Lecture 2

Model Selection
Textbook Sections: NA

Model Selection

The term *model selection* may refer to the choice of predictors to include in a model, or to the choice of a model from several candidates. It is a very important topic in statistics, and arises in most situations where model fitting takes place.

Notation

Suppose there are a total of $P - 1$ candidate predictors $(X_1, X_2, \cdots, X_{P-1})$.
For different values of p such that $1 \leq p \leq P$ we will consider subsets of $p - 1$ predictors.
The goal is to find a subset of $p - 1$ predictors that leads to a "good" model.

We assume throughout that all models include an intercept term β_0, and that $n > P$.
Ideally, we would like n to be significantly larger than P.

I will denote some quantities with the subscript p. This means that the given quantitiy is based on a model with p estimated coefficients, or $p - 1$ predictors. For example, SSE_p denotes the SSE of a model with $p - 1$ predictors, and $df(SSE_p) = n - p$.

Typical Approach

When choosing from several models for the same dataset, a typical approach is:

1. Choose an appropriate criterion to compare the models.
2. Evaluate the criterion for each model.
3. Choose the model that leads to the best criterion value.

Criteria That Don't Work

Recall that SSE_p decreases and $R_p^2 = 1 - \frac{SSE_p}{SSTO}$ increases whenever predictors are added to the model. (Unless the new predictors are perfectly correlated with the existing predictors, in which case SSE_p and R_p^2 will remain the same).

Because of this property, SSE_p and R_p^2 are unsuitable criteria for predictor selection. Observe:

Choosing the subset of predictors based on lowest SSE_p will always lead to choosing all predictors.
Choosing the subset of predictors based on highest R_p^2 will always lead to choosing all predictors.

Criteria That Do Work

It is possible to use mean squared error ($MSE_p = \frac{SSE_p}{n-p}$) or adjusted R^2 ($R^2_{adj,p} = 1 - \frac{MSE_p}{MSTO}$) for predictor selection. This is pretty common, but the following criteria may give better results.

Below are four well known criteria commonly used for model selection.

Akaike's FPE (final prediction error):	$FPE_p = \frac{n+p}{n-p} SSE_p$
Mallows' C_p:	$C_p = \frac{SSE_p}{MSE_{p\max}} - (n - 2p)$
AIC (Akaike's information criterion):	$AIC_p = n\ln(SSE_p) - n\ln(n) + 2p$
BIC (Schwartz's Bayesian criterion):	$BIC_p = n\ln(SSE_p) - n\ln(n) + [\ln(n)]p$
AICc (Corrected AIC):	$AICc = AIC + \frac{2p(p+1)}{(n-p-1)}$

Note: $MSE_{p\max}$ is the mean squared error of the largest model under consideration.

The criteria share the following features:

- The model with the smallest value of the criterion is "best."
- As SSE_p decreases, the value of the criterion decreases.
- As p increases, the value of the criterion increases.
- Each criterion "rewards" low error (SSE_p), and "penalizes" complexity (p).
- Given any sample, the term $n\ln(n)$ will be the same in the equations for all of the criteria. For this reason, sometimes shorter expressions are used.

AIC, AICc, and BIC are commonly used in time series analysis.

All-Subsets Selection

In this method every single subset of predictors is considered. The value of the desired criterion is computed for every possible subset of predictors, and the subset that leads to the best criterion value is chosen. The pros and cons of this method are easily summarized:

- Since all subsets are considered, the resulting model can really be considered the “best” according to the chosen criterion.
- Since all subsets are considered, it may be evident that instead of one clear winner, there are two or more models with similarly good values of the criterion. In this case, it is useful to compare these models in more depth, and choose the one that best meets the goals of the analysis.
- A set of $p-1$ candidate predictors results in 2^{p-1} possible subsets of predictors. This number can get very large, very quickly (20 predictors means over 1,000,000 possible subsets). It is often impractical or impossible to consider every single possibility.

Stepwise Regression

Stepwise regression is a very commonly used method of predictors selection. In this method, predictors are either sequentially added to an empty model, or sequentially removed from a full model. The process continues until further steps do not improve the model anymore, or until the improvement is negligible.

1. Starting with an empty model, predictors are added one by one. At each step, the predictor that leads to the most improvement is added to the model.
2. Starting with a full model, predictors are removed one by one. At each step, the predictor that gives the least improvement is removed.
3. Predictor addition and removal steps are alternated, which allows for more flexibility.

Forward Stepwise

This method is also called “Forward Selection.”

1. Start with an empty model (no predictors).
2. Consider all one-predictor models.
   Compare these based on the chosen criterion.
   Add to the model the predictor that leads to the best one-predictor model.
3. Consider all two-predictor models that include the predictor currently in the model.
   Compare these based on the chosen criterion.
   Add to the model the predictor that leads to the best two-predictor model.
4. Consider all three-predictor models that include the two predictors currently in the model.
   Compare these based on the chosen criterion.
   Add to the model the predictor that leads to the best three-predictor model.
5. Continue until the addition of new predictors no longer improves the model.

Backward Stepwise

This method is also called “Backward Elimination.”

1. Start with a full model (all predictors).
2. Consider removing one predictor. Fit all models with one predictor removed. Compare these based on the chosen criterion.
   Remove the predictor that leads to the best model.
3. Consider removing another predictor. Fit all models with one predictor removed. Compare these based on the chosen criterion.
   Remove the predictor that leads to the best model.
4. Consider removing another predictor. Fit all models with one predictor removed. Compare these based on the chosen criterion.
   Remove the predictor that leads to the best model.
5. Continue until the removal of predictors no longer improves the model.

Stepwise Selection in R

In R stepwise selection is done using the base package function $step()$. By default, the function chooses the predictors based on AIC, although it can easily be modified to use BIC (SBC). For example, if using forward selection, then at each step the function will add the predictor that leads to the greatest drop in AIC. This continues until the addition of any new predictor can only increase, and not decrease the AIC.

Keep in mind that $step()$ only considers whether a given predictor improves AIC or not. The function does not consider whether the improvement is significant or negligible. It is possible to modify the algorithm to include threshold values indicating what improvement is considered negligible.