Spring 2015 Statistics 151 (Linear Models) : Lecture Seventeen

Aditya Guntuboyina

April 1, 2015

1 Variable Selection

Consider a regression problem with a response variable y and p explanatory variables $x_1, \ldots, x_p$. Should we just go ahead and fit a linear model to y with all the p explanatory variables or should we throw out some unnecessary explanatory variables and then fit a linear model for y based on the remaining variables? One often does the latter in practice. The process of selecting important explanatory variables to include in a regression model is called variable selection. The following are reasons for performing variable selection:

1. Removing unnecessary variables results in a simpler model. Simpler models are always preferred to complicated models.
2. Unnecessary explanatory variables will add noise to the estimation of quantities that we are interested in. For example, the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + e_i$ is σ^2/n while the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + \beta_1 x_i + e_i$ is $\sigma^2 \sum_{i=1}^n x_i^2 / (\sum_i x_i^2 - n\bar{x}^2)$ where $\bar{x} := \sum_i x_i/n$.
3. Collinearity is a problem with having too many variables trying to do the same job.
4. We can save time and/or money by not measuring redundant explanatory variables.

There are two broad ways of performing variable selection in linear models:

1. Stepwise Regression
2. Criteria-based variable selection

2 Stepwise Regression Methods for Variable Selection

The two main stepwise regression methods are backward elimination and forward selection.

2.1 Backward Elimination

1. Start with all the explanatory variables in the model.
2. Remove the explanatory variable with highest p-value larger than a critical value.
3. Refit the model and go to the previous step.
4. Stop when all the p-values are less than the critical value.

The critical value is sometimes called the p-to-remove and does not have to be 0.05. If prediction performance is the goal, then a 0.15-0.20 cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

2.2 Forward Selection

This just reverses the backward method:

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than the critical value.
3. Continue until no new predictors can be added.

2.3 Other Stepwise Regression Methods

There are several other stepwise regression methods. These are all combinations of backward elimination and forward selection. These might be better than backward elimination or forward selection by addressing the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done.

2.4 Drawbacks of Stepwise Regression

Stepwise procedures are relatively cheap computationally but they do have the following drawbacks:

1. Because of the one-at-a-time nature of adding/dropping variables, it is possible to miss the optimal model.
2. The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest.
3. Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes.

3 Criteria Based Variable Selection

If there are p explanatory variables, then there are 2^p possible linear models. In criteria-based variable selection, we fit all these models and choose the best one according to some criterion. There are several criteria that one can use. Some of the common ones are given below.

3.1 R^2

For each candidate model m its $R^2(m)$ is defined as Recall the definition of R^2:

$$R^2(m) := 1 - \frac{RSS(m)}{TSS}$$

where $RSS(m)$ is the residual sum of squares for the model m and TSS is the total sum of squares.

$R^2(m)$ **should NOT be used** as a criterion for variable selection because then we will always pick the full model M which has the highest R^2 value among all the candidate models.

3.2 Adjusted R^2

Adjusted R^2 is defined as

$$(AdjR^2)(m) := 1 - \frac{RSS(m)/(n - p(m) - 1)}{TSS/(n-1)}$$

where $p(m)$ is the number of explanatory variables in the model m. This is very similar to R^2 but has the desirable property that when an explanatory variable is removed from a model, the value of $AdjR^2$ does not necessarily decrease. It might increase if the removed variable has no predictive power. This can therefore be used as a criterion for variable selection.

3.3 AIC

AIC stands for Akaike Information Criterion and is one of the most popular model selection techniques not just in linear models but in other contexts as well.

AIC for a model m is defined as

$$AIC(m) := -2\log(\text{maximum value of likelihood in } m) + 2(\text{number of parameters in } m) \tag{1}$$

We pick models with small AIC.

In the case of linear models, we can show that

$$AIC(m) = n\log\left(\frac{RSS(m)}{n}\right) + n\log(2\pi e) + 2(1 + p(m)) \tag{2}$$

To see this observe that the log-likelihood function in the linear model $Y = X\beta + e$ equals

$$\frac{-n}{2}\left(\log(2\pi) + \log\sigma^2\right) - \frac{||Y - X\beta||^2}{2\sigma^2}.$$

It is easy to see that this is maximized when

$$\beta = \hat{\beta} = (X^TX)^{-1}X^TY \text{ and } \hat{\sigma}^2_{mle} := \frac{RSS}{n}.$$

Plugging these values in the log-likelihood function and simplifying, we see that the maximized log-likelihood for the model is

$$-\frac{n}{2}\log(2\pi e) - \frac{n}{2}\log\left(\frac{RSS}{n}\right).$$

From here, it is easy to see that (1) implies (2).

AIC is used as a criteria to compare various models. The term $n\log(2\pi e)$ clearly is the same for all models and therefore, one often simply drops it and defines the AIC for linear models as

$$AIC(m) := n\log\left(\frac{RSS(m)}{n}\right) + 2(1 + p(m))$$

Note that if m_1 is a sub-model of m_2, then $RSS(m_1) \geq RSS(m_2)$ while $p(m_1) \leq p(m_2)$ so $AIC(m_1)$ may or may not be smaller than $AIC(m_2)$. If it is smaller, we would prefer m_1; otherwise, we prefer m_2.

3.4 BIC

BIC stands for Bayesian Information Criterion. BIC for a model m is defined as

$$BIC(m) := -2\log(\text{maximum value of likelihood in } m) + (\log n)(\text{number of parameters in } m). \quad (3)$$

In model selection via the BIC, one selects models with small BIC.

Note that the only difference between the formulae for AIC and BIC is the factor of the number of parameters term which is 2 for AIC and $\log n$ for BIC. Because $\log n$ is typically larger than 2, the size of models selected by BIC is smaller than those selected by AIC.

In the case of linear models, one has

$$BIC(m) := n\log\left(\frac{RSS(m)}{n}\right) + (\log n)(1 + p(m)).$$