Strategic Optimization From Qualifying to Pole Position in Formula 1*

An Analysis on Driver's Qualifying Lap Time for the Last 2 Decades (2006 - 2023)

Zhijun Zhong

April 2, 2024

In this study, I created predictive model from predicting the optimal lap time for Q2 eliminations that minimizes tire wear to using data from the first 2 round of the qualifying sessions to estimate the pole position lap time in the third qualifying round available in Kaggle. My finding revealed that it is possible to accurately estimate the pole position lap time and the optimal elimination time in the second qualifying. By using the predictive models, it can bring a better knowledge to teams and drivers on others while still own the ability to alter strategies in the following sessions. My aim is to be both competitive and preserving resources in qualifying in order to have the best outcome.

Table of contents

*Code and data are available at https://github.com/JerrZzzz/Strategic-Optimization-in-F1.

1 Introduction

Formula one had been the most advance and well-known auto sport around the world in the last few decades. We have witnessed the development of technology from the immense power of V12 to V6 engine to complex aerodynamics but have unlimited potential. Every single team on the grid have only one goal which is to win races and every single driver on the grid have only one aim which is to win Formula one driver world champion. Winning races in Formula one consists of skill, strategies, performances and luckiness. There are 15 to 20 Grand Prix every year taking place in all around of the globe. Every single one of the Grand Prix contain 3 practice sessions, 3 qualifying sessions and one big race session. Some people generally believe that only the race session matter in a Grand Prix since it is the only one which decide the points board. However, 3 qualifying sessions are equally important because it can directly decide the grid order. Dennis Wesselbaum and et al found in their research paper that the first grid position, AKA pole position, will give the driver an significant amount of advantage since the start of the race(Wesselbaum and Owen 2021). They also found that there is about 10% increase in probability for a pole position driver to win the race(Wesselbaum and Owen 2021). I decide to use a kaggle dataset published by Rohan Rao to build model and predict the lap time for second and third qualifying session(Rao 2020).

According to Formula 1 sporting regulations article 39.2(Fédération Internationale de l'Automobile (FIA) 2024), the slowest 5 cars in first qualifying session, aka Q1, will be "prohibited" to take further part of the qualifying session. The slowest 5 cars in second qualifying session, aka Q2, will be "prohibited" to take further part of the qualifying session

leaving 10 cars to the final qualifying session, aka Q3. Moreover, in article 30.2 (Fédération Internationale de l'Automobile (FIA) 2024), there is only 13 sets of dry weather tyres for every Grand Prix including 8 sets of soft compound tyres, 3 sets of median compound tyres and 2 sets of hard compound tyres. The harder the tyre is the longer the tyre is going to last and the slower the lap time is. As A. Tremlett and D. Limebeer stated that tyre saving whilst optimal the lap time is considered as the most important strategy in formula 1(Tremlett and Limebeer 2016). The teams have to make choices on what tyre they should use in practice and qualifying sessions in order to make sure that there are enough tyres for them to change in race sessions. Thus, I think that it is the best to save tyre in the second session of the qualifying for the race. I created my first model to find the slowest estimated lap time to enter Q3. It means that any driver who can be quicker than this lap time will be able to enter Q3 using Q1 as the predictor. Therefore, driver can both save tyres and enter Q3 safely. Furthermore, I believe that compete for the pole position is every important in Q3 for all teams. Thus, I created another model to find the estimate pole position time using Q1 and Q2 as the predictor to estimate pole position lap time. After using the models, we found that there is a significant linear relationship between Q1 and Q2 which means that we can use a linear model on Q1 to estimate Q2. Moreover, pole position lap time mainly rely on Q2 rather than Q1, but we still need Q1 variable on the model.

The remainder of this paper is structured as follows. Section 2 is a section about the dataset I used in this paper. I carefully examined every variable of in the dataset and created graphs to visualize every variable. In Section 3, I presented two significant models for estimating Q3 entering lap time and pole position lap time. I also explained the prediction of these models. Section 4 visualize these models with graphs which can be brought our understanding to a second level, While in the Section 5, I thoroughly discussed the application of my models and further action to strengthening using other currently unavailable variables like weather conditions and tyre usage. Moreover, in the Section A, it consists graphic analyze on the quality of my models, description of data columns and so on. Finally, Section C.2 includes all resources I used in this paper.

2 Data

2.1 Data sources

The data used in this paper is extracted from Kaggle created by Rohan Rao(Rao 2020). This dataset is created with real formula 1 data from real races. It contained from every circuit which had a formula one race in history to every race in history. In this paper, I only used races dataset and qualifying dataset to perform the model.

This paper is entirely made with R program (R Core Team 2023) along with other packages which help me with graphing, table building and so on. I used tidyverse(Wickham et al. 2019), dplyr(Wickham et al. 2023), knitr(Xie 2023) to clean data from original dataset.

By using ggplot2(Wickham 2016), modelsummary(Arel-Bundock 2022), stringr(Wickham 2023b), rstanarm(Goodrich et al. 2022), I am able to create the model and graph them. When I comes to downloading data for Kaggle, I used httr(Wickham 2023a), jsonlite(Ooms 2014), usethis(Wickham et al. 2024), devtools(Wickham et al. 2022), kaggler(Kearney 2024). readr(Wickham, Hester, and Bryan 2024) helped me to read csv files to my workspace. Some other used package are listed, janitor(Firke 2023), lubridate(Grolemund and Wickham 2011), gridExtra(Auguie 2017), here(Müller 2020).

2.2 Datasets

Table 1: Q1 Position 11 vs Q2 Position 11 Lap Time

raceId	q1sec	q2sec
Id number assigned to every unique grand prix in history Range: 1-1100	Q1 11th fast driver lap time in Q1 Range: 54.388-130.529	Q2 11th fast driver lap time in Q2 Range: 53.995-129.377
1	86.026	85.504
2	95.260	94.769
3	96.443	95.975
4	93.479	93.242

In race dataset, it contained the race information for every raceId. I can identify the year of the race which is very important to limited to recent races. In qualifying dataset, it consists of all the lap time for Q1 to Q3 for all races from 2006 to 2023. By using the qualifying dataset, I am able to create another dataset which involves the 11th fastest driver for Q1 and the 11th fastest driver in Q3. In this way, we are able to use the 11th fastest time in Q1 to predict the 11th fastest time in Q2(Table 1). There are 341 observations in this dataset where raceId go from 1 to 1110 and lap time varies from 54 seconds a lap to 130 seconds a lap. Moreover, I deviated from the original dataset to create a new one where it records the total pole position a each driver have in their career and collect those number of races where they turn their pole position to a race win. So, I can use these information to calcuate the ratio of pole position to win for each driver.

Table 2: Q1 fastest driver vs Q2 fastest driver vs Q3 fastest driver Lap Time

raceId	q1sec	q2sec	q3sec
Id number assigned to every unique grand prix in history Range: 1-1100	Q1 fastest driver lap time in Q1 Range: 53.904-127.130	Q2 fastest driver lap time in Q2 Range: 53.647-126.609	Q3 fastest driver lap time in Q3 Range: 53.377-125.591
1	78.143	77.328	76.609

Table 2: Q1 fastest driver vs Q2 fastest driver vs Q3 fastest driver Lap Time

raceId	q1sec	q2sec	q3sec
2	88.917	87.702	86.720
3	65.116	64.951	64.391
4	72.386	71.908	71.365

By using the same technique, I also create a dataset to record the fastest time in Q1, Q2 and pole position time in Table 2. There are in total of 340 observations in this dataset. Notice that I have changed the lap time of both dataset to a unit of second which will be easier to visualize the change in model section.

2.3 Visualizing Lap times

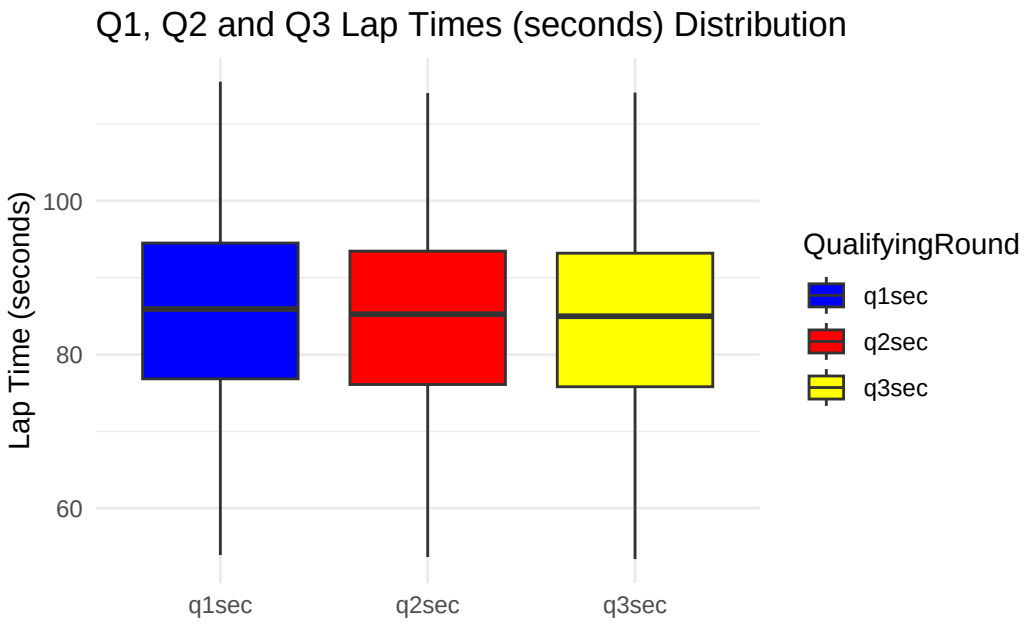

Figure 1

In order to take a close look at the distribution of the lap times, I created a box plot based on the q1, q2 and q3 lap times using the fastest driver from each qualifying session data(Table 2). We can see that the difference of the distribution is not obvious. The distribution of the three sessions are very similar. Since that the lap time between Q1, Q2 and Q3 are very close, it is normal to see a very close pattern between 3 sessions. If we look closely, we can still see that the distribution of lap time in Q3 is a tiny little faster compare to other 2 sessions.

2.4 Similar datasets

There are a lot of self made dataset in different statistical website like Kaggle where fans pull out the dataset all by themselves. I have done a lot of research and the dataset I used in this paper is by-far the most comprehensive and accurate. However, Formula one data are strictly limited to each team. Further information are confidential to team's own database. So, it is very hard for us to collect detailed information like tyre usage, car setup and engine output mode where these factors are crucial when we analyze specific lap times for pole position and so on. Since that the lack of those information, our predictions cannot be considered as accurate at a formula one team level. But it is sufficient for the use of television broadcast or building cliffhangers.

3 Model

The two models shown in this section below aim to predict lap times using different predictors and different type of model to fit. This is motivated by the fact that more and more teams didn't take qualifying sessions seriously since that overtaking in race on track is easier and easier. Ending on top nowadays seems having smaller and smaller advantage to others. Thus, By creating these models, I want to raise the importance of qualifying sessions. After building models, I created different graphs to verify that the model that I build is a good fit to most observations in Section C includes residual plots and distribution, and qq-plots.

3.1 Model set-up

I set up null hypothesis and alternative hypothesis to test if there is relationship between my predictor and response variable defined as following. I plan to use p-value to justify if we can reject null hypothesis. I plan to use p-value equal to 0.5 as a threshold where if p-value is below 0.5, then we can reject null hypothesis, and if p-value is above 0.5, then we do not have sufficient evidence to reject it meaning that if the p-value is above 0.5, the relationship between predictor and response variable are weak and vice versa.

- Null Hypothesis: There is no relationship between X and Y, the model between X and Y does not have meaning.
- Alternative Hypothesis: There is strong relationship between X and Y, the model between X and Y can predict general cases.

We can define our basic linear regression as Equation 1 while we use coefficients β_1 to create slope for our linear regression line and β_0 to be the value of y when x is zero. We can also add a x^2 to create Equation 2 in order to make it as a non-linear model so that we can compare it to the linear model and choose a better fit on our model. Moreover, by adding a second variable x_2(Equation 3), we can create a 3 dimensional linear model which we use 2 predictors

Table 3: Model of Elimination Time for Q2

	Linear Model Summary	Non-linear Model Summary
(Intercept)	0.0723	87.4653***
	(0.1232)	(0.0166)
q1sec	0.9937***	
	(0.0014)	
poly(q1sec, 2)1		207.4629***
		(0.2902)
poly(q1sec, 2)2		−0.1304
		(0.2902)
Num.Obs.	306	306
R2	0.999	0.999
R2 Adj.	0.999	0.999
AIC	114.4	116.2
BIC	125.6	131.1
Log.Lik.	−54.206	−54.104
RMSE	0.29	0.29

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

to estimate the response variable. With the application of these equations, we can properly measure our estimated lap time for both second qualifying elimination and pole position.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

$$y = \beta_0 + \beta_1 x + \epsilon \tag{3}$$

4 Results

4.1 Predict Q2 Elimination Time Model

When I use the 11th fastest driver in qualifying 1 to predict the elimination time for Q2, I introduced a linear model and a nonlinear model. In Table 3, the column on the left is the linear model while on the left is the nonlinear model. We define the linear model Equation 4 where $\beta_0 = 0.0723$ and $\beta_1 = 0.9937$. We define y as our response variable which is our

estimate goal while x as our predictor which is the 11th fastest time in qualifying 1, and ϵ is a random error occur when there is variability in the response variable that isn't explained by our predictor or the randomness in the real data.

$$y = 0.0723 + 0.9937x + \epsilon \tag{4}$$

The nonlinear model is defined by Equation 5 where where $\beta_0 = 87.4653$, $\beta_1 = 207.4629$, and $\beta_2 = -0.1304$. So the full equation becomes Equation 5 while x, y and ϵ is defined the same as above.

$$y = 87.4653 + 207.4629x + -0.1304x^2 + \epsilon \tag{5}$$

As suggested in the table Table 3, that any coefficients with 3 star behind it means that the p-value of this coefficient is smaller than 0.001. Any coefficient with p-value smaller than 0.001 suggest that there is very little chance of this coefficient to be zero. So in our model, the coefficient for the predictor as suggested with 3 stars tell us that the predictor will almost never be zero meaning that we can reject null hypothesis where there is no relationship between 11th fastest lap time in qualifying 1 and the elimination time. It is the same as the nonlinear model, where our predictor have p-value smaller than 0.001 suggesting strong relationship between 11th fastest lap time and elimination time.

In Table 3, the value for R^2 and adjR^2 are both 0.999 with a sample size of 306. R^2 and adjR^2 suggested the fit of the model on our data. It means that 99.9% of my observations can be explained by our model. Thus, it is a good fit for both linear model and nonlinear model. However, extreme high R^2 value often suggest lack of sample size or over-fitting. With a 306 sample size, we cannot conclude with a lack of sample size. In consideration of our qualifying 1 lap time in unit second, we expect R^2 value to be higher than normal since that the lap time difference are not significantly high often result in a 0.3 or a 0.5 second difference. So it is normal to have such high R^2 value because our prediction have to be accurate to the third decimal place.

We plan to compare two models using Bayesian Criterion and Akaike Criterion which is the AIC and BIC columns in Table 3 to find a better fit. Generally, if the value of AIC and BIC are low, it indicated a better model. However, there are no scale to measure how good the model is. It is often apply on both model and compare which one is lower suggesting a good fit. In Table 3, by comparing AIC and BIC, the difference is obvious suggesting linear model being a good fit. It matched the principle of model parsimony which stated that the all else being equal, the simple the model is the more complete the test of hypothesis is. So a linear model is what we prefer when it comes to predicting elimination time.

After creating a figure on the model for predicting second qualifying elimination time, we found that the relationship between the 11th fastest time in qualifying 1 and the elimination time in Q2 are very strong. For every one second gain in Q1 time, we expect to get a 0.99 second gain

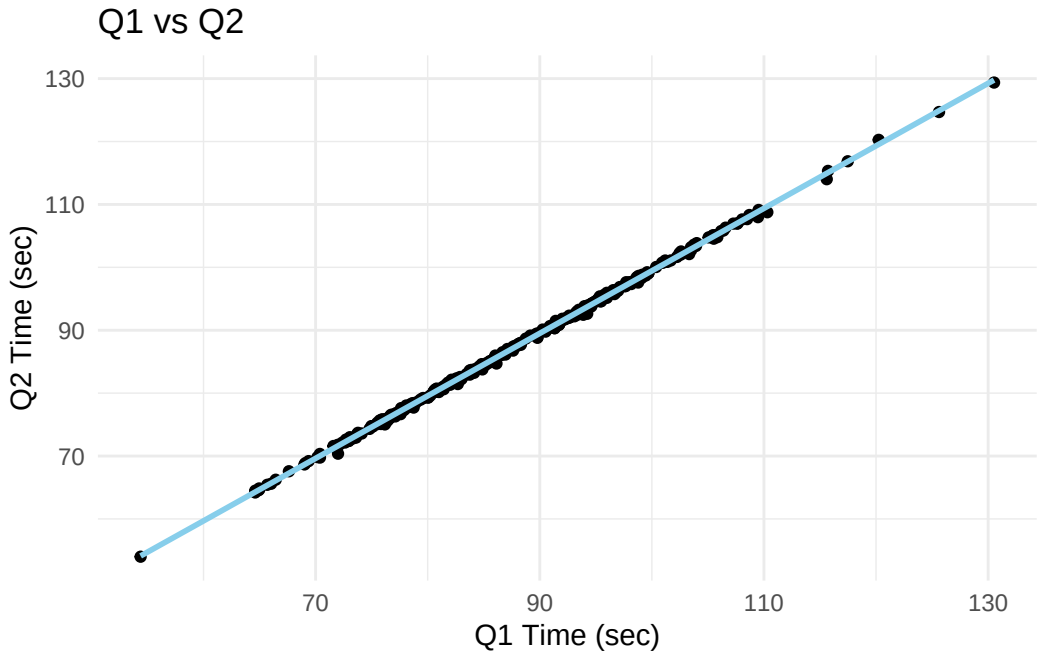

Figure 2: Fittness of the Model for Elimination Time in Q2

in Q2 time. However, our prediction is not always accurate due to the track condition and so on. We can create a table to measure the uncertainty of our model in Table 4.

Table 4: Uncertainty Calculation on Model of Elimination Time for Q2

fit	lower_bound	upper_bound	uncertainty	input
69.63007	69.05674	70.20339	0.5733268	70
74.59848	74.02616	75.17080	0.5723207	75
79.56689	78.99525	80.13853	0.5716392	80
84.53530	83.96402	85.10659	0.5712835	85
89.50372	88.93246	90.07497	0.5712542	90
94.47213	93.90058	95.04368	0.5715513	95
99.44054	98.86837	100.01272	0.5721744	100
104.40895	103.83583	104.98208	0.5731224	105
109.37736	108.80297	109.95176	0.5743937	110

In this Table 4, We have input a range of first qualifying time from 75 to 110 seconds. We discovered that the average uncertainty of predicted time is 0.572 seconds. It suggested that the all 95% of the points will lay in range of ± 0.572 seconds. So our predicted Q2 elimination time can be various in a range of 1 second or more. However, the fit of our model is reasonable since most of the difference of the position 11th between Q1 and Q2 are usually 0.6.

Table 5: Model of Pole Position Time

	Predicting pole position lap time model
(Intercept)	−0.0783
	(0.2386)
q1sec	−0.0541
	(0.0764)
q2sec	1.0529***
	(0.0770)
Num.Obs.	274
R2	0.998
R2 Adj.	0.998
AIC	411.8
BIC	426.3
Log.Lik.	−201.916
RMSE	0.51

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

4.2 Predict Pole Position Time Model

A more complex model in Table 5 is used when we are trying to predict pole position time for every race where we used 2 variables both the fastest time in the first qualifying and the fastest time in the second qualifying. This model is defined as a multivariable Equation 6 where $\beta_0 = -0.0783$, $\beta_1 = -0.0541$, and $\beta_2 = 1.0529$. x_2 is our main variable we use to form the regression line which is the fastest lap time in second qualifying while x_1 is defined as the support variable to x_2 to adjust its position which is the fastest lap time in first qualifying session, and ϵ is the random error term representing the variability and real-life errors.

$$y(x_1, x_2) = -0.0783 + -0.0541x_1 + 1.0529x_2 + \epsilon \tag{6}$$

From model summary Table 5, it suggest again like our first model that the R^2 and R^2Adj is very high at 99.8%. It means that 99.8% of our current data can be explained using this built model. With a sample size of 274, we cannot be confident that our sample size is big enough to prevent it from over-fitting. However, as mentioned above, we expect R^2 and R^2Adj to be big since the y scale is small. Moreover, 3 stars behind the coefficient of x_2 (fastest lap time in second qualifying) means its p-value is smaller than 0.001 which suggested a extremely weak possibility of it to be zero. It suggest a strong relationship between fastest lap time in second qualifying and our pole position time. However, there are no stars behind the coefficient of x_1 (fastest lap time in first qualifying session) showed us that the p-value is probably higher than 0.1 meaning there is a great chance it can be zero. Furthermore, the intercept row display no

star meaning the intercept value can probably be zero. Thus, to sum up, this model do suggest that there is a linear relationship between Q1, Q2 time with pole position time. However, it also tells us that the relationship between the fastest lap time in the first qualifying are weak to the pole position lap time. From the p-value of the interception, we understand that when we achieve a very quick lap in the second qualifying, the pole position lap is still predicted as a fast lap.

Table 6: Uncertainty Calculation on Model of Pole Position

fit	lower_bound	upper_bound	uncertainty	inputq1	inputq2
69.44370	68.43699	70.45041	1.006711	70	69.63007
74.40412	73.39935	75.40888	1.004769	75	74.59848
79.36453	78.36098	80.36809	1.003555	80	79.56689
84.32495	83.32188	85.32802	1.003072	85	84.53530
89.28537	88.28205	90.28869	1.003320	90	89.50372
94.24578	93.24148	95.25008	1.004300	95	94.47213
99.20620	98.20019	100.21221	1.006009	100	99.44054
104.16662	103.15817	105.17506	1.008442	105	104.40895
109.12703	108.11544	110.13863	1.011597	110	109.37736

Since that the linear model for estimating pole position used 2 variables, it is very hard to visualize it in the paper, but we can still look at the uncertainties applied on this model. I have simulated same time for first qualifying as our first model Section 4.1. I also assumed that the first position followed the same pattern discussed in the first model Section 4.1. Comparing it to the model above, we can see that the range for my uncertainty goes from approximately ± 0.572 to about ± 1.00. Even though that our first model is used to model 11th position lap time, it can also be possible to predict position 1. Thus, the uncertainty brought by multivariable model is significantly higher than a simple linear model in model 1 Table 4.

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

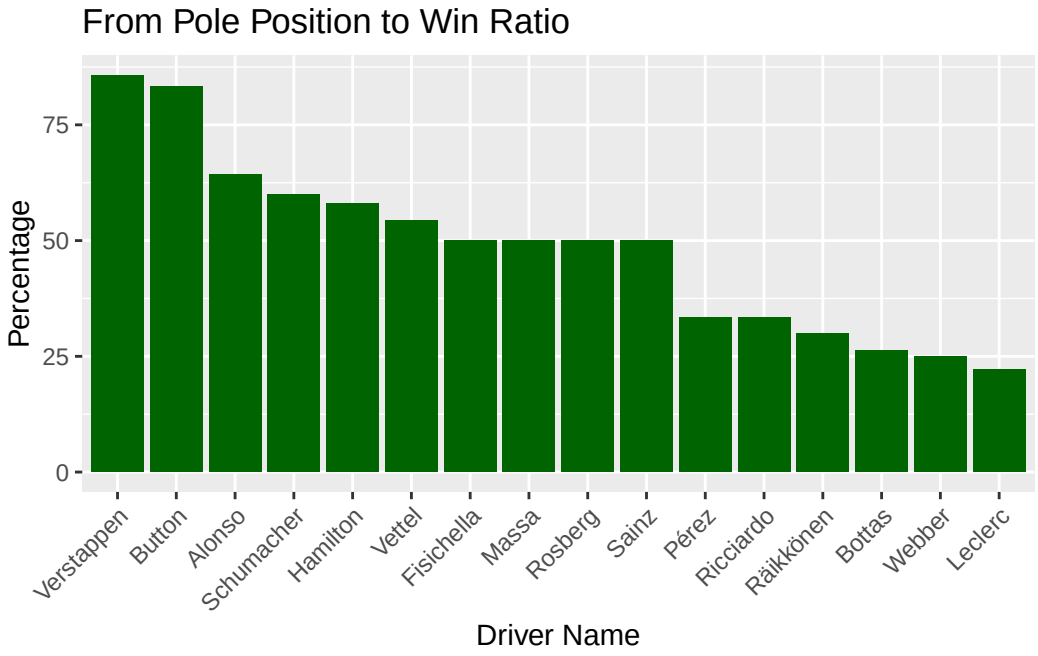

Figure 3

A Appendix

B Additional data details

C Model details

C.1 Residual distribution check

C.2 Diagnostics

is a trace plot. It shows... This suggests...

is a Rhat plot. It shows... This suggests...

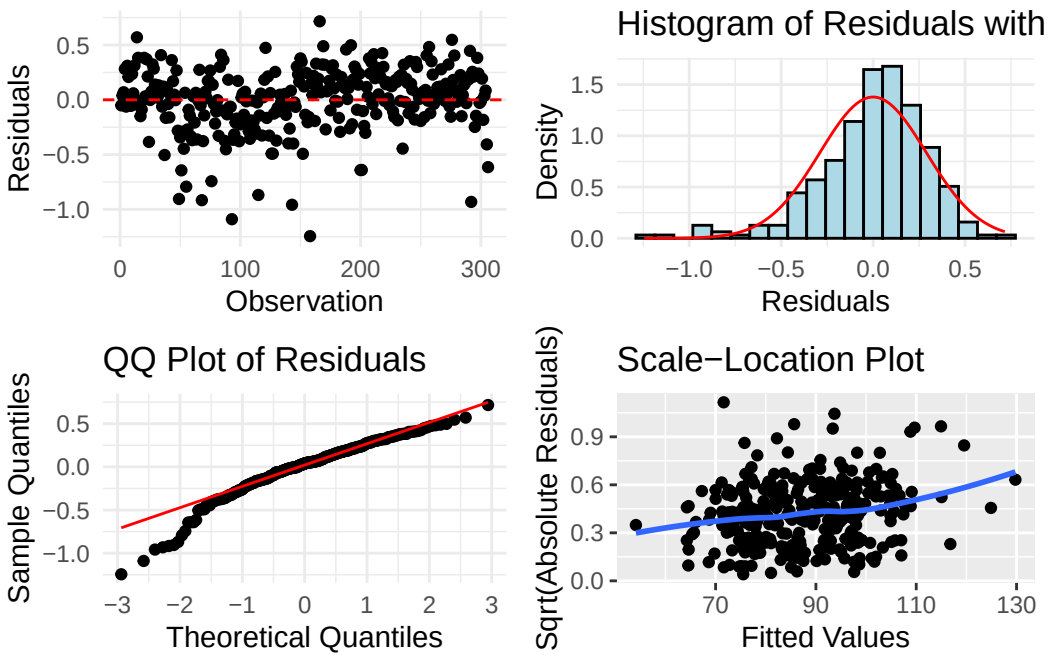

Figure 4: Examining how first model fits

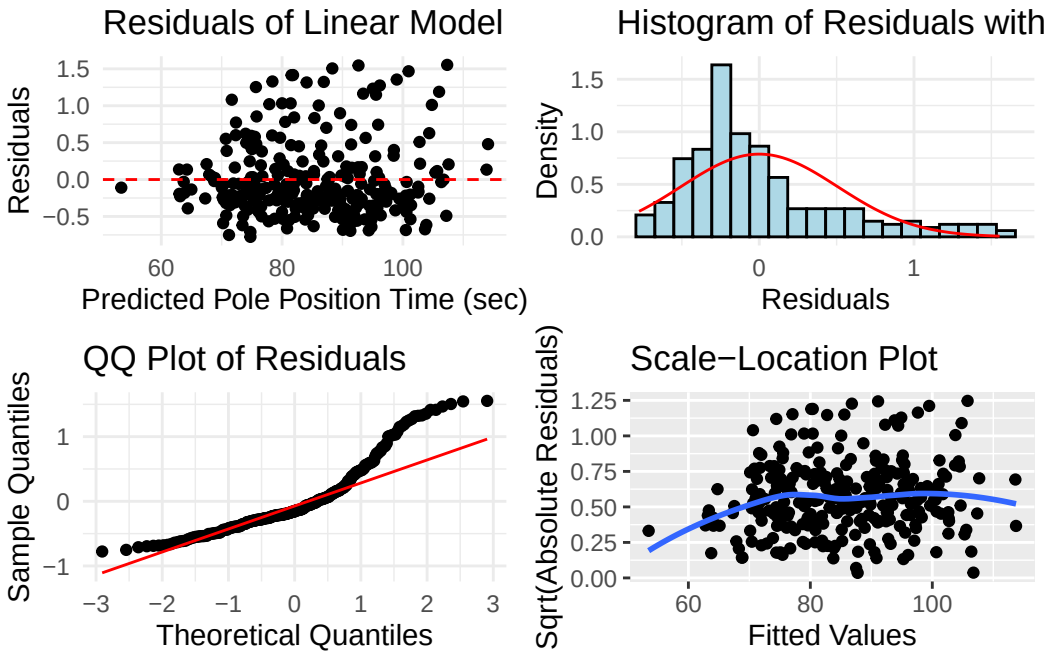

Figure 5: Examining how second model fits

Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

Fédération Internationale de l'Automobile (FIA). 2024. "2024 Formula One Sporting Regulations." https://www.fia.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Kearney, Michael Wayne. 2024. *Kaggler: Kaggle API Client*.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and r Objects." *arXiv:1403.2805 [Stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rao, Rohan. 2020. "Formula 1 World Championship (1950-2020) Dataset." https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020.

Tremlett, A. J., and D. J. N. Limebeer. 2016. "Optimal Tyre Usage for a Formula One Car." *Vehicle System Dynamics* 54 (10): 1448–73. https://doi.org/10.1080/00423114.2016.1213861.

Wesselbaum, Dennis, and P Dorian Owen. 2021. "The Value of Pole Position in Formula 1 History." *Australian Economic Review* 54 (1): 164–73.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023a. *Httr: Tools for Working with URLs and HTTP*. https://httr.r-lib.org/.

———. 2023b. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jennifer Bryan, Malcolm Barrett, and Andy Teucher. 2024. *Usethis: Automate Package and Project Setup*. https://usethis.r-lib.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. *Devtools: Tools to Make Developing r Packages Easier*. https://devtools.r-lib.org/.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.