constata ao efectuar o 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: $F = \frac{QMR}{QMRE} = (n-2)\frac{R^2}{1-R^2} \frown F_{(1,n-2)}$, sob H_0.

      Nível de significância: $\alpha = 0.05$.

      Região Crítica: Unilateral direita. Rejeitar H_0 se $F_{calc} > f_{0.05[1,53]}$. Este limiar está entre os valores tabelados $f_{0.05[1,40]} = 4.08$ e $f_{0.05[1,60]} = 4.00$. No que se segue utilizaremos o valor aproximado $f_{0.05[1,53]} \approx 4.03$.

      Conclusões: No enunciado não é dado o valor calculado da estatística, mas como vimos, $F_{calc} = (n-2)\frac{R^2}{1-R^2} = 53 \times \frac{0.5716}{1-0.5716} = 70.71615 \gg 4.03$. A rejeição de H_0 é muito clara, pelo que o modelo ajustado é muito significativamente diferente do Modelo Nulo (o correspondente valor de prova é $p = 2.503 \times 10^{-11}$).

      iii. A variável resposta é `ln(Evap)` e será indicada como Y^* (reservando Y para `Evap`). Por definição, $e_{51} = y^*_{51} - \hat{y}^*_{51} = y^*_{51} - (b_0 + b_1\, x_{51})$. O enunciado indica que $e_{51} = 2.01484$; $b_0 = -1.46881$; $b_1 = 0.63961$; e $x_{51} = 2.5175$. Substituindo, obtém-se $\hat{y}^*_{51} = 2.01484 + (-1.46881 + 0.63961 \times 2.5175) = 2.156248$. No entanto, deve recordar-se que este valor

está na escala logarítmica, pelo que o valor correspondente de `Evap` (a taxa máxima de evaporação) é $y_{51} = \mathrm{e}^{2.156248} = 8.638666$ mm h^{-1}.

iv. A recta ajustada tem equação $y^* = \ln(y) = b_0 + b_1\, x$. Exponenciando, obtém-se $y = \mathrm{e}^{b_0 + b_1\, x} = \mathrm{e}^{b_0}\mathrm{e}^{b_1\, x}$. Usando os valores de b_0 e b_1 constantes do enunciado, tem-se a curva exponencial de equação $y = 0.2302\ \mathrm{e}^{0.63961\, x}$. Viu-se nas aulas que as curvas exponenciais resultam de considerar que y é função de x (no nosso caso, a taxa máxima de evaporação, `Evap`, como função da velocidade do vento, `Vvento`) e admitir que a taxa de variação relativa de y (ou seja, quociente $\frac{y'(x)}{y(x)}$) é constante, sendo esse valor constante igual a β_1, o valor que, após a logaritmização, corresponde ao declive da recta de regressão. No nosso caso, esse valor é estimado por $b_1 = 0.63961$.

2. Considera-se a regressão linear múltipla de `ln(Stemflow)` sobre as 9 restantes variáveis, sendo o preditor `Pg` também logaritmizado. Os resultados indicam uma regressão bastante satisfatória, que explica cerca de $84,29\%$ da variância dos valores observados dos logaritmos do escorrimento.

   (a) É pedido um intervalo a $(1-\alpha)\times 100\% = 95\%$ de confiança para β_9, o coeficiente da última variável preditora (`Area.Copa`). Sabemos que este intervalo de confiança tem como ponto central o valor estimado $b_9 = -0.38015$ e semi-amplitude dada pelo produto do quantil de ordem $1-\frac{\alpha}{2} = 0.975$, numa distribuição $t_{n-(p+1)}$, ou seja, $t_{0.025(45)} \approx 2.015$; e do erro padrão do estimador de β_9, ou seja, $\hat{\sigma}_{\hat{\beta}_9} = 0.27904$ (dado no enunciado). Assim, tem-se 95% de confiança na veracidade da afirmação de que, na população, a variação esperada no log-escorrimento, por cada m^2 adicional na área da copa da árvore (e mantendo fixas as restantes variáveis preditoras) é um dos valores do seguinte intervalo:

   $$\left]\ b_9 - t_{0.025(45)} \times \hat{\sigma}_{\hat{\beta}_9}\ ,\ b_9 + t_{0.025(45)} \times \hat{\sigma}_{\hat{\beta}_9}\ \right[\ = \ ]\ -0.9424156\ ,\ 0.1821156\ [\ .$$

   Uma vez que o intervalo contém o valor zero, não se pode rejeitar que $\beta_9 = 0$. A ser esse o verdadeiro valor de β_9, a variável preditora `Area.Copa` não contribuiria em nada para a modelação dos valores de log-escorrimento. Assim, a variável `Area.Copa` é dispensável do modelo, sem perda significativa na qualidade do ajustamento.

   (b) O aumento médio no log-escorrimento (variável resposta) associado a aumentar uma unidade a sexta variável preditora (`alt.arv`), mantendo o resto igual, é dado pelo coeficiente β_6. O correspondente aumento médio no log-escorrimento associado a aumentar uma unidade o séptimo preditor (`alt.tronco`), mantendo o resto igual, é dado pelo coeficiente β_7. O enunciado pergunta se é possível rejeitar a igualdade destes dois coeficientes, o que pode ser feito através dum teste bilateral a $H_0 : \beta_6 - \beta_7 = 0$ ($\beta_6 - \beta_7$ é uma combinação linear dos parâmetros).

   Hipóteses: $H_0 : \beta_6 - \beta_7 = 0$ vs. $H_1 : \beta_6 - \beta_7 \neq 0$

   Estatística do teste $T = \frac{(\hat{\beta}_6 - \hat{\beta}_7) - (\beta_6 - \beta_7)|_{H_0}}{\hat{\sigma}_{\hat{\beta}_6 - \hat{\beta}_7}} \ \frown\ t_{n-(p+1)}$, sob H_0.

   Nível de significância: $\alpha = 0.05$

   Região Crítica: Bilateral. Rejeita-se H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(n-(p+1)} = t_{0.025(45)} \approx 2.015$.

   Conclusões: No nosso caso, o numerador da estatística calculada reduz-se a $b_6 - b_7 = 2.51834 - 0.94257 = 1.57577$. No que respeita ao denominador, e tratando-se do erro padrão duma diferença de dois estimadores, podemos começar por calcular a variância estimada dessa diferença:

   $$\widehat{V[\hat{\beta}_6 - \hat{\beta}_7]} = \widehat{V[\hat{\beta}_6]} + \widehat{V[\hat{\beta}_7]} - 2\,\widehat{Cov[\hat{\beta}_6, \hat{\beta}_7]} = \hat{\sigma}^2_{\hat{\beta}_6} + \hat{\sigma}^2_{\hat{\beta}_7} - 2\,\widehat{Cov[\hat{\beta}_6, \hat{\beta}_7]}\ .$$

   Os dois primeiros valores são os quadrados dos erros padrões de $\hat{\beta}_6$ e $\hat{\beta}_7$, que constam da listagem produzida aquando do ajustamento do modelo. A terceira e última parcela

é dada no enunciado. Logo, $\widehat{V[\hat{\beta}_6 - \hat{\beta}_7]} = 1.22131^2 + 2.02048^2 - 2 \times (-1.537745) = 8.649428$. Logo, a estatística calculada tem valor $T_{calc} = \frac{1.57577}{\sqrt{8.649428}} = 0.535796$. Assim, não se rejeita H_0, sendo de admitir a igualdade de β_6 e β_7.

(c) O primeiro passo num algoritmo de exclusão sequencial, baseado em testes *t-Student* bilaterais às hipóteses $H_0 : \beta_j = 0$, pode ser efectuado de forma rápida tendo em conta que os valores de prova (*p-values*) correspondentes a esses testes constam do enunciado. Utilizando o nível de significância $\alpha = 0.05$, todas as variáveis para as quais se verifique $p > 0.05$ são candidatas à exclusão do modelo (não se rejeitando H_0, admite-se que $\beta_j = 0$). Verifica-se que existem quatro preditores nessas condições: `int`, `dur`, `alt.tronco` e `Area.Copa`. Mas apenas há a garantia de que a exclusão de *um* único preditor não afecte significativamente a sua qualidade. A escolha recairá sobre a exclusão do preditor com o maior *p-value* associado, ou seja, a intensidade média da precipitação (`int`, cujo *p-value* é $p = 0.866430$). É possível calcular o valor de R^2 no submodelo resultante da exclusão do preditor `int`, uma vez que sabemos que o quadrado do valor calculado da estatística t associado ao teste a $H_0 : \beta_2 = 0$ (ou seja, o quadrado de $T_{calc} = -0.169$) é o valor calculado da estatística do teste F parcial correspondente à comparação do modelo completo com o submodelo sem esse preditor `int`. Usando a expressão da estatística do teste F parcial, tem-se:

$$\begin{aligned} (-0.169)^2 = F_{calc} &= \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1 - R_c^2} = \frac{45}{1} \times \frac{0.8429 - R_s^2}{1 - 0.8429} \\ \Leftrightarrow \qquad R_s^2 &= 0.8429 - \frac{0.028561}{45} \times 0.1571 = 0.8428 \, . \end{aligned}$$

Assim, a exclusão do preditor `int` apenas reduz a percentagem da variância do log-escorrimento que é explicada pelo modelo em 0.01%. Repare-se que este submodelo tem de ser o submodelo de $p-1 = 8$ preditores com maior valor de R_s^2. De facto, na expressão da estatística F (ver acima), a única quantidade que varia entre submodelos com $k = 8$ preditores é o valor de R_s^2. O submodelo de $k = 8$ preditores com o menor valor de F_{calc} será o submodelo cujo R_s^2 está mais perto de R_c^2 (ou seja, cuja diferença $R_c^2 - R_s^2$ é menor).

(d) Procedeu-se a ajustar o submodelo resultante de excluir os quatro preditores cujo *p-value* nos testes bilaterais a $H_0 : \beta_j = 0$ excedia $p = 0.05$ (ou seja, um submodelo com apenas $k = 5$ preditores). Como é dito no enunciado, o Coeficiente de Determinação resultante foi $R_s^2 = 0.8127$. É pedido um teste F parcial para verificar se se deve considerar que este submodelo tem uma qualidade de ajustamento significativamente mais baixa que a do modelo completo original.

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

Estatística do teste $F = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1 - R_c^2} \overset{\cdot}{\sim} F_{[p-k, n-(p+1)]}$, se H_0 verdade.

Nível de significância: $\alpha = 0.05$

Região Crítica: Unilateral direita. Rejeita-se H_0 se $F_{calc} > f_{\alpha[p-k, n-(p+1)]} = f_{0.05(4,45)} \approx 2.59$.

Conclusões: Tem-se $F_{calc} = \frac{45}{4} \times \frac{0.8429 - 0.8127}{1 - 0.8429} = 2.1626$. Assim, não se rejeita H_0. Não se pode concluir que os dois modelos tenham qualidade de ajustamento significativamente diferente.

Nota: *Não* se podia garantir isto a partir do facto de, na alínea anterior, os *p-values* individuais de cada um dos preditores excluídos serem superiores a 0.05.

(e) Por definição, o Coeficiente de Determinação usual é dado por $R^2 = \frac{SQR}{SQT} = \frac{SQT - SQRE}{SQT} = 1 - \frac{SQRE}{SQT}$. No formulário pode verificar-se que o Coeficiente de Determinação modificado é dado por $R_{mod}^2 = 1 - \frac{QMRE}{QMT}$. Uma vez que se trata dum submodelo com k preditores,

tem-se: $QMRE = \frac{SQRE}{n-(k+1)}$ e $QMT = \frac{SQT}{n-1}$. Logo, tem-se:

$$R^2_{mod} = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SQRE}{SQT} = 1 - \frac{n-1}{n-(k+1)} \cdot (1 - R^2) \ .$$

Sabemos que, neste submodelo, $R^2_s = 0.8127$, $k = 5$ e $n = 55$. Logo, $R^2_{mod} = 1 - \frac{54}{49} \times (1 - 0.8127) = 0.7936$. O facto de a diferença entre os valores das duas variantes do Coeficiente de Determinação ser menor no submodelo (cerca de 0.02) do que no modelo completo (cerca de 0.03) reflecte o facto de um valor semelhante do Coeficiente de Determinação usual no submodelo ser obtido com cerca de metade dos preditores.

III

1. A variável resposta (numérica) é o rendimento (em kg/planta). Há dois factores explicativos: `ambiente` (factor A, com $a = 4$ níveis) e `clone` (factor B, com $b = 6$ níveis). Havendo rendimentos observados para as $ab = 24$ possíveis combinações de ambientes e genótipos (clones), tem-se um *delineamento factorial*. Em cada uma das $ab = 24$ situações experimentais (associações ambiente/genótipo) existe um número idêntico de $n_c = 9$ repetições, pelo que se trata dum *delineamento equilibrado*, com um total de $n = ab\, n_c = 24 \times 9 = 216$ observações de rendimento. O facto de haver repetições nas células (situações experimentais) significa que é possível ajustar um modelo com efeitos de interacção, ou seja, que se vai considerar o seguinte modelo ANOVA:

   Equação do Modelo: $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$,
   com as restrições $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$, e onde:

   - $i = 1, 2, 3, 4$ indica o nível do Factor `ambiente` a que corresponde uma observação;
   - $j = 1, 2, 3, 4, 5, 6$ indica o genótipo (factor `clone`) a que corrersponde uma observação;
   - $k = 1, 2, ..., 9$ indica a repetição no seio da célula (i, j);
   - Y_{ijk} indica o rendimento da k-ésima repetição na célula (i, j), sendo ϵ_{ijk} o correspondente erro aleatório;
   - μ_{11} representa a média populacional na situação experimental $(1, 1)$, ou seja, do primeiro genótipo no primeiro ambiente;
   - α_i indica o efeito principal associado ao ambiente i;
   - β_j indica o efeito principal associado ao genótipo j;
   - $(\alpha\beta)_{ij}$ indica o efeito de interacção na célula (i, j).

   Distribuição dos erros: $\epsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.

   Independência dos erros: $\{\epsilon_{ijk}\}_{i,j,k}$ são variáveis aleatórias independentes.

2. Havendo três tipos de efeitos (principais do factor ambiente, principais do factor clone e de interacção) o quadro de síntese terá quatro linhas (uma para cada tipo de efeito, e ainda a linha associada à variabilidade residual), sem contar com a linha correspondente à variabilidade Total. Nesta tabela, há três valores directamente referidos no enunciado: o Quadrado Médio Residual, QMRE=1.8728, SQB=53.5437 e $F^A_{calc} = \frac{QMA}{QMRE} = 247.150$. Assim, tem-se $QMA = QMRE \times F^A_{calc} = 1.8728 \times 247.150 = 462.8625$. Os graus de liberdade são dados por: $a-1 = 3$ para o factor A; $b-1 = 5$ para o factor B; $(a-1)(b-1) = 15$ para os efeitos de interacção; e $n–ab = 216-24 = 192$ para a variabilidade residual. Logo, o Quadrado Médio associado ao Factor B tem valor $QMB = \frac{SQB}{b-1} = 10.70874$, de onde decorre que o valor calculado da estatística F^B para o teste à existência de efeitos principais do Factor B tem valor $F^B_{calc} = \frac{QMB}{QMRE} = 5.718037$.

A Soma de Quadrados Residual tem valor $SQRE = QMRE \times (n - ab) = 1.8728 \times 192 = 359.5776$ e a Soma de Quadrados associada ao Factor A tem valor $SQA = QMA \times (a - 1) = 462.8625 \times 3 = 1388.588$. Falta apenas calcular as três quantidades em falta na linha de valores correspondentes aos efeitos de interacção. Pode calcular-se a Soma de Quadrados associada à interacção a partir das fórmulas $SQT = (n-1)\, s_y^2 = 215 \times 8.756217 = 1882.587$ e $SQAB = SQT - (SQA + SQB + SQRE) = 1882.587 - (1388.588 + 53.5437 + 359.5776) = 80.8777$. Deste valor decorre que o respectivo Quadrado Médio é $QMAB = \frac{SQAB}{(a-1)(b-1)} = 5.391847$ e, finalmente, $F_{calc}^{AB} = \frac{QMAB}{QMRE} = 2.87903$. Juntando todos estes valores tem-se a seguinte tabela:

Fontes de Variação	gl	Somas de Quadrados	Quadrados Médios	F_{calc}
Factor Ambiente	3	1388.588	462.8625	247.150
Factor Genótipo	5	53.5437	10.70874	5.718037
Interacção	15	80.8777	5.391847	2.87903
Residual	192	359.5776	1.8728	—

3. Neste modelo há três testes F de interesse, um para cada tipo de efeito. Nos três casos, as hipóteses nulas correspondem à inexistência do tipo de efeitos referido (e a Hipótese alternativa à sua negação), e nos três casos a estatística do teste corresponde à razão entre o Quadrado Médio do tipo de efeito a ser testado e o Quadrado Médio Residual. As distribuições das estatísticas, sob as respectivas Hipóteses Nulas, são sempre F, com graus de liberdade associados aos Quadrados Médios que definem a estatística. Vejamos em pormenor o teste aos 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: $F^{AB} = \frac{QMAB}{QMRE} \frown F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(15,192)} \approx 1.75$.

   Conclusões: Como $F_{calc}^{AB} = 2.87903 > 1.75$, rejeita-se H_0. Pode concluir-se que existem efeitos de interacção significativos,ao nível $\alpha = 0.05$.

   No teste aos efeitos principais de ambiente, a Hipótese Nula $H_0 : \alpha_i = 0$, para todos os ambientes i, é rejeitada de forma clara. De facto, o limiar da Região Crítica será agora $f_{0.05(3,192)} \approx 2.68$, enquanto que $F_{calc}^{A} = 247.150$. Também se rejeita (embora de forma menos enfática) a Hipótese Nula da inexistência de efeitos principais do Factor B ($H_0 : \beta_j = 0$, para todos os genótipos j), já que a fronteira da Região Crítica é $f_{0.05(5,192)} \approx 2.29$, e $F_{calc}^{B} = 5.718037$. Assim, os três tipos de efeitos são significativos, ao nível de significância $\alpha = 0.05$, embora os efeitos de ambiente sejam bastantes mais pronunciados que os restantes tipos de efeitos.

4. Utilizamos testes de Tukey para comparar as médias de célula. A diferença de duas médias amostrais de célula pode ser considerada significativa se se verificar a desigualdade $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > \tau_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}}$. No cálculo do termo de comparação tem-se $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{1.8728}{9}} = 0.4561676$. O quantil de ordem $1-\alpha$ da distribuição de Tukey pode ser obtido (aproximadamente) nas tabelas de Tukey. Usando o nível global de significância $\alpha = 0.05$, tem-se $\tau_{0.05(24,192)} \approx 5.01$ (usando o valor tabelado correspondente aos parâmetros 20 e ∞). Logo, o limiar (aproximado) de significância é dado por $5.01 \times 0.4561676 = 2.2854$. Uma inspecção preliminar das médias de célula constantes do enunciado permite constatar que as diferenças entre as médias de célula do Ambiente 2 e do Ambiente 1 são sempre superiores a esse limiar. Para confirmar, basta considerar a menor média amostral de célula no Ambiente 2, que é $\overline{y}_{25.} = 5.823$, e a maior média

associada ao Ambiente 1, que é $\overline{y}_{14.} = 1.871$. A diferença entre essas duas médias amostrais de célula é $5.823 - 1.871 = 3.952 > 2.2854$. Qualquer outro par de médias (uma de cada ambiente) tem diferenças ainda maiores, logo mais significativas. A afirmação constante do enunciado é correcta.

5. Trata-se dum gráfico de interacção, com o Factor B (`clone`) associado ao eixo horizontal. Por cima do marcador de cada clone encontram-se 4 pontos, a alturas correspondentes às quatro médias amostrais de célula associadas a esse clone (um ponto por cada ambiente, ou seja, cada nível do Factor A). Segmentos de recta unem os pontos que correspondem a um mesmo ambiente (a legenda do gráfico indica o tipo de traço usado para construir os segmentos). O gráfico torna evidente que as células com os ambientes 2 e 3 têm, globalmente falando, médias amostrais bastante maiores que as dos ambientes 1 e 4. Estas diferenças justificam que os efeitos de ambiente sejam claramente significativos, como se verificou no respectivo teste F. Os efeitos de interacção (que foram considerados significativos ao nível $\alpha = 0.05$) correspondem à falta de paralelismo entre as "curvas" associadas aos ambientes de maiores rendimentos (2 e 3), e entre estas e as dos ambientes de menores rendimentos (1 e 4). Note-se em particular a forma como, no ambiente 3, o rendimento associado ao clone `AI2` é visivelmente superior aos rendimentos dos clones `AI1` e `AI3`, enquanto que no ambiente 2 se passa precisamente o contrário. Ao mesmo tempo, pouca diferença existe entre os rendimentos de todos os clones nos ambientes 1 e 4 (que são consistentemente baixos). Naturalmente, são os testes que nos indicam se estas diferenças sugeridas pelo gráfico devem, ou não, ser consideradas significativas.

IV

1. O estimador $\hat{\beta}_1$ pode ser escrito na forma $\hat{\beta}_1 = \sum_{i=1}^{n} c_i\, Y_i$, com $c_i = \frac{x_i - \overline{x}}{(n-1)\cdot s_x^2}$ (ver formulário). Sabemos que, no Modelo Linear, as observações Y_i da variável resposta são independentes e com distribuição Normal. Logo $\hat{\beta}_1$ é uma combinação linear de Normais independentes, pelo que também tem distribuição Normal. Falta determinar os respectivos parâmetros, $E[\hat{\beta}_1]$ e $V[\hat{\beta}_1]$. Ora, tendo em conta as propriedades dos valores esperados e variâncias, bem como os pressupostos do Modelo de Regressão Linear Simples, tem-se:

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\sum_{i=1}^{n} c_i\, Y_i\right] = \sum_{i=1}^{n} c_i\, E[Y_i] = \sum_{i=1}^{n} c_i\, E[\beta_0 + \beta_1\, x_i + \epsilon_i] = \sum_{i=1}^{n} c_i\, (\beta_0 + \beta_1\, x_i + \underbrace{E[\epsilon_i]}_{=0}) \\ &= \sum_{i=1}^{n} c_i\, \beta_0 + \sum_{i=1}^{n} c_i\, \beta_1\, x_i = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \underbrace{\sum_{i=1}^{n} c_i\, x_i}_{=1} = \beta_0 \sum_{i=1}^{n} \underbrace{\frac{x_i - \overline{x}}{(n-1)\cdot s_x^2}}_{=0} + \beta_1 = \beta_1. \end{aligned}$$

já que $\sum_{i=1}^{n} \frac{x_i - \overline{x}}{(n-1)\cdot s_x^2} = \frac{1}{(n-1)\cdot s_x^2} \sum_{i=1}^{n} (x_i - \overline{x}) = \frac{1}{(n-1)\cdot s_x^2} (\underbrace{\sum_{i=1}^{n} x_i}_{=n\,\overline{x}} - \underbrace{\sum_{i=1}^{n} \overline{x}}_{=n\,\overline{x}}) = 0.$

Por outro lado, e de novo tendo em conta a independência das observações Y_i e o Modelo RLS,

$$\begin{aligned} V[\hat{\beta}_1] &= V\left[\sum_{i=1}^{n} c_i\, Y_i\right] = \sum_{i=1}^{n} c_i^2\, V[Y_i] = \sum_{i=1}^{n} c_i^2\, V[\beta_0 + \beta_1\, x_i + \epsilon_i] = \sum_{i=1}^{n} c_i^2\, \underbrace{V[\epsilon_i]}_{=\sigma^2} \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{[(n-1)\cdot s_x^2]^2} = \frac{\sigma^2}{[(n-1)\cdot s_x^2]^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)\,s_x^2} = \frac{\sigma^2}{(n-1)\cdot s_x^2}\ . \end{aligned}$$

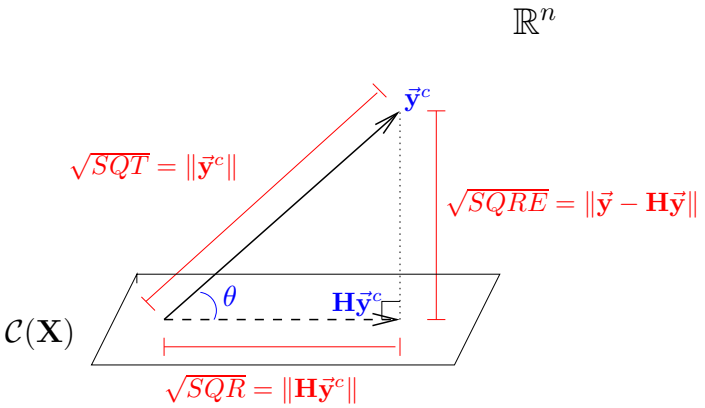
Logo, tem-se $\hat{\beta}_1 \frown \mathcal{N}\left(\beta_1\,,\,\frac{\sigma^2}{(n-1)\cdot s_x^2}\right)$.

2. (a) Seja $\vec{\mathbf{Y}}$ o vector aleatório com as n observações da variável resposta, e $\vec{\boldsymbol{\epsilon}}$ o vector aleatório dos correspondentes erros aleatórios. Seja $\mathbf{X}_{n\times(p+1)}$ a matriz (não aleatória) do modelo, cuja primeira coluna é constituída por n uns, e cujas colunas seguintes contêm as n observações de cada uma das p variáveis preditoras. Seja $\vec{\boldsymbol{\beta}}$ o vector (não aleatório) constituído pelos $p+1$ parâmetros do modelo: $\vec{\boldsymbol{\beta}}=(\beta_0,\beta_1,...,\beta_p)^t$. O Modelo de Regressão Linear Múltipla admite os seguintes pressupostos:

   - Equação do Modelo: $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$;
   - Pressupostos sobre os erros aleatórios: $\vec{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(\vec{\mathbf{0}}\,,\,\sigma^2\,\mathbf{I}_n)$.

   A equação do Modelo corresponde à relação linear de fundo entre os preditores e a variável resposta. Os erros aleatórios representam a variabilidade em torno dessa relação linear, admitindo-se a Multinormalidade, independência e variâncias homogéneas no segundo pressuposto do Modelo.

   (b) O triângulo ao qual se aplica o Teorema de Pitágoras tem como hipotenusa o vector centrado das observações de Y, o vector $\vec{\mathbf{y}}^c$, cujo elemento genérico é $y_i - \bar{y}$. Um dos catetos é a projecção ortogonal de $\vec{\mathbf{y}}^c$ sobre o subespaço gerado pelas colunas da matriz $\mathbf{X}$, o subespaço $\mathcal{C}(\mathbf{X})$. Essa projecção é o vector $\mathbf{H}\vec{\mathbf{y}}^c$, onde a matriz $\mathbf{H}$ é a matriz de projecção ortogonal no referido subespaço, dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$.



O outro cateto é dado pela diferença $\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c$. Sabemos que o comprimento da hipotenusa é $\|\vec{\mathbf{y}}^c\| = \sqrt{SQT}$. O do cateto que reside em $\mathcal{C}(\mathbf{X})$ é $\|\mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{SQR}$. O outro cateto tem comprimento $\|\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{SQRE}$ (que é também igual a $\|\vec{\mathbf{y}} - \mathbf{H}\vec{\mathbf{y}}\|$). Assim, o Teorema de Pitágoras garante que $SQT = SQR + SQRE$ (ver a figura).

   (c) Tem-se, tendo em conta a equação do modelo e o facto de $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$; e $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$:

$$\begin{aligned}\vec{\mathbf{E}} &= \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}-\mathbf{H})\vec{\mathbf{Y}} = (\mathbf{I}-\mathbf{H})(\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}) = \mathbf{X}\vec{\boldsymbol{\beta}} - \mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}} - \mathbf{H}\vec{\boldsymbol{\epsilon}} \\ &= \mathbf{X}\vec{\boldsymbol{\beta}} - \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X})}_{=\mathbf{I}}\vec{\boldsymbol{\beta}} + (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \cancel{\mathbf{X}\vec{\boldsymbol{\beta}}} - \cancel{\mathbf{X}\vec{\boldsymbol{\beta}}} + (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}\,.\end{aligned}$$

   A Soma de Quadrados dos Resíduos é dada por $SQRE = \|\vec{\mathbf{E}}\|^2 = \vec{\mathbf{E}}^t\vec{\mathbf{E}}$. Tendo em conta a alínea anterior, as propriedades de matrizes e ainda o facto das matrizes $\mathbf{I}$ e $\mathbf{H}$ serem simétricas e idempotentes, tem-se:

$$\begin{aligned}SQRE &= \vec{\mathbf{E}}^t\vec{\mathbf{E}} = [(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}]^t(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H})^t(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}^t-\mathbf{H}^t)(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} \\ &= \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H}-\mathbf{H}+\underbrace{\mathbf{H}\mathbf{H}}_{=\mathbf{H}})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t\vec{\boldsymbol{\epsilon}} - \vec{\boldsymbol{\epsilon}}^t\mathbf{H}\vec{\boldsymbol{\epsilon}}\,.\end{aligned}$$

   Ora a primeira parcela na expressão final ($\vec{\boldsymbol{\epsilon}}^t\vec{\boldsymbol{\epsilon}}$) é a soma de quadrados dos erros aleatórios. A segunda parcela não pode ser negativa, já que é uma norma (comprimento) ao quadrado: $\vec{\boldsymbol{\epsilon}}^t\mathbf{H}\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t\mathbf{H}\mathbf{H}\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\epsilon}}^t\mathbf{H}^t\mathbf{H}\vec{\boldsymbol{\epsilon}} = (\mathbf{H}\vec{\boldsymbol{\epsilon}})^t(\mathbf{H}\vec{\boldsymbol{\epsilon}}) = \|\mathbf{H}\vec{\boldsymbol{\epsilon}}\|^2$. Logo, tem de ter-se $SQRE = \vec{\mathbf{E}}^t\vec{\mathbf{E}} \leq \vec{\boldsymbol{\epsilon}}^t\vec{\boldsymbol{\epsilon}}$.

(d) O Modelo de Regressão Linear exige que $\vec{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(\vec{\mathbf{0}}\,,\, \sigma^2\,\mathbf{I}_n)$. Como se viu acima, o vector dos resíduos obtém-se multiplicando, à esquerda, o vector dos erros aleatórios $\vec{\boldsymbol{\epsilon}}$ por $(\mathbf{I}-\mathbf{H})$. Sabemos que esse tipo de transformação dum vector Multinormal não afecta a Multinormalidade. Fica por saber quais os dois parâmetros (vector esperado e matriz de (co-)variâncias) associados a $\vec{\mathbf{E}} = (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}$. Pelas propriedades dos vectores esperados e matrizes de (co-)variâncias, tem-se:

$$E[\vec{\mathbf{E}}] = E[(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}] = (\mathbf{I}-\mathbf{H})E[\vec{\boldsymbol{\epsilon}}] = (\mathbf{I}-\mathbf{H})\vec{\mathbf{0}} = \vec{\mathbf{0}}.$$

e

$$\begin{aligned} V[\vec{\mathbf{E}}] &= V[(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}] = (\mathbf{I}-\mathbf{H})V[\vec{\boldsymbol{\epsilon}}](\mathbf{I}-\mathbf{H})^t = (\mathbf{I}-\mathbf{H})\cdot\sigma^2\mathbf{I}\cdot(\mathbf{I}^t-\mathbf{H}^t) = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) \\ &= \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}+\underbrace{\mathbf{H}\mathbf{H}}_{=\mathbf{H}}) = \sigma^2(\mathbf{I}-\mathbf{H})\,. \end{aligned}$$

Assim, tem-se $\vec{\mathbf{E}} \frown \mathcal{N}_n\left(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}-\mathbf{H})\right)$.

(e) Têm-se duas expressões alternativas para o vector dos resíduos: $\vec{\mathbf{E}}=(\mathbf{I}-\mathbf{H})\vec{\mathbf{Y}}$ e $\vec{\mathbf{E}}=(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}}$. Logo, tem-se:

$$(\mathbf{I}-\mathbf{H})\vec{\mathbf{Y}} = (\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} \quad\Leftrightarrow\quad (\mathbf{I}-\mathbf{H})\vec{\mathbf{Y}}-(\mathbf{I}-\mathbf{H})\vec{\boldsymbol{\epsilon}} = \vec{\mathbf{0}} \quad\Leftrightarrow\quad (\mathbf{I}-\mathbf{H})(\vec{\mathbf{Y}}-\vec{\boldsymbol{\epsilon}}) = \vec{\mathbf{0}} \quad\Leftrightarrow\quad (\mathbf{I}-\mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} = \vec{\mathbf{0}}\,,$$

e, no entanto, nem $\mathbf{X}\vec{\boldsymbol{\beta}}=\vec{\mathbf{0}}$, nem a matriz $(\mathbf{I}-\mathbf{H})$ é uma matriz de zeros. Assim, em produtos matriciais/vectoriais, um produto nulo não obriga a que pelo menos um dos factores seja nulo.