

Modelos Lineares

Regressão Linear Múltipla

Susana Faria

Comparação de Modelos Aninhados

Neste caso, o que pretendemos é testar a hipótese:

- H_0 : As variáveis que estão presentes no modelo M_q mas não estão presentes no modelo M_p são todas irrelevantes para modelar Y
contra a hipótese alternativa
- H_1 : Pelo menos uma daquelas variáveis é relevante para modelar Y .

Esta hipótese corresponde a testar simultaneamente que $q - p$ parâmetros são nulos.

No R, tendo-se estimado dois modelos aninhados, o comando `anova()` realiza o teste descrito.

Comparação de Modelos Não Aninhados

AIC é uma medida de qualidade de ajustamento de um modelo estimado. De uma forma genérica, pode dizer-se que engloba a precisão e a complexidade do modelo. Para o modelo de regressão linear estudado:

$$AIC = n * \ln(SSE/n) + 2k$$

onde SSE representa a soma dos quadrados dos erros e k o número de parâmetros presentes no modelo.

- Quanto maior for o número de variáveis consideradas no modelo (e consequente mais parâmetros), menor será o valor de SSE. Por outro lado, porque um modelo mais complexo acarreta maiores custos (a todos os níveis), a introdução de variáveis no modelo é penalizada.
- A medida AIC é uma ferramenta para a selecção de modelos. Perante um conjunto de dados e vários modelos candidatos, estes podem ser ordenados de acordo com o AIC, **considerando-se o melhor modelo aquele que apresentar menor valor de AIC.**
- No entanto esta medida não dá qualquer informação sobre a significância dos modelos.

Seleção das variáveis explicativas

Considerando uma situação em que existem p variáveis explicativas, uma possibilidade seria ajustar:

- um modelo contendo as p variáveis;
- os $p(p-1)/2$ modelos contendo todas as combinações de $p-1$ das p variáveis;
- os C_k^p modelos contendo todas as combinações de k das p variáveis, $k = p-2, \dots, 1$
- e para terminar, ajustar o modelo sem variáveis explicativas.

Após ajustarmos os 2^p modelos, poderíamos escolher aquele que produzisse **menor erro quadrático médio** ou, de forma equivalente, maior coeficiente de determinação ajustado R_a^2 ou menor estimativa para o erro padrão.

Para valores pequenos de p , é possível analisar todos os possíveis subconjuntos. Mas para p médio ou grande, essa análise completa é inviável.

Métodos para selecção das variáveis explicativas

• Backward(Seleccção Regressiva)

- Construir o modelo contendo todas as variáveis disponíveis;
- Analisar o resultado do teste $H_0 : \beta_j = 0$ para cada $j = 1, \dots, m$. Se todos os coeficientes forem significativos, então conclui-se que todas as variáveis X_j são importantes para explicar Y e nenhuma deve ser eliminada do modelo.
- Se, pelo contrário, alguns coeficientes forem não significativos, retira-se do modelo aquela que apresentar o maior valor-p (essa variável é aquela à qual corresponde a estatística t com valor absoluto mais baixo) e ajusta-se um novo modelo considerando as variáveis restantes.
- Repetem-se os passos acima até que restem no modelo apenas variáveis consideradas significativas.

• Forward(Seleccção Progressiva):

- Neste procedimento, começa-se por considerar o modelo mais simples, com apenas uma variável. De seguida, passa-se a considerar o modelo com duas variáveis, depois três, e assim sucessivamente, parando-se quando as variáveis que se acrescentam ao modelo não são significativas.
- Esta metodologia tem o problema da determinação do melhor modelo em cada uma das fases, para além de ser muito dispendioso em termos de cálculo pois envolve a estimação de um número elevado de modelos durante o processo de selecção.

Métodos para selecção das variáveis explicativas

- **Stepwise:** Este método combina os anteriores. Basicamente é um procedimento forward pois vai adicionando variáveis uma a uma. No entanto, em cada passo é feita uma análise das variáveis já introduzidas até aí, por forma a garantir que permanecem relevantes após a introdução da nova variável. Este é o método mais completo dos três apresentados.

Nota: Pode-se comparar o modelo final com o modelo inicial aplicando um teste F-parcial.

Em R:

```
step()  
drop1()  
add1()
```

Métodos para selecção das variáveis explicativas

Exemplo: Para ilustrar os procedimentos de selecção de variáveis, consideremos o seguinte conjunto de dados *savings* do package *faraway*.