NBA 4920/6921 Lecture 10

Linear Model Best Subset Selection

Murat Unal

Johnson Graduate School of Management

09/30/2021

Agenda

Quiz 8

Linear regression
- Model performance
- Adjusted R^2

Model selection
- Best subset selection

Application in R

Linear regression

Recall the linear model assumes the relationship between the outcome Y and the inputs $X = X_1, X_2, \cdots, X_p$ is linear

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

We saw that we obtain estimates for the coefficients $\beta_0, \beta_1, \cdots, \beta_p$ by minimizing the **Residual Sum of Squares** (RSS)

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

The values $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ are the least squares coefficient estimates

Model performance

Recall to asses the fit of the linear model we compute

Residual Standard Error (RSE) and **R-squared**(R^2) using

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

$$RSE = \sqrt{\frac{RSS}{n-p-1}}, \quad R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Model performance

Adding more variables to the model always increases R^2, whereas RSE can increase or decrease

Therefore, we need to be careful about **overfitting**, especially if we aim for prediction

R^2 provides no protection against overfitting, quite opposite - **encourages** it because it is related to the training error

Model performance

We seek to find the model with the lowest test error, not the lowest training error

Recall also that training error is a poor estimate of test error

As such, R^2 should not be used for comparing models with different number of predictors

Adjusted R^2

One way to improve the test error estimates is by directly estimating the training error using **hold-out methods**

The other way is to indirectly estimating the test error by **adjusting** the training error to account for the bias due to overfitting

Adjusted R^2

Adjusted R^2 attempts to fix R^2 by paying a price for the inclusion of unnecessary variables

$$\text{Adjusted} R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

A large value of **Adjusted** R^2 suggests a model with a small test error

Now that the computational costs have become low, cross-validation is the preferred method for comparing model performance with different predictors.

Model selection

Best subset selection:

The idea is to estimate a model for every possible subset of variables; then compare their performances

Model selection

Best subset selection:

1. Let M_0 denote the null model, which contains no predictors.
2. For k in 1 to p:
   - Fit every possible model with k variables
   - Let M_k denote the **best** model with k variables
3. Select the **best** model from $M_0, \cdots, M_p$ using cross-validated prediction error
4. Train the chosen model on the full dataset

Model selection

Best subset selection:

Problem?

Model selection

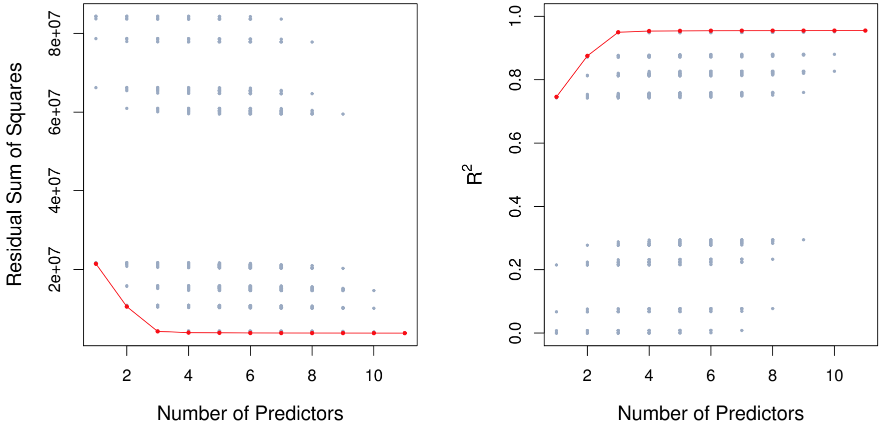
Best subset selection:

Problem?

- $p = 10 \rightsquigarrow$ fitting 1,024 models
- $p = 25 \rightsquigarrow$ fitting ≈ 33.5 mil models

Model selection

Best subset selection for Credit dataset



Source: ISL

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017)
An Introduction to Statistical Learning
Springer.
https://www.statlearning.com/

Ed Rubin (2020)
Economics 524 (424): Prediction and Machine-Learning in Econometrics
Univ, of Oregon.