Stat 215b (Spring 2004): Model Selection

B. M. Bolstad
bolstad@stat.berkeley.edu

Mar 4, 2004

Automated Model Selection procedures

When presented with a dataset containing a large number of possible regressors, it can be difficult to choose a good model (well fitting and parsimonious). As the number of regressors increases the number of possible models grows exponentially. To reduce the number of models that should be examined a number of procedures have been proposed. We discuss three below.

Backwards Deletion

Using this method we are looking to remove variables until we cannot remove anymore variables without significantly worsening the model. A procedure for carrying this out was described in Lab 2. One problem with this method is that once a variable has been removed from the model it can not re-enter.

Forward Selection

The idea behind this procedure is to keep adding variables to the model until no variable improves the model significantly. One problem with this procedure is that once a variable has been added to the model it can not be removed.

1. Start with an empty model
2. Try putting each term individually into the model. Are any significant? If yes then add the single best predictor into the model.
3. Test each term not currently in the model given what is all ready in the model. If any are significant then add the most significant to the model and repeat step 3. Otherwise go to step 4.
4. Stop.

Stepwise Regression

This combines both forward and backwards steps.

1. Start with an empty model

2. Try putting each term individually into the model. Are any significant? If yes then add the single best predictor into the model.

3. Test each term currently in the model. Are the all significant? If no, drop the most non-significant and repeat step 3. If yes goto step 4.

4. Test each term not currently in the model given what is all ready in the model. If any are significant then add the most significant to the model and goto step 3. Otherwise go to step 5.

5. Stop.

Other Criteria by which to judge models

Rather than use t statistics or F statistics in the stepwise procedure one could use another criteria to judge the different models.

AIC - Akaike Information Criterion

Due to Akaike (1974). It may be used to compare between models. A smaller value is better. K is the number of parameters.

$$AIC = -2 \log L(\hat{\theta}) + 2K$$

BIC - Bayesian Information Criterion

Due to Schwarz (1978). It may also be used to compare between models. A smaller value is better. K is the number of parameters and n is the number of observations.

$$BIC = -2 \log L(\hat{\theta}) + \log(n)K$$