INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2015/16

Algumas demonstrações de resultados nos acetatos de apoio às Teóricas

1. **Distribuição do estimador do declive, $\hat{\beta}_1$ (Teoremas dos acetatos 126 e 127).**

Foi visto (Acetato 124) que

$$\hat{\beta}_1 \;=\; \frac{cov_{xY}}{s_x^2} \;=\; \sum_{i=1}^{n} c_i\, Y_i \;, \qquad \left(c_i = \frac{x_i - \overline{x}}{(n-1)\, s_x^2}\right) \;.$$

Logo, $\hat{\beta}_1$ é uma **combinação linear das observações de Y, $\{Y_i\}_{i=1}^n$**. Mas essas observações têm distribuição Normal e são independentes (Acetato 123). **Qualquer combinação linear de Normais independentes é Normal** (Acetato 122). **Logo, $\hat{\beta}_1$ tem distribuição Normal**. Falta indicar com que parâmetros. Uma vez que o primeiro parâmetro duma Normal é o seu valor esperado (valor médio), temos (tendo em conta os acetatos 119 e 123):

$$E[\hat{\beta}_1] \;=\; E\left[\sum_{i=1}^{n} c_i\, Y_i\right] \;=\; \sum_{i=1}^{n} c_i\, E[Y_i] \;=\; \sum_{i=1}^{n} c_i\,(\beta_0 + \beta_1\, x_i) \;=\; \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i\, x_i \;.$$

Veremos de seguida que o primeiro destes somatórios tem soma nula, e o segundo tem soma 1, pelo que o valor esperado pretendido é β_1. Temos:

$$\sum_{i=1}^{n} c_i \;=\; \sum_{i=1}^{n} \frac{x_i - \overline{x}}{(n-1)\, s_x^2} \;=\; \frac{1}{(n-1)\, s_x^2} \sum_{i=1}^{n} (x_i - \overline{x}) \;=\; 0 \qquad \text{(ver Ex.3a) RLS)}$$

$$\sum_{i=1}^{n} c_i\, x_i \;=\; \sum_{i=1}^{n} \frac{(x_i - \overline{x})x_i}{(n-1)\, s_x^2} \;=\; \frac{1}{(n-1)\, s_x^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \;=\; \frac{s_x^2}{s_x^2} \;=\; 1$$

sendo a passagem intermédia na segunda linha resultado de aplicar o Exercício 3b) de RLS a uma situação em que $y_i = x_i$, para todo o $i = 1, ..., n$. Logo, $E[\hat{\beta}_1] = \beta_1$, pelo que o estimador é **centrado**.

Vejamos agora a expressão para o segundo parâmetro da Normal, que sabemos ser a variância. Recordem-se as propriedades das variâncias, e tenha-se presente que as observações $\{Y_i\}_{i=1}^n$ são v.a. independentes.

$$V[\hat{\beta}_1] \;=\; V\left[\sum_{i=1}^{n} c_i\, Y_i\right] \;=\; \sum_{i=1}^{n} c_i^2 \underbrace{V[Y_i]}_{=\sigma^2\,,\,\forall\, i} \;=\; \sigma^2 \sum_{i=1}^{n} \underbrace{\frac{(x_i - \overline{x})^2}{[(n-1)\, s_x^2]^2}}_{=c_i^2} \;=\; \frac{\sigma^2}{[(n-1)\, s_x^2]^2} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$=\; \frac{\sigma^2}{[(n-1)\, s_x^2]^2} \cdot (n-1)\, s_x^2 \;=\; \frac{\sigma^2}{(n-1)\, s_x^2} \;.$$

Temos assim que $\hat{\beta}_1 \cap \mathcal{N}\left(\beta_1 \,,\, \frac{\sigma^2}{(n-1)s_x^2}\right)$. Como sabemos (acetato 122), se a uma v.a. Normal subtrairmos a sua média e dividirmos pelo seu desvio padrão, obtemos uma Normal reduzida, $\mathcal{N}(0,1)$. Assim, e como referido no Acetato 128, $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \cap \mathcal{N}(0,1)$, sendo $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{(n-1)s_x^2}}$ o desvio padrão associado à estimação de β_1.

Os resultados para o outro estimador, $\hat{\beta}_0$, são análogos e são discutidos no Exercício 13 das aulas práticas.

2. **Distribuição de $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}}$ (Teorema do acetato 134).**

   Pretende-se mostrar que, se na quantidade $\frac{\hat{\beta}_1-\beta_1}{\sigma_{\hat{\beta}_1}}$ (que tem distribuição $\mathcal{N}(0,1)$) se substituir $\sigma_{\hat{\beta}_1}$ pelo seu estimador $\hat{\sigma}_{\hat{\beta}_1}$, a distribuição resultante é uma t-Student, com $n-2$ graus de liberdade. No acetato 128 viu-se que

   $$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \ \cap \ \mathcal{N}(0\,,\,1) \ .$$

   Por outro lado, no acetato 131 enunciou-se que

   $$W = \frac{SQRE}{\sigma^2} \ \cap \ \chi^2_{n-2} \ ,$$

   sendo $SQRE$ independente de $\hat{\beta}_1$. Logo, e relembrando a forma como surgem distribuições t-Student (acetato 133), sabemos que $\frac{Z}{\sqrt{W/(n-2)}} \cap t_{n-2}$.

   Tendo em conta a expressão para $\sigma_{\hat{\beta}_1}$ (acetato 128), a expressão para $\hat{\sigma}_{\hat{\beta}_1}$ (acetato 134) e a definição de $QMRE$ (acetato 132), temos:

   $$\frac{Z}{\sqrt{W/(n-2)}} = \frac{\frac{\hat{\beta}_1-\beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{SQRE}{\sigma^2\cdot(n-2)}}} = \frac{\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\cancel{\sigma^2}}{(n-1)\cdot s_x^2}}}}{\sqrt{\frac{QMRE}{\cancel{\sigma^2}}}} = \frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{QMRE}{(n-1)\cdot s_x^2}}} = \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \ .$$

   Assim, $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}}$ tem distribuição t-Student, com $n-2$ graus de liberdade.

3. **Intervalo de confiança para β_1 (Acetato 135).**

   Sabemos (acetato 134) que $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}$. 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-se escrever a seguinte equação:

   $$P\left[-t_{\alpha/2(n-2)} \ < \ \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \ < \ t_{\alpha/2(n-2)}\right] \ = \ 1-\alpha$$

   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_1) sozinho no meio duma dupla desigualdade. Tem-se (primeiro multiplicando a dupla desigualdade por $\hat{\sigma}_{\hat{\beta}_1}$, depois por -1 e finalmente somando $\hat{\beta}_1$):

   $$\begin{array}{rrcccl} & -t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_1} & < & \hat{\beta}_1-\beta_1 & < & t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_1} \\ \Leftrightarrow & t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_1} & > & \beta_1-\hat{\beta}_1 & > & -t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_1} \\ \Leftrightarrow & \hat{\beta}_1 - t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_1} & < & \beta_1 & < & \hat{\beta}_1 + t_{\alpha/2(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_1} \ . \end{array}$$

   Assim, a probabilidade de o verdadeiro valor do declive β_1 da recta populacional estar contido entre os dois extremos indicados é $1-\alpha$. Mas este intervalo é um *intervalo aleatório*: os seus extremos são constituídos por variáveis aleatórias ($\hat{\beta}_1$ e $\hat{\sigma}_{\hat{\beta}_1}$), que tomam diferentes valores para cada amostra concreta que seja extraída da população. Para *uma amostra concreta* obter-se-á um *intervalo concreto* substituindo o estimador $\hat{\beta}_1$ pela estimativa concreta b_1 e substituindo o erro padrão estimado $\hat{\sigma}_{\hat{\beta}_1}$ pelo seu valor concreto (que continuamos a designar por $\hat{\sigma}_{\hat{\beta}_1}$). O intervalo assim resultante chama-se um *intervalo de confiança a* $(1-\alpha) \times 100\%$ para β_1:

$$\Big] \, b_1 \, - \, t_{\alpha/2\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \quad , \quad b_1 \, + \, t_{\alpha/2\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \, \Big[\ .$$

Pode ser interpretado da seguinte forma: é um intervalo concreto (associado a uma amostra concreta) extraído ao acaso duma família de intervalos (que se obteriam se fossem extraídas todas as possíveis amostras concretas de tamanho n da nossa população). Nessa família, $(1-\alpha) \times 100\%$ dos intervalos conteriam o verdadeiro valor do declive populacional β_1. Não sabemos se o intervalo concreto por nós escolhido contém, ou não, esse valor. Mas temos uma *confiança* $(1-\alpha) \times 100\%$ em como contém, uma vez que essa era a proporção de intervalos na referida famiília de intervalos (de onde o nosso intervalo concreto foi escolhido ao acaso) que continham o valor β_1.