Chapters 6 and 7 Notes

Linear Regression

Linear regression models fit a response to a linear combination of predictors. These predictors are variables (or potentially functions of variables), and we assume that we have mean 0 iid normal residuals.

- Simple linear regression (single predictor and a constant term):

  $y = \beta_0 + \beta_1 x + \epsilon$

- Multiple linear regression (just means we have multiple non-constant predictors):

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \epsilon$

- Regression through the origin (force the model through 0 when all predictor variables are 0 by excluding a constant term):

  $y = \beta_1 x_1 + \beta_2 x_2 + ... + \epsilon$

- Polynomial regression (predictors are powers of a common variable):

  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \epsilon$

- Weighted linear regression (weights used to lessen the impact of some data points)

- ANOVA model (all non-constant predictors are nominal categorical variables, coded with dummy variables)

- ANCOVA model (some predictors are continuous and some are nominal categorical variables)

The models above are written in terms of the actual observations. We could write formulas for the predictions by adding hats (^) to the y's and β's and dropping the error terms.

We will talk about simple linear regression (Chapter 6 in the text) first, and then discuss multiple linear regression (Chapter 7 in the text).

Parameter Estimates

The parameter estimates (the $\hat{\beta}$'s) tell us the impact of the individual predictors on the model. If a parameter estimate is far enough from 0, it would be deemed statistically significant. We can look at t-statistics and their associated p-values to determine significance of individual terms in the model.

Confidence intervals tell us about where parameter estimates would be expected to be observed if the assumptions of the model are true. If 0 is not in the interval at a chosen confidence level, the parameter estimate would be considered significant at that confidence level.

Residuals

Residuals are the differences between the observed and predicted values. For a simple linear regression model, we can visualize these as lines between the data points and the fitted curve.

Residuals are very important in diagnostic checking for linear models. The residuals allow us to estimate the error variance, and measure how well the model captures the variation in the data (if the sum of squared residuals is small relative to the overall variation, it explains a lot of the variation and the model matches the data well), and we need the residuals for diagnosing problems with the assumptions of the model.

Trends in residual plots indicate problems with the assumptions of the model. Plotting residuals against predicted values or predictor variable values can show issues of non-constant variance, and nonlinearities. Studentized residuals are scaled versions of the residuals. They are scaled by the expected standard deviation of each residual $\hat{\sigma}_{(i)}\sqrt{1-h_i}$ where $\hat{\sigma}_{(i)}$ is the standard deviation estimate if the i^{th} data point is left out and h_i is the i^{th} diagonal element of the hat matrix, which is a measure of leverage of the the i^{th} data point.

Quantile plots can be used to check the assumption of normally distributed error terms (the residuals are realizations of error terms).

Influence Diagnostics

These diagnostics tell us something about the impact of individual data points on the model or predicted results. We want to model typical behavior. Points that are extremely influential on the fitting may pull our model away from typical behavior. Formulas for these and other model diagnostics are given in the SAS documentation under **The REG Procedure>Details>Model Fit and Diagnostic Statistics**.

The diagnostics we will discuss are deletion diagnostics. They provide measurements of how some aspect of the fitted model will change if individual points are left out or deleted from the data set (so we will have a value of each diagnostic for each data point).

Cook's distances tell us about influence of individual data points on the fitting. Data points with large Cook's distances (large being relative to the other Cook's distances) may have undue influence on the fitting. We might want to remove points with large Cook's distances or give them less weight to reduce their impact on the fitting.

DFFITS gives a measure of the influence of individual data points on the predicted values.

DFBETAS gives a measure of the influence of individual data points on the parameter estimates.

Large DFFITS and DFBETAS values may also be cause for removing or at least reducing the weight of a point.

Goodness of Fit

We will again have ANOVA tables. These will tell us if the variation described by the model is significantly greater than would be expected due to error.

We will again have R^2 values to tell us how much of the variation in the response is captured by the systematic portion of the model. As we get into multiple regression we will want to consider measures that penalize for adding more terms to the model (e.g. adjusted R^2, AIC, and BIC).

Linear Regression in SAS

In SAS we will use **proc reg** to perform linear regressions. It will generate many of the diagnostic results we want by default. The **Simple Linear Regression** example from the SAS documentation will give us a general idea of what results are generated and what they mean.

Exercises

Let's start with the **Alcohol and Cirrhosis** data set from the text book. It's not saved as a data set, so we will need to grab the code from the **programs** directory.

Simple Linear Trend

- Plot the data with **cirrhosis** related death rate (deaths per 100,000) as the y variable and **alcohol** consumption (litres per person per year) as the x variable.
- Does a linear trend seem reasonable?
- Do there appear to be any potentially problematic points?

Simple Linear Regression Model

- Fit a simple linear model for **cirrhosis** as a function of **alcohol**.
- Discuss the results and any problems which may need to be accounted for.
- Is the model significant?
- Are the terms in the model significant?
- What does the model tell us about the relationship between **alcohol** consumption and **cirrhosis** related death rate?

Accounting for Undue Influence

We can look at diagnostics to see general issues. We can also write out diagnostics to an output data set using the **output** statement in **proc reg**. We could then select particular data points based on diagnostic values (or weight data points based on diagnostic values).

- In the previous fitting, are there any points that may be too influential based on Cook's distance?
- We can read them off the plots, and we can also write out the values of interest using an **output** statement and then look at the Cook's distance values.
- If there are highly influential points (use a cutoff of 1), refit the model with those points removed.
- How do the results change?
- Do we see any remaining problems with the model?

Regression through the Origin

If we believe that cirrhosis-related deaths should go to 0 linearly when no alcohol is consumed, it would be reasonable to remove the constant term and force this to be true in our model.

- Fit the model containing the alcohol predictor but no constant term using the full data set.
- Compare your results with those from the model which contained an intercept.
- If there are highly influential points in the zero-intercept model, remove those points and re-fit the model.
- Of the models we've fit to the Alcohol and Cirrhosis data, which model do we think would be best and why?
- Are there any remaining concerns about the model and underlying assumptions?

Additional Issues for Multiple Regression

Multicollinearity

In multiple regression, the model involves multiple explanatory predictor variables (or functions of predictor variables). We assume that these variables are not highly related. If we have high correlation or dependencies between predictors, we have a problem called **multicollinearity**.

In the presence of multicollinearity, multiple predictors are telling us much of the same information. This is problematic for a few reasons:

- Having more terms makes a model more complicated to interpret in general.
- It will be difficult to interpret the impact of the highly correlated predictors because their impacts are confounded by each other.
- Variance estimates will be larger, so our predictions will be less reliable.

We will want to diagnose and remedy issues of multicollinearity. Graphically we can look for signs of high correlation between pairs of predictors by looking at a pairwise scatter plot via **proc sgscatter**. If we plot one variable against another and the scatter is roughly linear, there is a strong linear correlation between the two variables.

One useful quantitative measure for relationships among more than 2 predictors is called the variance inflation factor (**VIF**). The variance inflation factor for the j^{th} predictor is given by

$$VIF_j = \frac{1}{1 - {R_j}^2}$$

where ${R_j}^2$ is the R^2 value for predicting the j^{th} predictor by a linear combination of all the other predictors. If the VIF for a predictor is very high, that means that most of the information we get from that predictor is already described by the other predictors.

Note that while we can see correlation between two variables in the scatter plot, the variance inflation factors will pick up on more complex relationships. For instance, if x_4 is roughly a linear combination of x_3, x_2, and x_5, VIF_4 could point that out, but the relationship might not be as apparent from the scatter plots of x_4 against x_3, x_2, and x_5 individually.

In the multiple regression setting, we may want to sequentially remove predictors with very high VIF values and refit. The text suggests a cut-off of 10, which would correspond to 90% of the variation in

that predictor being explainable by the other predictors. Other possible solutions (which we won't go into detail on here…) include:

- principal component regression, where principal component analysis is used to get a new set of uncorrelated variables which are linear combinations of the original predictor variables, and the model is then fit using those new variables
- ridge regression, where a bias is introduced in the parameter estimates to reduce variances

Model Selection

We can still look at goodness of fit measures, such as R^2 as we mentioned before, but we want to keep the model simple and we want the terms to be significant. As we add more terms R^2 will get larger (or at least no smaller). We would want to use something like adjusted R^2, AIC, or BIC to compare models. These allow for comparing models with different numbers of parameters and include a penalty for adding more terms to the model.

Automatic Selection

With many possible predictors, we may want a way to automatically choose which terms to include or exclude from the model. Three typical approaches are:

- **Forward Selection** -- starts with no terms and sequentially adds terms if their inclusion would contribute significantly to the model
- **Backward Selection** -- starts with all possible predictors and sequential removes terms that do not contribute significantly
- **Stepwise Selection** -- starts with no terms and does a forward step and then performs a backward step when a term is added to see if any previous terms are no longer significant, then continues with forward and backward steps until the model doesn't change

The tests for significance at each step are once again F tests or based on finding Mallows' C_p close to p.

In SAS, we can add a **selection** option to the **model** statement in **proc reg** to define a selection method. The significance level for adding a term to the model is set by **sle**, and the significance level for keeping a term in the model is set with **sls**.

As an alternative to automated selection, we compare penalized measures such as adjusted R^2 for subsets of the predictors.

Exercises

To explore multiple regression and automatic model selection, let's start with the U.S. crime data in **uscrime.dat**. The variables are described on pages 129 and 130 of the text. We will be modeling **R** (crime rate) as a function of the other variables.

Visual Inspection

- Look at a pairwise scatter plot 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?

Starting with All Predictors

- Fit the linear regression for **R** as a function of all the other variables and also get the VIF values.
- Which predictors would exceed the VIF cutoff of 10?
- If there are terms above the cutoff, fit the linear regression model omitting the predictor with largest VIF and note any remaining terms with VIF above the cutoff.
- Which of the terms seem to be significant in this model?
- Are there any noticeable issues in the diagnostics (e.g. highly influential points, deviation from normality, etc.)?

Automated Selection

For the following exercises, it is useful to note that if we only want to see a summary of the selection (and not the entire history of results), we can get just the **SelectionSummary** ods table. We could then re-fit the model and look at diagnostics using only the terms from the final selection.

- Use stepwise selection starting with all the predictors and significance levels of .05 both for adding and for retaining terms.
- What is the final model?
- How much of the variation in crime rate does the model describe?
- Are there any indications of problems in the diagnostics for this model?
- Repeat the previous analysis using forward selection and an entry significance level of .05.
- Repeat using backward selection and significance level of .05 for keeping terms.
- Are there differences in the selected models, or do all selection methods agree in this particular case?