Final Written Report

Shresth Juyal

2023-06-18

Introduction

Research Question:

What is the impact of the number of laps and tyre degradation lap time in Formula 1, and how can linear regression be used to model and predict this relationship?

Introduction of the Study:

In the high-stakes world of Formula 1, tire degradation can be a key factor in determining whether you will finish the race first or last. During a 58-lap race, tires slowly start to wear out and lose grip, this leads to drivers not being able to drive through corners as fast as they could before and as a result, affects their lap time. Understanding the tires is a crucial area of development for all the teams. It is essential for teams to understand and effectively model tire degradation to predict lap times, optimize their strategy, and gain a competitive advantage.

We can use linear regression to predict and model tire degradation. By using linear regression in the context of Formula 1, we can analyze the relationship between lap number and lap time and model the tire degradation. The purpose of the research is to analyze and validate the effectiveness of applying linear regression to predict tire degradation.

Several studies have also tried to model tire degradation using several techniques. A study done by a student (Pontin) at the Luleå University of Technology models tire degradation using the Bayesian method, support vector machines, artificial NNs, logistic regression, decision trees, and a Fuzzy system. Each method mentioned has its advantages and disadvantages. A big issue with this study was the lack of specificity and the lack of data; the study does not seem to have a lot of data points, which makes it harder to create an accurate model.

Methods

Data Collection:

To conduct this research, the data must be accurate, reliable, and exhaustive. The “Formula 1 Race Data” on Kaggle contains contains several types of data such as tables describing constructors, race drivers, lap times, and pit stops. The dataset used in this study is “lapTimes.csv”. This spreadsheet contains 6 columns, which represent the raceID, driverID,

#lap, #position, time(in minutes), and time(in milliseconds). This study mainly use the lap time time and #lap to determine the rate of degradation of the tires.

Variable Selection:

The lap number will serve as the predictor variable since they give a good indication of how much the tires have aged and degraded. And lap time can be the response variable because it represents the performance of the tires; the lower the lap time, the better the performance. Analyzing the relationship between predictor and response variables will help us observe the change in lap time throughout a race to model tire degradation.

To model the relationship between the predictor and response variable, the ideal statistical tool to use would be linear regression. Linear regression can examine the linear relationship between the lap number (predictor variable) and lap time (response variable). A linear regression model to the data will allow us to evaluate and quantify the effect of the lap number on tire degradation. The model will be in the form of $y = \beta_0 + \beta_1 x + \epsilon$.

The coefficients of the linear regression model will give us an insight into the properties of the tire and help us interpret the relationship between lap number and tire degradation. A positive β_1 value would mean that as the number of laps increases, the lap time increases as well.

Model Validation:

I will use several techniques to validate my model to determine its accuracy. The first approach for model validation is using the R^2 and R^2_{adj} values to determine the goodness-of-fit of the model. This measures the proportion of the total sample variability in the Y axis is explained by the regression model. So, in the context of this study, a high R^2 would indicate that the model fits the data well and that the lap number is a significant predictor of the lap time. Another technique that will be used to validate the model is the splitting of the data. In this study, there will be a training set from which the linear model will be made and a testing set.

Some assumptions made while implementing linear regression are assuming linearity, independence, and common error variance (homoskedasticity). The independence assumption assumes that the performance of each lap is unaffected by the previous lap. And lastly, the common error variance assumption assumes that the variability of each lap timing is constant throughout the race.

Model Violation and Diagnostics:

To verify whether all our assumptions of linearity, constant variance, and uncorrelated errors are true, we can use three main types of residual scatter plots: Residuals versus predictor plots, Residuals versus fitted values plots, and Normal Quantile-Quantile (QQ) plots. If there are no discernible patterns seen in the residual plot and the residuals are scattered around zero, then that means assumptions hold.

Results

During a Formula 1 race weekend, Pirelli, the tire supplier of F1, allows teams to choose from 3 different compounds: soft, medium, and hard. The soft compound tires offer the most grip at the cost of longevity. The medium compound tires offer a more balanced approach with medium grip and medium longevity. And the hard compound tires on the other hand offer the least grip but are the most durable. Due to the unique properties of these tires, they must all have different tire models as well. We will model the tire degradation of each compound using linear regression.

Description of Data:

On 27th March 2011, the annual Australian Grand Prix was held at Albert Park Circuit. The thrilling race saw Red Bull's Sebastian Vettel finish 1st while McLaren's Lewis Hamilton finished a close 2nd. Let's look at the visualization of Lewis Hamilton's race.

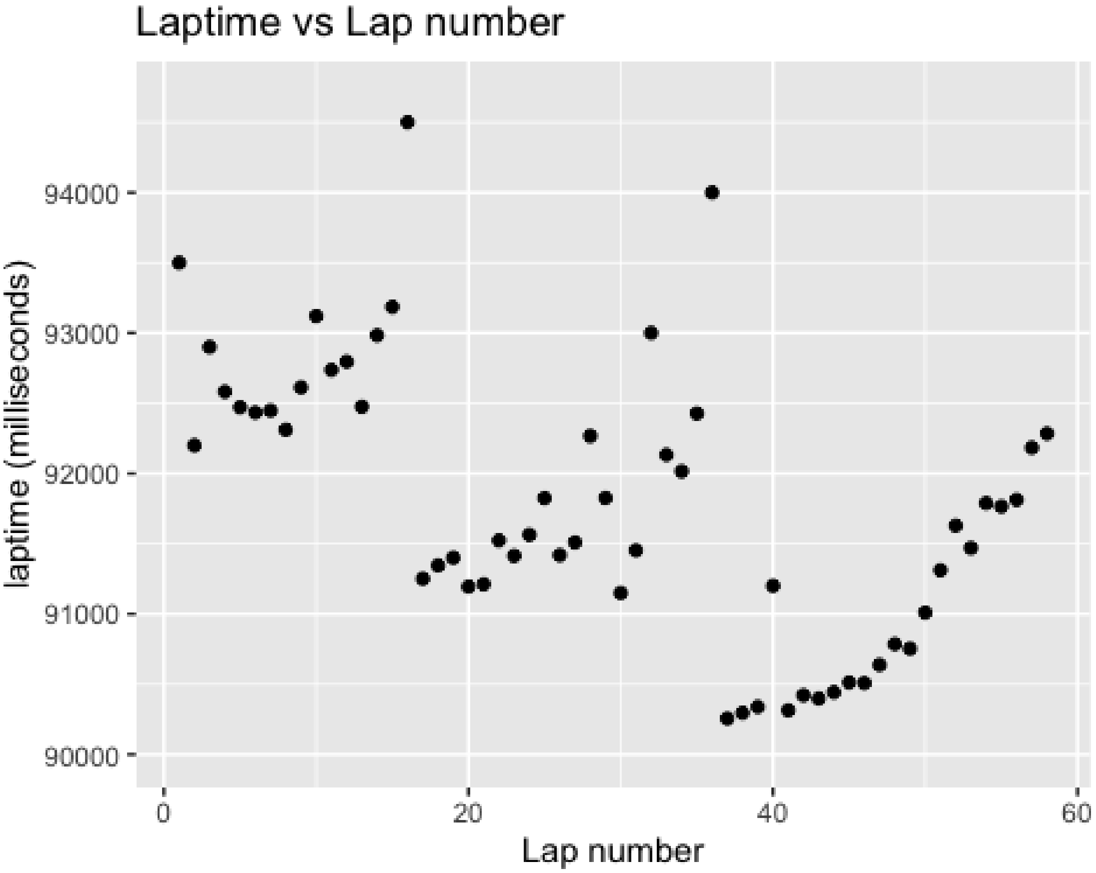

Figure 1: Scatter plot of Hamilton's race

This scatter plot visualizes the relation between the lap number and lap time. As mentioned before the predictor variable, which is the lap number shows the progression of the race. And the response variable, which is the lap time reflects the performance of the tires at any given lap. There are a total of 58 observations in our training set since that's how many laps the race lasted. This scatter plot helps us visualize and spot any patterns.

First, there are 3 distinct linear patterns in the scatterplot. This can be explained through pitstops. During the race, Hamilton made 2 pitstops. He started the race on hard compound tires and then pitted lap 16 for another set of hard tires and then later pitted on lap 36 for the soft compound tire. The 3 distinct linear patterns represent the 3 sets of tires Hamilton was on during the race. Modeling these 3 linear patterns will give us an insight into the rate of tire degradation throughout the race.

There are some outliers in this scatterplot. For instance, the very first outlier is on lap 1. This makes sense since lap 1 is the start of the race, therefore Hamilton started the lap from a standstill which slowed his lap time, and he was busy fighting with other drivers which further slowed him down. The next outlier is on lap 32, this can be explained through driver error. On lap 32, Hamilton went wide and traveled across the gravel which increased his lap time. And lastly, the final outlier in on the lap 40 when Hamilton was passing lapped cars; this also increased his lap time.

Presenting the Analysis Process and the Results

To model the 3 distinct stints Hamilton did on hard, hard and soft compounds respectively. We must look at these stints individually.

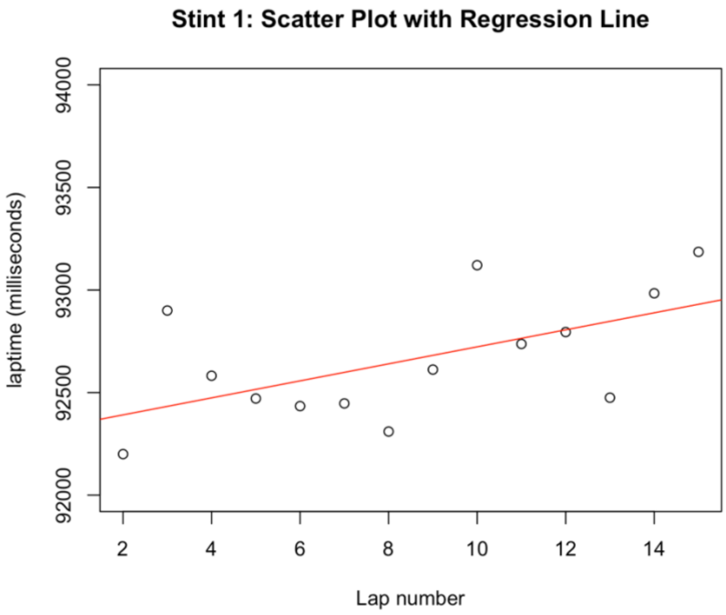

Figure 2: Scatterplot of Stint 1 (lap 2 to 15)

Similarly, this can be done for the other two stints as well.

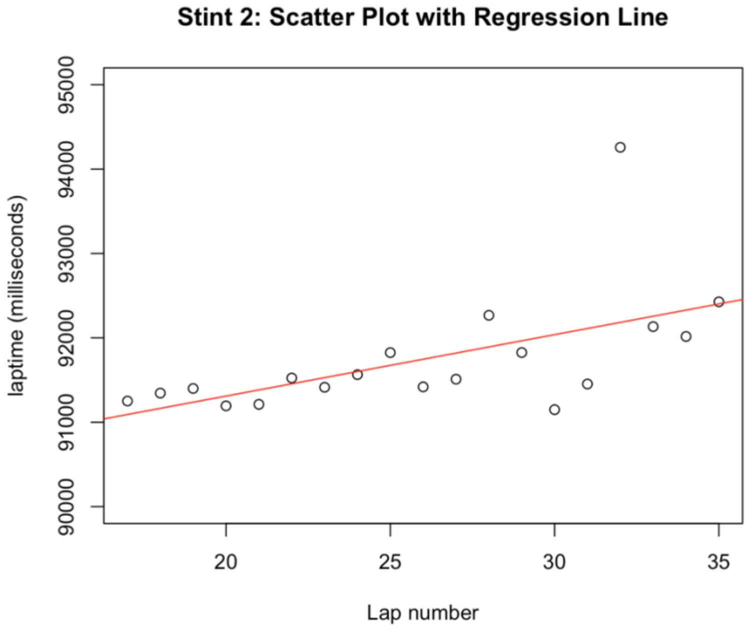

Figure 4: Scatterplot of Stint 2 (lap 17 to 35)

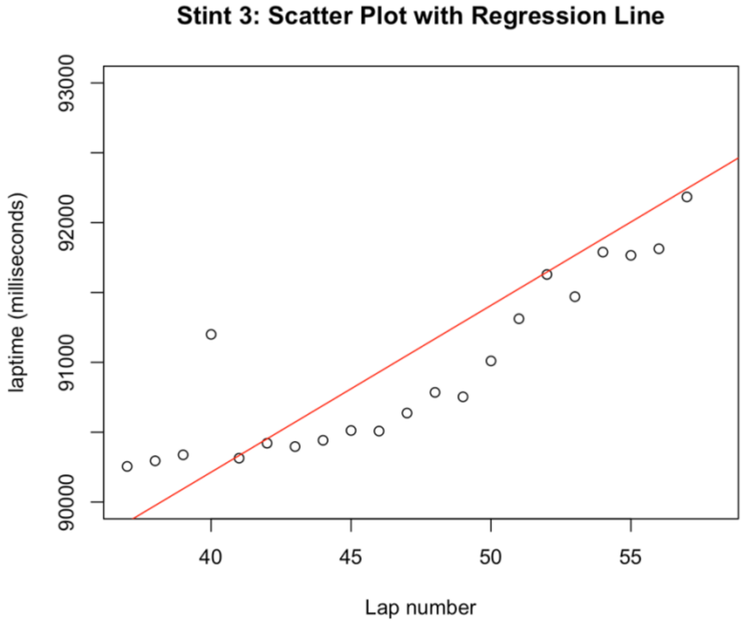

Figure 3: Scatterplot of Stint 3 (lap 37 to 58)

These 3 scatterplots just show each stint Hamilton did during the race. Stint 1 on softs was from lap 1 to 15, stint 2 also on softs was from lap 17 to 35, and stint 3 on the hards was from lap 37 to 58. The laps Hamilton pitted on have not been included in the scatter plot because they do not present an accurate lap time.

Stint 1 model: Using R studio, we can model the stint1 and add a regression line to it using. According to the summary provided by R), the β_1 coefficient value is 41.51 with the β_0 intercept value is 92308.18. Therefore, the regression line equation will be $y = 41.51x + 92308.18$. The value for β_1 coefficient shows us the rate at which the tires are degrading.

Stint 2 model: Using the same procedure as stint 1, we can also find the equation for the regression line for stint 2, which was $y = 72.81x + 89853.75$.

Stint3 model: Using the same procedure as before, the equation for the regression line for stint 3 is $y = 119.4x + 85437.74$

Goodness of the Final Model

The summaries of all the models also give us the $\boldsymbol{R^2}$ and $\boldsymbol{R^2_{adj}}$

	$\boldsymbol{R^2}$	$\boldsymbol{R^2_{adj}}$
Model 1	.3308	.2751
Model 2	.3286	.2891
Model 3	.6237	.6049

The R^2 and R^2_{adj} values for model 1 and model 2 suggest that there is some (although not strong) correlation and the variability in the response variable can be somewhat be explained by the predictor variable included in the model. However, model 3 has strong correlation and a larger portion of the variability can be explained by the predictor variable.

To verify all the assumptions, we made in the methods section, we need to create residuals versus fitted values plots, residuals versus predictor plots, and normal Quantile-Quantile (QQ) plots.

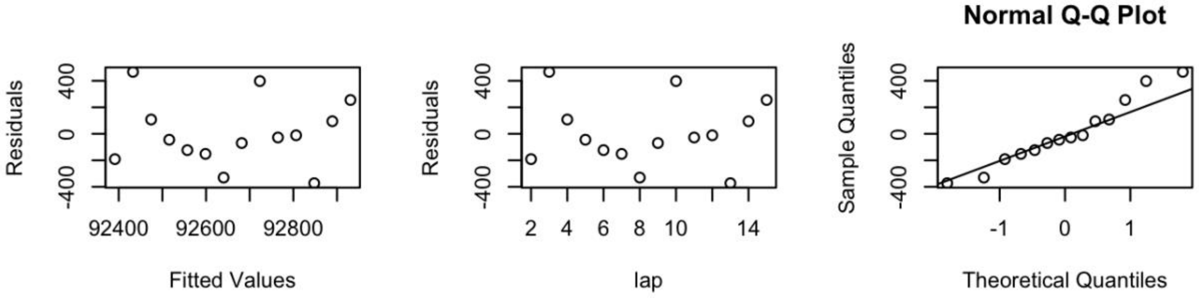

Figure 5: Residual vs Fitted values, Residuals vs Lap, and Normal Q-Q plot scatterplots.

These plots can help evaluate the assumptions of linearity, constant variance, and independence. Both the residual plots have points scattered around zero. However, there are some patterns that can be observed, which can be a sign that linearity is broken. However, we can verify the property of normality because the normal q-q plot follows a straight line with very little deviation. Since, we can only strongly verify one assumption, we cannot successfully validate this model.

Lastly, the appendix also includes a comparison against a test set of another driver during the same race. The model predicts the values of the test set fairly well, but there are some inconsistencies.

Discussion

Final Model Interpretation and Importance:

Therefore, the final model for stint 1, stint 2, and stint 3 were:

Stint 1 (hard compound): $y = 41.51x + 92308.18$

Stint 2 (hard compound): $y = 72.81x + 89853.75$

Stint 3 (soft compound): $y = 119.4x + 85437.74$

The first two models give us the regression line equation for the hard compound tire and the third equation is for the soft compound tire. In the context of the race, the coefficient of x represents the estimated increase in lap time in milliseconds per lap. So, it can be observed that the soft compound tire has the highest coefficient for x, which is 119.4. This confirms the fact that although the tires are fast, they have high degradation and are slower by approximately 119.4 milliseconds every lap. Both the hard compound tires on the other hand had coefficients of 41.51 and 72.81. This isn't as high and shows a lower tire degradation. For every lap the average increase in lap time was by 41.51 and 72.81 milliseconds respectively. This also verifies the fact that hard compound tires have lower tire degradation but are slower; whereas the soft compound tires have higher tire degradation but are faster. Regardless of what rate both the compound wear at, it is evident that there is a positive linear relationship between lap number and lap timing and that as the number of laps increases, the lap time increases as well.

Limitation of the Analysis:

Several other factors also affect the lap time, these variables were omitted in this study. For instance, track conditions, weather, driver performance, car setup, tire temperatures, and variations between the cars from team to team. All these factors can also be extremely influential in determining the lap time not just tire degradation. This affects our model's accuracy; these variable's not being factored in affects the model's ability to provide an extensive understanding of the relationship between the predictor and the response variable. However, taking into these variables in the model be extremely difficult as they can vary a lot, and having access to all these variables can be extremely difficult and impractical to analyze.

Appendix

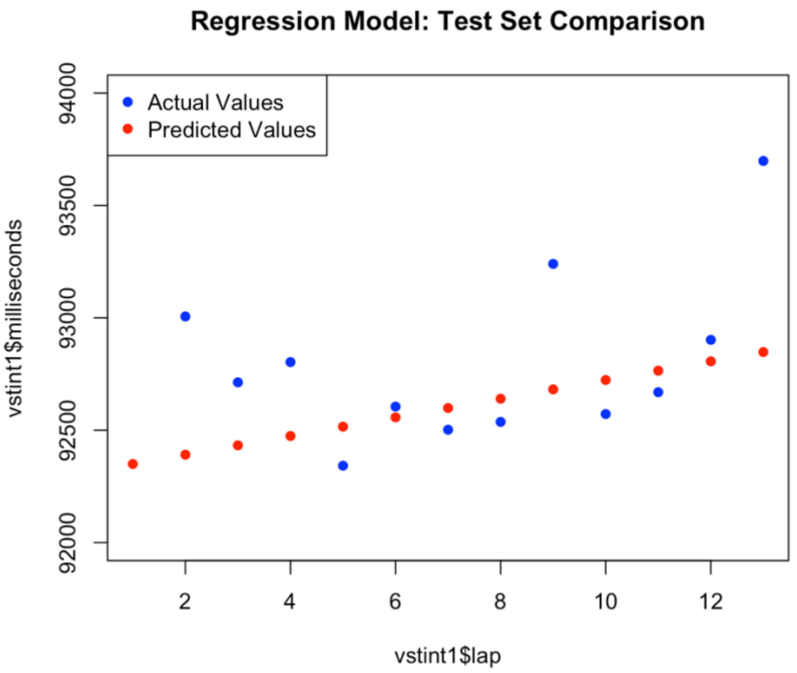

Figure 6: Test Set comparison

References

G, Chris. “Formula 1 Race Data.” *Kaggle,* 28 Nov. 2017, www.kaggle.com/datasets/cjgdev/formula-1-race-data-19502017.

Pirelli. “F1® Tires.” *Pirelli*, 1 Jan. 2023, www.pirelli.com/tires/en-us/motorsport/f1/tires.

Martin , Antonio Herrera. “Handout Lecture 6.” *Quercus*, 5 June 2023, q.utoronto.ca/courses/305501/files/26521235?module_item_id=4721460.

Pontin, Simon. “AIBased Race Strategy Assistant and Car Data Monitor.” *Diva Portal*, 1 Jan. 2023, https://ltu.diva-portal.org/smash/get/diva2:1756880/FULLTEXT01.pdf. Accessed 19 June 2023.