Advanced Computational Econometrics
Chapter 3: Model selection

1   Movie buzz data

Predicting the box office success of movies is a favorite exercise for econometricians. The common wisdom in Hollywood is "nobody knows". The file `movie_buzz.cls` (from Greene's book) contains the following variables about 62 movies :
— `Box` = First run U.S. box office ($),
— `MPRating` = MPAA Rating code, 1=G, 2=PG, 3=PG13, 4=R,
— `Budget` = Production budget ($Mil),
— `Starpowr` = Index of star power,
— `Sequel` = 1 if movie is a sequel, 0 if not,
— `Action` = 1 if action film, 0 if not,
— `Comedy` = 1 if comedy film, 0 if not,
— `Animated` = 1 if animated film, 0 if not,
— `Horror` = 1 if horror film, 0 if not,
— `Addict` = Trailer views at `traileraddict.com`,
— `Cmngsoon` = Message board comments at `comingsoon.net`,
— `Fandango` = Attention at `fandango.com`,
— `Cntwait3` = Percentage of Fandango votes that can't wait to see.

1. Split the data into a training set and a test set.

2. Using the training data, generate different regression models using the following methods :
   — Best subset selection
   — Forward and backward selection
   — Ridge
   — Lasso
   For subset selection methods, keep the best models according to adjusted R^2 and BIC. For ridge and lasso, select the best model using cross-validation. Evaluate the models selected in the previous step using the test data.

3. Repeat the previous steps without splitting the data into a training set and a test set. Instead, use two nested cross-validation loops.

2   `Default_credit_card` data

We consider again the `default_credit_card` data.

1. Split the data into a training set of 20,000 observations and a test set of 10,000 observations.

2. Using the training data, estimate the error rates of the LDA, QDA, naive Bayes and logistic regression classifiers using 10-fold cross-validation. Compute the standard error of the cross-validation error rate. Select the classifier with the smallest cross-validation error rate.

3. Compute the test error rate of the best classifier selected in the previous step.