BAS 320 - Assignment 4 - Simple Linear Regression Part 1

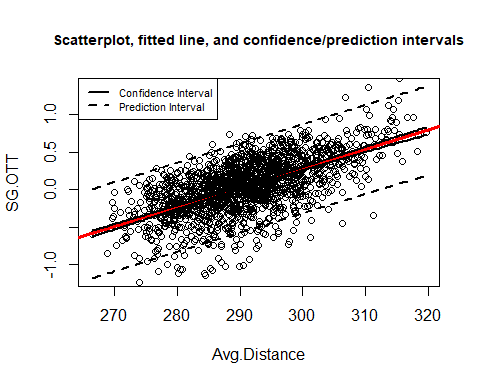
Jack Guptill

## Part 1: Analyzing Strokes Gained Off of the Tee and Average Distance

I’m using a simple linear regression to predict strokes gained off of the tee from the average distance a player will hit the ball off of a tee. Strokes gained is the amount of strokes a player will gain compared to what par is for a given course. Many tour pros have commented on how crucial distance now is to the overall game of golf. Before Tiger Woods and many long drive players, such as Bryson Dechambeau, driving the ball far wasn’t seen as too much of an advantage. With longer courses being developed and used on the tour, driving the ball far is now being seen as a competitive advantage. I am curious to see how many strokes you can potentially gain off of the tee by knowing what distance you can drive the ball. I am also curious to see how strong this linear relationship will be.

The data I am using for this analysis comes from Kaggle and it includes 18 columns with 2312 rows.

Link: <https://www.kaggle.com/datasets/jmpark746/pga-tour-data-2010-2018>



A simple linear regression model is appropriate for the relationship between strokes gained off the tee and average driving distance as the scatterplot shows roughly linear data that doesn’t display heteroscedasticity.

The regression equation is: y = -7.451 + .026x

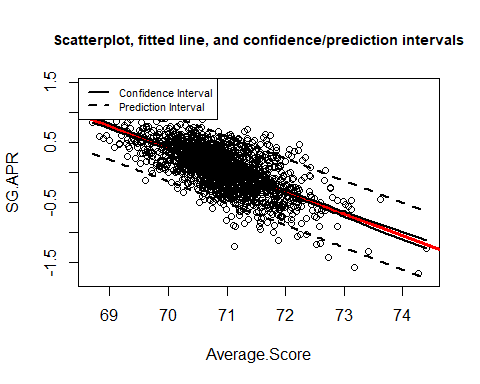
According to the model two players that differ by 1 yard are expected to differ by .026 strokes gained where the player with a higher yardage will have a higher strokes gained. The standard error, or the level of uncertainty with the model’s slope is .0008, meaning that the model might be off by .0008 strokes gained for every yard the ball is hit. We can be 95% confident that the true slope is contained in the interval .0241 to .0274. This is also meaningful as it tells us that there is some relationship between the variables as the slope is not zero.

The typical error when predicting strokes gained from off the tee with my model is .3027, compared to an error when predicting strokes gained off the tee from its overall average of .3817. To further prove usefulness of the model in predicting strokes gained off the tee, the r^2 value is .3654, or 36.54% of the variation in strokes gained off of the tee can be explained by differences in the average distance a player can hit the ball.

The p-value for the model is less than .05 indicating statistical significance. We can practically implicate that by knowing a players average distance, we could predict the players strokes gained off of the tee. For example, if the player has an average distance of 295 yards, then the predicted strokes gained is .219.

## Part 2: Analysis of Strokes Gained Approaching the Green and a Players Average Score

If you listen to any pro golfer, the most important shot range to master at a pro level is 100 yards and in. This is where most people mess up and start to shoot above par. We are going to analyze strokes gained from approach shots, which entails 100 yards and in, by a players average score. You might be thinking, well if a player shoots 72, or par for a course, then strokes gained from approach should zero by definition. This is not necessarily the case based on how the strokes gained by approach shots metric is calculated. It is also important to not that we are not dealing with a sample of “normal” golfers, in our case it is pro golfers in which it might take more (i.e. going well below par) to gain more strokes from approach shots.



A simple linear regression model is appropriate for the relationship between strokes gained approaching the green and a players average score in a given round as the scatterplot shows roughly linear data that doesn’t display heteroscedasticity. The distribution of the residuals is also fairly normal.

The regression equation is: y = 25.963 - .365x

According to the model two players that differ by 1 stroke in average score are expected to differ by .365 strokes gained in approaching greens where the player with a higher average score will have a lower strokes gained in approaching the greens. The standard error, or the level of uncertainty with the model’s slope is .0099, meaning that the model might be off by .0099 strokes gained/average score. We can be 95% confident that the true slope is contained in the interval -.385 to -.346. This is also meaningful as it tells us that there is some relationship between the variables as the slope is not zero.

The typical error when predicting strokes gained in approaching the green with my model is .2831, compared to an error when predicting strokes gained from its overall average of .3817. To further prove usefulness of the model in predicting strokes gained off the tee, the r^2 value is .4481, or 44.81 % of the variation in strokes gained in approach shots can be explained by differences in the average score by a player.

The model is statistical significant due to the p-value of the slope being less than .05. Knowing this, we can practically implicate that by knowing a players average score we could predict the strokes gained from approach shots to the greens. For example, a player with an average score of 72 (scratch golfer) will have an average strokes gained in approach shots of -.317. In other words, you need to be better than par on average to gain stokes when it comes to approach shots. With our data and sample, it implies that you roughly need to shoot at least below 71 to predict a strokes gained in approach shots of greater than 0.