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.
3. Evaluate the models selected in the previous step using the test data.

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. 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.
4. Apply regularized discriminant analysis to this data (use function rda in package klaR.

3 USArrests

The USArrests dataset contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. This dataset is included in R. You can load it by the command load(USArrests).

1. Analyze these data using principal component analysis (PCA).
2. Plot the data in the space spanned by the first two components using the biplot function.
3. Interpret the first two components. Which proportion of the variance do they account for ?

4 Movie buzz data (continued)

Using the movie_buzz data, apply PCA to the four variables Addict, Cmngsoon, Fandango and Cntwait3. Repeat the analysis of Exercise 1, replacing these four predictors by their first principal component. Do this operation improve the prediction results ?