INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

3 de Novembro, 2014 **PRIMEIRO TESTE 2014-15** **Uma resolução possível**

I

Tem-se uma tabela de contingências de dimensão 4×2.

1. O problema colocado corresponde a um teste de homogeneidade, em que se procura saber se as proporções de frutos defeituosos, ou com valor comercial, são iguais nas quatro variedades.

   Hipóteses: Representando por π_i a probabilidade de um fruto ser defeituoso, se fôr da variedade i, e tendo em conta que os frutos apenas se classificam em duas categorias (defeituosos ou com valor comercial) podemos escrever as hipóteses de teste como

   $$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 \qquad vs. \qquad H_1 : \exists\, i, j \in \{1,2,3,4\} \text{ tal que } \pi_i \neq \pi_j$$

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

   sendo $a = 4$, $b = 2$, O_{ij} o número de observações na célula (i, j) e $\hat{E}_{ij}$ os valores esperados ao abrigo da hipótese de homogeneidade, estimados a partir das frequências relativas marginais de coluna. O enunciado diz que podemos admitir a validade do critério de Cochran, ou seja, a validade da distribuição assintótica acima indicada.

   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 $\chi^2_{\text{calc}} > \chi^2_{\alpha[(a-1)(b-1)]} = \chi^2_{0.05(3)} = 7.815$.

   Conclusões: É dito no enunciado que $X^2_{calc} = 66.3942$. Logo há uma clara rejeição de H_0, i.e., conclui-se que não há homogeneidade na distribuição dos frutos de cada variedade pelas categorias "defeituoso" e "com valor comercial".

   O resultado do teste não surpreende. Uma análise da tabela de valores observados indica que há duas variedades (41A, 35) em que o número de frutos defeituosos é superior, e num caso (35) muito superior, ao de frutos com valor comercial, enquanto que noutras duas variedades (40C e 15A) passa-se o contrário, com diferenças acentuadas. Assim, a hipótese de que a probabilidade dum fruto ser defeituoso fosse igual em todas as variedades dificilmente seria admissível.

2. Pede-se o valor da soma das duas parcelas associadas à variedade 35, correspondentes às células $(4, 1)$ e $(4, 2)$. Ora,

   $$\begin{aligned} \hat{E}_{41} &= \tfrac{N_{4.} \times N_{.1}}{N} = \frac{181 \times 394}{764} = 93.34293 \\ \hat{E}_{42} &= \tfrac{N_{4.} \times N_{.2}}{N} = \frac{181 \times 370}{764} = 87.65707\ . \end{aligned}$$

   Logo,

   $$\begin{aligned} \frac{(O_{41} - \hat{E}_{41})^2}{\hat{E}_{41}} + \frac{(O_{42} - \hat{E}_{42})^2}{\hat{E}_{42}} &= \frac{(135 - 93.34293)^2}{93.34293} + \frac{(46 - 87.65707)^2}{87.65707} \\ &= 18.59071 + 19.79659 = 38.3873\ . \end{aligned}$$

Assim a variedade 35 é responsável por mais de metade do valor calculado da estatística do teste, o que só por si coloca o valor da estatística muito acima do limiar da região crítica. Este valor elevado reflecte a singularidade desta variedade, que é a única em que a grande maioria dos frutos não tem valor comercial.

3. 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 ($i = 1$, com $N_{1.} = 120$) e à coluna com menor frequência ($j = 2$, com $N_{.2} = 370$). Mas este menor valor esperado estimado é muito superior a 5: $\hat{E}_{12} = \frac{120 \times 370}{764} = 58.11518 \gg 5$. Logo, a dimensão da amostra pode considerar-se suficientemente grande, permitindo usar a distribuição assintótica.

II

1. (a) Sabemos que numa regressão linear simples, o coeficiente de determinação é o quadrado do coeficiente de correlação amostral entre o preditor (x, no nosso caso `temperatura`) e a variável resposta (y, no nosso caso `dias`). Logo, o coeficiente de correlação amostral r_{xy} é uma das raízes quadradas do coefciente de determinação, que é indicado na listagem: $R^2 = 0.6080$. Falta saber o sinal dessa raíz. Mas o gráfico é claro em indicar que estamos perante uma relação decrescente entre x e y, pelo que tem de ter-se: $r_{xy} = -\sqrt{R^2} = -\sqrt{0.6080} = -0.77974$. Alternativamente, o sinal do coeficiente de correlação pode ser determinado recordando que, numa regressão linear simples, é sempre igual ao sinal do declive da recta de regressão.

   (b) O coeficiente de determinação é $R^2 = 0.6080$. Não sendo um valor muito elevado, mesmo assim significa que esta regressão linear explica mais de 60% da variabilidade nos valores observados da variável resposta (`dias`). Em particular, é um modelo significativamente melhor do que o Modelo Nulo, como se pode verificar pelo 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 $F_{calc} > f_{\alpha[1,54]} \approx 4.03$.

   Conclusões: Tem-se $F_{calc} = 54 \times \frac{0.6080}{1-0.6080} = 83.7551 \gg 4.03$. Logo há uma clara rejeição de H_0, i.e., a recta de regressão não é inútil para prever o número de dias entre postura e emergência, a partir da temperatura.

   (c) Os erros aleatórios do modelo são os ϵ_i cuja variância é dada por $V[\epsilon_i] = \sigma^2$ (para todo o i). Esta variância é estimada pelo Quadrado Médio Residual, cuja raíz quadrada é dada nas listagens produzidas pelo programa `R`, com a designação *Residual standard error*. Assim, $QMRE = (3.369)^2 = 11.35016$. Este valor tem unidades de medida. De facto, $QMRE = \frac{SQRE}{n-2}$. O denominador não tem unidades de medida, mas o numerador tem as unidades de medida do quadrado dum resíduo. Uma vez que as unidades dum resíduo são as unidades de medida da variável resposta Y, trata-se no nosso caso de $11.35016\,\text{dias}^2$.

   (d) Pede-se um intervalo de predição para um valor de Y associado ao valor $x = 22.9$. Este intervalo de predição tem extremos: $(b_0 + b_1\,x) \pm t_{0.025(n-2)} \sqrt{QMRE\,\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}\right]}$,

sendo conhecidos a partir do enunciado os seguintes valores: $b_0 = 91.5285$, $b_1 = -2.5284$, $\sqrt{QMRE} = 3.369$, $n = 56$ (porque $n-2 = 54$), $\overline{x} = 22.96$, $s_x^2 = 2.7046$. Pelas tabelas, tem-se $t_{0.025(54)} \approx 2.01$. Substituindo, obtém-se o intervalo:

$$] \, 26.796 \; , \; 40.460 \, [\; .$$

(e) 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. Sabemos (ver formulário) que a distância de Cook duma observação i é função do resíduo internamente estandardizado (R_i) e do valor do efeito alavanca (h_{ii}) dessa mesma observação. Mais concretamente, numa regressão linear simples tem-se (já que $p=1$): $D_i = R_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{1}{2}$. Ora os valores de R_i e h_{ii} definem os eixos vertical e horizontal, respectivamente, do gráfico do enunciado. Para a observação 37 tem-se $R_{37} \approx 2$ e $h_{37,37} \approx 0.08$. Logo $D_{37} \approx 4 \times \frac{0.08}{0.92} \times 0.5 = 0.174$ (o verdadeiro valor é $D_{37} = 0.191$). É um valor relativamente elevado, mas ainda distante do limiar 0.5.

2. Considera-se agora a regressão linear resultante duma dupla transformação logarítmica de x e y.

(a) A transformação utilizada corresponde à transformação linearizante dum modelo potência $y = c\,x^d$. De facto, logaritmizando esta equação do modelo potência obtém-se $\underbrace{\ln(y)}_{=y^*} = \underbrace{\ln(c)}_{=b_0^*} + \underbrace{d}_{=b_1^*} \underbrace{\ln(x)}_{=x^*}$. Assim, o declive da recta 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 assim $y = \mathrm{e}^{b_0^*}\, x^{b_1^*}$ Logo, a equação potência ajustada no nosso caso é $y = \mathrm{e}^{8.8404}\, x^{-1.7058} = 6907.755\, x^{-1.7058} = \frac{6907.755}{x^{1.7058}}$. O facto da potência ser negativa (isto é, do declive da recta na transformação linearizada ser negativo) indica que se trata duma relação decrescente, o que é coerente com a nuvem de pontos dada no enunciado.

(b) O enunciado pergunta se é admissível considerar que $y = \frac{\alpha}{x^2}$. 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 $\beta_1 = -2$ (sendo $b_1^* = -1.7058$ a estimativa amostral desse valor). Eis o teste de hipóteses pedido:

Hipóteses: $H_0 : \beta_1 = -2 \quad vs. \quad H_1 : \beta_1 \neq -2.$

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: $\alpha = \mathrm{P}[$ Erro do tipo I $] = \mathrm{P}[$ Rej. $H_0 \mid H_0$ verdade $] = 0.05$.

Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(54)} \; \approx \; 2.01$.

Conclusões: Tem-se $T_{calc} = \frac{-1.7058-(-2)}{0.1715} = 1.715452$. Este valor não pertence à região crítica, logo não se rejeita H_0. Não se pode excluir a hipótese referida no enunciado.

(c) Pede-se um intervalo de predição (a 95%) para Y, dado $x = 22.9$. Apenas sabemos determinar intervalos de predição no contexto dum modelo linear, pelo que teremos de começar por determinar um intervalo de predição para $\ln(Y)$, dado o valor $\ln(x) = \ln(22.9) = 3.1312$. Ora, a forma do intervalo de predição já foi dada em cima, tendo por extremos neste caso: $(b_0^* + b_1^*\, x^*) \pm t_{0.025(n-2)} \sqrt{QMRE \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x^*})^2}{(n-1)s_{x^*}^2}\right]}$, onde x^* indica que se trata dos valores

de temperaturas logaritmizadas. Em geral, será preciso conhecer as média ($\overline{x^*}$) e variância ($s^2_{x^*}$) dos valores transformados $x^*_i = \ln(x_i)$. Mas o enunciado indica um facto importante, que dispensa o conhecimento desses valores: a temperatura $x = 22.9$ pedida tem por logaritmo a média dos logaritmos dos x_i ($\overline{x^*}$). Assim, a última parcela debaixo da raíz desaparece, ficando apenas os extremos $(b^*_0 + b^*_1\, x^*) \pm t_{0.025(n-2)} \sqrt{QMRE\ \left[1 + \frac{1}{n}\right]}$. O Quadrado Médio Residual referido também diz respeito aos valores transformados, mas é conhecido a partir da listagem: $\sqrt{QMRE} = 0.09072$. Por outro lado, $\sqrt{1 + \frac{1}{n}} = 1.008889$. Finalmente, $b^*_0 + b^*_1\, x^* = 8.8404 - 1.7058 \times \ln(22.9) = 3.499307$. Logo, o intervalo de predição para $Y^* = \ln(Y)$ é $]\, 3.315339\, ,\, 3.683275\, [$. Tal significa que o intervalo de predição (95%) para Y tem por extremos a exponencial destes dois valores, ou seja, é] 27.532 , 39.776 [. Trata-se dum intervalo de predição muito semelhante ao obtido com a regressão linear original, embora este intervalo seja marginalmente mais preciso (de menor amplitude).

III

1. Da definição de variância duma variável aleatória X tem-se $\mathrm{V}[X] = \mathrm{E}[(X – E(X))^2] = \mathrm{E}[X^2] – \mathrm{E}^2[X]$. No nosso caso, tem-se a partir do enunciado que o valor esperado da variável aleatória E_i é zero, logo fica apenas $\mathrm{V}[E_i] = \mathrm{E}[E_i^2]$.

2. Por definição, $SQRE = \sum_{i=1}^{n} E_i^2$. Logo, e tendo também em conta a informação sobre $\mathrm{V}[E_i]$ constante do enunciado e as propriedades do valor esperado, tem-se:

$$\mathrm{E}[SQRE] = \mathrm{E}\left[\sum_{i=1}^{n} E_i^2\right] = \sum_{i=1}^{n} \mathrm{E}[E_i^2] = \sum_{i=1}^{n} \mathrm{V}[E_i] = \sum_{i=1}^{n} \sigma^2\,(1 - h_{ii}) = n\,\sigma^2 - \sigma^2 \sum_{i=1}^{n} h_{ii}\ .$$

   Mas, considerando a expressão para h_{ii}, tem-se $\sum_{i=1}^{n} h_{ii} = 2$, o que completa a demonstração:

$$\sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2}\right] = \cancel{n}\frac{1}{\cancel{n}} + \frac{1}{(n-1)\, s_x^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)\, s_x^2} = 1 + 1 = 2\ .$$

3. Pede-se para mostrar que $\mathrm{E}[QMRE] = \sigma^2$. Ora, por definição, numa regressão linear simples, $QMRE = \frac{SQRE}{n-2}$. Logo, e tendo em conta a alínea anterior,

$$\mathrm{E}[QMRE] = \mathrm{E}\left[\frac{SQRE}{n-2}\right] = \frac{1}{n-2}\,\mathrm{E}[SQRE] = \frac{1}{\cancel{n-2}}\,\cancel{(n-2)}\,\sigma^2 = \sigma^2\ .$$

4. Por definição, $E_i = Y_i - \hat{Y}_i$, o que equivale a dizer que $Y_i = E_i + \hat{Y}_i$. Logo, aplicando a propriedade relativa à variância duma soma de variáveis aleatórias, tem-se: $V[Y_i] = V[E_i] + V[\hat{Y}_i] + 2\,Cov[E_i, \hat{Y}_i]$. Sabemos pelo enunciado que $V[E_i] = \sigma^2\,(1 - h_{ii})$ e que $Cov[E_i, \hat{Y}_i] = 0$. Sabemos ainda que do modelo RLS decorre directamente que $V[Y_i] = \sigma^2$. Substituindo e isolando $V[\hat{Y}_i]$, vem: $V[\hat{Y}_i] = V[Y_i] - V[E_i] = \sigma^2 - \sigma^2(1 - h_{ii}) = \sigma^2\, h_{ii}$.

5. Tendo em conta a alínea anterior, apenas é necessário provar duas coisas: (i) que $\hat{Y}_i$ tem distribuição Normal; e (ii) que $\mathrm{E}[\hat{Y}_i] = \beta_0 + \beta_1\, x_i$. Começando por esta última questão, e aproveitando os resultados anteriores, tem-se:

$$\mathrm{E}[\hat{Y}_i] = \mathrm{E}[Y_i - E_i] = \underbrace{\mathrm{E}[Y_i]}_{=\beta_0+\beta_1\, x_i} - \underbrace{\mathrm{E}[E_i]}_{=0} = \beta_0 + \beta_1\, x_i \ ,$$

como se queria mostrar. Embora $\hat{Y}_i$ seja a diferença de duas v.a.s Normais (Y_i e E_i) não é possível concluir daí que tenha distribuição Normal, uma vez que falta uma condição essencial: a independência dessas v.a. Normais (que não se verifica - recordar que no Exercício 20b) da RLS se viu que $Cov[E_i, Y_i] \neq 0$). Mas é sempre possível recorrer ao raciocínio usado nas aulas para este tipo de situações: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1\, x_i = \sum_{j=1}^{n} (d_j + c_j\, x_i)\, Y_j$ é uma combinação linear das observações Y_i que, essas sim, sabemos serem v.a. Normais e independentes. Logo $\hat{Y}_i$ tem distribuição Normal, o que completa a demonstração.