Chapter 7

Linear Regression

(Multiple Regression Case)

Additional Considerations

- Have multiple possible explanatory variables
- Assume that explanatory variables are (roughly) independent
- Will need to select best subset of explanatory variables to use

Multicollinearity

Problems with highly correlated predictors:

- Model is more complicated to interpret
- Predictors confound each other
- Variance estimates will be larger
- Predictions will be less reliable
- Want predictors to not be highly dependent on each other

Checking for High Correlation

- Pairwise scatter plot for correlation between pairs of variables
- Use variance inflation factors (VIFs)
- $VIF_j = \frac{1}{1-{R_j}^2}$
- VIF_j>10 means at least 90% of x_j explained by other predictors

Model Selection

- Penalized goodness of fit measures (adjusted R^2, AIC, or BIC) for comparing models
- Other penalized measures like Mallows' C_p

Automatic Selection

- **Forward Selection** -- start with no terms, sequentially add significant terms
- **Backward Selection** -- start with all terms, sequentially remove insignificant terms
- **Stepwise Selection** – start with no terms, alternate between forward and backward steps
- **selection** option with **sle** and **sls** values in **reg**

Exercises: U.S. Crime Data

- Response: crime rate **R**
- Thirteen possible explanatory variables
- Will want to choose best subset of these 13 variables for modeling crime rate

Exercise: Visual Inspection

- Create a pairwise scatter plot matrix for all of the variables using **proc sgscatter**
- What does this plot tell us about relationships among the various predictors?
- What does the plot tell us about possible predictors for crime rate?

Exercise: All Predictors

- Fit **R** as a function of all the other variables and obtain the VIF values
- Predictors exceeding the VIF cutoff of 10?
- Omit the predictor with largest VIF and refit
- Any terms with VIF above the cutoff now?
- Which terms seem to be significant in this model?
- Any noticeable issues in the diagnostics?

Example: Stepwise Selection

- Start with all the predictors and significance levels of .05 for adding and for retaining terms
- What is the final model?
- Amount of variation in crime rate described by model?
- Problems in the diagnostics for this model?

Exercise: Forward Selection

- Use forward selection and entry significance level of .05
- Compare steps of the selection process
- What is the final model?
- Amount of variation in crime rate described by model?
- Problems in the diagnostics for this model?

Exercise: Backward Selection

- Use backward selection and significance level of .05 for keeping terms
- Compare steps of the selection process
- What is the final model?
- Amount of variation in crime rate described by model?
- Problems in the diagnostics for this model?