**SEIS 631 Foundations of Data Analysis**

**Assignment 8**

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 assignment, we’ll be looking 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 will be to summarize these relationships both graphically and numerically in order to find which variable, if any, helps us best predict a team’s runs scored in a season.

The data is stored in the **mlb11.RData** file. Load the data.

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 newer variables: on-base percentage, slugging percentage, and on-base plus slugging. We’ll consider the seven traditional variables.

|  |
| --- |
| **Q1)** 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.

**cor(mlb11$runs, mlb11$at\_bats)**

|  |
| --- |
| **Q2)** What correlation did you get? Describe the relationship between *at\_bats* and *runs*? Make 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. Use the following interactive function to select the line that you think does the best job of going through the cloud of points.

**plot\_ss(x = mlb11$at\_bats, y = mlb11$runs)**

After running the above command, you’ll be prompted to click two points on the plot to define a line. Once you’ve done that, the line you specified will be shown in black and the residuals in blue. Note that there are 30 residuals, one for each of the 30 observations. Recall that the residuals are the difference between the observed values and the values predicted by the line:

ei=yi-yhati

The most common way to do linear regression is to select the line that minimizes the sum of squared residuals. To visualize the squared residuals, you can rerun the plot command and add the argument showSquares = TRUE.

**plot\_ss(x = mlb11$at\_bats, y = mlb11$runs, showSquares = TRUE)**

Note that the output from the plot\_ss function provides you with the slope and intercept of your line as well as the sum of squares.

**The linear model:** It is rather cumbersome manually to try to get the correct least squares line, i.e. the line that minimizes the sum of squared residuals, through trial and error. Instead we can use the **lm** function in R to fit the linear model (a.k.a. regression line).

**m1 <- lm(runs ~ at\_bats, data = mlb11)**

First argument can be read as we want to make a linear model of runs as a function of at\_bats. The second argument specifies that R should look in the mlb11 data frame to find the runs and at\_bats variables.

The output of lm is an object that contains all of the information we need about the linear model that was just fit. We can access this information using the summary function.

**summary(m1)**

Let’s consider this output piece by piece. First, the formula used to describe the model is shown at the top. After the formula you find the five-number summary of the residuals. The “Coefficients” table shown next is key; its first 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:

yhat = -2789.2429 + 0.6305\*atbats

One last piece of information we will discuss from the summary output is the Multiple R-squared, or more simply, R2. The R2 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.

|  |
| --- |
| **Q3)** Fit a new model that uses homeruns to predict runs. Using the estimates from the R 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:** Let’s create a scatterplot with the least squares line laid on top.

**plot(mlb11$runs ~ mlb11$at\_bats)**

**abline(m1)**

The function abline plots a line based on its slope and intercept. Here, we used a shortcut by providing the model m1, which contains both parameter estimates. This line 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’re also used to compute the residuals.

|  |
| --- |
| **Q4)** 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 5,579 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 if the relationship between runs and at-bats is linear using a scatterplot. We should also verify this condition with a plot of the residuals vs. at-bats. Recall that any code following a # is intended to be a comment that helps understand the code but is ignored by R.

**plot(m1$residuals ~ mlb11$at\_bats)**

**abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0**

|  |
| --- |
| **Q5)** Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats? |

1. Nearly normal residuals: To check this condition, we can look at a histogram, **hist(m1$residuals)**, or a normal probability plot of the residuals.

**qqnorm(m1$residuals)**

**qqline(m1$residuals) # adds diagonal line to the normal prob plot**

|  |
| --- |
| **Q6)** Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met? |

1. Constant variability:

|  |
| --- |
| **Q7)** Based on the plot in (1), does the constant variability condition appear to be met? |

**Submission:**

Submit the following two files (no need to compress into a folder):

* **Pdf or word document** with answers to all the questions asked in the assignment.
* **.R file** with relevant code. It is useful to add comments within the code with a pound sign.