Aula Lab – R – Capítulo 6

26/11/2020

In the regression setting, the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Relação entre p e n deve ser observada:

- n >> p
- n < p

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Lembrando que:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned} \qquad (3.22)$$

Capítulo 6:

- Não gostei da função que o livro indicou para fazer a análise dos modelos.

A saída é muito confusa.

- Achei melhor usar a função que usamos para Regressão Linear Simples, e que também pode ser usada para Regressão Linear Múltipla.
- Lembrando da função: lm(Y ~ X) para Regressão Linear Simples.
- Para Regressão Linear Múltipla com duas variáveis independentes fica assim: lm(Y ~ X_1+X_2) .

Exemplo para a base de dados "Boston", que já usamos.

```
> library(MASS)
> lm.fit =lm(medv~lstat+age ,data=Boston )
```

Comandos do R usados

Para calcular a Soma dos Resíduos ao Quadrado (RSS em inglês), comecei fazendo um teste com os seguintes vetores de dados de entrada (x) e saída (y):

y<- c(2.01,4.03,6.02,7.98,9.96)
> y
[1] 2.01 4.03 6.02 7.98 9.96

- x<-c(1,2,3,4,5)
- fit.reta<- lm(y~x)

O comando "**residuals( )" ou "resid( )" (abreviado)** mostra o vetor de diferenças entre a reta regredida e os dados fornecidos de y.

> **residuals(fit.reta)**
    1   2   3   4   5
-0.020 0.015 0.020 -0.005 -0.010

> fit.reta
Call: lm(formula = y ~ x) Coefficients: beta_0= 0.045 beta_1= 1.985

E usei a seguinte fórmula para calcular RSS:

> **rss<-sum(resid(fit.reta)^2)**
> rss
[1] 0.00115

Variáveis Consideradas - Notação

1. **X_1 -> zn:** proporção de terrenos residenciais zoneados para lotes com mais de 25.000 metros quadrados.
2. **X_2 -> Indus:** proporção de acres de negócios não varejistas por cidade
3. **X_3 - > Age:** idade das construções
4. **X_4 -> Lstat:** porcentagem de moradores com status inferior
5. **Y= medev:** valor médio de casas ocupadas pelo proprietário em \ $ 1000s

```
> library(MASS)
```

comando para ver a descrição do conjunto de dados “Boston”:

- ?Boston

A partir daqui foram feitas as regressões simples para cada uma das 4 variáveis e calculados os respectivos RSS.

```
> reta_x_1=lm(medv~zn,data=Boston)
> rss<-sum(resid(reta_x_1)^2)
> rss
[1] 37166.56
> reta_x_2=lm(medv~indus,data=Boston)
> rss<-sum(resid(reta_x_2)^2)
> rss
[1] 32721.11
> reta_x_3=lm(medv~age,data=Boston)
> rss<-sum(resid(reta_x_3)^2)
> rss
[1] 36646.53
> reta_x_4=lm(medv~lstat,data=Boston)
> rss<-sum(resid(reta_x_4)^2)
> rss
[1] 19472.38
```

Organizando os valores calculados temos:

Variável	Modelo	RSS
Zn	Y ~X_1	37167
Indus	Y~X_2	32721
Age	Y~X_3	36647
Lstat	Y~X_4	19472

Então o melhor modelo, que considera apenas uma variável , é Y ~ X_4 pois tem o menor valor de RSS.

Fazer o mesmo para o modelo com duas variáveis:

```
> reta_x_1_x_2=lm(medv~zn+indus,data=Boston)
> rss<-sum(resid(reta_x_1_x_2)^2)
> rss
[1] 32096.89
> reta_x_1_x_3=lm(medv~zn+age,data=Boston)
> rss<-sum(resid(reta_x_1_x_3)^2)
> rss
[1] 35303.35
```

Completar com as demais combinações de modelos com 2 duas variáveis em 4.

Pode ser criada uma função para fazer os cálculos de RSS de forma mais rápida.

```
> reta_x_1_x_2_x_3=lm(medv~zn+indus+age,data=Boston)
> reta_x_1_x_2_x_3

Call:
lm(formula = medv ~ zn + indus + age, data = Boston)

Coefficients:
(Intercept)          zn       indus         age
   29.04377     0.04842    -0.50530    -0.02091

> rss<-sum(resid(reta_x_1_x_2_x_3)^2)
> rss
[1] 32007.1
> reta_x_1_x_2_x_4=lm(medv~zn+indus+lstat,data=Boston)
> rss<-sum(resid(reta_x_1_x_2_x_4)^2)
> rss
[1] 19282.81
> reta_x_1_x_3_x_4=lm(medv~zn+age+lstat,data=Boston)
> rss<-sum(resid(reta_x_1_x_3_x_4)^2)
> rss
[1] 18653.46
> reta_x_2_x_3_x_4=lm(medv~indus+age+lstat,data=Boston)
> rss<-sum(resid(reta_x_2_x_3_x_4)^2)
> rss
[1] 18781.43
```

Aqui o mesmo foi feito para a combinação de 3 variáveis.

Vocês podem montar uma tabela para organizar os valores de RSS; ou simplesmente olhar os valores e escolher o melhor modelo.

Calcular o RSS para o modelo com as 4 variáveis

Você já tem os 5 (p+1) valores para executar o passo 3?