Lab 7: Introduction to Linear Regression

Batter Up

The movie *Moneyball* focuses on the "quest for the secret of success in baseball." It follows a low-budget team, the Oakland Athletics, who believed that underused statistics, such as a player's ability to get on base, better predict the ability to score runs than typical statistics like home runs, RBIs (runs batted in), and batting average. Obtaining players who excelled in these underused statistics turned out to be much more affordable for the team.

In this lab, we look at data from all 30 Major League Baseball teams and examining the linear relationship between runs scored in a season and a number of other player statistics. Our aim is to summarize these relationships both graphically and numerically in order to find which variable, if any, helps us best predict the runs scored in a season by a team.[1]

The Data

Let's look at the data for the 2011 season. First, we need to load the data from the OpenIntro website.

If you are using SAS University Edition, you need to ensure that interactive mode is turned off. To do this, click the button to the right of **Sign Out** in the upper right corner of the window and then click **Preferences**. In the Preferences window, on the General tab, the bottom check box (located next to the text **Start new programs in interactive mode**) should ***not*** be selected. If the box is selected, you need to clear it and save your change.

```
filename mlb11 url
         'http://www.openintro.org/stat/data/mlb11.csv';

proc import datafile=mlb11
            out=mlb11
            dbms=csv
            replace;
   getnames=yes;
run;
```

[1] This is a product of OpenIntro that is released under a Creative Commons Attribution-ShareAlike 3.0 Unported (http://creativecommons.org/licenses/by-sa/3.0/). This lab was adapted for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel from a lab written by the faculty and TAs of UCLA Statistics. The lab was then modified by SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries (® indicates USA registration) and are not included under the CC-BY-SA license.

Though it's not necessary for this lab, if you'd like a refresher in the rules of baseball and a description of these statistics, visit http://en.wikipedia.org/wiki/Baseball_rules and http://en.wikipedia.org/wiki/Baseball_statistics.

In addition to runs scored, there are seven traditionally used variables in the data set: at-bats, hits, home runs, batting average, strikeouts, stolen bases, and wins. There are also three new variables: on-base percentage, slugging percentage, and on-base plus slugging. For the first portion of the analysis, we consider the seven traditional variables. At the end of the lab, you work with the new variables on your own.

Exercise 1: What type of plot would you use to display the relationship between **runs** and one of the other numerical variables? Plot this relationship using the variable **at_bats** as the predictor. Does the relationship look linear? If you knew a team's at-bats, would you be comfortable using a linear model to predict the number of runs?

If the relationship looks linear, we can quantify the strength of the relationship with the correlation coefficient.

```
proc corr data=mlb11;
   var runs at_bats;
run;
```

The CORR procedure begins with a DATA= option to tell SAS that we are working with the **mlb11** data set. The VAR statement lists the variables for which we want to obtain a correlation table (**runs** and **at_bats**).

Plotting a Regression Line

Think back to how we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It's also useful to be able to describe the relationship of two numerical variables, such as **runs** and **at_bats** above.

Exercise 2: Looking at your plot from the previous exercise, describe the relationship between these two variables. Be sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

Just as we used the mean and standard deviation to summarize a single variable, we can summarize the relationship between these two variables by finding the line that best follows their association. The most common way to do linear regression is to select the line that minimizes the sum of squared residuals. Use the following syntax to plot the observations and the regression line that minimizes the sum of squared residuals.

```
proc sgplot data=mlb11;
   reg x=at_bats y=runs;
run;
```

The SGPLOT procedure begins with a DATA= option to tell SAS that we are working with the **mlb11** data set. The REG statement requests that SAS produce a scatter plot with a least squares regression line overlaid. The X= specifies the variable to be plotted on the x-axis, and the Y= specifies the variable to be plotted on the y-axis. Note that using the REG procedure to regress **at_bats** on **runs** would produce a similar graph by default, but with the addition of confidence and prediction limits.

The Linear Model

The syntax above provided us with a plot that showed the correct least squares line, but it did not provide us with any further information. Instead, we can use the REG procedure to get estimates of the intercept and slope of the regression line.

```
proc reg data=mlb11;
   model runs=at_bats;
run;
quit;
```

The REG procedure begins with a DATA= option to tell SAS that we are working with the **mlb11** data set. The MODEL statement takes the form *y*=*x*. Here it can be read that we want to estimate a regression model of **runs** as a function of **at_bats**. The default output from the REG procedure contains all of the information that we need about the linear model that was just fit, as well as an assortment of plots that are useful for model diagnostics.

Let's consider this output piece by piece. First, the dependent variable is shown at the top. Next, the output displays the number of observations read and the number of observations used. These would differ if observations were excluded from the analysis due to missing data. Next comes the Analysis of Variance table. This table provides information about the overall significance of the model. After the ANOVA table comes a table containing the root mean square error, the mean of the dependent variable, the coefficient of variation, the R-square, and the adjusted R-square.

The R-square value is also called the multiple R-squared or, more simply, R^2. The R^2 value represents the proportion of variability in the response variable that is explained by the explanatory variable. For this model, 37.3% of the variability in **runs** is explained by **at_bats**.

The Parameter Estimates table shown next is key. The Parameter Estimate column displays the linear model's y-intercept and the coefficient of **at_bats**. With this table, we can write down the least squares regression line for the linear model:

$$\hat{y} = -2789.2429 + 0.6305 * atbats$$

Exercise 3: Fit a new model that uses **homeruns** to predict **runs**. Using the estimates from the SAS output, write the equation of the regression line. What does the slope tell us in the context of the relationship between success of a team and its home runs?

Prediction and Prediction Errors

We previously created a scatter plot with the least squares line laid on top using the following syntax:

```
proc sgplot data=mlb11;
   reg x=at_bats y=runs;
run;
```

The line generated by PROC SGPLOT can be used to predict y at any value of x. When predictions are made for values of x that are beyond the range of the observed data, it is referred to as *extrapolation* and is not usually recommended. However, predictions made within the range of the data are more reliable. They are also used to compute the residuals.

Exercise 4: If a team manager saw the least squares regression line and not the actual data, how many runs would he or she predict for a team with 5,578 at-bats? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

Model Diagnostics

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

1. Linearity: You already checked whether the relationship between **runs** and **at_bats** is linear using a scatter plot. We should also verify this condition with a plot of the residuals versus **at_bats**. A plot of the residuals versus **at_bats** is produced by the REG procedure by default, and is labeled "Residuals for runs."

Exercise 5: Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between **runs** and **at_bats**?

2. Nearly normal residuals: To check this condition, we can look at a histogram or a normal probability plot of the residuals. Both of these plots are produced by the REG procedure by default. However, these graphs are output as part of a panel and might be smaller than we would like. We can request that the graphs be output separately by modifying the regression syntax as follows:

```
proc reg data=mlb11 plots=diagnostics(unpack);
   model runs=at_bats;
run;
quit;
```

The histogram is labeled "Distribution of Residuals for runs" and the normal probability plot is labeled "Q-Q Plot of Residuals for runs."

Exercise 6: Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

3. Constant variability:

Exercise 7: Based on the plot in (1), does the constant variability condition appear to be met?

On Your Own

1. Choose another traditional variable from **mlb11** that you think might be a good predictor of **runs**. Produce a scatter plot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?
2. How does this relationship compare to the relationship between **runs** and **at_bats**? Use the R^2 values from the two model summaries to compare. Does your variable seem to predict **runs** better than **at_bats**? How can you tell?
3. Now that you can summarize the linear relationship between two variables, investigate the relationships between **runs** and each of the other five traditional variables. Which variable best predicts **runs**? Support your conclusion using the graphical and numerical methods that we have discussed (for the sake of conciseness, include only output for the best variable, not all five).

4. Now examine the three new variables. These are the statistics used by the author of *Moneyball* to predict a team's success. In general, are they more or less effective at predicting runs than the old variables? Explain using appropriate graphical and numerical evidence. Of all 10 variables that we have analyzed, which seems to be the best predictor of **runs**? Using the limited (or not so limited) information that you know about these baseball statistics, does your result make sense?
5. Check the model diagnostics for the regression model with the variable that you decided was the best predictor for runs.
6. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere (for example, lecture, discussion section, previous labs, or homework problems)? Be specific in your answer.