INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

5 de Janeiro, 2016 **SEGUNDO TESTE 2015-16** **Uma resolução possível**

I

1. Em qualquer modelo linear, os graus de liberdade residuais são dados pela diferença entre o número total de observações e o número de parâmetros do modelo. No modelo de regressão linear múltipla em apreço, tem-se $g.l.(QMRE) = n-(p+1) = 28$. Uma vez que existem $p = 8$ variáveis preditoras, tem-se que o ajustamento foi feito com base em $n = 28+9 = 37$ observações.

2. Em geral, os coeficientes β_j associados a variáveis preditoras numa regressão linear múltipla representam a variação no valor esperado da variável resposta Y, associado a um aumento duma unidade na respectiva variável preditora X_j, e mantendo constantes os valores dos restantes preditores. Assim, a estimativa $b_1 = -14.11886$ corresponderia a dizer que a transpiração diária média diminui $14.11886\, mm/dia$, caso o preditor `BLUE` aumente uma unidade (para valores fixos dos restantes preditores). Mas, como é sublinhado no enunciado, no nosso caso não se verificam aumentos de uma unidade nesse preditor, tendo em conta que a gama de valores desse preditor associados ao problema varia entre 0.04 e 0.12. Pode repetir-se o raciocínio que conduziu à interpretação acima referida de β_j, mas considerando aumentos de k unidades em X_j (para qualquer constante real k). Nesse caso, a diferença nos valores esperados de Y, após e antes um tal aumento (e em igualdade das restantes circunstâncias) é dado por:

$$\begin{aligned} E[Y \mid x_1, ..., x_j+k, ..., x_p] &= \cancel{\beta_0} + \cancel{\beta_1 x_1} + ... + \cancel{\beta_{j-1} x_{j-1}} + \underbrace{\beta_j (x_j+k)}_{=\cancel{\beta_j x_j} + k\,\beta_j} + \cancel{\beta_{j+1} x_{j+1}} + ... + \cancel{\beta_p x_p} \\ - \quad E[Y \mid x_1, ..., x_j, ..., x_p] &= \cancel{\beta_0} + \cancel{\beta_1 x_1} + ... + \cancel{\beta_{j-1} x_{j-1}} + \cancel{\beta_j x_j} + \cancel{\beta_{j+1} x_{j+1}} + ... + \cancel{\beta_p x_p} \\ \hline = & \quad k\,\beta_j \end{aligned}$$

   Tomando, por exemplo, $k = 0.01$, podemos afirmar que a estimativa da variação esperada na média da transpiração diária, quando a reflectância `BLUE` aumenta 0.01 unidades (mantendo outros preditores constantes) é $kb_1 = -0.1411886$ mm/dia, ou por outras palavras, a transpiração diária média diminui 0.1411886 mm/dia.

3. Pede-se um teste em que a hipótese alternativa seja $\beta_8 > 0$. Tem-se o seguinte teste t:

   Hipóteses: $H_0 : \beta_8 \leq 0 \quad vs. \quad H_1 : \beta_8 > 0.$

   Estatística do Teste: $T = \frac{\hat{\beta}_8 - \beta_{8|H_0}}{\hat{\sigma}_{\hat{\beta}_8}} \cap t_{n-(p+1)}$, 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 $T_{calc} > t_{\alpha(28)} = 1.70113$.

   Conclusões: O valor de T_{calc} é dado na listagem, sendo $T_{calc} = \frac{0.04466-0}{0.02403} = 1.858 > 1.70113$. Logo, rejeita-se H_0 e é possível concluir pela afirmação do enunciado: aumentos de temperatura estão associados a aumentos de transpiração diária (sendo as restantes condições observadas constantes). Note-se, no entanto, que não se pode afirmar que aumentos de temperatura *provoquem* aumentos de transpiração: a associação estatística não é sinónimo duma relação de causa e efeito.

4. Trata-se dum gráfico de resíduos (internamente) estandardizados R_i (eixo vertical) contra efeitos alavanca h_{ii} (eixo horizontal). Dada a relação existente entre as distâncias de Cook e estas duas quantidades ($D_i = R_i^2 \frac{h_{ii}}{1-h_{ii}} \frac{1}{p+1}$, como indicado no formulário) é possível traçar curvas de igual distância de Cook, e neste gráfico essas curvas são dadas para as distâncias de Cook 0.5 e 1. As principais conclusões do gráfico são as seguintes:

   - Os resíduos estandardizados estão compreendidos no intervalo aproximado [-2,2], ou seja, nenhuma observação tem resíduos estandardizados de destaque.
   - Os efeitos alavanca de algumas observações assumem valores bastante elevados. Mesmo tendo em conta que para este caso, o efeito alavanca médio é relativamente elevado ($\overline{h} = \frac{p+1}{n} = \frac{9}{37} = 0.243$), tem-se $h_{ii} > 0.6$ em duas observações ($i=4$ e $i=7$) e para a observação $i=7$ o efeito alavanca é provavelmente superior a 0.7. São duas observações com um elevado "poder de atracção"do hiperplano ajustado.
   - As duas observações com elevado efeito alavanca têm também distâncias de Cook elevadas, acima do limiar de alerta 0.5 e, no caso da observação 7, tem-se mesmo $D_7 > 1$. Trata-se, pois, de duas observações muito influentes, ou seja, a exclusão duma dessas observações do conjunto de dados provocaria alterações importantes no hiperplano ajustado. Saliente-se que, embora os dois resíduos estandardizados correspondentes sejam os de maior magnitude, não são resíduos assinaláveis. Assim, as distâncias de Cook muito elevadas reflectem os elevados efeitos alavanca destas observações, como indicado acima.

5. É dado um submodelo com apenas $k=4$ preditores.

   (a) Pede-se um teste F parcial para comparar o submodelo com o modelo completo original, de $p=8$ preditores. Tem-se:

   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 $\mathcal{R}_c^2$ e $\mathcal{R}_s^2$ indicam os coeficientes de determinação populacional, respectivamente do modelo completo e do submodelo.

   Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \frac{R_c^2-R_s^2}{1-R_c^2} \cap F_{(p-k,n-(p+1))}$, 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 $F_{calc} > f_{0.05[4,28]} = 2.71$.

   Conclusões: Tem-se $F_{calc} = \frac{28}{4} \frac{0.9032-0.8746}{1-0.9032} = 2.068$. Logo, não se rejeita H_0, i.e., considera-se que a qualidade de ajustamento do modelo completo não é significativamente melhor (ao nível $\alpha = 0.05$) do que o da do submodelo, que é assim um modelo mais parcimonioso. A perda de menos de 3% na variabilidade explicada da variável resposta parece ser compensada pela redução em metade no número de variáveis preditoras.

   (b) As variáveis `BLUE` e `RED` formam, juntamente com o preditor `GREEN`, um grupo de variáveis muito fortemente correlacionadas entre si. A mais baixa correlação entre pares destas três variáveis é 0.97. Assim, é natural que a informação presente em dois destes três preditores seja bem substituída pelo terceiro destes preditores, sem grande prejuízo no valor de R^2.

   (c) Já foi referida antes a fórmula de cálculo duma distância de Cook: $D_i = R_i^2 \frac{h_{ii}}{1-h_{ii}} \frac{1}{p+1}$. Conhecemos os valores $p+1 = 5$ e $h_{77} = 0.39501$. Falta calcular o resíduo estandardizado que, como é também indicado no formulário, é dado por $R_i = \frac{E_i}{\sqrt{QMRE\,(1-h_{ii})}}$. Tendo em conta que $e_7 = -0.1333$ e $\sqrt{QMRE} = 0.3384$, tem-se $R_7 = \frac{-0.1333}{0.3384 \times \sqrt{1-0.39051}} = -0.5045$ e $D_7 = (-0.5045)^2 \times \frac{0.39051}{1-0.39051} \times \frac{1}{5} = 0.0326$. Trata-se dum valor muito baixo, que é tanto mais de assinalar quanto esta mesma observação tinha, no modelo completo, uma distância de Cook elevadíssima, superior a 1. Não sendo disponibilizados os dados originais, não é

possível identificar as causas deste facto, mas é possível afirmar que a enorme influência da observação 7 no modelo completo está associada aos seus valores nas quatro variáveis preditoras entretanto excluídas no submodelo, ou seja, `BLUE`, `RED`, `SWIR1` e `SWIR2`.

II

1. Trata-se dum delineamento factorial a dois factores, sendo a variável resposta Y o rendimento (em kg/planta); o primeiro factor (A) o genótipo, com $a=12$ níveis (uma vez que os correspondentes graus de liberdade, indicados na tabela resumo, são a–$1=11$); e o segundo factor (B) os anos, com $b=5$ níveis (referidos no enunciado). A estrutura da tabela confirma tratar-se dum delineamento factorial, tendo sido ajustado o modelo que prevê efeitos de interacção. Neste modelo, os graus de liberdade são dados pela diferença entre o número total de observações (n) e o número de células (situações experimentais), $ab=12\times 5=60$. Como a tabela indica que $n-ab=420$, tem-se ao todo $n=480$ observações. Tratando-se dum delineamento equilibrado (como é referido no enunciado), houve $n_c=\frac{n}{ab}=\frac{480}{60}=8$ observações em cada célula. O modelo ajustado é:

   - O rendimento da k-ésima parcela associada ao genótipo i, no ano j, é dado por $Y_{ijk} = \mu_{11}+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\epsilon_{ijk}$, $\forall\, i,j,k$, sendo μ_{11} o rendimento esperado do primeiro genótipo em 1994; α_i o efeito principal (acréscimo ao rendimento) associado ao genótipo i (com a restrição $\alpha_1=0$); β_j o efeito principal (acréscimo ao rendimento) associado ao ano j (com a restrição $\beta_1=0$); $(\alpha\beta)_{ij}$ o efeito de interacção (acréscimo ao rendimento) associado à combinação do genótipo i com o ano j (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.

2. A variância dos erros aleatórios é, pelo segundo ponto do modelo, dada por $V[\epsilon_{ijk}]=\sigma^2$. Em qualquer modelo linear, uma tal variância é estimada pelo Quadrado Médio Residual. Logo, $\hat{\sigma}^2 = QMRE = \frac{198.44}{420} = 0.4725$. As unidades de medida dos resíduos são iguais às unidades de medida da variável resposta, e tendo em conta que esses resíduos são elevados ao quadrado no cálculo de $SQRE$, tem-se que as unidades de medida desta QMRE são $(\text{kg/planta})^2$.

3. Vai-se efectuar em pormenor o teste aos efeitos principais do Factor B (Ano), e descrever sinteticamente os testes aos efeitos principais do Factor A (genótipo) e aos efeitos de interacção.

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

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

   Nível de significância: $\alpha=0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,420)}\approx 2.4$ (entre os valores tabelados 2.37 e 2.45).

   Conclusões: Como $F_{calc}=\frac{QMB}{QMRE}=110.180\gg 2.45$, rejeita-se claramente H_0, sendo possível concluir pela existência de efeitos principais de ano (ao nível $\alpha=0.05$ e, presumivelmente para todos os níveis usuais de α). No nosso contexto, tal corresponde a afirmar que, em pelo menos alguns anos, há uma mudança no rendimento médio, quando comparado com outros anos.

No teste aos efeitos de interacção, com hipóteses $H_0 : (\alpha\beta)_{ij} = 0$, para todo o i e j, contra H_1 : existe pelo menos uma célula (i, j) onde $(\alpha\beta)_{ij} \neq 0$, o *p-value* muito elevado (p=0.354688) indica a não rejeição de H_0 para os habituais níveis de significância, pelo que se pode concluir pela inexistência de efeitos significativos de interacção.

Finalmente, no teste aos efeitos principais do factor genótipo, com hipóteses $H_0 : \alpha_i = 0\ ,\ \forall\ i$ vs. $H_1 : \exists i$ tal que $\alpha_i \neq 0$, tem-se um valor de prova muito baixo ($p = 0.000904$) e para qualquer dos níveis de significância usuais rejeita-se H_0, ou seja, conclui-se pela existência de efeitos principais de genótipo nos rendimentos médios. Assim, vale a pena escolher (ou excluir) alguns dos genótipos estudados.

4. A tabela-resumo da ANOVA correspondente ao modelo com o único factor, genótipo (ou seja, ao Modelo M_A), tem apenas 2 linhas: a correspondente à variabilidade associada ao factor e a correspondente à variabilidade residual. Por definição, a Soma de Quadrados, grau de liberdade e, por conseguinte, o Quadrado Médio associado ao factor genótipo são calculados como na tabela-resumo do modelo ANOVA a dois factores, com efeitos de interacção (o modelo M_{A*B}, onde o factor genótipo desempenhava o papel de Factor A). Logo, os correspondentes valores são iguais nas duas tabelas. Uma vez que a soma de todas as Somas de Quadrados em cada modelo ANOVA tem de ser sempre igual a $SQT = (n-1)s_y^2$, e uma vez que os valores da variável resposta Y_i com que se ajusta os dois modelos são os mesmos, tem de verificar-se $SQRE_A = SQB + SQAB + SQRE_{A*B} = 208.25 + 22.29 + 198.44 = 428.98$. De forma análoga, os graus de liberdade das duas tabelas têm de somar $n-1$, pelo que $g.l.(SQRE_A) = 4+44+420 = 468$. Tem-se então $QMRE = \frac{SQRE}{n-k} = \frac{428.98}{468} = 0.9166239$ e $F = \frac{QMF}{QMRE} = \frac{1.39}{0.9166239} = 1.51643$. A tabela-resumo vem assim:

	g.l.	SQs	QMs	F_{calc}
Factor genótipo	11	15.31	1.39	1.51643
Residual	468	428.98	0.9166239	

A principal conclusão a extrair é que ao nível de significância $\alpha = 0.05$, os efeitos de genótipo seriam agora considerados *não significativos*, uma vez que o limiar da região crítica seria agora $f_{0.05(11,468)}$ a que, pelas tabelas, corresponde um valor entre 1.91 e 1.75. Assim, os efeitos de genótipo deixariam de ser considerados significativos a esse nível de α. O facto de ignorar a existência de efeitos presentes na realidade levou ao inflacionamento da variabilidade residual, e consequentemente à diminuição do valor da estatística F, apesar de a variabilidade associada ao factor genótipo ter permanecido igual.

III

1. (a) A projecção ortogonal de qualquer vector $\vec{z} \in \mathbb{R}^n$ sobre $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$ é o vector $\mathbf{H}\vec{z}$, onde $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ é a matriz de projecção ortogonal correspondente. Se $\vec{z}$ já pertence a $\mathcal{C}(\mathbf{X})$, pode escrever-se na forma $\vec{z} = \mathbf{X}\vec{a}$ (para algum vector $\vec{a} \in \mathbb{R}^m$), como indicado no enunciado. Logo a projecção é dada por $\mathbf{H}\vec{z} = \mathbf{H}\mathbf{X}\vec{a} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\cdot\mathbf{X}}_{=\mathbf{I}_m}\vec{a} = \mathbf{X}\vec{a} = \vec{z}$.
   Assim, o vector $\vec{z}$ permanece invariante sob a acção da projecção.

   (b) A matriz quadrada $\mathbf{H}$ ser simétrica significa que $\mathbf{H}^t = \mathbf{H}$. Ora, tendo em conta as propriedades de produtos matriciais (também constantes do formulário, nomeadamente

$(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$; $(\mathbf{A}^t)^t = \mathbf{A}$; e $(\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1}$), tem-se:

$$\begin{aligned}\mathbf{H}^t &= [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t = (\mathbf{X}^t)^t[(\mathbf{X}^t\mathbf{X})^{-1}]^t\mathbf{X}^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1}\mathbf{X}^t \\ &= \mathbf{X}[\mathbf{X}^t(\mathbf{X}^t)^t]^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\ ,\end{aligned}$$

como se queria mostrar. Por outro lado, $\mathbf{H}$ ser idempotente significa que $\mathbf{HH} = \mathbf{H}$. Ora,

$$\mathbf{HH} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \cdot \mathbf{X}}_{=\mathbf{I}_m}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\ .$$

2. Tendo em conta que, no contexto duma ANOVA a um factor, a tradicional Soma de Quadrados associada ao ajustamento do modelo (que na regressão linear se designa SQR) é chamada SQF, tem-se $R^2 = \frac{SQF}{SQT}$.

   (a) A condição $R^2 = 0$ equivale a $SQF = 0$. Ora, no contexto ANOVA a um factor tem-se (ver formulário e tendo em conta que o delineamento é equilibrado):

   $$SQF = \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2 = n_c \sum_{i=1}^{k}(\overline{Y}_{i.} - \overline{Y}_{..})^2 = 0\ .$$

   Ora, uma soma de quadrados só se pode anular se *todas* as suas parcelas se anulam o que, neste contexto, significa que $\overline{Y}_{i.} = \overline{Y}_{..}$, para todo o i. Por outras palavras, $R^2 = 0$ se e só se todas as médias amostrais de nível forem iguais à média amostral da totalidade das observações (e portanto iguais entre si). Assim, a informação proveniente da amostra aponta de forma clara em abono da hipótese de igualdade de todas as médias populacionais de nível ($\mu_1 = \mu_2 = ... = \mu_k$), que é a hipótese nula no teste F duma ANOVA a um único factor. Este resultado é inteiramente coerente com a não rejeição da hipótese nula do teste que resulta do facto de $R^2 = 0 \Leftrightarrow F_{calc} = 0$. Repare-se ainda que a condição $SQF = 0$ é equivalente a dizer que $SQT = SQF + SQRE = SQRE$, ou seja, toda a variabilidade de Y é residual, ou seja, interna aos níveis do factor.

   (b) A condição $R^2 = 1$ equivale a $SQF = SQT$, ou seja, $SQRE = 0$. Ora, no contexto ANOVA a um factor tem-se (ver formulário e para um delineamento equilibrado):

   $$SQRE = \sum_{i=1}^{k}(n_i - 1)S_i^2 = (n_c - 1)\sum_{i=1}^{k} S_i^2 = 0\ .$$

   De novo, uma soma de quadrados só pode ser nula se *todas* as suas parcelas forem nulas, pelo que $SQRE = 0$ equivale a $S_i^2 = 0$, para todo o nível i, ou seja, não existe variabilidade das observações de Y no seio dum mesmo nível do factor. Neste caso tem-se também $QMRE = \frac{SQRE}{n-k} = 0$. Embora não seja possível construir a estatística do teste $F = \frac{QMF}{QMRE}$, a divisão por zero sugere um valor limite infinito, que corresponderia sempre à rejeição da hipótese nula de igualdade das médias populacionais de nível μ_i, o que é coerente com o referido facto de, neste caso, toda a variabilidade nas observações de Y corresponder à mudança entre níveis do factor.