INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

2 de Novembro, 2015 **PRIMEIRO TESTE 2015-16** **Uma resolução possível**

I

Tem-se uma tabela de contingências de dimensão 3×4. Apenas o número total de observações foi fixado pelo experimentador, sendo as duas margens da tabela de contingência livres.

1. O problema colocado corresponde a um teste de independência, em que se procura saber se os dois critérios de classificação (composição da vegetação e percentagem de rocha) estão relacionados, ou se são independentes.

   Hipóteses: Represente-se por π_{ij} a probabilidade de um local observado ser do Grupo i ($i = 1, 2, 3$) e com percentagem de rocha na classe j ($j = 1, 2, 3, 4$). Seja $\pi_{i.}$ a probabilidade marginal de uma localidade ser do Grupo i e $\pi_{.j}$ a probabilidade marginal dum local ter percentagem de rocha na classe j. As hipóteses em confronto são:

   $$H_0 : \pi_{ij} = \pi_{i.} \times \pi_{.j} \quad \forall\, i, j \qquad vs. \qquad H_1 : \exists\, i, j \text{ tais que } \pi_{ij} \neq \pi_{i.} \times \pi_{.j} \ .$$

   H_0 corresponde à hipótese de independência entre os dois factores de classificação.

   Estatística do Teste: É a estatística de Pearson, $X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$, sendo $a = 3$, $b = 4$, O_{ij} o número de observações na célula (i, j) e $\hat{E}_{ij}$ os valores esperados ao abrigo da hipótese de independência, estimados a partir das frequências relativas marginais de linhas e colunas. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$ com, no nosso caso, $(a-1)(b-1) = 6$ graus de liberdade.

   Região Crítica: (Unilateral direita) Para um nível de significância α = P[Erro do tipo I] = P[Rej. H_0 | H_0 verdade], a regra de rejeição deve ser a de rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{\alpha[(a-1)(b-1)]} = \chi^2_{\alpha(6)}$.

2. Pede-se para verificar a validade das condições de Cochran, que indicam condições suficientes de dimensão da amostra para que se possa considerar válida a distribuição assintótica da estatística do teste. Essas condições exigem que, em nenhuma das células da tabela, o valor esperado seja inferior a 1 e que não seja inferior a 5 em mais de 20% das células. Ora, tendo em conta que $\hat{E}_{ij} = \frac{N_{i.} \times N_{.j}}{N}$, o mais pequeno valor esperado estimado tem de corresponder à linha com menor dimensão de amostra ($i = 3$, com $N_{3.} = 29$) e à coluna com menor frequência ($j = 2$, com $N_{.2} = 23$). Este menor valor esperado estimado é $\hat{E}_{32} = \frac{29 \times 23}{175} = 3.811429 < 5$. O segundo menor valor esperado estimado corresponde à célula $(3, 3)$, com $\hat{E}_{33} = \frac{29 \times 30}{175} = 4.971429 \approx 5$. Assim, verifica-se que em apenas duas das 12 células se verifica $\hat{E}_{ij} < 5$, o que corresponde a uma proporção de $0.16667 < 0.20$ (e numa das referidas células o valor esperado estimado é praticamente 5). Assim, pode considerar-se que a dimensão da amostra é suficientemente grande para aceitar a validade da distribuição assintótica.

3. Pede-se para verificar as conclusões do teste com o conjunto de dados no enunciado. Tem-se $X^2_{calc} = 24.675$, que tem de ser comparado com a fronteira duma região crítica ao nível

$\alpha = 0.05$, numa distribuição χ^2_6. Pelas tabelas, tem-se $\chi^2_{0.05(6)} = 12.592$, pelo que se rejeita a hipótese nula, ou seja, rejeita-se a hipótese de independência. Para um nível de significância $\alpha = 0.001$, a fronteira da região crítica correspondente seria $\chi^2_{0.001(6)} = 22.458$. Isto significa que $P[\chi^2_6 > 22.458] = 0.001$. Com esta fronteira de região crítica continua-se a verificar a rejeição da hipótese nula, uma vez que $X^2_{calc} > \chi^2_{0.001(6)} = 22.458$. E conclui-se também que área à direita de $X^2_{calc} = 24.675$ tem de ser inferior a 0.001, ou seja, $P[\chi^2_6 > 24.675] < 0.001$, o que equivale a dizer que o valor de prova tem de verificar $p < 0.001$. Assim, pode concluir-se com convicção que a composição da vegetação está associada à percentagem de rocha nas margens dos rios estudados.

4. Pede-se o valor, na estatística do teste, da soma das três parcelas associadas à classe $]40, 80]$ de percentagem de rocha nas margens dos rios, ou seja, à soma das parcelas correspondentes às células $(1, 4)$, $(2, 4)$ e $(3, 4)$. Ora,

$$\begin{aligned}
\hat{E}_{14} &= \tfrac{N_{1.} \times N_{.4}}{N} = \frac{100 \times 33}{175} = 18.85714 \\
\hat{E}_{24} &= \tfrac{N_{2.} \times N_{.4}}{N} = \frac{46 \times 33}{175} = 8.674286 \ . \\
\hat{E}_{34} &= \tfrac{N_{3.} \times N_{.4}}{N} = \frac{29 \times 33}{175} = 5.468571 \ .
\end{aligned}$$

Logo, a soma das três parcelas é:

$$\begin{aligned}
\sum_{i=1}^{3} \frac{(O_{i4} - \hat{E}_{i4})^2}{\hat{E}_{i4}} &= \frac{(O_{14} - \hat{E}_{14})^2}{\hat{E}_{14}} + \frac{(O_{24} - \hat{E}_{24})^2}{\hat{E}_{24}} + \frac{(O_{34} - \hat{E}_{34})^2}{\hat{E}_{34}} \\
&= \frac{(8 - 18.85714)^2}{18.85714} + \frac{(14 - 8.674286)^2}{8.674286} + \frac{(11 - 5.468571)^2}{5.468571} \\
&= 6.25108 + 3.269806 + 5.59501 = 15.1159 \ .
\end{aligned}$$

Assim, a classe das margens mais rochosas é responsável por mais de metade do valor calculado da estatística do teste, o que só por si coloca o valor da estatística acima do limiar da região crítica, ao nível $\alpha = 0.05$. Note-se como, para a globalidade das observações, o número total de locais no Grupo 1 (de composição de vegetação) é bastante superior ao dos Grupos 2 e 3, enquanto que na classe de margens mais rochosas esta relação é a inversa. Está identificada uma causa importante para a inexistência de independência entre os dois critérios de classificação.

II

1. Sabemos que numa regressão linear simples, os graus de liberdade associados ao quadrado médio residual são $n - 2$. Pelo enunciado verifica-se que este valor é 93, logo o estudo baseou-se em $n = 95$ observações.

2. Numa regressão linear simples, o coeficiente de determinação é o quadrado do coeficiente de correlação amostral entre o preditor (x, no nosso caso `log(LCP)`) e a variável resposta (y, no nosso caso `log(q)`). Logo, o coeficiente de correlação amostral r_{xy} é uma das raízes quadradas do coeficiente de determinação, que é indicado na listagem: $R^2 = 0.775$. Falta saber o sinal dessa raíz. O gráfico indica que estamos perante uma relação decrescente entre q e LCP, logo uma relação decrescente entre `log(q)` e `log(LCP)` (o logaritmo é uma função crescente). Assim, a raíz relevante de R^2 é a raíz negativa: $r_{xy} = -\sqrt{R^2} = -\sqrt{0.775} = -0.88034$. Alternativamente, o sinal do coeficiente de correlação pode ser determinado recordando que, numa regressão linear simples, o sinal de r_{xy} é sempre igual ao sinal do declive da recta de regressão, que no nosso caso é negativo ($b_1 = -0.94603$).

3. O coeficiente de determinação é $R^2 = 0.775$, o que significa que a regressão linear explica $77,5\%$ da variabilidade nos valores observados do log-rendimento quântico da fotossíntese (`log(q)`). Este valor é relativamente bom em si mesmo, e é significativamente melhor do que o valor $R^2 = 0$ associado ao Modelo Nulo, como se pode verificar através dum teste F de ajustamento global:

   Hipóteses: $H_0 : \mathcal{R}^2 = 0 \quad vs. \quad H_1 : \mathcal{R}^2 > 0.$

   Estatística do Teste: $F = \frac{QMR}{QMRE} = (n-2)\,\frac{R^2}{1-R^2} \;\cap\; F_{(1,n-2)}$, 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 $F_{calc} > f_{\alpha[1,93]} \;\approx\; 3.95$.

   Conclusões: Tem-se $F_{calc} = 320.3 \gg 3.95$. Logo há uma clara rejeição de H_0, i.e., a recta de regressão não é inútil para prever o log-rendimento quântico da fotossíntese, a partir dos valores de log-`LCP`.

4. A transformação utilizada corresponde à transformação linearizante dum modelo potência $y = c\,x^d$. De facto, a equação de base do modelo estudado é $\ln(y) = b_0 + b_1 \ln(x)$. Exponenciando os dois lados desta equação, e tendo em conta as propriedades de exponenciais e logaritmos, obtém-se: $y = \mathrm{e}^{b_0 + b_1 \ln(x)} = \mathrm{e}^{b_0}\,\mathrm{e}^{b_1 \ln(x)} = \underbrace{\mathrm{e}^{b_0}}_{=c}\,\mathrm{e}^{\ln(x^{b_1})} = c\,x^{b_1} = c\,x^d$. Assim, o declive da recta ajustada corresponde à potência ($b_1 = d$), enquanto que a ordenada na origem da recta corresponde ao logaritmo natural da constante multiplicativa na equação potência ($b_0 = \ln(c)$). A equação potência que relaciona directamente as variaveis originais é, no nosso caso, $y = \mathrm{e}^{0.87106}\,x^{-0.94603} = 2.389442\,x^{-0.94603} = \frac{2.389442}{x^{0.94603}}$.

5. O enunciado pergunta se é admissível considerar que $y = \frac{\alpha}{x}$. Tendo em conta a resposta da alínea anterior, vemos que um tal modelo corresponde a admitir que, na transformação linearizada, o declive teórico da recta de regressão entre $\ln(y)$ e $\ln(x)$ seria $\beta_1 = -1$ (sendo $b_1 = -0.94603$ a estimativa amostral desse valor). Eis o teste de hipóteses pedido:

   Hipóteses: $H_0 : \beta_1 = -1 \quad vs. \quad H_1 : \beta_1 \neq -1.$

   Estatística do Teste: $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \;\cap\; t_{n-2}$, 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: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(93)} \;\approx\; 1.985$.

   Conclusões: Tem-se $T_{calc} = \frac{-0.94603-(-1)}{0.05286} = 1.021$. Este valor não pertence à região crítica, logo não se rejeita H_0. Não se pode excluir a hipótese de proporcionalidade inversa entre `q` e `LCP` referida no enunciado.

6. Pede-se um intervalo de predição (95%) para um valor de `q` associado ao valor `LCP`$= 100$. Com base na recta de regressão entre as variáveis logaritmizadas pode construir-se um intervalo de predição para `log(q)`, quando `log(LCP)`$= \ln(100) = 4.60517$. Indicando os valores de log-`LCP` por x^*, este intervalo de predição tem extremos (ver formulário): $(b_0 + b_1\,x^*) \pm t_{0.025(n-2)}\sqrt{QMRE\,\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x^*})^2}{(n-1)s^2_{x^*}}\right]}$, sendo conhecidos a partir do enunciado os seguintes valores: $b_0 = 0.87106$, $b_1 = -0.94603$ (logo $\hat{\mu}_{Y|x^*} = b_0 + b_1 \ln(100) = -3.485569$), $\sqrt{QMRE} = 0.7618$, $n = 95$, $\overline{x^*} = 3.5152$, $s^2_{x^*} = 2.2098$. Já se viu que $t_{0.025(93)} \approx 1.985$. Substituindo, obtém-se o intervalo de predição para uma observação de `log(q)` correspondente a `log(LCP)`$= \ln(100)$: $]\,-5.010022\;,\;-1.961116\,[$. Exponenciando, obtém-se o correspondente intervalo de predição

para o rendimento quântico da fotossíntese (q), se LCP= 100:] 0.006671 , 0.1407013 [. Este intervalo é coerente com a nuvem de pontos de q *vs.* LCP no enunciado.

7. O diagrama ao lado da listagem de resultados da regressão é um diagrama em que ao eixo horizontal correspondem os valores h_{ii} do efeito alavanca de cada observação, e ao eixo vertical correspondem os respectivos resíduos (internamente) estandardizados, R_i. No canto inferior direito do gráfico surge uma curva correspondente a distâncias de Cook iguais a 0.5. As distâncias de Cook são uma medida da influência duma observação, ou seja, do impacto que a exclusão dessa observação teria no ajustamento da regressão. Quanto maior fôr a distância de Cook, maior a influência da observação e convenciona-se considerar o valor 0.5 como um "limiar de alarme" para observações excessivamente influentes. No nosso caso, os resíduos estandardizados estão quase todos no intervalo $]-3,3[$, indicando que se trata de resíduos sem magnitude especialmente elevada. Apenas uma observação (no canto inferior direito do gráfico, com o número $i = 87$) tem um resíduo abaixo de -3, mas mesmo assim, pouco abaixo desse limiar ($R_{87} \approx -3.5$). Os efeitos alavanca estão todos compreendidos no intervalo entre o seu menor valor possível, $\frac{1}{n} = \frac{1}{95} = 0.01052632$ e aproximadamente 0.08. Este valor não é muito superior ao valor médio dos efeitos alavanca, $\overline{h} = \frac{2}{n} = \frac{2}{95} = 0.02105263$, e está muito longe do valor máximo possível, que é 1. Assim, nada de especial se destaca no que respeita aos efeitos alavanca. No entanto, há uma observação que, ao conjugar o maior valor do efeito alavanca ($h_{ii} \approx 0.08$) e o maior (em módulo) resíduo internamente estandardizado ($R_i \approx -3.5$) acaba por ter também um valor elevado da distância de Cook, e - facto mais importante - esse valor é tão elevado que supera o limiar de guarda (0.5) usualmente associado a este indicador. Trata-se novamente da observação 87 no canto inferior direito do gráfico. É possível identificar o respectivo valor de LCP, uma vez que sabemos tratar-se da observação com maior efeito alavanca e sabemos que, numa regressão linear simples, o maior efeito alavanca está associado ao indivíduo cujo valor na variável preditora está mais afastado da média desses mesmos valores. Adaptando a fórmula para os efeitos alavanca (ver formulário) à notação usual para a existência de transformações de variáveis, temos $h_{ii} = \frac{1}{n} + \frac{(x_i^* - \overline{x^*})^2}{(n-1)\, s_{x^*}^2}$. Ora a média dos log(LCP) é, pelo enunciado, $\overline{x^*} = 3.5152$. O valor de log(LCP) mais afastado desta média só pode ser um dos dois valores extremos de LCP. Sendo o logaritmo natural uma função crescente, e uma vez que são dados no enunciado os valores extremos de LCP, é possível identificar esse valor. Tem-se $\ln(0.7336) = -0.30979$ e $\ln(849.0622) = 6.74413$. Entre estes dois valores, o mais distante da média $\overline{x^*} = 3.5152$ é $\ln(0.7336)$, que corresponde ao menor LCP, pelo que a resposta à pergunta do enunciado é LCP= 0.7336.

III

1. A partir do formulário sabemos que $\sigma^2_{\hat{\beta}_0} = V[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]$. Sabemos que o desvio padrão é a raíz quadrada da variância, e que o valor desconhecido σ^2 é, na regressão linear simples, estimado pelo Quadrado Médio Residual, $QMRE = \frac{SQRE}{n-2}$. Logo, $\hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]}$.

2. Trata-se dum procedimento totalmente análogo ao usado nas aulas teóricas para o caso do intervalo de confiança do declive β_1, e que segue os passos indicados no texto de apoio *Algumas demonstrações de resultados nos acetatos das Teóricas*. Concretamente, e já que $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \cap t_{n-2}$, designando por $t_{\alpha/2(n-2)}$ o valor que, numa distribuição t_{n-2} deixa à sua direita uma região de probabilidade $\alpha/2$, e uma vez que o simétrico desse valor, $-t_{\alpha/2(n-2)}$, será (dada a simetria da

distribuição t-Student em torno de zero) o valor que deixa à sua *esquerda* uma área $\alpha/2$, pode escrever-se:

$$P\left[-t_{\alpha/2(n-2)} \;<\; \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \;<\; t_{\alpha/2(n-2)}\right] \;=\; 1-\alpha \;.$$

Substituindo a dupla desigualdade por outras duplas desigualdades equivalentes não altera a probabilidade $1-\alpha$. Vamos efectuar essas substituições com o objectivo de deixar o parâmetro para o qual se pretende construir o intervalo de confiança (β_0) sozinho no meio duma dupla desigualdade. Tem-se (primeiro multiplicando a dupla desigualdade por $\hat{\sigma}_{\hat{\beta}_0}$, depois por -1 e finalmente somando $\hat{\beta}_0$):

$$\begin{array}{rcccl} & -t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} & < & \hat{\beta}_0-\beta_0 & < t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} \\ \Leftrightarrow & t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} & > & \beta_0-\hat{\beta}_0 & > -t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} \\ \Leftrightarrow & \hat{\beta}_0 - t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} & < & \beta_0 & < \hat{\beta}_0 + t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0}\;. \end{array}$$

Substituindo neste *intervalo aleatório* as variáveis aleatórias dos extremos ($\hat{\beta}_0$ e $\hat{\sigma}_{\hat{\beta}_0}$), pelas correspondentes estimativas na *amostra concreta* obtém-se o intervalo a $(1-\alpha)\times 100\%$ de confiança para β_0: $\left]\; b_0 \,-\, t_{\alpha/2\,(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} \quad , \quad b_0 \,+\, t_{\alpha/2\,(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0}\;\right[$.

3. Sabemos que $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\,\overline{x}$. Logo,

$$\begin{array}{rcl} \mathrm{Cov}[\hat{\beta}_1,\hat{\beta}_0] &=& \mathrm{Cov}[\hat{\beta}_1,\overline{Y}-\hat{\beta}_1\overline{x}] \;=\; \mathrm{Cov}[\hat{\beta}_1,\overline{Y}] - \mathrm{Cov}[\hat{\beta}_1,\hat{\beta}_1\,\overline{x}] \\ &=& 0 \,-\, \overline{x}\mathrm{Cov}[\hat{\beta}_1,\hat{\beta}_1] \;=\; -\overline{x}V[\hat{\beta}_1] \;=\; -\overline{x}\frac{\sigma^2}{(n-1)s_x^2}\;. \end{array}$$

Duas variáveis aleatórias só podem ser independentes se a sua covariância fôr nula (embora isso não seja condição suficiente para garantir a independência). Ora as variáveis aleatórias $\hat{\beta}_0$ e $\hat{\beta}_1$ só têm covariância nula se $\sigma^2 = 0$ ou $\overline{x} = 0$. A primeira destas condições equivale a dizer que os erros aleatórios têm sempre variância nula, ou seja que existe uma relação linear (afim) exacta entre Y e X, situação que não existe num estudo *estatístico*, onde se pressupõe existir oscilação de valores em torno duma tendência de fundo. A única condição eventualmente controlável pelo utilizador e que permite a independência dos dois estimadores é que a média dos valores de x_i observados seja nula.

4. Tendo em conta as propriedades das variâncias, a independência das observações Y_i e o enunciado da alínea anterior, tem-se:

$$\begin{array}{rcl} V[\hat{\beta}_0] &=& V[\overline{Y}-\hat{\beta}_1\,\overline{x}] \;=\; V[\overline{Y}] + V[\hat{\beta}_1\,\overline{x}] - 2\,\mathrm{Cov}[\overline{Y},\hat{\beta}_1\,\overline{x}] \\ &=& V\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] + \overline{x}^2 V[\hat{\beta}_1] - 2\overline{x}\underbrace{\mathrm{Cov}[\overline{Y},\hat{\beta}_1]}_{=0} \\ &=& \frac{1}{n^2}V\left[\sum_{i=1}^{n} Y_i\right] + \overline{x}^2\frac{\sigma^2}{(n-1)s_x^2} \\ &=& \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{V[Y_i]}_{=\sigma^2} + \frac{\overline{x}^2\sigma^2}{(n-1)s_x^2} \\ &=& \frac{1}{n^2}\,n\,\sigma^2 + \frac{\overline{x}^2\sigma^2}{(n-1)s_x^2} \;=\; \sigma^2\left[\frac{1}{n}+\frac{\overline{x}^2}{(n-1)s_x^2}\right] \end{array}$$

Note-se que para garantir que $V\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} V[Y_i]$ é necessário que as variáveis Y_i sejam independentes. Esta passagem é válida porque se exigiu no modelo de regressão linear simples que os erros aleatórios sejam independentes. Da forma análoga, é usada a condição $V[Y_i] = \sigma^2$, para qualquer observação Y_i, que também é válida porque o modelo RLS exige variâncias homogéneas dos erros aleatórios. Trata-se de exemplos de passagens que seriam diferentes para modelos com outros pressupostos.