Linear models, part II Fall 2009

Lab 3: Model selection.

In this lab you will compare various model selection criteria, and examine the stability of model selection in regression.

The data set you will be working with pertains to the effect of pollution on mortality rates in different districts/cities. The outcome is the total age-adjusted mortality in rate per 100,000 (MORT). You will find a table with the explanatory variable descriptions below.
Note, the data set is quite old (1960s) so some of the variables may seem off-scale to you.

Variable name	Description
PREC	Average annual precipitation in inches
JANT	Average January temperature in degrees F
JULT	Same for July
OVR65	% of population aged 65 or older
POPN	Average household size
EDUC	Median school years completed by those over 22
HOUS	% of housing units which are sound and with all facilities
DENS	Population per sq. mile in urbanized areas
NONW	% non-white population in urbanized areas
WWDRK	employed in white collar occupations
POOR	% of families with income $< \$3000$
HC	Relative hydrocarbon pollution potential
NOX	Same for nitric oxides
SO2	Same for sulphur dioxide
HUMID	Annual average % relative humidity at 1pm

As you can see, the explanatory variables can be roughly divided into a couple of categories; (1) climate (temperature, precipitation); (2) socio-economic factors (household size, education level, etc); and (3) pollution statistics (e.g. nox).

1 Modeling the pollution data

Everyone will be working on different data sets for this lab. That is, you should start the lab by splitting the data into a training data set and a test data set. The original data set is of size 60. Choose your own data set size. Perhaps 30, 40 or 50 for training. Make sure to state what you do.

Before attempting model selection you have to explore the data set in full. Make sure you check the basic assumptions, presence of outliers, transformations, etc. Note again, since you all have chosen a random subset of data to work on, everyone will have slightly different results here -

different set of outliers etc.

Come up with an initial model for which all basic assumptions have been verified. Remember to check the residual plots, the normal-normal plots, and leverage and Cook's D plots for outliers. Discuss the coefficient magnitudes as best possible, bearing in mind that you may have correlated variables in your data set. In fact, compute the correlation matrix of the data `cor()` and comment. Given this correlation structure, which parameters in the model are particularly difficult to interpret?

1.1 Model selection

Once you are satisfied that you have a working model, consider model selection via $F - test$, AIC, BIC or C_p. You can use all-subset or backward selection. Compare at least two selection criteria in your report. Do you identify the same model using both criteria?

Please provide a table of models selected with different criteria. It is also a good idea to report the "trace" of the selection (i.e. the order in which the variables were dropped or added).
While our ultimate goal is prediction, try to come up with a reasonably interpretable model for the data.

1.2 Reality check

What if I told you the point of the exercise was to examine the relative impact/importance of socio-economic factors versus pollution on mortality. Think of (and apply) simple model selection type strategies to answer the following questions; (1) which is more important for mortality prediction - economic factors or pollution?; (2) given the socio-economic factors, do pollution data significantly contribute to explaining the variation of mortality rates between districts?; given the pollution factors, do socio-economic factors appear to contribute significantly to mortality?
Remember to check individual pairwise relationships between the outcome and each variable. Keep in mind that you might be dealing with collinear explanatory variables.

1.3 Selection stability

Going back to model selection via selection criteria. Create a new data set by selecting observations at random yet again. Do this repeatedly. I have provided code for you to do this, but I want you to choose another size for training and testing for example.

Examine the variability of model selection. Is the same model always selected? Is there a set of variables that are always selected. Summarize the (in)stability of model selection and discuss.

OPTIONAL: If time permits, you can try this exercise with a known model. That is, simulate some data for which your favourite model is true. Does model selection identify this model? All

the time? Most of the time? This depends on the noise level and sample size of course, which is something else you can play with.

Another thing you might want to try is to create a data set without collinearity problems. Look at the correlation matrix for the data and pick just one variable from each of the highly correlated sets. These variables for a new data set. Run model selection on this reduced data set. Is model selection more or less stable now?

2 Writing the report

Write a data analysis report with an introduction, a results section (selected models, validation), and a summary/conclusions.
Tables and figures should go in the main body of the text, and captions should be informative and complete. Don't intermix code and text in your report. Focus on the results and conclusion. Focus on answering the following questions; (1) Can you interpret your selected model - what have you learnt about the relationship between mortality and the pollution or socio-economic variables in the data set?; (2) Has model selection clearly identified the most important predictors of mortality, or are there open questions that the statistical analysis has not been able to resolve?; (3) which is more "important" (you decide what is meant by that term here) - pollution or socio-economic factors in explaining variation of mortality rates?