Joseph McGill
112840126
Math 4753
25 April 2016

Math 4753 Linear Regression Project

In hockey a common question arises: "how do we win more games?" Unfortunately, the answer is generally not very simple. There are many statistics recorded for players and teams. Is there a correlation between any of these variables? A popular saying is "defense wins games," but is this really true? This study will look at two statistics for hockey teams – goals against and wins – in an effort to answer the question "does defense really win games?"

The data for this study was collected during the 2014-15 NHL season for all 30 teams in the league. We are concerned with goals against and wins, both discrete variables with no units. Goals against (GA) is the total number of goals that a team's opponent scored against them during the season. Wins (W) is the total number of wins for a team during the season.

This study will use a simple linear regression analysis (SLR) to determine and quantify the relationship between goals against and wins. Simple linear regression creates a linear model that fits a trend in the data in a way that minimizes the sum squared of the residuals. A SLR analysis attempts to explain the connection between the independent variable (x) and the dependent variable (y) of the data. The linear model takes the form

$$y_1 = \hat{\beta}_0 + \hat{\beta}_1 * x_i + \varepsilon_i$$

where $\hat{\beta}_i$ is estimated using the least squares method, so

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Data retrieved from http://www.hockey-reference.com/leagues/NHL_2015.html

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

Since this study is attempting to quantify a relationship between two variables, a simple linear regression analysis is appropriate. In order to use a SLR analysis the variables must satisfy four assumptions:

1. There is a linear relationship between x and y (GA and W)
2. The errors are independent – that is, y is independent identically distributed
3. The errors are distributed normally
4. The errors have a mean of 0 and a constant variance

To prove that there is a linear trend in the data a Lowess Smoother plot will be used. A Lowess Smoother plot fits a line to the overall trend in the data. In R, a Lowess Smoother plot is created using the trendscatter( ) function from the s20x package. **Figure 1** shows the Lowess Smoother plot created for the data using R. We can see that the data follows a roughly linear trend, which satisfies our assumption.

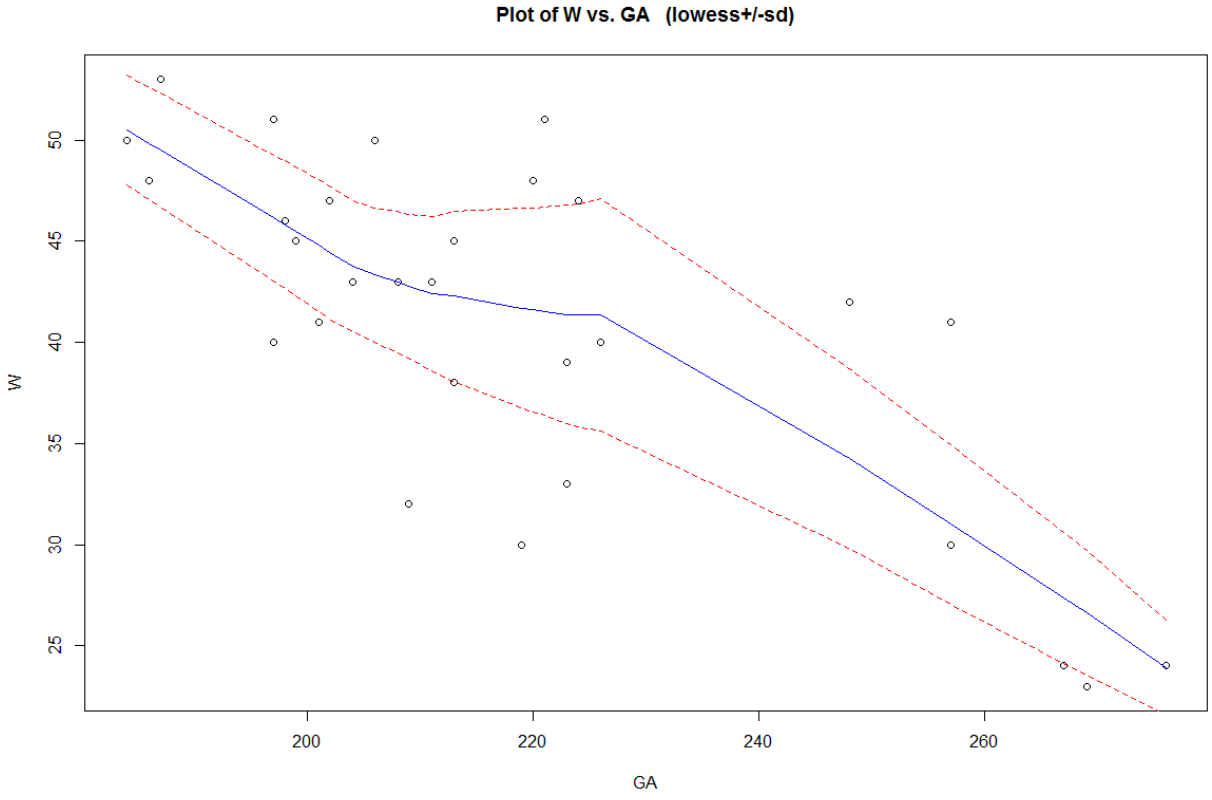

Figure 1: Lowess Smoother plot for Wins vs Goals Against

For the second assumption that the errors are independent identically distributed, we only need to examine the data itself. For a given hockey team, the goals against for a team are not affected by another hockey team's goals against. Similarly, a team's wins is not influenced by another team's wins. This is sufficient enough to say that the errors are distributed independently.

To prove that the errors are distributed normally, it would be useful to look at a Normal Q-Q plot of the residuals, as well as a histogram of the residuals to try and identify a normal distribution. **Figure 2** shows the Normal Q-Q plot and Histogram from a Shapiro-Wilk test of the residuals created in R

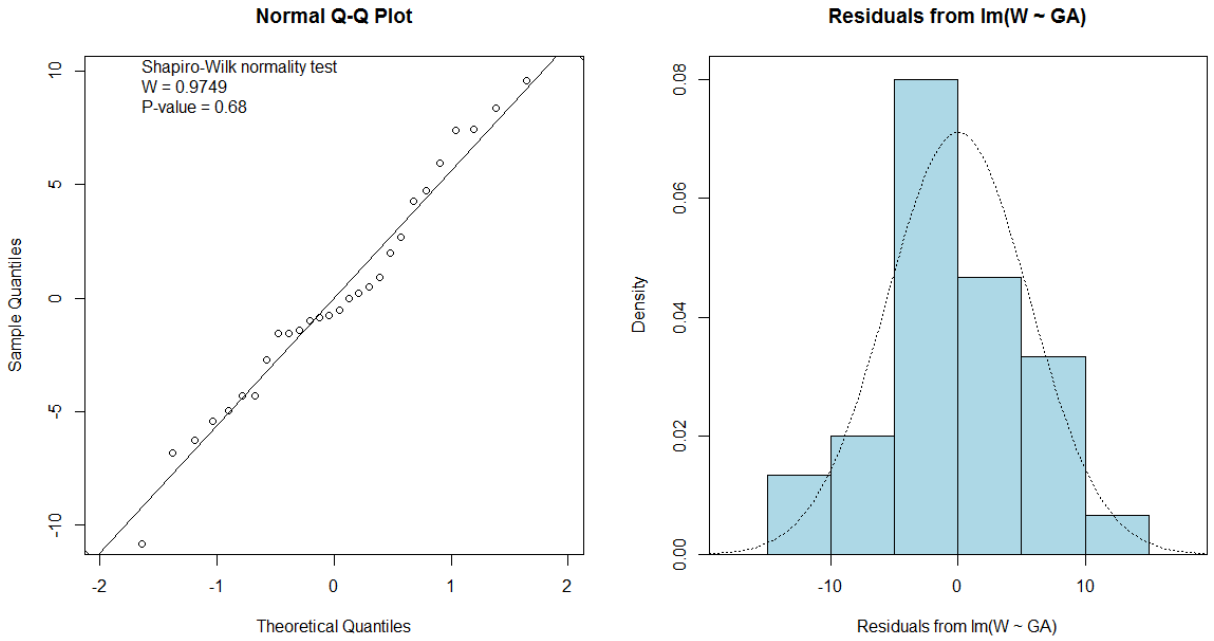

Figure 2 Q-Q Plot and Histogram from the Shairo-Wilk normality test

From the histogram above, we can see that the residuals are distributed fairly normally. Additionally, the p-value in the Q-Q plot is 0.68, which means we should not reject the null hypothesis that the residuals are distributed normally.

The fourth assumption is that the residuals have a constant variance. To check this assumption we will plot the residuals vs. the fitted values of the linear model. If the data is constantly distributed about the origin, we can say that the residuals have a constant variance. **Figure 3** shows the residuals vs. the fitted values.

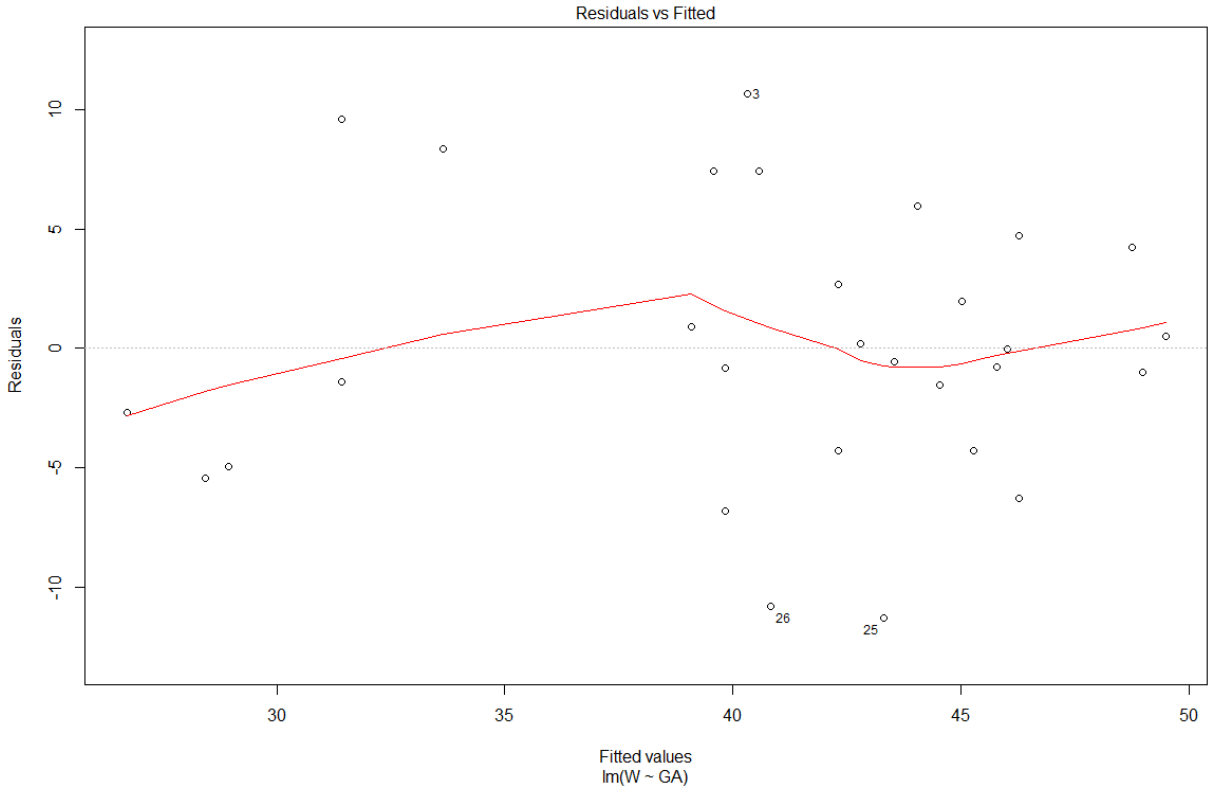

Figure 3 Residuals vs. Fitted values

We can see in the plot above that the data is distributed constantly about the origin except for the first part of the data. This supports the assumption that the residuals are distributed with a constant variance. With all four of the assumptions satisfied, we can say that a SLR analysis is appropriate for this data.

Now that we have proven that the model is appropriate, we can actually analyze the model. Upon examining the summary of the linear model, it can be seen that the model only accounts for 55.58% of the variance in the data. Ideally a better model would account for more variance, but we are using a SLR analysis so it will do. Using R, we can get point estimates for the parameters $\hat{\beta}_0$ and $\widehat{\beta}_1$:

```
> hockey.lm$coefficients
(Intercept)          GA
 95.0639723  -0.2476591
```

$\hat{\beta}_0 = 95.06397$
$\hat{\beta}_1 = -0.24766$

We can also generate 95% confidence intervals for both parameters:

```
> ciReg(hockey.lm)
             95 % C.I.lower    95 % C.I.upper
(Intercept)       76.23426         113.89368
GA                -0.33336          -0.16196
```

So our confidence intervals are

$\hat{\beta}_0 = (76.23426, 113.89368)$
$\hat{\beta}_1 = (-0.33336, -0.16196)$

From the confidence intervals, we can say with 95% confidence that the underlying change in y over the change in x is at least -0.33336 and at most -0.16196. The point estimate for $\hat{\beta}_1$ is within this interval, therefore it can be used to make predictions about the number of games won given goals against in a hockey season.

Predictions were made in R using 3 values (185, 200, 215) for Goals Against:

```
GA = 185, W = 49.24705
GA = 200, W = 45.53216
GA = 215, W = 41.81727
```

These predicted values are plotted on **Figure 4** below. The plot contains the actual data and the line from the linear model. The blue dots on the red line represent the predicted values.

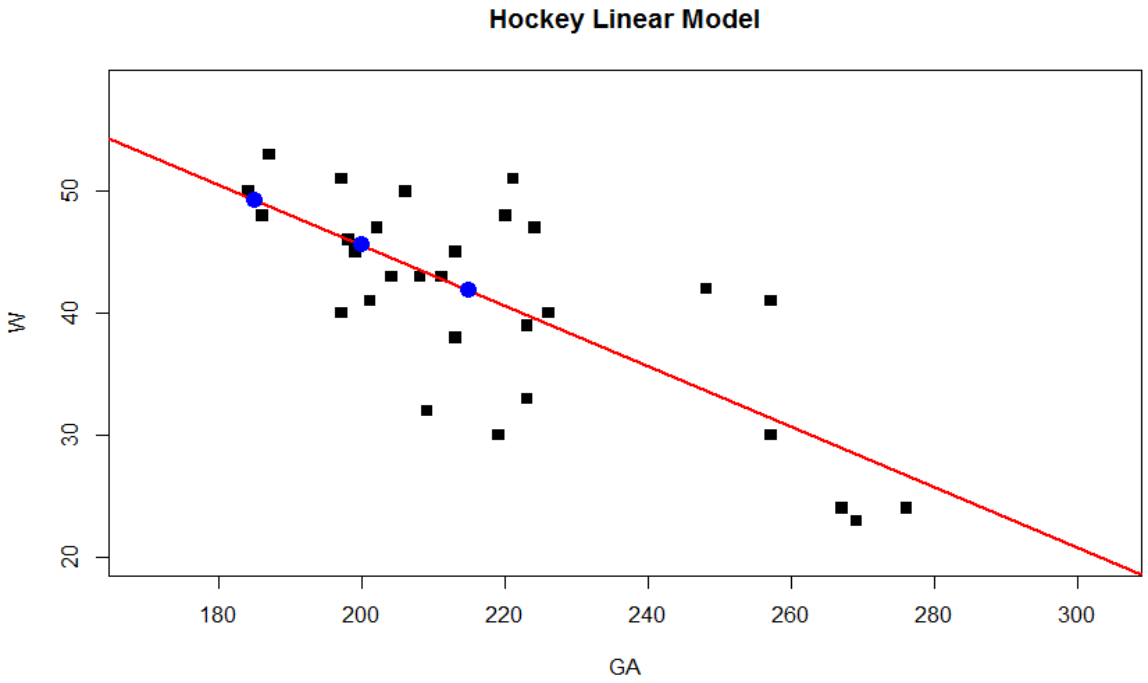

Figure 4 Predicted values

From the plot above, we can see there are some possible outliers in the data. Using a Cook's Distance plot we can determine possible outliers and their relative influence on the data set. When using a Cook's Distance plot, a data point with a distance greater than 0.1 is considered a possible outlier. **Figure 5** below shows the Cook's Distance plot for the hockey data.

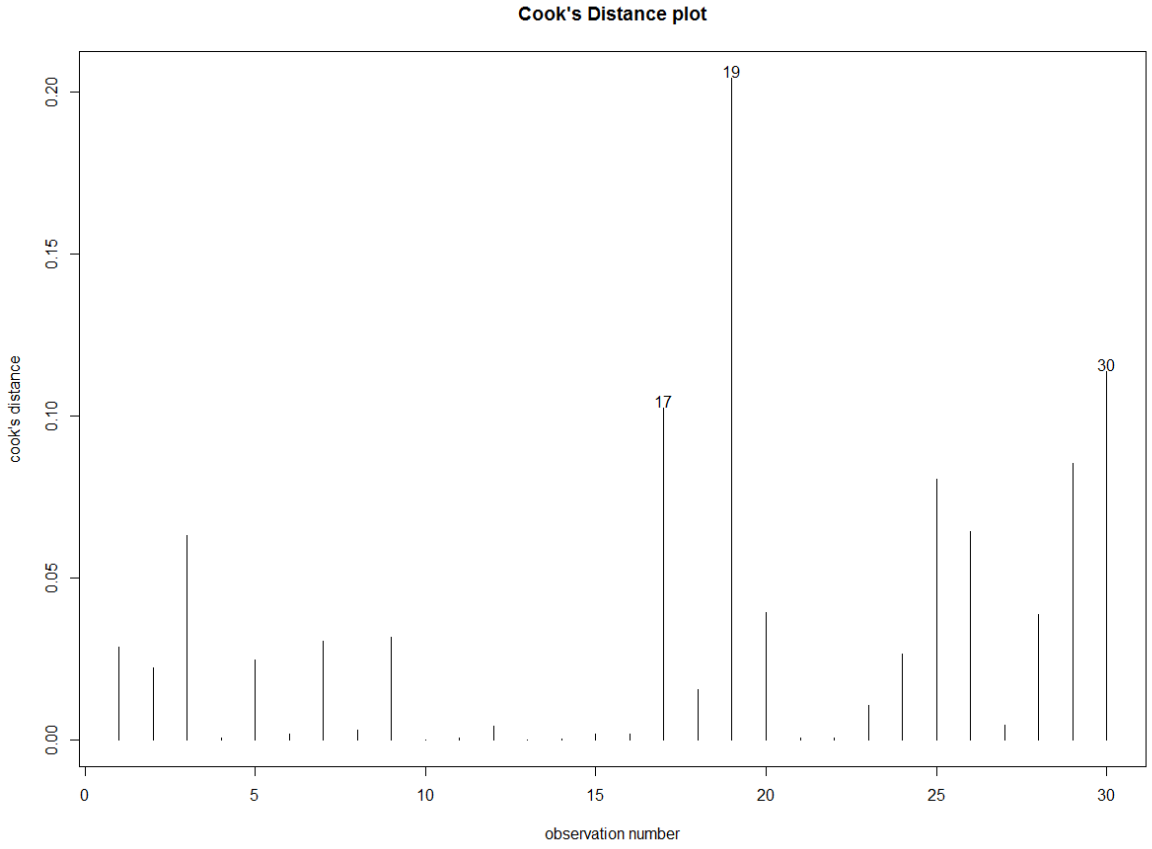

Figure 5 Cook's Distance plot

There are 3 possible outliers in the data at indices 17, 19, and 30. These indices correspond to the Columbus Blue Jackets, the Dallas Stars, and the Buffalo Sabres, respectively. The largest (and most influential) distance is the Dallas Stars at 0.20. Upon examining the data entry for the Dallas Stars we have

```
Rk          Team  W  L OL PTS  PTS.  GF  GA TG.G   PK.    S S.   SA   SV.
19 Dallas Stars 41 31 10  92 0.561 257 257 6.35 80.71 2558 10 2454 0.895
```

The Dallas Stars have a higher win count then their Goals Against would suggest. This can be explained by the Dallas Stars scoring the 2nd most goals in the league with 257 Goals For.

Now that an analysis has been done, the question "Does defense win games in hockey?" can be answered. Using a linear model it can be said with 95% confidence that after between 76.23426 and 113.89368 goals against a team, every goal allowed reduces the number of wins by at least 0.16196 and at most 0.33336. In other words, for every 4 goals allowed after (approximately) 95, the number of wins for a team is reduced by 1. So, does defense win games? According to SLR analysis performed – yes, it does.