Chapter 6

Linear Regression

(Simple Linear Case)

Review: ANOVA Models

In Chapters 4 and 5:

- A continuous response
- One or more categorical explanatory variables
- Errors assumed to be iid N(0, $\hat{\sigma}^2$)
- Interested in differences of expected values for response between groups

Linear Regression Model

- Continuous response
- One or more explanatory variables
- Errors assumed to be iid N(0, $\hat{\sigma}^2$)

Common Linear Models

- Simple linear regression (Chapter 6)
- Multiple linear regression (Chapter 7)
- Regression through the origin (mention in Chapter 6)
- Polynomial regression
- Weighted linear regression
- ANOVA model
- ANCOVA model

proc reg

- Procedure for linear regression in SAS
- Similar setup to **proc anova** and **proc glm**
- Only continuous predictors
- Covers special case of general linear model
- Will provide many useful diagnostics

Parameter Estimates

- Obtained by minimizing sums of squared errors (e.g. least squares estimates)
- Tell us impact of predictor on response
- How much response is expected to change if predictor increases by 1
- Significant if significantly different from 0
- T statistics and confidence intervals

Residuals

- Differences between observed and predicted responses
- Assumed iid N(0, $\hat{\sigma}^2$)
- Quantile plots to visually check normality
- Plot against predictors or fitted value to see trends

Influence Diagnostics

- Leverage – measure of impact of data point on fitting
- Cook's distances – influence of individual data points on the fitting
- DFFITS – influence of individual data points on the predicted values
- DFBETAS – influence of individual data points on the parameter estimates

Goodness of Fit

- ANOVA tables
- R^2 value

Penalized Measures (examples):

- Adjusted R^2 value
- AIC (Akaike Information Criterion)
- BIC (Bayesian Information Criterion)

Example

- Simple Linear Regression example for **proc reg**

Exercise: Cirrhosis and Alcohol

Data set:

- Data from 15 countries
- Cirrhosis deaths per 100,000 people
- Annual alcohol consumption (in litres per person per year)

Example: Linear Trend?

- Considering cirrhosis deaths as a function of alcohol consumption
- Create a scatter plot of the data
- Linear trend reasonable?
- Any indications of problems with the data?

Exercise: Linear Regression

- Fit linear regression model with cirrhosis deaths as response
- Comment on quality of model
- Comment on any problems noticed in diagnostics
- Relationship between alcohol consumption and cirrhosis related death rate?

Exercise: Undue Influence

- Points too influential based on Cook’s distance?
- Use **output** statement to write Cook’s distance values to data set
- Remove points with Cook’s distance greater than 1 and refit the model
- How do the results change?
- Any remaining problems with the model?

Exercise: Zero-Intercept Model

- Fit the model containing alcohol but no constant term using the full data set (see **noint** option)
- Compare results with those for model with intercept
- Remove highly influential points and re-fit zero-intercept model
- Which of the models would be best and why?
- Are there any remaining concerns about the model and underlying assumptions?