INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

27 Janeiro 2020 **Segunda Chamada de Exame 2019-20** **Uma resolução possível**

I

1. Dado o total de $N=2501$ observações, mas sem que tenham sido previamente fixados os totais marginais de qualquer das margens, a pergunta pode ser respondida através dum teste de independência à tabela de contingência com $a=3$ linhas e $b=4$ colunas. A Hipótese Nula é a hipótese de independência e admite que a probabilidade (conjunta) duma observação recair em cada célula da tabela é o produto das probabilidades (marginais) de recair na respectiva linha e coluna, isto é, $H_0 : \pi_{ij} = \pi_{i.} \times \pi_{.j}$, para todo o i e j. A Hipótese Alternativa H_1 é a respectiva negação: existe pelo mens uma célula da tabela para a qual $\pi_{ij} \neq \pi_{i.} \times \pi_{.j}$. A estatística de Pearson é dada por $X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$, com distribuição assintótica $\chi^2_{(a-1)(b-1)}$, caso haja independência (H_0). Rejeita-se H_0 (ao nível $\alpha=0.05$) se $X^2_{calc} > \chi^2_{0.05(6)} = 12.5916$.

2. A dimensão da amostra é adequada se permitir usar a distribuição assintótica. O critério de Cochran visa legitimar essa distribuição assintótica, garantindo uma dimensão mínima para os valores *esperados*: Nenhum $\hat{E}\_ij$ deve ser inferior a 1, e não mais de 20% devem ser inferiores a 5. Para verificar o Critério de Cochran, basta escolher a célula com o menor *valor esperado* estimado e verificar que este excede 5 (**Nota:** O critério de Cochran não diz respeito aos valores *observados*, mas sim aos *esperados*). Essa célula corresponde ao cruzamento da linha (espécie) e coluna (orientação) com menos observações. Trata-se da célula $(3,2)$, para a qual $\hat{E}_{32} = \frac{N_{3.} \times N_{.2}}{N} = \frac{466 \times 366}{2501} = 68.19512 \gg 5$. Logo, é seguro admitir a validade da distribuição assintótica da estatística de Pearson.

3. A contribuição da célula $(3,3)$ para o valor de X^2_{calc} é $\frac{(O_{33}-\hat{E}_{33})^2}{\hat{E}_{33}}$. Tem-se $O_{33} = 243$ e $\hat{E}_{33} = \frac{N_{3.} \times N_{.3}}{N} = \frac{466 \times 484}{2501} = 90.18153$. Logo, a parcela tem valor 258.9608. Este valor é superior ao da soma das restantes 11 parcelas da estatística (que é dada no enunciado: 229.6256). Esse valor enorme resultado duma associação positiva: o número observado de indivíduos nesta célula é muito superior ao que seria de esperar ao abrigo da hipótese de independência. O valor final da estatística do teste é $X^2_{calc} = 488.5864$, pelo que a hipótese de independência é claramente rejeitada (e já o seria apenas com o valor das 11 parcelas dado no enunciado). Esta rejeição é de esperar: uma inspecção visual da tabela indica que a espécie *Zygophyllum simplex* tem uma clara apetência pela orientação a Sul, ao contrário das outras duas espécies que preferem uma orientação a Norte.

II

1. A regressão linear múltipla com $n=109$ observações e $p=4$ preditores.

   (a) Como $R^2 = 0.7363$, o modelo explica 73.63% da variância observada na variável resposta (teor `brix`). Trata-se dum valor razoavelmente bom.

(b) É pedido um teste a que β_3 seja *negativo*. Não dando o benefício da dúvida a essa hipótese, tem-se $H_0 : \beta_3 \geq 0$ vs. $H_1 : \beta_3 < 0$. Como o valor fronteira das duas hipóteses é $\beta_3 = 0$, o valor da estatística do teste é dado no enunciado: $T_{calc} = -3.512$ (**Nota:** o *p-value* ao lado diz respeito a um teste com região crítica bilateral, pelo que não é utilizável aqui). Dada a natureza das hipóteses, a região crítica associada a este teste é unilateral esquerda, rejeitando-se H_0 se $T_{calc} < -t_{0.01(104)} = -2.362739$. Assim, rejeita-se H_0 pelo que $b_3 = -0.61539$ pode ser considerado significativamente inferior a zero, sendo legítima a afirmação do enunciado.

(c) O gráfico tem os valores dos resíduos (internamente) estandardizados (R_i) no eixo vertical. Em nenhum caso excedem o valor absoluto 3 (embora duas observações se aproximem desse limiar). Assim, não se pode falar em observações atípicas. No entanto, três observações têm efeito alavanca (cujos valores definem o eixo horizontal e medem o grau de atracção de cada ponto sobre a hipersuperfície ajustada) superior a 0.15, mais de três vezes superior ao valor médio $\overline{h} = \frac{p+1}{n} = 0.04587$. Entre estas observações, apenas uma (a 102) tem valor de R_i distante de zero, razão pela qual a sua distância de Cook é elevada (ver no formulário a expressão de D_i) e está já próxima do limiar 0.5. A influência mede o impacto que a exclusão da observação em questão tem sobre a hipersuperfície ajustada, e tende a crescer com o afastamento duma observaçao em relação ao centro de gravidade da nuvem de pontos. Ora, a observação 102 é extrema em três das variáveis preditoras (tendo o menor rendimento e acidez, e o maior pH, nas 109 observações), e nas restantes duas variáveis tem valores num dos quartis extremos (entre o mínimo e o primeiro quartil no peso do bago e entre o terceiro quartil e o máximo na variável resposta `brix`). A observação 102 tem um impacto importante no ajustamento do modelo, sendo conveniente inspeccioná-la com mais atenção.

2. Regressão linear simples de `brix` (y) sobre `pH` (x).

(a) Pede-se um teste F parcial para comparar o modelo completo do ponto anterior com o submodelo de regressão linear simples (logo $k = 1$) de `brix` sobre `pH`. A Hipótese Nula do teste corresponde à igualdade dos dois modelos, $H_0 : \mathcal{R}_c^2 = \mathcal{R}_s^2$ e $H_1 : \mathcal{R}_c^2 > \mathcal{R}_s^2$. A estatística do teste pode ser escrita como $F = \frac{n-(p+1)}{p-k} \frac{R_c^2 - R_s^2}{1-R_c^2}$, cuja distribuição sob H_0 é $F_{[p-k,n-(p+1)]}$. Rejeita-se H_0 se $F_{calc} > f_{0.05(3,104)} \approx 2.7$. Para calcular o valor da estatística, será necessário conhecer o valor do coeficiente de determinação do submodelo, R_s^2. Tratando-se dum modelo de regressão linear simples, esse valor é o quadrado do coeficiente de correlação linear entre variável resposta e preditor, que consta do enunciado. Assim, $R_s^2 = 0.8305^2 = 0.6897$. Tem-se $F_{calc} = 6.1222$, pelo que se rejeita H_0 ao nível $\alpha = 0.05$. O submodelo ajusta-se significativamente pior do que o modelo completo.

(b) Do formulário consta a expressão para o efeito alavanca duma observação numa regressão linear simples: $h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2}$. Tem-se $n = 109$; $x_{102} = 3.93$; $\overline{x} = 3.684495$; e $s_x^2 = 0.075136^2 = 0.005645418$. Logo, $h_{102,102} = 0.1080$, cerca de metade do valor correspondente no modelo de regressão linear múltipla do ponto anterior. No entanto, a distância de Cook é de novo próxima do limiar 0.5. De facto, pela expressão para D_i (ver formulário), $D_{102} = R_{102}^2 \cdot \frac{h_{102,102}}{1-h_{102,102}} \cdot \frac{1}{2} = 0.404$, que continua a ser assinalável.

3. Regressão linear simples de `brix` (y) sobre `acidez` (x).

(a) Tratando-se duma regressão linear simples, o coeficiente de correlação entre x e y é uma das raízes quadradas do coeficiente de determinação. É a raíz negativa, pois o declive negativo da recta ($b_1 = -0.9263$) indica que se trata duma relação decrescente. Assim, tem-se $r_{xy} = -\sqrt{R^2} = -\sqrt{0.1005} = -0.3170$.

(b) O teste de ajustamento global tem como Hipótese Nula $H_0 : \mathcal{R}^2 = 0$ (com $H_1 : \mathcal{R}^2 > 0$). A estatística do teste (no contexto duma regressão linear simples) é $F = (n-2) \cdot \frac{R^2}{1-R^2}$, com distribuição $F_{[1,n-2]}$ se H_0 verdadeira. A região crítica é unilateral direita, rejetando-se H_0 se $F_{calc} > f_{0.05(1,107)} \approx 3.94$. Ora $F_{calc} = 11.95497$, pelo que se rejeita H_0, apesar do valor muito baixo de R^2. Tal facto não é contraditório, uma vez que o teste de ajustamento global apenas permite afirmar que $R^2 = 0.1005$ é significativamente diferente de zero, e não que o modelo ajustado seja um bom modelo.

III

1. Uma vez que nada permite associar terrenos de ambientes diferentes, este delineamento deve ser considerado hierarquizado (a cada ambiente, os seus terrenos), com dois factores: ambiente (Factor dominante A, com $a = 8$ níveis) e terrenos (Factor subordinado B, com $b_i = 9$ níveis para todos os ambientes). O delineamento é equilibrado, com $n_c = 6$ repetições em cada uma das $\sum_{i=1}^{a} b_i = 72$ situações experimentais, para um total de $n = 6 \times 72 = 432$ observações.

   Equação do Modelo: $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, onde $i = 1, ..., 8$ indica ambiente; $j = 1, ..., 9$ terreno (dentro do ambiente); $k = 1, ..., 6$ repetição (dentro da situação experimental); Y_{ijk} indica o rendimento da k-ésima repetição no terreno j do ambiente i; ϵ_{ijk} é o correspondente erro aleatório. Com as restrições $\alpha_1 = 0$ e $\beta_{1(i)} = 0$ para qualquer i, μ_{11} representa o rendimento médio populacional no primeiro terreno do primeiro ambiente; α_i indica o efeito associado ao ambiente i; e $\beta_{j(i)}$ indica o efeito associado ao j-ésimo terreno do ambiente i.

   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 dois tipos de efeitos (do factor ambiente e do factor terreno) o quadro de síntese terá três linhas (uma para cada tipo de efeito, e ainda a linha associada à variabilidade residual), sem contar com a linha correspondente à variabilidade Total. Há dois valores dados no enunciado: o Quadrado Médio Residual, QMRE=2.2347 e SQA=1666.2. Os graus de liberdade são: $a-1=7$ (Factor A); $\sum_{i=1}^{a}(b_i–1) = 64$ (Factor B) e $n–\sum_{i=1}^{a} b_i = 432–72 = 360$ (Residual). Assim, tem-se $QMA = \frac{SQA}{a-1} = 238.0286$, donde $F^{A}_{calc} = \frac{QMA}{QMRE} = 106.5148$; $SQRE = \left(n - \sum_{i=1}^{a} b_i\right) \times QMRE = 804.492$. A Soma de Quadrados associada ao factor B resulta do facto de $SQB(A) = SQT - (SQA + SQRE) = (n-1)\, s_y^2 - (1666.2 + 804.492) = 431 \times 6.05404 - 2470.692 = 2609.291 - 2470.692 = 138.5992$. O respectivo Quadrado Médio é $QMB(A) = \frac{SQB(A)}{\sum_{i=1}^{a}(b_i-1)} = 2.165612$. Finalmente, a estatística do teste aos efeitos do factor subordinado é $F^{B(A)}_{calc} = \frac{QMB(A)}{QMRE} = 0.969084$. Eis a tabela-resumo:

Fontes de Variação	gl	Somas de Quadrados	Quadrados Médios	F_{calc}
Factor Ambiente (Factor A)	7	1666.2	238.0286	106.5148
Factor Terreno (Factor B(A))	64	138.5992	2.165612	0.969084
Residual	360	804.492	2.2347	—
Total	431	2609.291	–	–

3. Neste modelo há dois testes F de interesse, um para os efeitos de cada factor. No teste aos efeitos de ambiente, as hipóteses são $H_0 : \alpha_i = 0, \forall\ i$ e $H_1 : \exists\, i,$ tal que $\alpha_i \neq 0$. A estatística de teste é

$F^A = \frac{QMA}{QMRE} \frown F_{[a-1\,,\,n-\sum_{i=1}^a b_i]}$, sob H_0. A regra de rejeição ao nível de significância $\alpha = 0.05$ é rejeitar H_0 se $F_{calc} > f_{0.05(7,360)} \approx 2.02$. Como $F^A_{calc} = 106.5148$, há uma claríssima rejeição de H_0, ou seja, conclui-se claramente pela existência de efeitos dos ambientes sobre o rendimento. No teste aos efeitos de terreno, a Hipótese Nula $H_0 : \beta_{j(i)} = 0$ para todos os terrenos e ambientes (sendo H_1 que existe i, j tal que $\beta_{j(i)} = 0$) não é rejeitada. O valor calculado da estatística, $F^{B(A)} = 0.969084$, é inferior a 1, logo inferior a qualquer valor tabelado que possa constituir a fronteira da região crítica (que, para $\alpha = 0.05$, é $f_{0.05(64,360)} \approx 1.32$). Assim, conclui-se que a variabilidade de rendimentos ao longo dos terrenos não é significativa, uma vez considerada a variabilidade ao longo dos ambientes estudados, pelo que a cosnideração do factor subordinado não parece justificar-se.

4. Duas médias populacionais de rendimento, em dois diferentes terrenos (de qualquer ambiente) podem ser consideradas diferentes (ou seja, rejeita-se $\mu_{ij} = \mu_{i'j'}$ a favor de $\mu_{ij} \neq \mu_{i'j'}$) se se verificar a desigualdade $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > \tau_{\alpha(\sum\limits_i b_i\,,\,n-\sum\limits_i b_i)} \sqrt{\frac{QMRE}{n_c}}$. No cálculo do termo de comparação tem-se $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{2.2347}{6}} = 0.6102868$. Usando o nível global de significância $\alpha = 0.05$, tem-se $\tau_{0.05(72,360)} = 5.939$ (valor dado no enunciado, já que corresponde a parâmetros muito distantes dos disponíveis nas tabelas). Logo, o limiar de significância é dado por $5.939 \times 0.6102868 = 3.624493$. O menor rendimento médio amostral no Ambiente 2 corresponde ao terreno 1, e é $\overline{y}_{21.} = 4.873$. O maior rendimento médio corresponde ao terreno 6, e é $\overline{y}_{26.} = 8.617$. A diferença entre essas duas médias amostrais é $8.617 - 4.873 = 3.744 > 3.624493$, logo trata-se duma diferença significativa (embora por pouco) ao nível $\alpha = 0.05$. Esta conclusão parece contraditória com o resultado do teste F à existência de efeitos de terreno. Tal facto é possível, uma vez que os resultados teóricos subjacentes aos testes de Tukey e aos testes F são diferentes. Além disso, a significância agora detectada é signficativa por pouco (ao nível $\alpha = 0.05$).

5. No caso de haver nove diferentes tipos de terrenos previamente definidos, e de em cada ambiente se seleccionaram os nove terrenos de forma a que cada tipo de terreno esteja representado, estaríamos perante um delineamento de tipo factorial, já que os 8 ambientes estariam cruzados com os 9 tipos de terreno. Uma vez que há repetições para cada uma das 72 situações experimentais resultantes, pode ajustar-se o modelo ANOVA *com* efeitos de interacção, a que corresponde a equação $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, que difere da equação para o modelo hierarquizado na substituição das antigas parcelas $\beta_{j(i)}$ pela soma de dois novos tipos de parcelas: os efeitos de terreno β_j (que correspondem aos efeitos dos $b = 9$ diferentes tipos de terrenos, mas que com a restrição $\beta_1 = 0$ reduzem-se a oito); e os efeitos de interacção $(\alpha\beta)_{ij}$ que correspondem a cada situação experimental (e que com as restrições $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$, serão em número de $(a-1)(b-1) = 56$).

IV

1. Tem-se $y = \frac{1}{1+e^{-(c+d\,x)}}$.

   (a) Logo $1 - y = 1 - \frac{1}{1+e^{-(c+d\,x)}} = \frac{\cancel{1}+e^{-(c+d\,x)}-\cancel{1}}{1+e^{-(c+d\,x)}} = \frac{e^{-(c+d\,x)}}{1+e^{-(c+d\,x)}}$. Dividindo y por $1-y$ tem-se:

$$\frac{y}{1-y} = \frac{\frac{1}{\cancel{1+e^{-(c+d\,x)}}}}{\frac{e^{-(c+d\,x)}}{\cancel{1+e^{-(c+d\,x)}}}} = \frac{1}{e^{-(c+d\,x)}} = e^{c+d\,x} \; .$$

Logaritmizando, fica $\ln\left(\frac{y}{1-y}\right) = c + d\,x$, ou seja, o *logit* de y está linearmente relacionado com o preditor x.

(b) A taxa de variação relativa pedida no enunciado é o quociente $\frac{y'(x)}{y(x)}$, sendo por isso necessário calcular a derivada $y'(x)$. Ora,

$$\begin{aligned} y'(x) &= [(1+\mathrm{e}^{-(c+dx)})^{-1}]' \;=\; (-1)[1+\mathrm{e}^{-(c+dx)}]^{-2}(1+\mathrm{e}^{-(c+dx)})' \\ &= (-1)[1+\mathrm{e}^{-(c+dx)}]^{-2}\mathrm{e}^{-(c+dx)}(-d) \;=\; \frac{d\,\mathrm{e}^{-(c+dx)}}{(1+\mathrm{e}^{-(c+dx)})^2}\,. \end{aligned}$$

Dividindo por $y(x)$ obtém-se a taxa de variação relativa:

$$\frac{y'(x)}{y(x)} = \frac{\frac{d\,\mathrm{e}^{-(c+dx)}}{(1+\mathrm{e}^{-(c+dx)})^{\cancel{2}}}}{\frac{1}{\cancel{1+\mathrm{e}^{-(c+dx)}}}} = \frac{d\,\mathrm{e}^{-(c+dx)}}{1+\mathrm{e}^{-(c+dx)}} = d\,[1-y(x)]\;,$$

tendo em conta a expressão para $1-y(x)$ deduzida na alínea anterior.

2. (a) O vector $(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = \vec{\mathbf{Y}} - \hat{\vec{\mathbf{Y}}}$ tem como elemento genérico $y_i - \hat{y}_i$, ou seja, o resíduo de cada observação (por outras palavras, $(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{E}}$ é o vector dos resíduos). Ora, a norma de qualquer vector é a raíz quadrada da soma dos quadrados dos elementos do vector. Logo, $\|(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}\|^2 = \|\vec{\mathbf{E}}\|^2 = \sum\limits_{i=1}^{n} e_i^2 = SQRE$.

(b) Qualquer produto duma matriz por (à direita) um vector calcula a combinação linear das colunas da matriz, cujos coeficientes são dados pelos elementos do vector. Assim, o vector $\vec{\mathbf{1}}_n$, que é a primeira coluna da matriz do modelo $\mathbf{X}$, resulta do produto $\mathbf{Xv}$ onde $\mathbf{v}^t = (1,0,0,...,0)$ é o vector cujo único elemento não nulo é o 1 na primeira posição. Logo, tem-se $\mathbf{H}\vec{\mathbf{1}}_n = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \cdot \mathbf{Xv} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X})}_{=\mathbf{I}}\mathbf{v} = \mathbf{Xv} = \vec{\mathbf{1}}_n$. (**Nota:** Nas aulas e apontamentos justifica-se que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$ de forma diferente, igualmente aceitável).

O produto $\mathbf{H}\vec{\mathbf{1}}_n$ também define uma combinação linear das colunas da matriz $\mathbf{H}$, sendo todos os coeficientes associados a essas colunas iguais a 1 (o elemento comum a todas as posições do vector $\vec{\mathbf{1}}_n$). Logo, $\mathbf{H}\vec{\mathbf{1}}_n$ é o vector da soma das colunas de $\mathbf{H}$. Em cada posição do vector $\mathbf{H}\vec{\mathbf{1}}_n$ estará a soma dos elementos da linha correspondente de $\mathbf{H}$. Como $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, essas somas são todas iguais e 1.

(c) A média das observações de y pode ser calculada como $\overline{Y} = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i = \frac{1}{n}\vec{\mathbf{1}}_n^t\vec{\mathbf{Y}}$, já que o produto interno do vector $\vec{\mathbf{1}}_n$ com qualquer outro vector soma os elementos desse outro vector. De forma análoga, a média dos valores ajustados ($\hat{Y}_i$) resulta de considerar $\overline{\hat{Y}} = \frac{1}{n}\vec{\mathbf{1}}_n^t\hat{\vec{\mathbf{Y}}} = \frac{1}{n}\vec{\mathbf{1}}_n^t\mathbf{H}\vec{\mathbf{Y}}$. Mas $\vec{\mathbf{1}}_n^t\mathbf{H} = (\mathbf{H}\vec{\mathbf{1}}_n)^t$, já que $(\mathbf{H}\vec{\mathbf{1}}_n)^t = \vec{\mathbf{1}}_n^t\mathbf{H}^t$ e a matriz de projecção ortogonal $\mathbf{H}$ é uma matriz simétrica. Logo, $\overline{\hat{Y}} = \frac{1}{n}(\mathbf{H}\vec{\mathbf{1}}_n)^t\vec{\mathbf{Y}} = \frac{1}{n}(\vec{\mathbf{1}}_n)^t\vec{\mathbf{Y}} = \overline{Y}$.

(d) Tem-se $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$, logo cada valor ajustado $\hat{Y}_j$ é dado pelo elemento correspondente do produto $\mathbf{H}\vec{\mathbf{Y}}$, que corresponde ao produto interno da linha j de $\mathbf{H}$ com o vector das observações $\vec{\mathbf{Y}}$, ou seja, $\hat{Y}_j = \sum\limits_{i=1}^{n} h_{ji}Y_i$. Viu-se na alínea (b) que a soma dos h_{ji} em qualquer linha j é 1, logo $\sum\limits_{i=1}^{n} h_{ji} = 1$, pelo que $\hat{Y}_j$ é uma média ponderada de todas as observações Y_i, sendo os pesos dados pelos coeficientes h_{ji}. A contribuição da própria observação Y_j para o correspondente valor ajustado $\hat{Y}_j$ tem peso h_{jj}, que é o efeito alavanca associado à observação Y_j.