Aula – Capítulo 6

25/11/2020

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

Vamos usar o método dos mínimos quadrados para:

- os modelos lineares simples (p=1 “já sabemos fazer”);
- para o modelo com 2 preditores (p=2), ou seja são “p duas a duas combinações”;
- para o modelo com 3 preditores (p=3), ou seja são “p três a três combinações”;
- E assim por diante.
- Vamos ter 2^p (2 elevado a p modelos)

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.

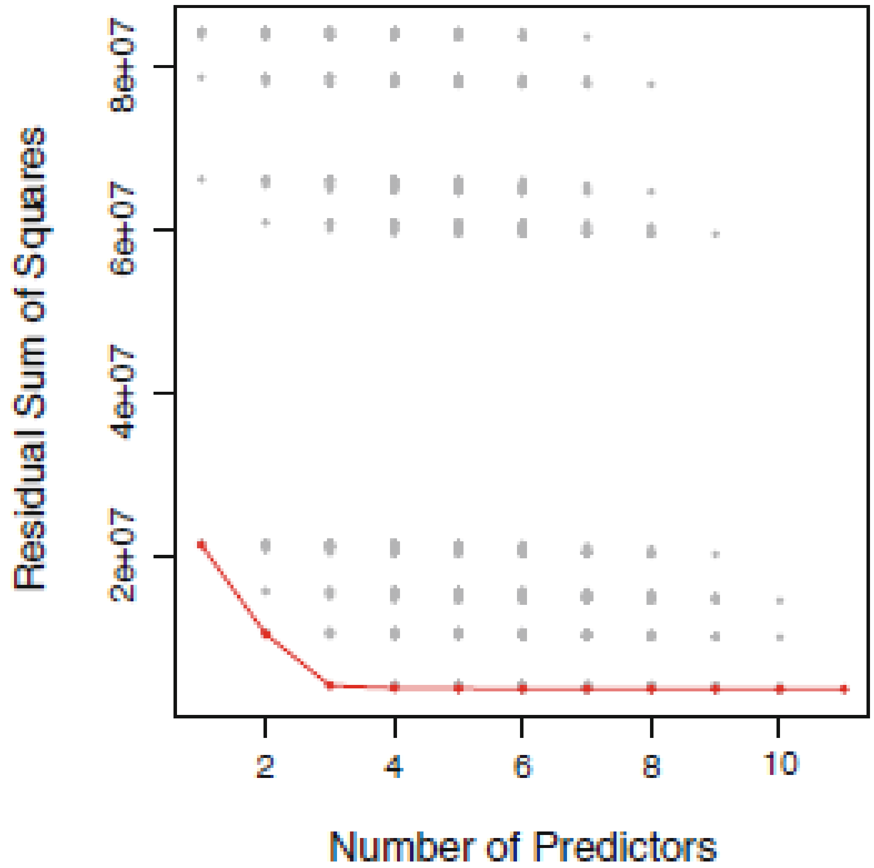
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Lembrando que:

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

- **O que o passo 2 faz?**
- **Quantos modelos temos antes de aplicar o Algoritmo 6.1?**
- **Quantos temos depois? p+1**

p=10



Residual Sum of Squares
2e+07
4e+07
6e+07
8e+07
2
4
6
8
10
Number of Predictors

For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right), \tag{6.2}$$

The AIC criterion is defined for a large class of models fit by maximum likelihood. In the case of the model (6.1) with Gaussian errors, maximum likelihood and least squares are the same thing. In this case AIC is given by

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right),$$

BIC is derived from a Bayesian point of view, but ends up looking similar to C_p (and AIC) as well. For the least squares model with d predictors, the BIC is, up to irrelevant constants, given by

$$\mathrm{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right). \tag{6.3}$$

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 )
```

```
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)
> 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:
(Intercept)            x
      0.045        1.985

> rs<-residuals(fit.reta)
> rs
     1      2      3      4      5
-0.020  0.015  0.020 -0.005 -0.010
> rss<-sum(resid( fit.reta )^2)
> rss
[1] 0.00115
> library(MASS)
> fix(Boston)
```

Zn Y ~X_1 (37166.56)	Zn+indus
Indus Y~X_2 (32721.11) Age Y~X_3 (36646.53) Lstat Y~X_4 (19472.38)	Zn (X_1)+ indus (X_2) (32096.89) Zn +age (35303.35)
Lstat Y~X_4	

```
reta_x_1_x_2=lm(medv ~ zn+indus,data=Boston)
```

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