INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2017-18

Uma breve nota a propósito do teste F parcial

No contexto do estudo do modelo de regressão linear múltipla, o teste F parcial foi apresentado como ferramenta para comparar um modelo completo, com p variáveis preditoras, e um seu submodelo, em que apenas se retêm k dos p preditores originais, ambos ajustados com o mesmo conjunto de n observações.

Na realidade, o teste F parcial é de aplicação mais geral do que a situação considerada nas aulas. O teste é aplicável na comparação de dois modelos para os quais os subespaços de $\mathbb{R}^n$ gerados pelas colunas das respectivas matrizes do modelo, $\mathbf{X}$, estejam contidos um no outro. Em concreto, considere-se um modelo de RLM com p preditores, cuja matriz associada é $\mathbf{X}_c$, e outro modelo, com k preditores, cuja matriz associada é $\mathbf{X}_s$. Se $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$, então pode aplicar-se o teste F parcial para testar a hipótese nula de que os dois modelos coincidem (contra a alternativa de que não coincidem), sendo a estatística dada pela mesmas expressões vistas nas aulas:

$$F \;=\; \frac{\frac{SQRE_s - SQRE_c}{p-k}}{\frac{SQRE_c}{n-(p+1)}} \;=\; \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \;.$$

Caso os dois modelos sejam equivalentes, esta estatística tem uma distribuição $F_{(p-k,n-(p+1))}$[1].

No caso de as colunas da matriz do modelo $\mathbf{X}_s$ serem um subconjunto das colunas da matriz do modelo $\mathbf{X}_c$ (o caso discutido nas aulas, correspondente a ter-se um submodelo constituído apenas por algumas das variáveis preditoras do modelo completo), a condição $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$ verifica-se sempre, uma vez que qualquer combinação linear das colunas de $\mathbf{X}_s$ ($\mathbf{X}_s\vec{\mathbf{a}}$) também se pode escrever como combinação linear das colunas de $\mathbf{X}_c$, bastando associar às colunas da matriz $\mathbf{X}_c$ que não sejam colunas de $\mathbf{X}_s$ o coeficiente zero, e às colunas comuns às duas matrizes os mesmos coeficientes (dados pelos elementos do vector $\vec{\mathbf{a}}$). Mas a condição $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$ é de aplicação mais geral, como se verá de seguida.

Vamos começar por exemplificar a aplicação desta generalização ilustrando uma forma alternativa de resolver a alínea f) do Exercício 7, onde se pede para testar a igualdade de dois parâmetros β_j num modelo de regressão linear múltipla. Seguidamente, veremos como se pode usar a mesma ideia para estudar a hipótese de igualdade entre três ou mais parâmetros β_j.

1. No Exercício 7 de regressão linear múltipla estudam-se os dados relativos a $n = 600$ folhas de videira, nas quais se observam a área foliar (variável resposta, `Area`, em cm^2) e os comprimentos de três nervuras (as variáveis preditoras): a nervura principal (`NP`), a nervura lateral esquerda (`NLesq`) e a nervura lateral direita (`NLdir`), todas em cm. A equação do modelo é:

$$Area \;=\; \beta_0 + \beta_1\, NP + \beta_2\, NLesq + \beta_3\, NLdir + \epsilon. \tag{1}$$

   Assim, a matriz do modelo $\mathbf{X}_c$ é composta por quatro colunas: uma coluna de n uns, uma coluna com os n valores observados da variável `NP`, uma terceira coluna com os n valores observados de `NLesq`, e uma coluna final com os n valores observados de `NLdir`. O modelo ajustado tinha um coeficiente de determinação $R_c^2 = 0.8649$.

[1] Na realidade, $p-k$ indica a diferença nas dimensões dos subespaços encaixados, $\mathcal{C}(\mathbf{X}_c)$ e $\mathcal{C}(\mathbf{X}_s)$.

Na alínea f) do exercício era pedido para estudar a hipótese $H_0 : \beta_2 = \beta_3$. Esse estudo foi feito considerando a hipótese equivalente $H_0 : \beta_2 - \beta_3 = 0$, e utilizando os resultados relativos a combinações lineares $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ dos parâmetros do modelo. Usando a estatística de teste $T = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3}} \cap t_{(n-(p+1))}$, obteve-se o valor calculado $T_{calc} = -0.3636027$. O valor de prova respectivo pode ser calculado (dado tratar-se dum teste com Região Crítica bilateral, e dum valor de T_{calc} na parte esquerda da distribuição) como $p = 2 \times P[T_{506} < T_{calc}]$. Com o auxílio do `R`, obtém-se:

```
> 2*pt(-0.3636027, 596)
[1] 0.7162836
```

2. No entanto, poder-se-ia proceder da seguinte forma alternativa. A hipótese nula $H_0 : \beta_2 = \beta_3$ corresponde ao modelo de regressão linear múltipla de equação:

$$Area \;=\; \beta_0 + \beta_1\, NP + \beta_2\,(NLesq + NLdir) + \epsilon. \tag{2}$$

Trata-se dum modelo com $k = 2$ variáveis preditoras, as variáveis `NP` e `NLesq+NLdir`. A matriz deste modelo, $\mathbf{X}_s$, tem três colunas: uma coluna de n uns, uma coluna com os n valores observados da variável `NP`, e uma coluna final com as n somas de valores das duas nervuras laterais, `NLesq`+`NLdir`. Ora, qualquer combinação linear destas três colunas se pode escrever também como combinação linear das quatro colunas da matriz $\mathbf{X}_c$, bastando igualar, nesta última, os coeficientes individuais de `NLesq` e `NLdir`. Assim, o subespaço das colunas da matriz $\mathbf{X}_s$ está contido no subespaço das colunas da matriz $\mathbf{X}_c$, ou seja, $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$. Será então possível efectuar um teste F parcial para comparar os modelos (1) e (2). Com o auxílio do `R`, tem-se:

```
> videiras.lm <- lm(Area ~ NP + NLesq + NLdir, data=videiras)
> vid2Betas.lm <- lm(Area ~ NP + I(NLesq+NLdir), data=videiras)
> anova(vid2Betas.lm, videiras.lm)
Analysis of Variance Table

Model 1: Area ~ NP + I(NLesq + NLdir)
Model 2: Area ~ NP + NLesq + NLdir
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    597 365391
2    596 365310  1    81.001 0.1322 0.7163
```

Ou seja, a estatística tem valor calculado $F_{calc} = 0.1322$, com valor de prova $p = 0.7163$. O valor F_{calc} poderia ser calculado a partir duma das expressões da estatística F, por exemplo a que utiliza os coeficientes de determinação (tem-se $R^2_s = 0.864873$). Não é uma coincidência que o valor de prova seja o mesmo que foi obtido na resolução alternativa efectuada nas aulas, e usando o teste t. Tal como não é uma coincidência que o quadrado do valor então calculado da estatística T seja o valor agora calculado da estatística F: $T^2_{calc} = (-0.3636027)^2 = 0.1322069$. Esta relação ilustra que também se generaliza a relação que sabíamos existir na aplicação dum teste F parcial para comparar um modelo com p preditores e um seu submodelo com apenas $p-1$ preditores, ou seja, resultante da exclusão dum único preditor.

3. Consideremos agora o exemplo de se querer testar a igualdade de três ou mais coeficientes β_j num modelo RLM. Este problema já não poderia ser estudado considerando a teoria de combinações lineares dos β_j dada nas aulas. Mas pode ser abordado através dum teste F parcial, de forma análoga à acima ilustrada. Continuemos com o exemplo dos dados do Exercício 7, e consideremos

a única hipótese deste tipo possível, a hipótese de que, no modelo 1, os coeficientes populacionais dos comprimentos das três nervuras sejam iguais, ou seja, $H_0 : \beta_1 = \beta_2 = \beta_3$. A essa hipótese corresponde um novo modelo, de equação:

$$Area \;=\; \beta_0 + \beta_1 \left(NP + NLesq + NLdir\right) + \epsilon. \tag{3}$$

Neste novo modelo há apenas $k = 1$ preditor: a soma dos três comprimentos de nervura. A matriz do modelo $\mathbf{X}_s$ correspondente tem agora apenas duas colunas: a coluna de n uns, e a coluna destas n somas das três nervuras. Qualquer combinação linear destas duas colunas pode também escrever-se como combinação linear das quatro colunas da matriz $\mathbf{X}_c$ do modelo original, bastando usar o coeficiente de `NP+NLesq+NLdir` nas três colunas de $\mathbf{X}_c$ correspondentes a estas três variáveis individuais. Logo, de novo, $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$. Podemos efectuar um teste F parcial para testar a igualdade dos modelos (1) e (3). Com o auxílio do `R`:

```
> vid3Betas.lm <- lm(Area ~ I(NP+NLesq+NLdir), data=videiras)
> anova(vid3Betas.lm, videiras.lm)
Analysis of Variance Table

Model 1: Area ~ I(NP + NLesq + NLdir)
Model 2: Area ~ NP + NLesq + NLdir
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    598 365766
2    596 365310  2    456.13 0.3721 0.6895
```

Também neste caso, não se rejeita H_0 para nenhuma dos níveis de significância habituais, pelo que se considera admissível a hipótese $\beta_1 = \beta_2 = \beta_3$.

Nota: Em todos os exemplos considerados, não se discute o problema da curvatura que parece existir na relação de fundo, e que é visível nos gráficos de resíduos estudados na alínea h) do Exercício 7.