Lecture 28

Model Selection I

In this lecture, we are going to complete our discussion from the last class. Then we will discuss methods for building models; that is, for selecting the best predictor variables to include in a linear regression model. The data we will use are data about predicting baseball salaries from player statistics and are available at:
http://www.amstat.org/publications/jse/jse_data_archive.html

Highly Correlated Input Variables

For highly correlated input variables, there are three (different) standard things to do.

- Take out the variables that are causing the problem.
- Leave the variables in but caution the reader that the input variables were highly correlated and predictions are only valid if new inputs follow the same correlation. This is quite reasonable if, say, height and the square of height are variables in the model and are causing the problem.
- Replace the dependent variables by a single combination which captures their dependency. This involves computing principal components and can lead to issues of interpreting what the new variable means. It retains the same problem that the new variable is a specific linear combination of the original variables and any new value of the original variables that does not follow the original trend may be poorly estimated. I'll draw a picture in class to describe what the principal components do - but they find independent directions that explain the variance.

Covariates and Control

For the purposes of the GLM dialog box in Minitab, the word "covariate" means any continuous predictor variable in the model. A lay definition is a variable (continuous or categorical) that you want to control for. Any variables (continuous or categorical) that you want to control for goes into your GLM. Your interpretation of the effect of any one variable (or categorical variable group) is then "the effect given that all the other variables are in the model." That is, you have controlled for all of the other variables that have been included in the model.

Methods for Model Building

If two models are nested, they can be compared directly using a nested models F-test as discussed earlier. This is the method used for forward and backward selection procedures where variables are added (removed) from the model one at a time until the model does not become better (worse). Minitab can do something better than this - it compares all possible models using a best subset approach and spits out statistics about the models that allows you to choose your own best model. There is, however, no set way to compare non-nested models. There are many, many ways. A few of the common ways are described in the following table. In this table $\boldsymbol{k}$ is the number of predictors included in the model and $\boldsymbol{n}$ is the number of individuals (the sample size). The full model includes all possible predictors.

Criterion	Formula	Comments
$\boldsymbol{C_p}$	$\boldsymbol{(k+1)+(n-(k+1))\dfrac{\hat{\sigma}^2-\hat{\sigma}^2_{\text{Full}}}{\hat{\sigma}^2_{\text{Full}}}}$	Mallow's criterion. Want the value to be close to $\boldsymbol{k+1}$.
AIC	$\boldsymbol{n\log(\hat{\sigma}^2)+2(k+1)}$	Aikake Information Criterion. Larger penalty for more predictors. Want the value as small as possible.
BIC	$\boldsymbol{n\log(\hat{\sigma}^2)+(k+1)\log n}$	Bayesian Information Criterion. Even larger penalty for more predictors, especially if the sample size is large too. Tries to correct for over-fitting. Want the value as small as possible.
GCV	$\boldsymbol{\dfrac{n^2\hat{\sigma}^2}{(n-(k+1))^2}=\dfrac{n\text{SSE}}{(n-(k+1))^2}}$	Generalized Cross Validation. Approximates the leaving-one-out routine for cross-validation. You want this value as small as possible too.

There is no one way of selecting variables and there may be several nearly equally good models. You also want to use common sense and convenience when constructing a model. For example, variables that are expensive to collect should only go in the model if absolutely necessary, for instance. Also, if all else is equal, models with fewer variables are preferred to models with more variables.

We will go over how to use best subsets in Minitab with the baseball data. With that data, we should first figure out what the right transformations of the variables should be before we do variable selection procedures as described above.

Exercises for Lecture 28

1. –

2. –