I wanted to do a basic linear regression project as my first demonstration of my data analysis skills. I decided to explore the relationship between passing attempts and passing yards in the NFL in 2018. I compiled the passing statistics of players who attempted at least one pass in a given game this season (including playoffs.) There were 267 total games and there were 646 instances of players throwing at least 1 pass in those games. This is a very large sample size which is great for analysis. I wanted to determine and interpret the linear regression model to explain how much of the response variable passing yards can be explained by the explanatory variable passing attempts.

My data set can be found here:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

I began by creating a scatterplot of the 646 data points.

(Pic of scatterplot)

It appears clear that there is a positive correlation in the data which makes sense because more passing attempts should lead to more passing yards. The correlation coefficient of this data is 0.872 which is very close to 1. This means that there is a very strong correlation between attempts and yards. Although the scatterplot and the correlation coefficient both useful tools, they need to be examined together in order to arrive at any conclusions.

I ran a linear regression analysis and got the following results.

(Pic of linear regression coefficients and scatterplot with line)

The y intercept of the equation is 9.55. This can be interpreted as the starting point for passing yards when passing attempts = 0. However, since my data for passing attempts has the range of 1 through 62, this y-intercept does not have practical meaning.

The slope of the equations is 7.017. This can be interpreted as the amount that you can expect passing yards to increase when you increase passing attempts by 1. This makes sense because the league average for yards per attempt is 7.0.

Using this equation, you can make estimates for passing yards based on a given amount of passing attempts. If you want to look for a one-point estimate of passing yards if someone has 25 attempts, you multiply 25 attempts by the slope (7.017 yards per attempt) and add the y-intercept. You find that if someone makes 25 passing attempts, the average result would be approximately 185 yards.

Now it is time to take variability into account. Although the correlation is strong and the sample size is large, the results would be different if you took a different sample. This is where margin of error and confidence intervals come in. The confidence interval gives you a range of values you might find for the slope and y-intercept if you were to take a new sample.

To take margin of error into account for the slope, you need to add and subtract a certain number of standard errors (based on sample size and desired confidence) from the slope of the regression line. Since my sample size is >30 and I am looking for 95% confidence, the t\* value is 1.96. As you can see if the linear regression coefficient chart, the standard error coefficient is 0.155. Because of these values, you can say that a 95 percent confidence interval for the slope of the best-fitting regression line is 7.017 + or – 1.96 \* 0.155, which goes from 6.71 to 7.32. I also found the 95% confidence interval for the y-intercept goes from -0.19 to 19.29.

Here is what the regression line looks like with the 95% confidence interval built in.

(Pic of scatterplot with confidence interval)

Along the same lines as the confidence interval, you can also find a predication interval that can be used to take a guess at an amount of passing yards someone might throw for based a future value of passing attempts. I found the following interval for predicting with 95% confidence the amount of passing yards someone will throw for if they have 35 passing attempts.

(Prediction chart)

The reason for such a large difference between a 95% confidence interval and a 95% prediction interval is that confidence intervals are a range for the average of a sample and prediction intervals are a range for an individual observation.

The last element of basic linear regression I want to explore in this project is the value of r^2 also known as the coefficient of determination. This value will be the percentage of the variability in passing yards that can be explained due to passing attempts. To find r^2, I took the correlation coefficient of 0.872 and multiplied it by itself, which is 0.76. In terms of r^2, this is a fairly high value. This means that 76% of the variability in passing yards is explained by passing attempts.

To dig in deeper and apply football concepts, I would guess that the rest of passing yards can be explained by variables such as quarterback skill, defensive efficiency, and drop rate, among others. A multiple variable regression would be able to expand on these variables, but that is above my paygrade.

I hope you enjoyed my first project. I hope there are many more that are much better in the future. Cheers!