O PROBLEMA DOS TESTES MÚLTIPLOS

Manuel Pita, Universidade Lusófona. November 12, 2023

Esta é a versão 0.2 deste caderno. Envie qualquer correção ou sugestão para
`manuel.pita@ulusofona.pt`

Um dos aspetos que consideramos quando treinamos um modelo de regressão linear é a validação estatística dos coeficientes do modelo. Quando obtemos os coeficientes β_0, β_1, etc. testamos a hipótese nula que o seu valor é zero. Para testar um dos coeficientes usamos um teste T com a hipótese nula $H_0 : \beta_i = 0$ (onde i identifica o coeficiente específico que estamos a testar, por exemplo β_1). Na teoria, o procedimento é, portanto, fazer m testes, um para cada coeficiente.

Todos estes m testes calculam a sua estatística T com base nos mesmos dados que foram usados para treinar o modelo. E è aqui que nos deparamos com um conhecido problema estatístico: o *problema dos testes múltiplos* porque os testes usam como base a mesma evidência (dados).

Para entender o problema de testes múltiplos temos de ter em linha de conta o significado do nível de confiança escolhido para um determinado teste estatístico. Por exemplo com confiança 95% temos um valor $\alpha_{\text{crit}} = 0.05$ (5%). Quando rejeitamos a hipótese nula (isto é, quando o valor p é menor que α_{crit}), há uma probabilidade de 5% de que esta conclusão esteja errada. A hipótese nula poderá ter sido rejeitada por erro. Isto é precisamente o que significa o termos 95% confiança no resultado do teste.

Agora considera um modelo de regressão linear com trinta coeficientes... Qual é a confiança que temos que pelo menos um dos trinta testes tenha rejeitado a hipótese nula por erro?

$$0.95 \times 0.95 \times ... \times 0.95 = 0.95^{30} = 0.21$$

Portanto se estivermos a reportar o conjunto dos 30 testes como uma família dentro dum mesmo estudo, a nossa confiança real é 21%. Nesta situação queremos fazer alguma coisa que nos permita recuperar o nível de confiança que queriamos originalmente.

Uma das formas de resolver o problema é corrigir os valores p que rejeitaram a hipótese nula, para aceitá-la (de forma a eliminar as prováveis rejeições errada). A pergunta é agora: quais valores p? e quantos?

Existe uma técnica chamada *False Discovery Rate* que implementa uma política moderada (não demasiado estrita) e cujo método é bastante simples. Podes implementar em Pandas.

1. Ordena todos os (m) valores p do em ordem crescente numa coluna
2. Adiciona uma coluna (i) de valores de ordem onde o primeiro valor p tem o valor 1 na coluna (i), o segundo terá o valor 2 e assim sucessivamente até m.
3. Adiciona uma terceira coluna `crit` cujos valores são `crit` $= (i/m)Q$. O parâmetro Q é um valor entre zero e um, o qual corresponde a percentagem de falsas rejeições da

hipótese nula que queremos que o procedimento estime. Existem métodos matemáticos para identificar o melhor valor Q, mas por norma usamos $Q = 10\%$

4. Identifica o maior valor p que é menor ou igual que o correspondente valor `crit`. Esse valor p, e todos os inferiores a ele rejeitam a hipótese nula, independentemente dos resultados obtidos nos correspondentes testes individuais.

Exemplo

Supõe que fizemos um conjunto de vinte testes estatísticos com $\alpha = 0.05$ (o que significa que temos 95% confiança em cada resultado independentemente). A tabela abaixo contem o conjunto de valores p obtidos. Nota que os valores p já estão ordenados de menor a maior, a coluna (i) foi adicionada (chamada *rank*) e finalmente foram calculados os valores `crit` de acordo com a fórmula descrita anteriormente. Os valores `crit` realçados correspondem com os valores p que vamos considerar para rejeitar hipóteses nulas neste conjunto de testes. Nota que, ainda que temos outros valores p menores que $alpha$ (nomeadamente os valores com *rank* 5-8), os mesmos não são usados para rejeitar as correspondentes hipóteses nulas. Esta é de facto a correção feita pelo método.

Atividades

1. Implementa o método em Pandas e verifica que obténs os mesmos valores da tabela
2. investiga o uso da biblioteca `statsmodels.stats.multitest.fdrcorrection` e os seus parâmetros para obter o mesmo resultado (ou aproximado). Investiga se o parâmetro α usado na biblioteca corresponde ou não com o parâmetro Q definido neste caderno. Documenta as tuas observações.

valor p	rank	crit
0,0012	1	**0,005**
0,0090	2	**0,01**
0,0100	3	**0,015**
0,0120	4	**0,02**
0,0250	5	0,025
0,0395	6	0,03
0,0428	7	0,035
0,0451	8	0,04
0,0696	9	0,045
0,0710	10	0,05
0,0750	11	0,055
0,0820	12	0,06
0,0950	13	0,065
0,1200	14	0,07
0,1320	15	0,075
0,2100	16	0,08
0,5282	17	0,085
0,6710	18	0,09
0,7774	19	0,095
0,8509	20	0,1