Formula for the Podium: A Linear Model Approach to Predicting Formula 1 Championship Outcomes

GPH-GU 2353: Linear Regression and Modeling

Sanjana Battula

Farah Beche

Kajal Gupta

Shreya Ramella

Introduction

Formula One stands as the pinnacle of modern-day motorsport, the herald of pushing limits of speed, precision and technological innovation. Since its inception, Formula One has garnered an international audience with thrilling races, epochal venues, and acclaimed drivers. Beginning around early March, Formula One launches its eight-month season. Packed with Grand Prixs on weekends, practice sessions during the week, and strategies being implemented both on and off-track, the season is fast paced, intricate, glamorous and ruthless all at once.

Several factors are active when deciding on a world champion in Formula One. Every decision is crucial. With every engineering feat, strategy discussed, upgrade done to the car, the sport is famed for its delicacy and complexity. This investigation compares these factors to determine which collection of predictors has the greatest influence on winning a Formula One World Championship. Looking at the data from the 2023 season, the goal is to determine the statistical influence and significance of various predictors using regression analysis and to ultimately create a model to predict race outcomes for future seasons.

Every position won from tenth position (referred to as P10 in the sport) earns the driver (and the team) an amount of points. Winning P1 earns the driver 25 points. P2 amounts to 18 points, while 15 points go to P3. From there, 12, 10, 8, 6, 4, 2, and 1 point(s) are awarded to the remaining finishers of the top ten. Taking in account other nuances such as one extra point added to the driver with the fastest lap of the race and the Sprint race points, the world championship is decided by the number of points a driver scores throughout the season.

Methods

With the point system established, the total number of points scored over the season acts as the outcome variable. The predictors, contingent on the datasets taken from GitHub, includes 2023 data detailing the year, driver names, race circuit, qualifying session results and timings, driver starting and finishing positions, number of laps at each circuit, fastest lap times for each race, pit stop times, status (finished, disqualified, collision etc.) of the race, and overall points scored for each driver at each race. In addition to these variables, we included the conversion of fastest lap and qualifying times to milliseconds. Merged, the final dataset contains 389 unique entries to conduct statistical analysis to predict a world championship.

A univariable analysis was conducted to examine the distribution and summary statistics of each predictor variable. An initial simple linear regression model was fitted to explore the relationship between our outcome variable (points) and all possible predictors. In order to plot a correlation matrix between our predictor and outcome variables, the models must be of equal size. To accomplish this, a sub data frame was created removing all NA values, and a correlation matrix was plotted. A linear regression model was plotted comparing standardized residuals with the fitted values to determine nonlinearity, unequal error variance and outliers. A QQ plot and a histogram were also plotted to establish if the data is normally distributed or skewed in a direction. Outliers and leverage points within the dataset were identified, and a Bonferonni test was additionally conducted to reflect the presence of outliers. A Cook's distance was calculated to identify influential points, which were then plotted against half-normal quantiles. Variance inflation factors (VIF) were calculated for each variable to test for severity of collinearity. Highly correlated predictors - which had a VIF value greater than 10 - were each removed from the model.

Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R-squared methods were implemented to evaluate how well the model fits the data by concluding which predictors are significant. AIC and BIC values are indicative of the best model fit and parsimony, while the adjusted R-squared value establishes the goodness of fit. The step function is useful in deciding which combination of predictors sets the best fit model. To handle the non-linearity in the model and to improve the non-constance, a Box-Cox was used to find the best transformation of the response variable. The variables that were significant in the AIC, BIC, and the adjusted R-squared models were squared in new linear models and each lambda was calculated. The lambda's were used in new transformation models. However, as the transformation models did not result in better fits, the model remained as it was for predictions.

At the end of the regression, confidence intervals and prediction intervals were calculated based on the model selected. With this, we can be statistically confident that our model has the best fit with the most significant combination of predictors to accurately predict how many points a driver may earn in a Formula One race.

Results

Running a linear regression with points as our outcome variable and all other race factors as the predictors, showed the statusId, rank and position variables to be statistically significant at the 0.001 level. Plotting the linear regression model with these predictors as a scatter plot shows mild heteroskedasticity, where the residuals are scattered unequally across the fitted values. Increasing variance among the residuals is a condition of heteroskedasticity. As each race results in different statuses for each driver, a different amount of points, and different starting positions and ending positions for each driver, this variance supports our dataset. Similarly, the Q-Q plot of the standardized residuals against the theoretical quantiles of a normal distribution shows a trend of the data that violates normality. The histogram, plotted to show the shape of the data, is skewed slightly to the right, with a peak around 0 with a longer tail on the positive side. Both plots (Figure 1, Figure 2) suggest and support the conclusions of the regression scatter plot, in that the data is not normally distributed.

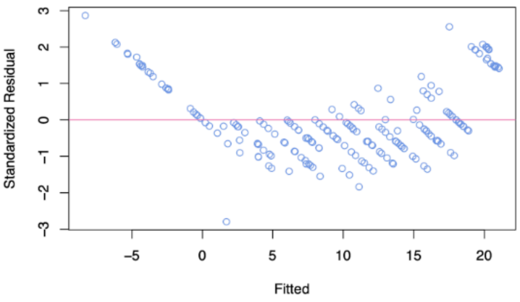

Figure 1: Fitted vs. Residual Standardized Plot

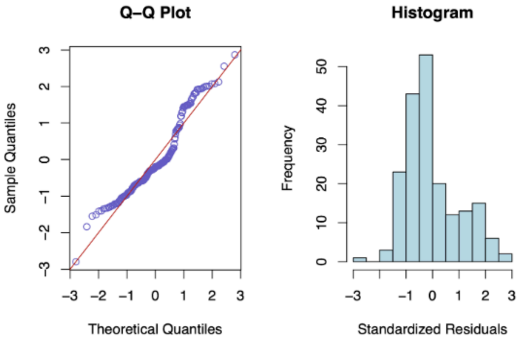

Figure 2: Q-Q Plot; Histogram of Standardized Residuals

Correlation between each predictor variable was conducted to determine the strength of each relationship. Figure 3 shows that each round of qualifying times is positively correlated with the following round. Qualifying in Formula One is a tiered competition. Drivers must be fast enough to qualify into the following round, and the final round determines where the driver will begin the race. Therefore, the strong correlation is verifiable. There is a strong relationship between qualifying times and the number of laps in each race. While the correlation dictates a negative relationship, it is important to note that the number of laps in each Grand Prix is a randomized value and qualifying results and timings do not impact how many laps a Grand Prix has. Other relationships to note are between position and rank, which are strongly and positively related. Where a driver begins a race does impact how they will finish a race.

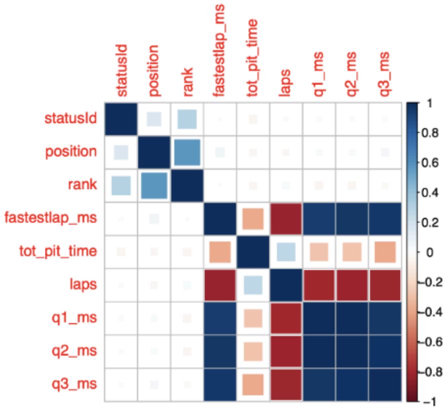

Figure 3: Correlation Plot

After determining correlation, outliers and leverage points were tested for. Both Bonferroni and a R studentized test confirm that the dataset has no outliers, but the data does include several, high, leverage points according to the hat values. Our data contains 27 high leverage points, which are all at least two times the average leverage. The half-norm plot supports this with the slight incline of the leverages across the half-normal quantiles. Most other observations have low to moderate leverage values given their straight line shape from the origin. The observations of 104 and 82 are noted to be the highest points in the plot (Figure 4), characterized by the significant deviation from the straight line pattern.

To evaluate the impact these high leverage values have on the model, a Cook's Distance test is used to measure its influence. With a threshold value of 0.5, the Cook's Distance for the model resulted in one influential point; point 174 measured an influential value of 17.86349. The plot is representative of that, with a single observation at 82 which has a Cook's Distance value of 174 shown in Figure 5. When a model without the influential point is run, there is a slight difference in the resulting p-values, which is indicative of how sensitive the estimates are in the presence of just one observation.

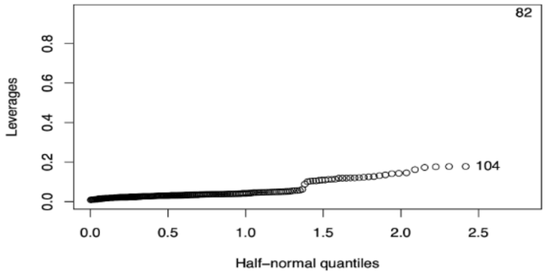

Figure 4: Leverage Points Plot

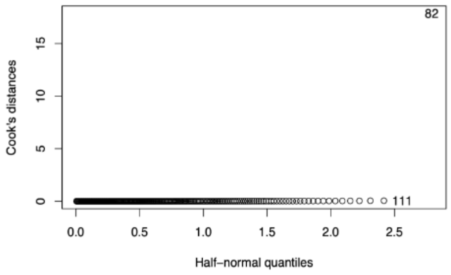

Figure 5: Plot of Cook's Distance

VIF test for the severity in any multicollinearity a model may have. A value greater than 10 signifies that the predictor(s) has multicollinearity, and that the variables are highly correlated with each other. Testing for multicollinearity in the model produced fastestlap_ms, q1_ms, q2_ms, and q3_ms as the predictors with high dependency on other variables. To account for high multicollinearity, several models were created, each with one of the listed predictors missing to determine any VIF value changes. Ultimately, the model that removed the q3_ms predictor resulted in a significantly lower VIF value for q2_ms at 8.993870. As such, this model with the q3_ms predictor removed is used for further analysis.

The model, now without outliers, influential points, or leverage points, can be tested for the best combination of predictors. AIC was calculated on different model sizes. Our AIC calculation shows that the third model (statusId, position and rank) is the lowest value (807.0182), and therefore, the best model to use. Similarly, the smallest BIC value (822.8725) favored the third model, with the same predictors as AIC. Finally, an adjusted R^2 value was tested. Our R^2 calculations (highest value being 0.8630793) resulted in the fifth model, with the same predictors as the AIC and BIC models but with the inclusion of fastestlap_ms and q2_ms.

In addition, we applied the step function to select the best-fitting model from the full model. The output indicates that the seventh model, with an AIC value of 417.1, is the most optimal model. The seventh model includes the statusId, position, and rank predictors, supporting the AIC and BIC calculations. These predictors were consistently selected across multiple model selection criteria.

To investigate if the model can be further optimized, a Box-Cox transformation was attempted. With the data pre-processed to handle any zeroes, three models were created, each one containing one squared predictor in addition to the entire set of predictors. The lambda value was calculated (0.989899) and a new, transformed, model was fitted. The lambda graph shows that the value is very close to 1, signifying very little change to the original model. However, each transformed model yielded the same lambda value and therefore, the model was not a better fit. The model with the statusId predictor, complete with its calculated lambda and final transformation, was kept for reference, and all other predictor models were removed.

Once we established the best model to use, a prediction and confidence interval were calculated to consider the uncertainty associated with the model's prediction. Our testing values came from the last line of the dataset, and were used in the prediction interval. The prediction interval [3.80, 17.96], signifies a 95% probability that a future observation of points will be contained in the interval, given the values of position = 3, rank = 6, and statusId = 1. At the same time, there is a 5% probability that the next observation of points will not land between the intervals. A confidence interval measures how confident we can be in the model that the true average points of the outcome will fall within the confidence interval range. With a 95% confidence interval of [10.30, 11.49], we can be 95% confident that the true average points earned by a driver with the given predictor values will fall within this range.

Summary

Winning a Formula One World Championship is the result of skill, luck, and making decisions that are statistically sound. This analysis measures and tests the contributing factors to the number of points earned by drivers in Formula One. The selected linear regression model, including the status of race completion, a driver's starting grid position, and their final rank predictors explain that these variables have the greatest impact on the outcome of a race, which is measured in points. Stripping the data of its outliers, the leverage points, and influential points sets up the model for testing the goodness of fit. AIC, BIC, adjusted R^2, and the step function all determine that the best model to use includes a combination of status, position and rank as the predictors.

While we are confident that the prediction and confidence intervals will accurately measure the model's predictive abilities, we must be wary of the model's limitations. The presence of multicollinearity among certain predictor variables related to lap times and qualifying times raises concerns about unstable coefficient estimates. Additionally, the assumption of linearity may be overly simplistic, as exploring potential non-linear relationships between predictors and the outcome variable could reveal more nuanced patterns and improve predictive accuracy.

The generalization of our model and analysis is a limitation itself, as we could not include every factor that may impact the outcome of the race. Other factors that are not as easily measurable, such as driver skill, tire strategies, weather conditions and the temperature of each track were also excluded. For simplicity, our data also left out sprint data, which is an additional race of fewer laps during the Grand Prix weekend, and contains its own set of points, all of which would have impacted the outcome of this analysis. Moving forward, future research could focus on addressing the identified limitations. Exploring techniques to mitigate multicollinearity, such as ridge regression or principal component analysis, could enhance the understanding and stability of the model's predictors. Additionally, incorporating additional non-linear modeling approaches could capture potential non-linear relationships between predictors and the outcome variable.