INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

12 Julho 2021 Segunda Chamada de EXAME Uma resolução possível

NOTA: Este exame esteve inicialmente previsto para dia 25 de Janeiro de 2021, mas apenas se realizou a 12 de Julho devido à pandemia Covid-19.

I

1. Regressão Linear Simples

   (a) A recta de regressão de `antocianas` (y) sobre `fenois` (x) tem equação $y = b_0 + b_1\, x$, com $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x}$ e $b_0 = \overline{y} - b_1\,\overline{x}$. Tendo em conta os valores do enunciado, $b_1 = 0.89735 \times \frac{53.6160}{101.9808} = 0.4717782$ e $b_0 = 275.0065 - 0.4717782 \times 564.3342 = 8.765927$. Logo, a equação pedida é $y = 8.765927 + 0.4717782\, x$.

   (b) O coeficiente de determinação é $R^2 = r_{xy}^2 = 0.89735^2 = 0.805237$. A recta explica cerca de $80,5\%$ da variabilidade observada no teor de antocianas, um valor razoavelmente bom e muito significativamente diferente de zero, já que num teste F de ajustamento global (com hipóteses $H_0 : \mathcal{R}^2 = 0$ vs. $H_1 : \mathcal{R}^2 > 0$), a região crítica unilateral direita tem fronteira $f_{0.05(1,50)} = 4.03431$ e o valor calculado da estatística é $F_{calc} = \frac{(n-2)R^2}{1-R^2} = 206.7223$. A rejeição de H_0 é muito clara.

   (c) Pede-se um intervalo de confiança para o declive β_1 da recta populacional, que é da forma $\left]\ b_1 - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\ ,\ b_1 + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\ \right[$. O valor de b_1 foi calculado na alínea a). Para um IC a 95% de confiança, tem-se: $t_{0.025(50)} = 2.00856$. Usando a fórmula para a variância de $\hat{\beta}_1$ (formulário), tem-se a *estimativa* $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{QMRE}{(n-1)\,s_x^2}$. O enunciado dá os valores de $QMRE = 571.186$ e $s_x = 101.9808$. Assim, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{571.186}{51 \times (101.9808)^2}} = 0.03281597$. Logo, o IC a 95% de confiança para β_1 é $\left]\ 0.40587\ ,\ 0.53769\ \right[$. É admissível afirmar que $\beta_1 = 0.5$, que é um valor pertencente ao intervalo.

   (d) Por definição, o resíduo usual é dado por $e_i = y_i - \hat{y}_i$. Dado o posicionamento do ponto no gráfico, sabemos que se trata da observação com os maiores valores observados, quer de y, quer de x. Logo, $y_i = y_{max} = 448.615$ e $\hat{y}_i = b_0 + b_1\, x_{max} = 8.765927 + 0.4717782 \times 832.909 = 401.7142$. Assim, o resíduo é $e_i = 448.615 - 401.7142 = 46.9008$.

   (e) Numa regressão linear simples, o maior efeito alavanca está associado à observação cujo valor de preditor x_i mais se afasta da média dos valores do preditor, $\overline{x}$, o que terá forçosamente de ser a observação com o menor, ou o maior, valor observado x_i. No nosso caso, $\overline{x} = 564.3342$. O extremo que mais se afasta deste valor médio é $x_{max} = 832.909$. O valor do efeito alavanca correspondente é (formulário) $h_{ii} = \frac{1}{n} + \frac{(x_1 - \overline{x})^2}{(n-1)\,s_x^2} = \frac{1}{52} + \frac{(832.909 - 564.3342)^2}{51 \times 101.9808^2} = 0.1552259$.

2. Regressão polinomial (cúbica).

   (a) O polinómio de terceiro grau ajustado tem equação $y = b_0 + b_1\, x + b_2\, x^2 + b_3\, x^3 = -811.4 + 4.906\, x - 0.007789\, x^2 + 0.000004445\, x^3$.

   (b) O modelo de regressão linear simples é um submodelo deste modelo de regressão polinomial, pelo que podem ser comparados através dum teste F parcial. A Hipótese Nula é $\mathcal{R}^2_s = \mathcal{R}^2_c$, onde os índices s e c indicam, respectivamente, o submodelo (regressão linear) e modelo completo (regressão polinomial). Sabemos que a estatística de teste é $F = \frac{n-(p+1)}{p-k}\,\frac{R^2_c - R^2_s}{1-R^2_c}$,

que sob H_0 tem distribuição $F_{[p-k,n-(p+1)]}$. A fronteira da região crítica unilateral direita, ao nível de significância $\alpha = 0.05$ é $f_{0.05[2,48]} \approx 3.20$. O valor calculado da estatística de teste é $F_{calc} = \frac{48}{2} \times \frac{0.8238-0.805237}{1-0.8238} = 2.528445$. Assim, não se rejeita H_0 ao nível $\alpha = 0.05$: o ganho na precisão não é significativo, não havendo justificação para trocar a regressão linear simples pela regressão cúbica.

3. Regressão linear múltipla, com $p=5$ preditores.

   (a) Pede-se um teste unilateral ao sinal do coeficiente do preditor `acidez`, mais concretamente um teste com as hipóteses $H_0 : \beta_4 \geq 0$ vs. $H_1 : \beta_4 < 0$. A estatística de teste é $t = \frac{\hat{\beta}_4 - \beta_{4|H_0}}{\hat{\sigma}_{\hat{\beta}_4}}$ que, sob H_0, tem distribuição $t_{n-(p+1)}$. A região crítica é unilateral esquerda, com limiar $-t_{0.05(46)} \approx -1.68$. O valor calculado da estatística de teste é dada na listagem do enunciado (já que $\beta_{4_{|H_0}} = 0$), sendo $t_{calc} = -0.654$. Logo, não se rejeita H_0, pelo que não é válida a afirmação do enunciado.

   (b) Nas duas colunas finais da tabela de coeficientes na listagem, encontram-se os valores de t_{calc} e respectivos valores de prova (*p-values*) para os testes bilaterais a $H_0 : \beta_j = 0$. Há dois preditores em que nunca se rejeita H_0, qualquer que seja o nível de significância (sensato) usado, pelo que a sua exclusão não afecta de forma significativa o ajustamento, sendo a não rejeição mais pronunciada no caso do teste a $\beta_4 = 0$ ($p = 0.51655$). Logo, a exclusão do preditor `acidez` é a que menos afecta a qualidade de ajustamento do modelo.

   (c) O valor desconhecido R_s^2 pode ser calculado sabendo que a estatística do teste F parcial que compara o modelo completo e o submodelo resultante de excluir *um único* preditor é o quadrado da estatística *t-Student* a que o coeficiente desse preditor seja nulo (as estatísticas consideradas na alínea anterior). No caso da exclusão do preditor `fenois`, tem-se $F_{calc} = t_{calc}^2 = 10.247^2 = 105.001$. Mas pela expressão genérica da estatística do teste F parcial, tem-se $F_{calc} = \frac{n-(p+1)}{p-k} \frac{R_c^2 - R_s^2}{1-R_c^2} = \frac{46}{1} \times \frac{0.8439-R_s^2}{1-0.8439}$. Igualando, tem-se $105.001 = 46 \times \frac{0.8439-R_s^2}{0.1561}$, logo $0.8439 - R_s^2 = 0.35632$ e portanto $R_s^2 = 0.4876$.

   Nota: o modelo de regressão linear múltipla com 4 preditores, sem o preditor `fenois`, explica uma proporção da variância observada de `antocianas` muito inferior à explicada pelo modelo de regressão linear simples apenas com o preditor `fenois`. Não se trata duma contradição, uma vez que não estamos perante um modelo e submodelo.

II

1. Trata-se dum delineamento experimental factorial, com dois factores: `ano` (com $a = 3$ níveis) e `genotipo` (com $b = 9$ níveis). Em cada uma das $ab = 27$ situações experimentais existem $n_{ij} = n_c = 5$ observações, pelo que se trata dum delineamento equilibrado, com um total de $abn_c = 135$ observações. Havendo repetições, é possível ajustar o modelo ANOVA com efeitos de interacção, a seguir indicado:

   Equação do Modelo: $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, onde $i = 1, 2, 3$ indica ano; $j = 1, ..., 9$ indica genótipo; $k = 1, ..., 5$ repetição (dentro de cada combinação ano/genótipo); Y_{ijk} indica o rendimento da k-ésima repetição do genótipo j, no ano i; ϵ_{ijk} é o correspondente erro aleatório. Com as restrições $\alpha_1 = 0$, $\beta_1 = 0$ e $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$, a constante aditiva comum a todas as observações, μ_{11}, representa o rendimento médio populacional do primeiro genótipo no primeiro ano (MV1 em 2013); α_i indica o acréscimo associado ao ano i; β_j indica o acréscimo associado ao genótipo j e $(\alpha\beta)_{ij}$ indica o efeito de interacção entre ano i e genótipo 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. A tabela de síntese tem a estrutura correspondente ao modelo indicado, com três linhas associadas aos tipos de efeitos no modelo (factor A, factor B, e interacção) e ainda a linha associada à variabilidade residual (sem contar com a linha correspondente à variabilidade total). Há oito valores omissos, entre os quais os graus de liberdade, que são dados por $a-1=2$ (Factor A); $b-1=8$ (Factor B), $(a-1)(b-1)=16$ (Interacção); e $n-ab=135-27=108$ (Residual). O valor da estatística do teste aos efeitos do Factor A é dada por $F^A_{calc} = \frac{QMA}{QMRE} = \frac{93.16}{0.57} = 163.4386$. O Quadrado Médio associado ao factor `genotipo` é dado por $QMB = \frac{SQB}{b-1} = \frac{11.69}{8} = 1.46125$. Finalmente, a Soma de Quadrados Residual é dada por $SQRE = QMRE \times (n - ab) = 0.57 \times 108 = 61.56$. Com estes valores fica completa a tabela-resumo:

```
              Df Sum Sq Mean Sq F value
ano            2 186.33   93.16 163.439
genotipo       8  11.69    1.46   2.566
ano:genotipo  16  10.17    0.64   1.117
Residuals    108  61.56    0.57
```

3. Eis o teste aos efeitos do Factor B (de resultado mais incerto) em pormenor. As hipóteses são $H_0 : \beta_j = 0$ para todo o $j > 1$ e $H_1 : \beta_j \neq 0$ para pelo menos um $j > 1$. A estatística de teste é $F^B = \frac{QMB}{QMRE}$, com distribuição $F_{[b-1,\, n-ab]}$, sob H_0. A regra de rejeição ao nível de significância $\alpha = 0.05$ é rejeitar H_0 se $F^B_{calc} > f_{0.05(8,108)} \approx 2.02$. Como $F^B_{calc} = 2.566$, rejeita-se H_0 (ao nível $\alpha = 0.05$) e conclui-se pela existência de efeitos significativos de genótipo sobre o rendimento. Nos outros dois testes, as Hipóteses Nulas correspondem sempre à inexistência dos efeitos de cada tipo. Pelos valores de prova (*p-values*) disponíveis no enunciado, é evidente que haverá uma claríssima rejeição de H_0 no que respeita à existência de efeitos de ano e uma clara não rejeição de H_0 para efeitos de interacção, para qualquer dos valores usuais de α. No teste aos efeitos do Factor A, a fronteira da Região Crítica é $f_{0.05(2,108)} \approx 3.07$ e o enorme valor $F^A_{calc} = 163.439$. No teste aos efeitos de interacção, a fronteira da Região Crítica é $f_{0.05(16,108)} \approx 1.75$ e o valor $F^{AB}_{calc} = 1.117$ não permite rejeitar H_0. Assim, conclui-se que existem efeitos muito significativos de ano e efeitos significativos (mas menos claros) de genótipos, não existindo efeitos significativos de interacção ano/genótipo.

4. As comparações das médias de célula pedidas serão feitas através do termo de comparação de Tukey, que é dado por $\tau = q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}}$. Usando o nível de significância $\alpha = 0.05$ e o valor $q_{0.05(27,108)} = 5.369266$ dado no enunciado, tem-se o termo de comparação $\tau = 5.369266 \times \sqrt{\frac{0.57}{5}} = 1.812873$. Mesmo sem efectuar contas rigorosas, é possível constatar que a diferença entre as médias amostrais dos rendimentos do genótipo MV7 em 2013 e 2014 é inferior a este limiar (logo, não é significativo), enquanto que o rendimento do referido genótipo em 2017 é significativamente diferente do dos outros dois anos (as diferenças das respectivas médias amostrais excedem esse termo de comparação).

III

Trata-se duma hipótese completamente especificada de probabilidades associadas a cada uma das quatro células da tabela ($a = 2$, $b = 2$), que pode ser avaliada através dum teste Qui-quadrado baseado na estatística de Pearson. Concretamente, tem-se:

Hipóteses: $H_0 : \pi_{11} = \frac{9}{16}; \pi_{12} = \frac{3}{16}; \pi_{21} = \frac{3}{16}; \pi_{22} = \frac{1}{16}$ vs. H_1 : pelo menos uma das igualdades não se verifica.

Estatística do Teste: $X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{ab-1}$, sob H_0.

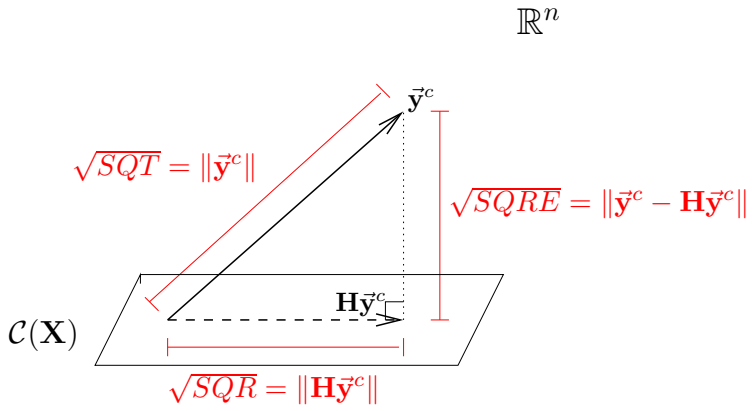
Nível de significância: $\alpha = 0.05$

Região Crítica: Unilateral direita, com a rejeição de H_0 se $X^2_{calc} > \chi^2_{0.05(3)} = 7.815$.

Conclusões: Os valores esperados em cada célula são dados por $E_{ij} = N \times \pi_{ij}$, sendo $N = 408$ (a soma de todas as contagens). Por outras palavras, $E_{11} = 229.5$; $E_{12} = 76.5$; $E_{21} = 76.5$; e $E_{22} = 25.5$. Estes valores esperados são todos muito superiores a 5, pelo que se encontram amplamente verificadas as condições de Cochran que permitem considerar válida a distribuição assintótica $\chi^2_{(3)}$ da estatística de teste. Tem-se $X^2_{calc} = \frac{(248-229.5)^2}{229.5} + \frac{(56-76.5)^2}{76.5} + \frac{(56-76.5)^2}{76.5} + \frac{(48-25.5)^2}{25.5} = 1.491285 + 5.493464 + 5.493464 + 19.85294 = 32.33115$. A rejeição da Hipótese Nula é claríssima, pelo que o mecanismo genético descrito no enunciado não é admissível. A rejeição resulta em grande medida da célula (2,2), que revela uma associação positiva muito maior do que seria de esperar ao abrigo das probabilidades associadas ao mecanismo genético hipotizado no enunciado ($48 \gg 25.5$), gerando uma parcela de X^2_{calc} muito grande (19.85294) que só por si já seria suficiente para rejeitar H_0. Assim, o mecanismo genético proposto não é compatível com os dados observados.

IV

1. O triângulo rectângulo pedido é definido no espaço das variáveis ($\mathbb{R}^n$), onde cada um dos n eixos é definido por um dos n indivíduos observados e onde cada variável corresponde (através dos seus valores nos n indivíduos observados) a um vector. O triângulo de interesse mais imediato para a regressão tem por hipotenusa o vector *centrado* das n observações da variável resposta, ou seja, o vector $\vec{\mathbf{y}}^c$, cujo elemento genérico é $y_i - \overline{y}$. Esse vector tem norma $\|\vec{\mathbf{y}}^c\| = \sqrt{SQT}$ (já que a norma dum vector é a raíz quadrada da soma de quadrados dos seus elementos). A projecção ortogonal desse vector sobre o subespaço $\mathcal{C}(\mathbf{X})$, gerado pelas colunas da matriz do modelo $\mathbf{X}$, é o vector $\mathbf{H}\vec{\mathbf{y}}^c$ (onde $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ é a matriz de projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$), cujo elemento genérico é $\hat{y}_i - \overline{y}$, e cuja norma será portanto $\|\mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{SQR}$. A diferença entre esses dois vectores (que define o outro cateto do triângulo rectângulo resultante da projecção ortogonal de $\vec{\mathbf{y}}^c$) é o vector de elemento genérico $y_i - \hat{y}_i$, ou seja, o vector dos resíduos das n observações. A sua norma será portanto $\|\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{SQRE}$. O triângulo referido é ilustrado em baixo. A fórmula fundamental da regressão ($SQT = SQR + SQRE$) resulta de aplicar o Teorema de Pitágoras a este triângulo.



A proporção da variabilidade de Y explicada pela regressão é o Coeficiente de Determinação $R^2 = \frac{SQR}{SQT}$. A proporção da variabilidade de Y não explicada pela regressão é $1 - R^2 = 1 - \frac{SQR}{SQT} = \frac{SQRE}{SQT}$. A razão referida no enunciado é assim $\frac{R^2}{1-R^2} = \frac{SQR}{SQRE}$. Este quociente (que aparece na estatística do teste F de ajustamento global) é a razão entre o quadrado do comprimento do cateto adjacente ao ângulo θ, formado pelo vector $\vec{\mathbf{y}}^c$ e a sua projecção ortogonal $\mathbf{H}\vec{\mathbf{y}}^c$, e o quadrado do comprimento do cateto oposto. Assim, a razão referida é o quadrado da co-tangente desse ângulo θ.

2. Submodelo com k preditores.

   (a) Pelo formulário sabemos que $R^2_{mod} = 1 - \frac{QMRE}{QMT}$ (sendo $QMT = \frac{SQT}{n-1} = s^2_y$ igual nos dois modelos). Logo, o R^2 modificado do submodelo só pode ser maior que o do modelo completo se $\frac{QMRE_s}{QMT} < \frac{QMRE_c}{QMT}$, ou seja, se $QMRE_s < QMRE_c$. Uma vez que o Quadrado Médio Residual estima a variância σ^2 dos erros aleatórios, está mostrado o que era pedido no enunciado.

   (b) A estatística do teste F parcial é $F = \frac{\frac{SQRE_s - SQRE_c}{p-k}}{\frac{SQRE_c}{n-(p+1)}}$ (formulário). Logo, $F < 1$ equivale a dizer que $\frac{SQRE_s - SQRE_c}{p-k} < \frac{SQRE_c}{n-(p+1)} \Leftrightarrow (SQRE_s - SQRE_c)\,[n-(p+1)] < SQRE_c\,[p-k]$. Juntando as Somas de Quadrados Residuais de cada modelo em lados opostos da desigualdade, fica $SQRE_s\,[n-(p+1)] < SQRE_c\,[\underbrace{(\not{p}-k) + n - (\not{p}+1)}_{=n-(k+1)}]$. Dividindo pelas constantes obtém-se o resultado pedido: $\frac{SQRE_s}{n-(k+1)} < \frac{SQRE_c}{n-(p+1)} \Leftrightarrow QMRE_s < QMRE_c$. Assim, são equivalentes as três condições referidas nesta alínea e na anterior: $F < 1$; $QMRE_s < QMRE_c$; e $R^2_{mod(c)} < R^2_{mod(s)}$.

   (c) Sabemos que nos algoritmos de exclusão sequencial apenas são comparados modelos e submodelos que diferem numa única variável preditora. Sabemos também que nessas comparações, os testes *t-Student* são equivalentes a testes F parciais, verificando-se a relação entre as respectivas estatísticas de teste $T^2 = F$. Uma consulta rápida às tabelas da distribuição F revela que as potenciais fronteiras das regiões críticas dos testes F parciais são *sempre* maiores ou iguais a 1, quaisquer que sejam os valores de n, p e α (sensato) considerados. Logo, se $F_{calc} < 1$, é garantido que não haverá rejeição da Hipótese Nula, ou seja, o submodelo e o modelo não serão considerados significativamente diferentes e, por conseguinte, optar-se-á sempre pelo submodelo. Das alíneas anteriores decorre então que, caso no decurso do algoritmo se verifique estarmos perante um submodelo com $QMRE$ menor que o modelo anterior, ou com um valor superior do R^2 modificado, o algoritmo irá sempre optar por esse submodelo. **Atenção:** No caso de mais de um submodelo verificar estas condições, a opção irá naturalmente recair apenas sobre um deles. Por outro lado, $F < 1$ é apenas condição suficiente, mas não necessária, para que não haja rejeição de H_0. Podem não se verificar as três condições equivalentes referidas e, mesmo assim, não se rejeitar H_0, optando-se pelo submodelo.