Optimizing Formula 1 Strategies: A Statistical Modelling Approach

DATA 410: Regression and Generalized Linear Models Final Project

Alejandro Builes (48801039)
Sharan Srinivasan (84003490)
Nozomu Hirama (88210729)

The University of British Columbia

May 11, 2024

1 Introduction and Study Aim

This project outlines a statistical approach to analyzing Formula 1 racing data to predict race strategies and outcomes. Formula 1 (F1) racing began in 1950 and is recognized as the world's most prestigious motor racing competition, as well as the world's most popular annual sporting series. The championship features a series of races, known as Grands Prix, held in various countries across four continents.

The motivation behind this project is to learn about the strategy of pit stops in F1. Pit stops are a critical and strategic component of F1 racing, offering teams the opportunity to change tires, refuel (when regulations permit), and make necessary adjustments to the car to adapt to the race conditions, track, and competition. It is not just about the speed of the car or the skill of the driver; it also involves making smart decisions during the race. Therefore, we are interested in understanding how these decisions are made and how they can influence the outcome of a race.

The study aims to identify whether there is some relationship between the position of the driver, i.e. their standing, and their pit stop strategy for the race. Pit stop strategy can be described through a number of variables. Whether it be the number of pit stops a driver takes during a race, the lap number at which a driver decides to take a pit stop, or how close a driver takes consecutive pit stops to each other, the pit stops offer various advantages that must be incorporate with great care. Thus, our over-arching goal is to explore various aspects of these pit stops to come up with a statistical model that best predicts the drivers' positions. At the same time, numerous other variables such as the driver status, circuit conditions, and driver performance prior to a race could heavily impact the outcome of a race. Therefore, we ultimately aim to construct a multivariate model that can account for some of the other potential relationships that can be explained by existing data. One highly plausible model is a generalized linear mixed effects model with an added random effect of the race circuit, as we hypothesize pit stop strategies to differ depending on the circuit a race takes place on.

2 Scientific Question

How can we predict Formula 1 race outcomes using pit stop strategies that can be modelled from collected data, while accounting for a range of external predictors that may have a simultaneous effects on the final standings of drivers?

3 Data Description

The dataset we are studying for this project is maintained by ergast.com (a popular web API that provides cleaned motor racing data), and it has been continually added to since 2009. Later, a retrospective study was done to compile older Formula 1 data dating back to 1950, but many parts of this data are incomplete (mostly particular metrics in races from 1950-1989) due to the nature of the study and the drastic changes made to Formula 1 rules over the years that made past data (before 1989) largely irrelevant for predicting future outcomes. As such, we subset the data to entries only after 2011 as pit stop recordings initiated in 2011.

This dataset contains information on 280 races from 2011. There are 59 unique variables recorded, where multiple measurements are possible for some variables for each of the approximately 20 drivers per race. As it is not plausible to incorporate all features of the dataset, we select only the variables that align with our interests. With this we end up utilizing the following 6 out of 14 total CSV files: `circuits.csv`, `pit_stops.csv`, `races.csv`, `results.csv`, `status.csv`, and `lap_times.csv`. After data cleaning and processing, we also created some new variables to complete the data we are going to use in this project. These are explained in more detail in Table 1.

Table 1: Description of variables for F1 racing pit stop strategy analysis.

Variable Name	**Data Type**	**Units**	**Description**
raceId	Categorical	Unique ID	Unique identifier for each race
driverId	Categorical	Unique ID	Unique identifier for each driver
year	Continuous	Year	Year when the race took place
stops_total	Continuous	Count	Total number of pit stops made by a driver in a race
sd_from_mean	Continuous	Standard units	Number of standard deviations a driver's pit stop is from the mean
var_standardized	Continuous	Variance units	Variance of standardized lap numbers for pit stops
var_raw	Continuous	Variance units	Variance of raw pit stop positions
positionOrder	Ordinal	Position	The final race position of a driver
grid	Ordinal	Position	The starting position of the driver on the grid
improvement	Continuous	Position change	Change in position from the start to the end of the race
circuit_name	Categorical	Text	Name of the circuit where the race took place
statusId	Categorical	Unique ID	Code representing the driver's status at the end of the race

4 Regression Analysis

4.1 Data Visualization: Dealing with our Response

Plotting the densities of driver placements from all the races as a histogram, it can be seen that our response of driver placements is not normally distributed (see Fig 1.). Due to this, we can expect to compute unreliable estimates if we proceed with a Gaussian GLM. Thus, we decide to fit use a family of Poisson distributions. The other aspect of our response is that it is a count variable, in the sense that we are counting the number of drivers that are ahead of a driver (self-inclusive) at the end of a race. This makes both Poisson and Negative-binomial distributions suitable for our analysis. In the subsequent analyses, we proceed to fit regression models with these two families of distributions, while comparing their performances.

The observed distribution of driver standings can also be conceptually understood, since under the assumption that every driver ends up in some placement at the end of the race, there will be one driver standing in each position. If the number of drivers per race is fixed, each race will contribute a single data point for every placement ranging from 1 to the maximum number of drivers. In the case of F1 the maximum number of drivers sits at 20 for most races, with some exceptions. Considering this, even assuming a Poisson distributed response would be slightly difficult. In theory, we would be seeing a uniform distribution instead. However, we will simply acknowledge this as a limitation for the time

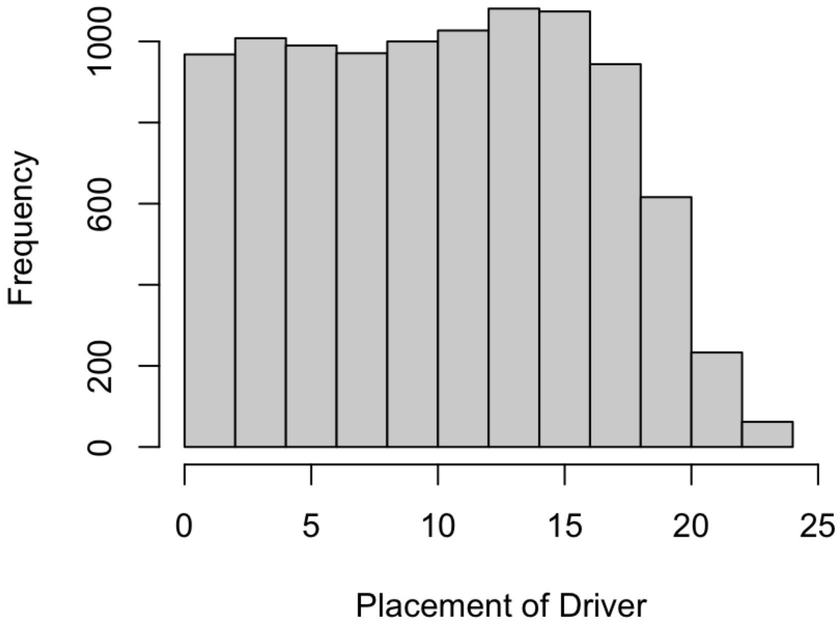

Figure 1: Histogram of driver placements shows Uniform-like distribution.

being, in order to assess our model adequacy in latter sections.

Another potential approach that can be taken in terms of dealing with the response variable, is to take a different measure of ”race performance”. Instead of predicting the raw placement measures, we could also quantify the improvement score of a driver in any given race. This improvement score was calculated as the starting position - the final position, where a positive value indicates an improvement in the placement during the race and a negative score would be a drop in placement. This new measure takes into account the advantage a driver may get from starting from a higher position. Unlike the raw placement values, this variable resembles a Gaussian distribution (see Fig 2.). However, looking at the quantile-quantile plot, it can be seen that both left and right tails are much lighter, with the majority of the points located around the mean. Nonetheless, this does allow for a possibility for a Gaussian GLM, with a continuous response. The performance using the three families of distributions (Uniform, Negative-binomial, and Gaussian) will be compared after the base model is constructed.

4.2 Variable Selection

In order to predict the driver placements from some predictor variable(s), we first need a suitable representation of ”pit stop strategy”. Since the lap number of each pit stop is difficult to compare across races and drivers, it may not function as a useful predictor. Not only does the number of pit stops vary from driver to driver, but the number of total laps also differ depending on the race/circuit. This causes the raw lap number of pit stops to be dependent on such external variables, making comparisons unreliable. To overcome this challenge, we decided to investigate how a driver’s tendency to stay with or diverge from the norm would impact their final standing. During a race in F1, one of the key events is surpassing another driver. With high speed cars competing for the slightest of differences in lap times, it is crucial that a driver attempts a take-over at an optimal timing. Thus, the most ideal scenario for a driver would be to have already taken a pit stop, while the other drivers have not had the chance to do so. Therefore, we hypothesize that being slightly desynchronized from the majority of drivers in pit stop timing, could

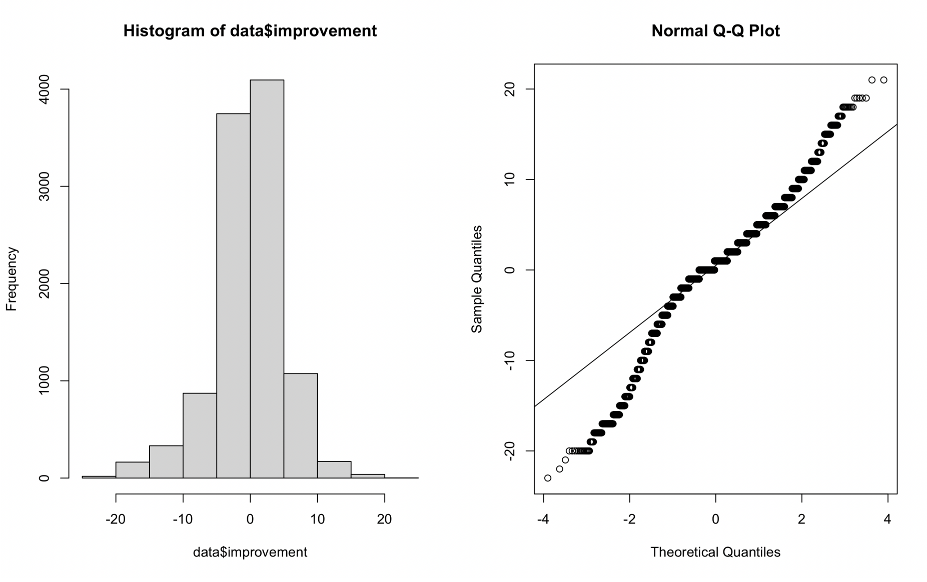

Figure 2: Improvement score is distributed as a bell-shape with light tails on both ends as indicated by the Q-Q plot.

give one a better chance to place higher in the end.

To quantify the difference from the norm, we took the mean lap number of all the drivers' for a given pit stop in a given race (e.g., the mean lap number of all drivers' second pit stop during race 841). We then used this, along with the standard deviation, to compute how many standard deviations a given driver's pit stop location was away from the mean. The standard deviation values obtained are A negative value indicates that the driver took the pit stop earlier than the others, and similarly for positive values, the drivers were late. Using this measure, we fit both Poisson and Negative-binomial regression models to predict driver placement.

An initial fit of a Poisson regression model on the number of standard deviations away from the mean returned significant associations between the two variables. A over dispersion test was conducted, which estimated a dispersion parameter of 3.14, indicating slight over dispersion picked up by our model, although not significantly high. The model was then refitted, this time however, with the adjustment for over dispersion using the estimated parameter. Adjusting the model by the dispersion factor of 3.14, however, did not affect the model performance by any significant amount, providing the summary below. The residual deviance was the same at 33962 on 9696 degrees of freedom for both the adjusted and non-adjusted model, with the same p-value for both the intercept and the slope in both cases. The Poisson model indicates that there is a slight decrease in placement (indicating an improvement in performance) as a driver diverges away from the mean pit stop location.

```
Call:
glm(formula = positionOrder ~ sd_from_mean, family = poisson,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4028  -1.6617  -0.0018   1.3216   4.0607

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.350351   0.005570  421.97   <2e-16 ***
sd_from_mean -0.090086   0.005689  -15.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for poisson family taken to be 3.144219)

    Null deviance: 34744  on 9697  degrees of freedom
Residual deviance: 33962  on 9696  degrees of freedom
  (276 observations deleted due to missingness)
AIC: 72650

Number of Fisher Scoring iterations: 5
```

The Negative-binomial model also yielded similar results, although it seems to slightly out-perform the Poisson model. The estimated slope was -0.08 positions as opposed to the -0.09 positions in the Poisson model, which can be deemed as very similar. However, the Negative-binomial model had lower residual deviance (10493) than the Poisson model (33962). This indicates some outperformance by the Negative-binomial model. Comparing the AIC values also indicates that the Negative-Binomial model may be performing slightly better than the Poisson model (Poisson: 72650, Negative-binomial: 61548). However, there is no significant difference in the adequacy of the two models.

```
Call:
glm.nb(formula = positionOrder ~ sd_from_mean, data = data, init.theta = 3.721204648,
    link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.61421  -0.91574  -0.00134   0.64305   1.93999

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.350624   0.006129  383.50   <2e-16 ***
sd_from_mean -0.083530   0.006345  -13.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.7212) family taken to be 1)

    Null deviance: 10681  on 9697  degrees of freedom
Residual deviance: 10493  on 9696  degrees of freedom
  (276 observations deleted due to missingness)
AIC: 61548

Number of Fisher Scoring iterations: 1


              Theta:  3.7212
          Std. Err.:  0.0763

 2 x log-likelihood:  -61541.5530
```

A possible explanation for the poor performance in both the models could be that the number of standard deviations away from the mean may be quadratically related to the driver placement. Since there could be an advantage for being both early or late compared to other drivers' pit stops, perhaps there is some divergent effect where the driver's performance increases as you move away from the mean, whereas there is a minimum peak in performance around the mean. In order to test this, we transformed our predictor to take the absolute value of the number of standard deviations. With this, we can purely quantify how far away a driver is from the mean, regardless of whether they are early or late.

Results of the adjusted fits indicated that although the model performance did not seem to change significantly, which was confirmed by comparing the residual deviance, AIC, and p-values, the estimated slope was now positive, being approximately 0.13 for both distributions. This reveals that as a driver moves away from the mean pit stop location, their placement actually falls down. Thus, combining our conclusions from both measures, it may be that it is better to stay within ranges of the average timing of other drivers' pit stops, but perhaps there is some benefit in being slightly late. This would be beneficial information for a driver, as they could use other drivers as an indicator for when to start planning on entering for a pit stop.

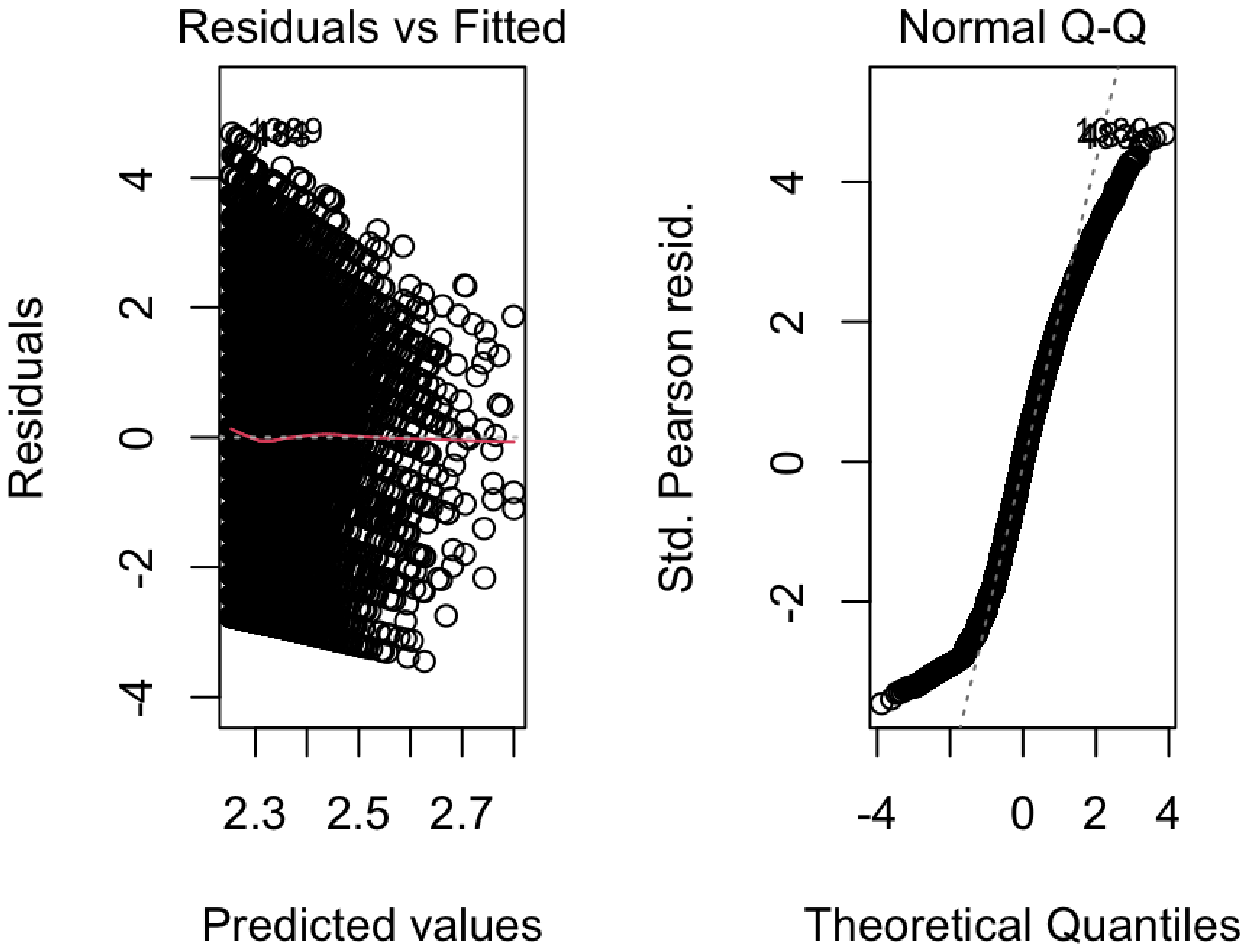

Figure 3: Diagnostic Plots for the Poisson model on standard deviations away from the mean pit stop location and driver placement. Residuals seem to be relatively well-behaved, although the Q-Q plot indicates non-normal residuals with deviance from normality at the tails.

The diagnostic plots reveal a somewhat adequate fit with the Poisson model (similar performance in the Negative-binomial model, although we are still violating the normal assumption of errors, which may simply be a result of our data and its structure. Since one of our major assumptions in GLMs is that errors are i.i.d, our models fail to perform to the optimal level as soon as our errors are inherently not i.i.d. As all of our response data are slightly co-dependent, we cannot fully assume that there is no association between the different placements. If one driver places in 1^{st} place, no other driver will be able to place in 1^{st} place after that. Similarly all the drivers are constantly influencing each others' performance. However, we are unable to alter the system of study and therefore will proceed with caution.

4.3 Model Extensions

In order to better predict the driver placements, we can attempt to fit a multivariate GLM with combinations of potentially significant predictors. As the pit stop strategy may differ based on the circuit and type of track the drivers are on, here we add a random effect of circuits. Different lengths of circuits, typical weather conditions, number of turns, are all aspects of F1 racing that can be extremely important and could greatly contribute to when an ideal pit stop becomes.

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: Negative Binomial(3.7256)  ( log )
Formula: positionOrder ~ abs(sd_from_mean) + (1 | circuit_name)
   Data: data

     AIC      BIC   logLik deviance df.resid
 61558.2  61586.9 -30775.1  61550.2     9694

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5879 -0.7874  0.0126  0.7161  2.4562

Random effects:
 Groups       Name        Variance Std.Dev.
 circuit_name (Intercept) 0.001088 0.03299
Number of obs: 9698, groups:  circuit_name, 34

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.251150   0.011702  192.37   <2e-16 ***
abs(sd_from_mean) 0.130564   0.009927   13.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr)
abs(sd_fr_) -0.650
```

From the summary output, we can see that the standard deviation from the mean pit stop time (sd_from_mean) carries a significant negative coefficient (0.130564), indicating that increased variability in pit stops is associated with lower placement in races. The random effect of the circuit shows a small standard deviation (0.03299), suggesting circuit characteristics exert a consistent influence on race positions across different tracks. With a notably high z value for the intercept (192.37) and sd_from_mean (13.15), both are statistically significant, pointing to a robust model despite the scaled residuals hinting at unexplained variance. The model's AIC of 61558.2 reveals no improvement compared to our base model. The diagnostic plots still reveal issues regarding overestimation around the mean, underestimation slightly above the mean, and overestimation again towards the right tail.

Thus, we switch to predicting the improvement score with multiple variables, which is approximately normally distributed.

```
Linear mixed model fit by REML ['lmerMod']
Formula: improvement ~ sd_from_mean + (1 | circuit_name)
   Data: data

REML criterion at convergence: 62123.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2968 -0.4143  0.0412  0.5738  4.2247
```

```
Random effects:
 Groups       Name        Variance Std.Dev.
 circuit_name (Intercept)  0.03474 0.1864
 Residual                 25.64757 5.0643
Number of obs: 10211, groups:  circuit_name, 35

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.35170    0.06176   5.694
sd_from_mean  0.87110    0.05197  16.761

Correlation of Fixed Effects:
           (Intr)
sd_from_men 0.000
```

Although the LMM fits the improvement score reasonably well, it may not fully capture the light-tailed distribution observed at the ends of the data. The residual variance is mostly homogeneous, with a peak around the mean indicating a concentration of values.

Now, a linear model was analyzed to determine the relationship between a driver's starting position (grid) and their improvement by the end of the race.

Table 2: Summary statistics for Linear Regression Model with Starting Grid Position.

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
(Intercept)	-4.12304	0.08687	-47.46	$< 2e-16$
Grid	0.41226	0.00693	59.52	$< 2e-16$

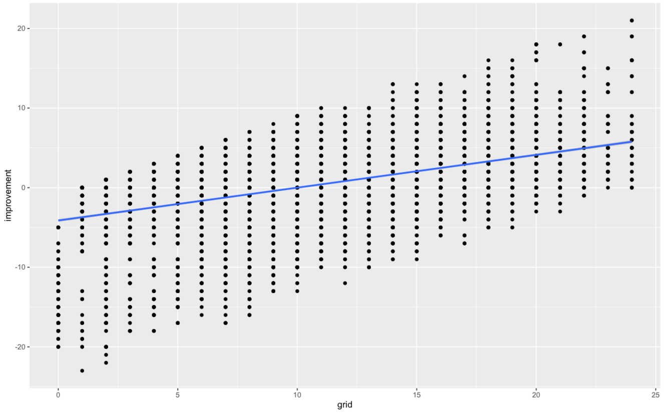

Figure 4: Linear Model of Starting Position on Improvement.

Despite appearing somewhat spread out, there may be some indication of a linear relationship between start position and improvement. There is less data for the combinations of high starting positions with high improvement and low starting positions with high improvement. This is logical, as moving up a position is less likely if you are already starting at a high rank. To account for this, we include the starting position as a secondary predictor.

4.4 Full Model

The full model now incorporates both the standard deviation from the mean pit stop time (sd_from_mean) and the starting grid position (grid) as fixed effects, along with a random effect for the circuit name.

```
Linear mixed model fit by REML ['lmerMod']
Formula: improvement ~ sd_from_mean + grid + (1 | circuit_name)
   Data: data
```

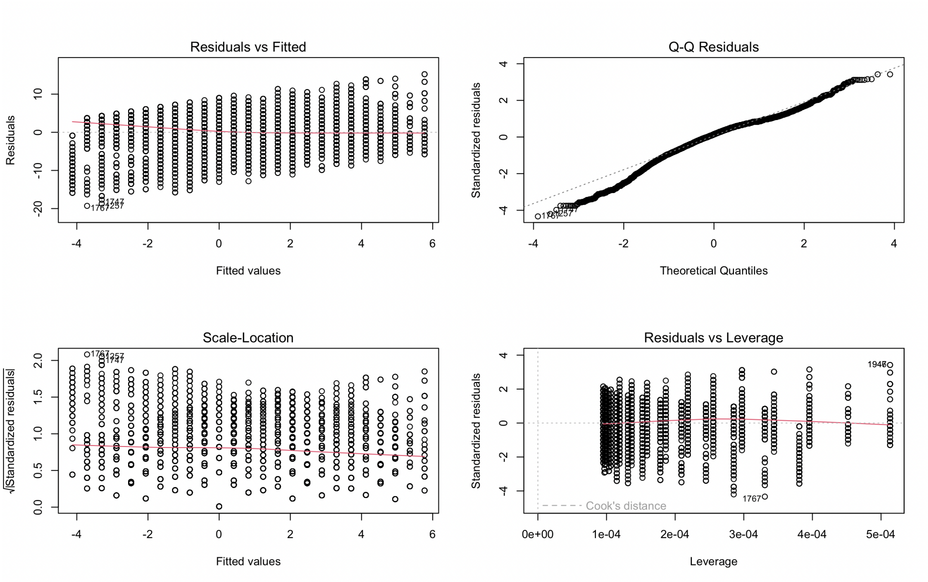

Figure 5: Diagnostic Plots for the Linear Model of Starting Position on Improvement.

```
REML criterion at convergence: 59050.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.1623 -0.5684  0.1519  0.6805  3.6711

Random effects:
 Groups       Name        Variance Std.Dev.
 circuit_name (Intercept)  0.06349 0.252
 Residual                 18.94837 4.353
Number of obs: 10211, groups:  circuit_name, 35

Fixed effects:
             Estimate Std. Error t value
(Intercept)  -4.12455    0.09829  -41.96
sd_from_mean  0.90719    0.04468   20.31
grid          0.41269    0.00688   59.98

Correlation of Fixed Effects:
            (Intr) sd_fr_
sd_from_men -0.010
grid        -0.758  0.013
```

The linear mixed-effects model reveals that the variability in pit stop timing (sd_from_mean coefficient = 0.90719) and the starting grid position (grid coefficient = 0.41269) significantly predict a driver's race improvement. The model's random effects suggest circuit characteristics contribute to performance variability, though to a smaller extent (circuit random effect standard deviation = 0.252). The model fits the data with a REML criterion of 59050.7 and shows that starting further back provides more scope for improvement during a race.

5 Discussion and Conclusions

Both Poisson and Negative-binomial regression models seemed to perform moderately well in predicting a F1 driver's placement from how many standard deviations away from the mean pit stop lap number they were. Modification of model variables improved our performance, especially when the starting position information was added. This is most likely due to the heavy association between one's starting position and ending position. In most cases, it seems to be that drivers will not move too far away from their starting position. In other words, if a driver starts at a high placement since they performed well prior to the present race, they are more likely to continue in a higher position throughout the race as well. Thus, comparing the two models, one with and one without the starting position information, to F1 racing viewers who watch from the beginning and possess the prior knowledge on their starting positions as opposed to those who only have viewer knowledge from mid-race, similar to the statistical models, those with prior knowledge are expected to predict the final outcomes better. As such, there is much improvement to be made to both the structure of this dataset and model construction. With more untested variables, there may be higher performing models that could provide more useful information. Nonetheless, it seems to be that it is somewhat possible to attempt to predict F1 race outcomes, whether it be from a strategic point-of-view such as through pit stop strategies, or from a logistics point-of-view such as through prior performance.