BAS 320 - Assignment 5 - Simple Linear Regression Part 2

Jack Guptill

A map of a race track

Description automatically generated

## Analysis of Lap Times by Straight Line Speed

I am using a simple linear regression to predict the total time in seconds of a race lap by the straight line speed. This data comes from an API that tracked car telemetry from the Qatar Grand Prix last weekend. Overall, I am curious if and to what extent you could predict the lap time by the speed of the car. While multiple different speeds are recorded, straight line speed seemed to be the most important given the configuration of the circuit. See above for the circuit.

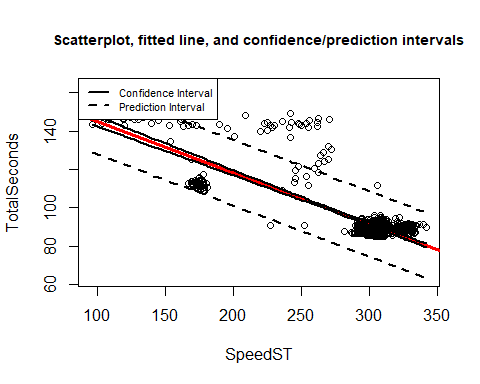
As mentioned previosuly, the data comes from the FastF1 API which is an open source API to gather f1 telemetry data.

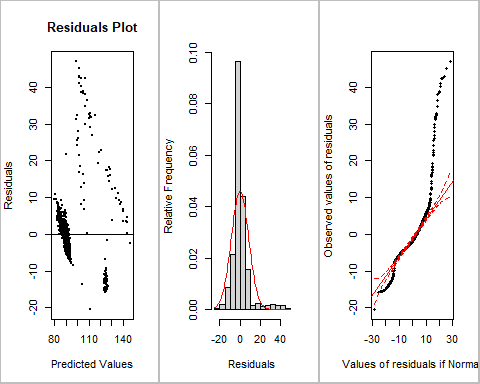
FastF1 Documentation: <https://docs.fastf1.dev/>

The overall data was gathered and loosely cleaned in python to create the CSV file.

For the CSV file and Python Notebook please go to: <https://github.com/jguptil1/FastF1-Work>

The dataset is 15 columns with 1006 observations.





TO.PREDICT <- data.frame(SpeedST = c(96,296,302,309,342))#Set up a dataframe so you can predict at the 25th, 50th, and 75th percentiles of x  
TO.PREDICT

## SpeedST  
## 1 96  
## 2 296  
## 3 302  
## 4 309  
## 5 342

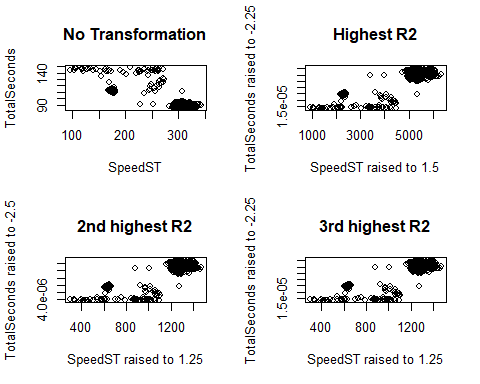
predict(m,newdata=TO.PREDICT,interval="prediction") #Get 95% prediction intervals

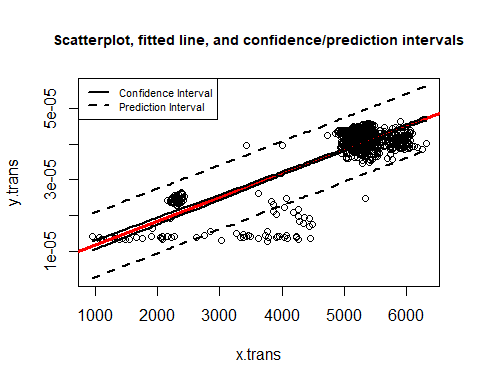
## fit lwr upr  
## 1 146.10885 128.82434 163.39336  
## 2 92.59296 75.51056 109.67537  
## 3 90.98749 73.90454 108.07043  
## 4 89.11443 72.03037 106.19849  
## 5 80.28431 63.18791 97.38072

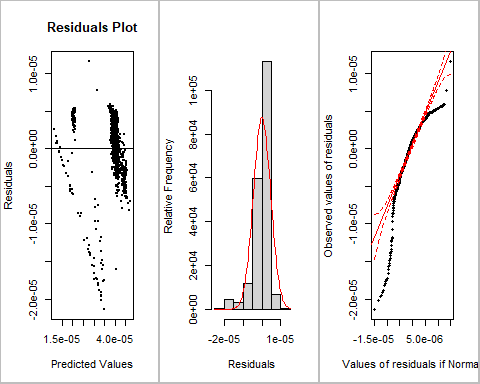
The regression equation is:

The median speed observed in Qatar was 302 kph or 187 mph. At this speed we can predict that the total time in seconds to complete a lap is between 73.905 and 108.070 seconds with 95% confidence. With a best guess and within this interval we can predict a lap time of 90.987 seconds.

Unfortunately, a simple linear regression does not appear to be the appropriate tool when trying to model the relationship between straight line speed and total lap time in seconds. We reject linearity, equal spread, and the normality assumption for this model as all values are zero. We can then look at the plots given that our sample size is larger than 25. We can see that in the QQ plot shows data that is extending past both the upper and lower confidence bands, telling us that there are significant outliers and skew to the overall distribution. We can see also in the predicted values chart that the data is not exactly split evenly down the middle with unequal spread. The residuals themselves have skew and do not seem to have a normal distribution.







TO.PREDICT <- data.frame(x.trans= c(96,296,302,309,342)^1.5)  
predict(m.trans,newdata=TO.PREDICT,interval="prediction")

## fit lwr upr  
## 1 1.147016e-05 2.422452e-06 2.051786e-05  
## 2 3.901492e-05 3.005406e-05 4.797578e-05  
## 3 4.004735e-05 3.108621e-05 4.900850e-05  
## 4 4.126488e-05 3.230307e-05 5.022668e-05  
## 5 4.718893e-05 3.821881e-05 5.615905e-05

#Sticking with the median  
c(4.004735\*10^-5,3.108621\*10^-5,4.900850\*10^-5)^(1/-2.25)

## [1] 90.03506 100.76349 82.30665

#fit = 90.03506 lwr = 100.76349 Upper = 82.30665

To hopefully find an even more reasonable reflection of reality, I’m going to explore the relationship between straight line speed raised to the 1.5 power and the lap time in seconds raised to the -2.25 power which has an R^2 of .6766. The prior model with untransformed data has an r^2 value of 0.638.

The updated equation becomes:

Using the same median value of 302 KpH we get a total lap time in seconds of 90.035.

For these models I would be comfortable using speeds within the range of 96 to 342 kph. Anything outside of this range would be an extrapolation as the data doesn’t support speeds outside of that range.

By the r^2 value alone, this model is more useful, however there are still gross violations of assumptions. While the p-value of the slope is statistically signifant, we fail the assumptions for equal spread, normality, and linearity as all values are zero. For the residuals plot we do not see equal spread, there seems to be skew and outliers within the QQ plot, and the line does not cut through the middle within the residuals plot. Although this regression was not fixed, the overall path of the equation seems to more closely ressemble the relationship

## Next Steps/Considerations

In order to build a better linear model for the future it would be ideal to refine the data. There seems to be tons of outliers within the data which could be due to pit stops or safety cars in which these would need to be removed from the data set. These outliers can really mess up the data and most likely are heavily affecting the linearity and skew of the data. I was able to find additional data externally to when there were safety cars and pit stops, however it would take more time than allotted to factor this into the dataset here. In making these changes, it would most likely take out all of the straggler data points. I also considered subsetting the data further to just one driver to cut away a lot of the bad data.