Farrukh Anwar, Connor Horan, Elishel Sason

Professor Simone

STA 4164-0002

2 December 2024

# A Regression Analysis of Lead Changes in NASCAR

**Abstract**

This study uses multiple linear regression to investigate the factors influencing the number of lead changes in the National Association for Stock Car Auto Racing (NASCAR) Cup Series. Key motivations included finding out how to increase the number of lead changes, one of the most exciting aspects of the sport, and investigating the impact of the Next Gen car design implemented from the 2022 season onwards on lead changes. The dataset comprises 164 races from 2020 to 2023, taken from the NASCAR website and processed for use. Each race entry includes quantitative predictors: number of lead changes, completed race distance (in miles), number of cautions, and average speed (in miles per hour); and categorical predictors: restrictor plate usage, playoff status, and car generation. The analysis was conducted in R to assess outliers, collinearity, and assumption violations. Exploratory data analysis added a natural logarithmic transformation of speed, as well as interactions between distance and cautions and the transformation of speed and restrictor plate usage, to the maximum model. After fixing the maximum model's violations of Normality and Homoscedasticity with a fourth root transformation on the response, the final model was then determined using a backward selection algorithm with $\alpha = 0.10$. No outliers were removed as all observations contained plausible values and none violated Cook's Distance. After fixing collinearity and assessing assumptions, the fitted sample model had $r^2 = 0.668$ and became

$$\widehat{\sqrt[4]{CHANGES}} = 1.5573 + 0.0021 \cdot DISTANCE - 0.0027 \cdot SPEED + 0.3445 \cdot RESTRICTED.$$

The final model implies that longer race distances, lower average speeds, and restrictor plate usage increase the number of lead changes, whereas the number of cautions, playoff status, and the Next Gen implementation have no impact. However, these findings should be taken with caution as correlation does not necessarily indicate causation and no training-testing split was performed.

**Introduction and Motivation**

The National Association for Stock Car Auto Racing (NASCAR) is one of the most popular motorsports-sanctioning bodies in the United States, with the Cup Series Championship being its highest offered level of competition. While much emphasis is placed on who wins each event, an arguably more substantial portion of the racing spectacle stems from the thrilling changes in first as drivers battle over the lead track position.

This paper attempts to tackle the research question "Is there a relationship between the number of lead changes in a race and one or more key predictors?" For example, the introduction of the newly designed Next Gen car from the 2022 season onwards was partially intended to promote more overtaking opportunities and create better "raceability in traffic" (Taranto). One application of research into multiple linear regression involving the number of lead changes and the usage of these Next Gen cars, alongside other predictors, is yielding insight into whether the design overhaul has genuinely correlated with more changes in position up front. This study may be useful to those in the motorsports entertainment business by shedding light on what factors are correlated with lead changes, which could potentially drive viewership and profits if these factors are implemented in correspondence with the results of this research.

**Data Description**

The dataset used for this research was compiled from the official NASCAR website's data backend found at the cf.nascar.com domain. This race data (comprised of in-car sensor readings and manual entries) is collected by NASCAR for use in their website's front end. Therefore, the data is stored in a JavaScript Object Notation (JSON) format, which requires automated means for further data transformation and processing. The Python script used to extract the relevant data and transform it into a usable CSV file is found in the Appendix.

Descriptive information, such as the name of the track, was stripped from the original dataset in order to provide cleaner data to work with. Unquestionably collinear data was also excluded from extraction in order to avoid potential collinearity issues in future analysis. For example, completed race distance was extracted from the original dataset whereas the number of laps completed was not. This decision happened because the completed race distance is a linear multiple of laps completed as the track length remains constant for each lap. So for a 200-lap race on a 2.5-mile track, the race distance will usually be 500 miles. The reason race distance was used instead of laps completed is because race distance is a more reliable measure for comparison between data points. Comparing a 200-lap race to a 300-lap race requires knowledge of the track length of each event in order to make a valid comparison. On the other hand, it is much easier to tell the difference between a 200-mile race and a 300-mile race to draw logical inferences.

After performing the described data cleanup, the final dataset contains observations from all 2020-2023 NASCAR Cup Series races (including exhibition races). It has 164 race entries with each race entry containing 7 variables (4 quantitative and 3 categorical), which are described in the following table.

| Variable Name | Type | Description |
|---|---|---|
| CHANGES | Quantitative | Number of lead changes (new leader at the end of a lap) |
| DISTANCE | Quantitative | Completed race length distance in miles |
| CAUTIONS | Quantitative | Number of caution/yellow flags thrown for accidents and planned breaks (cars must slow down to a set speed for the duration of the caution) |
| SPEED | Quantitative | Average speed traveled by cars throughout the entire race in miles per hour |
| RESTRICTED | Categorical | 1 if cars are equipped with a restrictor plate in the engine that limits top speed; 0 otherwise |
| PLAYOFF | Categorical | 1 if the race is a playoff race; 0 otherwise |
| NEWGEN | Categorical | 1 if the race season is 2022/2023 meaning drivers are driving the Next Gen cars; 0 otherwise |

All of these variables appear useful in predicting lead changes. Longer race distances could provide more overtaking opportunities for the lead as drivers have more time to make a move. More caution flags can also help create lead changes as all cars are bunched up before resuming racing, eliminating any preexisting gaps between the leader and the pack behind. Additionally, because cars are at a slower pace during a caution, it encourages drivers to try alternative tire and fuel strategies to gain an advantage. As such, the leader may pull into the pitlane to momentarily stop to perform tire changes and refueling, allowing someone who did not pit the ability to take the lead. A lower average speed can also generate more lead changes, as

more skilled drivers with better-tuned cars have more opportunities to make a move for the lead in the longer braking zones. Restrictor plates, which limit top speed, can help with overtaking as cars are forced to bunch up together, preventing the leader from pulling away and allowing cars behind to slipstream past them as the leader is slowed down by being the first car punching a hole through the air. If the race is a playoff race, drivers may perform more attempts to take the lead as winning races in the playoffs provides a lot more value in the overall championship fight. As mentioned previously, because the Next Gen cars are intended to improve overtaking, they may lead to more lead changes in comparison to the previous generation cars used in the 2020-2021 seasons.

**Exploratory Data Analysis**

Because of the reasons outlined above, all variables in the final dataset will be explored using R. After importing the dataset into R using `races <- read.csv("races.csv")`, exposing variable names using `attach(races)`, printing out the summary statistics for each predictor using `summary(races)`, and graphically analyzing the distribution of each quantitative variable using the `boxplot` and `hist` (histogram) functions, it appears that no impossible values are found in this dataset, although some observations appear to be potential outliers.

For example, one observation has a particularly low race distance of 37.5 miles in comparison to the lower quartile of race distance at 252.8 miles. That same observation also has a low average speed of just 21.83 MPH, when the lower quartile of average speed is at 92.54 MPH. To determine the plausibility of this observation, it is important to take into account the context of these measurements. This observation is recorded from the 2023 Busch Light Clash at the Coliseum, which was a race held inside the Los Angeles Memorial Coliseum. As such, these

low race distance and average speed values are expected from such a small track that is essentially the size of a football field with a track length of just 0.25 miles. Considering that the dataset contains several other observations with sub-50-mile race distances and sub-50-MPH average speeds, it would be fair to label this particular observation as plausible. Therefore, it will currently be included in the development of the final model.

Another observation that warranted a look was one with 70 lead changes in comparison to the upper quartile of lead changes being just 21. This measurement was taken from the 2023 YellaWood 500 at the Talladega Superspeedway. Again, context is key for ascertaining the likelihood of this observation. Talladega is a 2.5-mile tri-oval track that requires full throttle throughout. The track's simple design, combined with the implementation of restrictor plates at superspeedways, promotes consistent side-by-side pack racing across multiple laps. Because a lead change is measured as a new driver being first at the end of a lap, it is possible for two lead drivers to repeatedly exchange position lap after lap without one ever completing an overtake on the other, potentially inflating this statistic. Given that other superspeedways in this dataset also yield 50+ lead changes, this particular observation will currently be kept. While all of the observations highlighted above appear plausible, if future diagnostics indicate trouble, the inclusion of these measurements may need to be re-examined.

Because a full model has not yet been constructed, it is only possible to examine the Existence assumption. By looking at the summary of the data using `summary(races)`, all variables appear to have considerable non-zero variance, thus indicating that the Existence assumption is not violated.

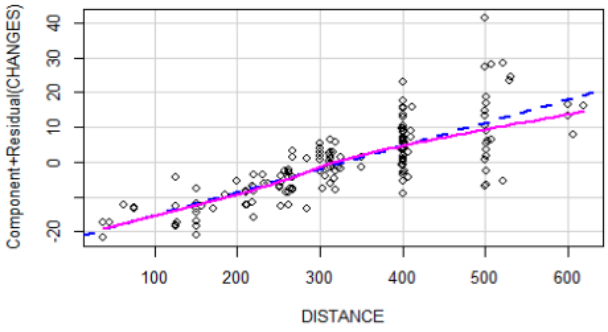
| CHANGES | DISTANCE | CAUTIONS | SPEED | RESTRICTED | PLAYOFF | NEWGEN |
|---|---|---|---|---|---|---|
| Min. : 2.00 | Min. : 37.5 | Min. : 0.000 | Min. : 21.83 | Min. :0.0000 | Min. :0.0000 | Min. :0.0 |
| 1st Qu.:11.00 | 1st Qu.:252.8 | 1st Qu.: 5.000 | 1st Qu.: 92.54 | 1st Qu.:0.0000 | 1st Qu.:0.0000 | 1st Qu.:0.0 |
| Median :15.00 | Median :317.7 | Median : 7.000 | Median :118.73 | Median :0.0000 | Median :0.0000 | Median :0.5 |
| Mean :17.93 | Mean :329.8 | Mean : 7.146 | Mean :113.39 | Mean :0.1768 | Mean :0.2439 | Mean :0.5 |
| 3rd Qu.:21.00 | 3rd Qu.:400.5 | 3rd Qu.: 9.000 | 3rd Qu.:134.87 | 3rd Qu.:0.0000 | 3rd Qu.:0.0000 | 3rd Qu.:1.0 |
| Max. :70.00 | Max. :619.5 | Max. :18.000 | Max. :191.97 | Max. :1.0000 | Max. :1.0000 | Max. :1.0 |

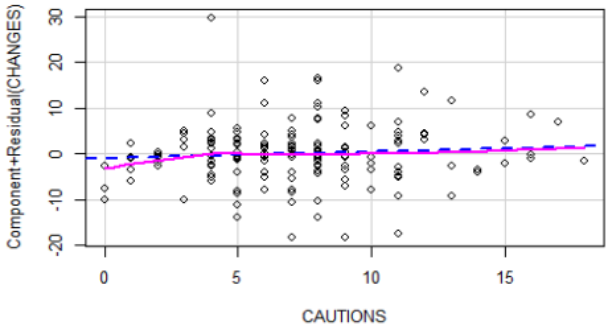
To assess potential collinearity issues between quantitative predictors that could lead to future model instability, the following correlation matrix was analyzed by running `cor_matrix <- cor(races[, c("DISTANCE", "CAUTIONS", "SPEED")])`.

| | DISTANCE | CAUTIONS | SPEED |
|---|---|---|---|
| DISTANCE | 1.00000000 | 0.32825409 | 0.54161536 |
| CAUTIONS | 0.32825409 | 1.00000000 | -0.31253822 |
| SPEED | 0.54161536 | -0.31253822 | 1.00000000 |

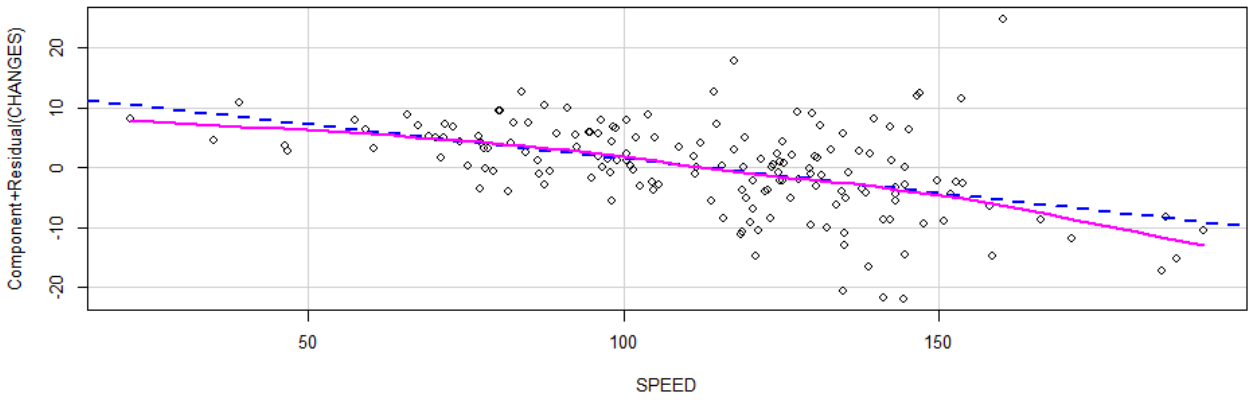
As none of the absolute correlation values ($|r|$) between two of the predictors above exceed 0.8, it is currently assumed that collinearity is not an issue in this dataset. The subject of collinearity will be revisited once a full model is constructed.

To determine the relationships between CHANGES and each of the quantitative predictors, partial regression plots were constructed by creating a temporary model with all predictors using `temp_model <- lm(CHANGES ~ DISTANCE + CAUTIONS + SPEED + RESTRICTED + PLAYOFF + NEWGEN, data = races)`, importing the car library which enables the creation of these plots with `library(car)`, and then finally graphing these plots using `crPlots(temp_model)`. DISTANCE appears to have a strong, positive linear relationship with CHANGES, while CAUTIONS has a more subtle, positive relationship.

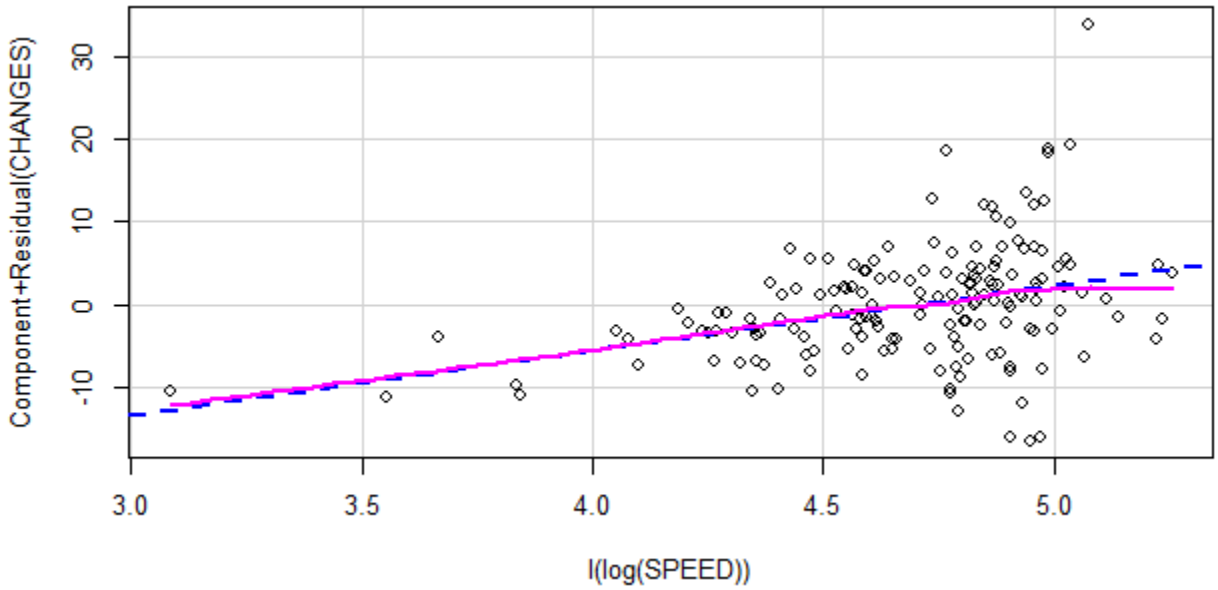


Besides demonstrating the relationships between the response and each predictor when accounting for all other predictors, these plots can also shed light on whether certain predictors need to be transformed. While DISTANCE and CAUTIONS do not appear to need transformations, the SPEED predictor strays off of the ideal blue line at both ends, indicating that it may require a curvilinear relationship to better correlate with CHANGES.




After performing a natural logarithmic transformation on SPEED and adding it to the initial temporary model, it appears that $ln(SPEED)$ better approximates the ideal line and has a moderate, positive linear relationship with CHANGES.



It is also important to consider potential interactions between predictors that could affect their relationships with the response. One possible interaction term can be between DISTANCE and CAUTIONS. Because having more cautions means less of the race distance is run under conditions where cars can freely overtake each other, a higher caution count can minimize the race distance's impact on the number of lead changes. Another interaction term can be between $ln(SPEED)$ and RESTRICTED. Because restrictor plates are usually used on superspeedways and superspeedways tend to have higher average speeds, the usage of restrictor plates can increase the effect of average speed on lead changes as slipstream overtaking is easier at higher speeds (as the car in front is slowed down more by air resistance) and the pack is condensed closer together because of the impact of the restrictor plates, thus making these overtaking opportunities more frequent.

**Future Direction**

This project aims to uncover what race factors most significantly influence lead changes in NASCAR, with a partial focus on understanding whether the Next Gen car design, introduced to improve overtaking and race dynamics, has achieved its goal. By developing and refining a multiple linear regression model, the analysis will examine variables like race distance, thrown caution flags, average speed, usage of restrictor plates, and playoff conditions, and the implementation of the Next Gen design overhaul to assess their roles in creating more overtaking opportunities for the lead. The full model will likely include the main effects of DISTANCE, CAUTIONS, SPEED, RESTRICTED, PLAYOFF, and NEWGEN, a natural logarithmic transformation of SPEED, and the interactions between DISTANCE and CAUTIONS and $ln(SPEED)$ and RESTRICTED. This model may reveal key motivators behind lead changes and provide insight into how these factors can enhance race excitement. Moving forward, key steps

will include creating this model in R, validating model assumptions, cleaning up outliers, carrying out model selection, performing more model diagnostics, testing for collinearity, and evaluating the final model's reliability. Once complete, the findings could offer actionable recommendations for race design that amplify engagement, making NASCAR events even more thrilling for fans.

**Maximum Model Selection and Diagnostics**

Based on the previous exploratory data analysis, the initial model was decided to include the variables DISTANCE, CAUTIONS, SPEED, RESTRICTED, PLAYOFF, and NEWGEN as well as $ln(SPEED)$, the interaction between $ln(SPEED$ and RESTRICTED, and the interaction between DISTANCE and CAUTIONS. This model takes the form of

$$CHANGES = \beta_0 + \beta_1 DISTANCE \cdot CAUTIONS + \beta_2 ln(SPEED) \cdot RESTRICTED + \beta_3 ln(SPEED) + \beta_4 DISTANCE + \beta_5 CAUTIONS + \beta_6 SPEED + \beta_7 RESTRICTED + \beta_8 PLAYOFF + \beta_9 NEWGEN + E$$

and was built using `full_model <- lm(CHANGES ~ DISTANCE * CAUTIONS + I(log(SPEED)) * RESTRICTED + I(log(SPEED)) + DISTANCE + CAUTIONS + SPEED + RESTRICTED + PLAYOFF + NEWGEN, data = races).`

Next, the model was assessed for outliers. Using Leverage, $h_i$ was calculated as $\frac{2(k+1)}{n} = \frac{2(8+1)}{164} = 0.110$. Checking the top 20 highest Leverage values of the observations, 19 appeared to be in violation, suggesting they were outliers.

```
> tail(sort(hatvalues(full_model)), n = 20)
      1      59      90     134     117      12     159      77      84      85      36     125
0.1042074 0.1125553 0.1126869 0.1131003 0.1211844 0.1272285 0.1301992 0.1321861 0.1546503 0.1593441 0.1613427 0.1668042
     94      43     106      83     142     147     101     124
0.1704460 0.1915834 0.1977895 0.2013833 0.2121286 0.2172681 0.3372278 0.5618516
```

Jackknife Residuals were then used to further assess for outliers. Compared to a cutoff of $t_{0.975}^{df=n-k-2=164-9-2=153} = 1.976$, observations 45, 4, 79, 131, and 155 appeared to be outliers in

the upper half, and observations 159, 138, 16, 118, 136, and 36 appeared to be outliers in the lower half.

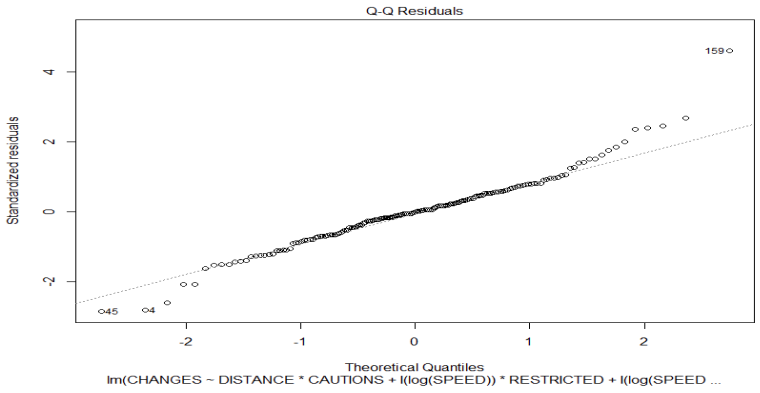
```
> head(sort(studres(full_model)), n = 20) # 45, 4, 79, 131, 155 appear to be outliers
       45         4        79       131       155       154        50       147        10        84       106        73
-2.914522 -2.880470 -2.656727 -2.109601 -2.098310 -1.637040 -1.536480 -1.526808 -1.512623 -1.449444 -1.432326 -1.406500
        8        57         1       160       107       122        19       121
-1.287903 -1.269186 -1.253296 -1.244386 -1.233014 -1.213080 -1.119820 -1.115738
> tail(sort(studres(full_model)), n = 20) # 159, 138, 16, 118, 136, 36 appear to be outliers
      133        23        83        75        49       117         5       151        42       127        72       128
0.9565066 0.9619529 0.9781708 1.0295184 1.0519465 1.2468726 1.2742545 1.3980214 1.4122577 1.5057151 1.5077541 1.6378413
       48        87        36       136       118        16       138       159
1.7734162 1.8573619 2.0174473 2.3957460 2.4331338 2.4851767 2.7361540 4.9485873
```

Cook's Distance was then assessed with the 10 largest observations using `tail(sort(cooks.distance(full_model)), n = 10)`. None of them appeared to be outliers when using a cutoff of 1.

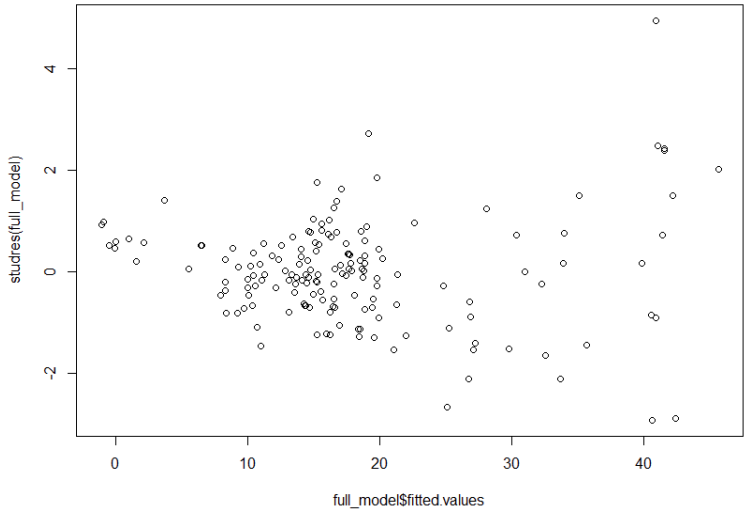
```
> tail(sort(cooks.distance(full_model)), n = 10) # Print 10 largest values
        16         79        106        124        118         45          4        147         36        159
0.04245874 0.04414669 0.05023924 0.05582155 0.05735585 0.05812837 0.05979252 0.06415261 0.07677087 0.31805457
```

With knowledge of the data gained from exploratory analysis, these detected outliers appeared to contain plausible values. Combined with the fact that none of these observations have an adverse impact on the formation of the model according to Cook's Distance, it has been decided to not remove any potential outlier values from the data at this time.

While checking the plots of the full model using `plot(full_model)` to assess assumptions, the Q-Q Residuals plot revealed what seemed to be a violation of the Normality assumption.



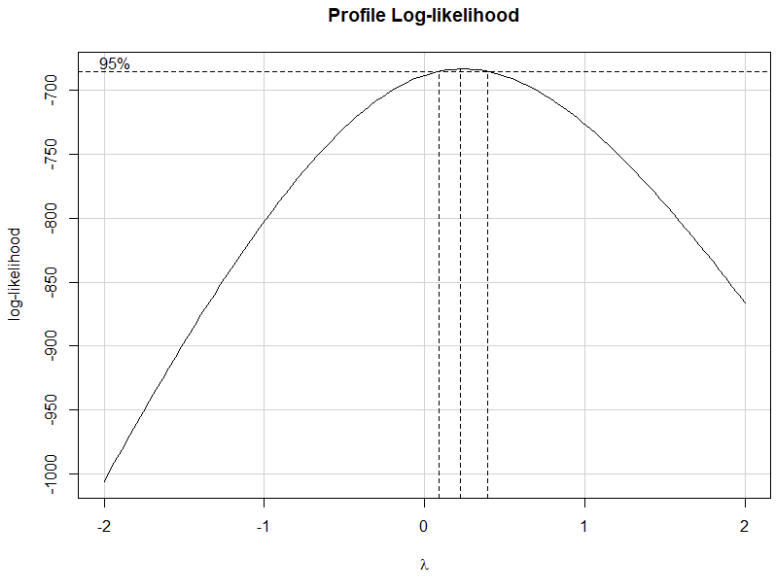
A Shapiro-Wilk normality test was also used to further assess Normality with `shapiro.test(full_model$residuals)`. The result of this test was a p-value of 0.0004. With this p-value being significantly low, we can conclude that Normality is violated at a significance level of $\alpha = 0.05$. Homoscedasticity was also found to be violated using a plot of Jackknife Residuals against Fitted Values with `plot(full_model$fitted.values, studres(full_model))`.



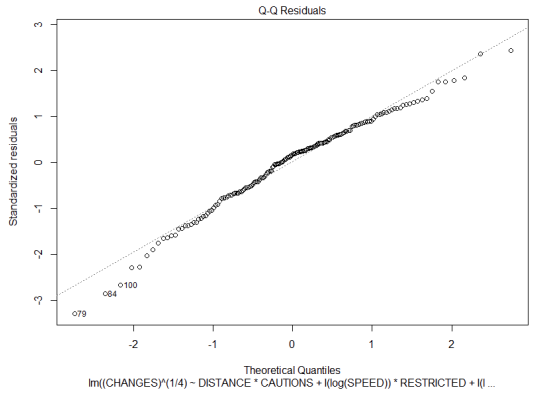

The plot appears to form a fan shape, indicating a violation of Homoscedasticity. This assumption was further assessed using the Breusch Pagan Test for Heteroskedasticity with `ols_test_breusch_pagan(full_model)`. With a p-value of $3.468 \times 10^{-21}$, we can assume that the Homoscedasticity assumption was violated at a significance level of $\alpha = 0.05$. Besides these, no other assumptions seem to be violated.

To fix violations of Normality and Homoscedasticity, the model was first subjected to a square root transformation on CHANGES. After this transformation, the Shapiro-Wilk normality test was performed and resulted in a p-value of 0.4957, showing that Normality was no longer

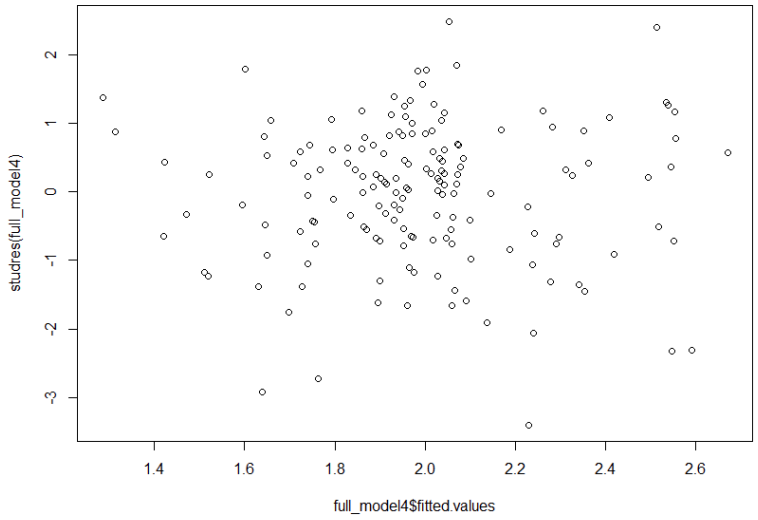
violated. However, using another Breusch Pagan Test for Heteroskedasticity, Homoscedasticity was still violated with a p-value of $3.500 \times 10^{-6}$. A cubic root transformation was also applied but ended with similar failed results. To determine a better power transformation to fix these violations, a Box-Cox plot was formed using `boxCox(full_model)`.




Looking at the Box-Cox plot, a fourth root transformation is suggested. With a fourth root transformation on CHANGES, the Normality and Homoscedasticity assumptions were fixed according to the Q-Q Residuals plot, Shapiro-Wilk normality test, Jackknife Residuals vs Fitted Values plot, and Breusch Pagan heteroskedasticity test results below.



Shapiro-Wilk normality test: $p - value = 0.0523 > 0.05$




Breusch Pagan test for heteroskedasticity: $p - value = 0.0605 > 0.05$

After applying the fourth root transformation to the response, the full model became

$$\sqrt[4]{CHANGES} = \beta_0 + \beta_1 DISTANCE \cdot CAUTIONS + \beta_2 ln(SPEED) \cdot RESTRICTED + \beta_3 ln(SPEED) + \beta_4 DISTANCE + \beta_5 CAUTIONS + \beta_6 SPEED + \beta_7 RESTRICTED + \beta_8 PLAYOFF + \beta_9 NEWGEN + E$$

and was built using

```
full_model <- lm((CHANGES)^(1 / 4) ~ DISTANCE * CAUTIONS +
I(log(SPEED)) * RESTRICTED + I(log(SPEED)) + DISTANCE + CAUTIONS
+ SPEED + RESTRICTED + PLAYOFF + NEWGEN, data = races).
```

**Final Model Selection and Diagnostics**

A backward selection algorithm with a significance cutoff of $\alpha = 0.10$ was used to determine an initial reduced model by running `ols_step_backward_p(full_model, p_val = 0.10, details = TRUE)` on the maximum model (predictors with a utility $p - value > 0.10$ when added last to the model are removed). This algorithm removed the

CAUTION, PLAYOFF, and NEWGEN main effects, as well as the interactions between DISTANCE and CAUTIONS and $ln(SPEED)$ and RESTRICTED, leaving the final model to be $\sqrt[4]{CHANGES} = \beta_0 + \beta_1 ln(SPEED) + \beta_2 DISTANCE + \beta_3 SPEED + \beta_4 RESTRICTED + E$. After this model was built with `final_model <- lm((CHANGES)^(1/4) ~ I(log(SPEED)) + DISTANCE + SPEED + RESTRICTED, data = races)`, collinearity was checked by running `ols_coll_diag(final_model)` to display a Variance Inflation Factor (VIF) table. Because the VIF values of $ln(SPEED)$ and SPEED are greater than 10, there is a potential collinearity issue between the two, which logically makes sense as one is a transformation of the other.

| | Variables | Tolerance | VIF |
|---|---|---|---|
| 1 | I(log(SPEED)) | 0.05301122 | 18.863932 |
| 2 | DISTANCE | 0.60832485 | 1.643859 |
| 3 | SPEED | 0.05003593 | 19.985639 |
| 4 | RESTRICTED | 0.58339274 | 1.714111 |

To combat this issue that could cause model instability, the SPEED data was standardized with `races$SPEED <- ((SPEED - mean(SPEED)) / sd(SPEED)) + (min(SPEED) + 1)` and the same model was rebuilt. Displaying the VIF table again shows that $ln(SPEED)$ and SPEED are still collinear with each other.

| | Variables | Tolerance | VIF |
|---|---|---|---|
| 1 | I(log(SPEED)) | 0.0006315078 | 1583.511692 |
| 2 | DISTANCE | 0.5221085687 | 1.915310 |
| 3 | SPEED | 0.0006336539 | 1578.148612 |
| 4 | RESTRICTED | 0.5597377368 | 1.786551 |

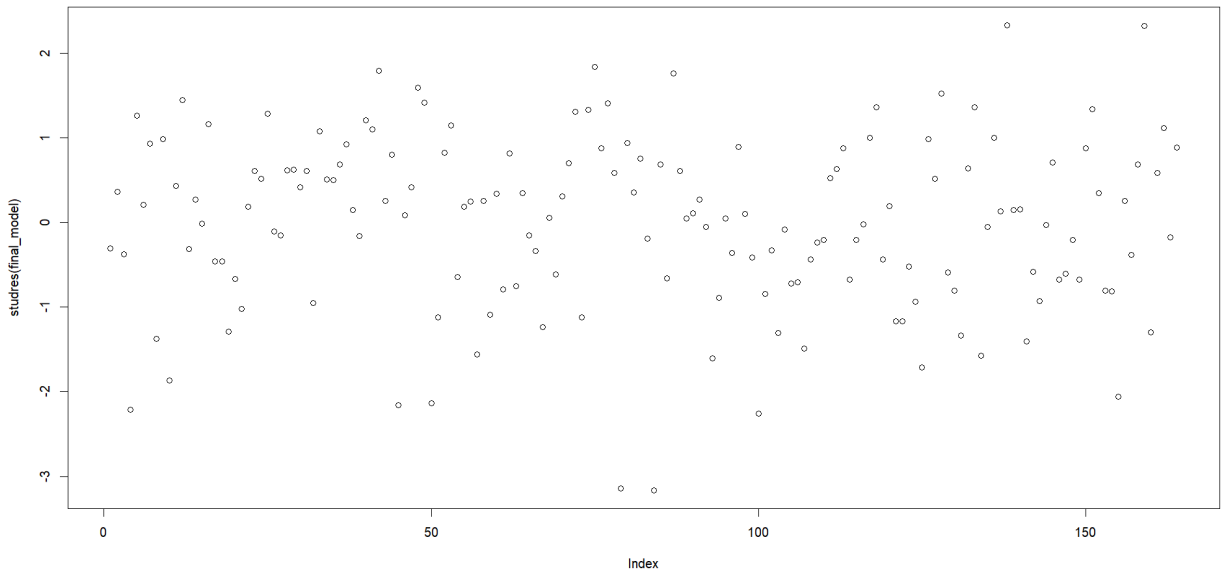
Because standardizing the data did not fix the collinearity issue, the $ln(SPEED)$ predictor was removed from the model as it has the highest VIF value, making the new model

$\sqrt[4]{CHANGES} = \beta_0 + \beta_1 DISTANCE + \beta_2 SPEED + \beta_3 RESTRICTED + E$. Reverting the standardization of SPEED and building this new model by running `races <- read.csv("races.csv")` and `final_model <- lm((CHANGES)^(1/4) ~ DISTANCE + SPEED + RESTRICTED, data = races)` reveals no more collinearity issues after rechecking the VIF table as no predictors have a value above 10.

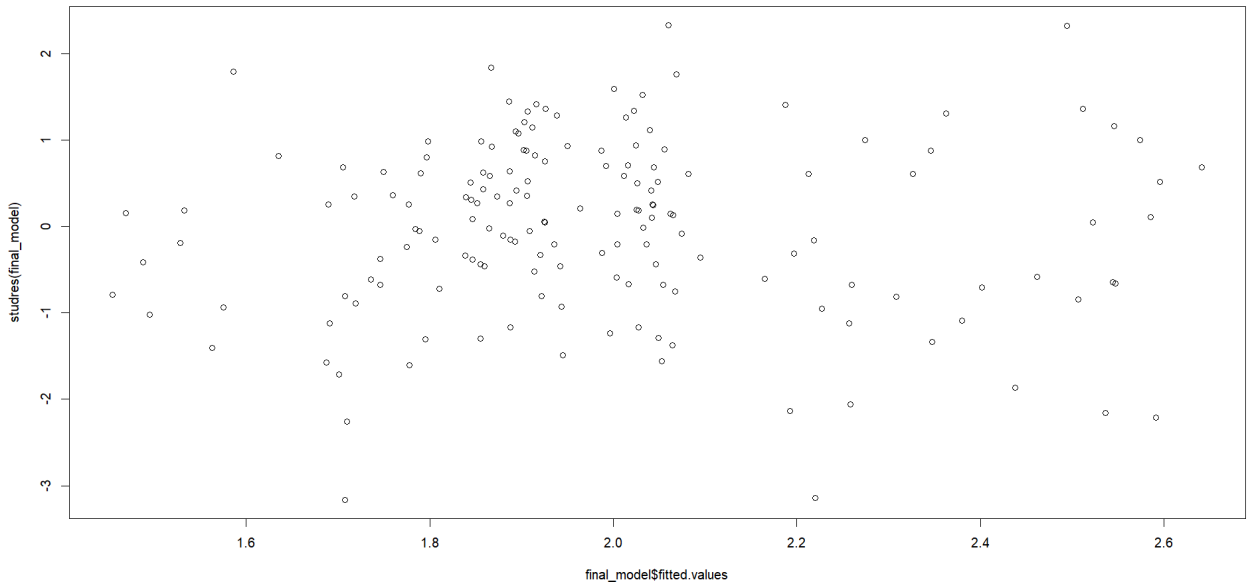
| | Variables | Tolerance | VIF |
|---|---|---|---|
| 1 | DISTANCE | 0.6623441 | 1.509789 |
| 2 | SPEED | 0.4622339 | 2.163407 |
| 3 | RESTRICTED | 0.6432333 | 1.554646 |

After combatting collinearity, several outlier detection methods were applied to this new model to determine whether any observations needed to be removed. Observations 4 and 84 were detected as outliers both when using a Leverage cutoff of $\frac{2(k+1)}{n} = \frac{2(3+1)}{164} = 0.0488$ and a Jackknife Residual cutoff of $t_{0.975}^{df=n-k-2=164-3-2=159} = 1.975$ and running `tail(sort(hatvalues(final_model)), n = 20)` alongside `tail(sort(abs(studres(final_model))), n = 20)`. While the recorded values of these potential outliers lie well in the realm of plausibility, it is important to consider whether the inclusion of these observations drastically affects the construction of the linear model. Running `tail(sort(cooks.distance(final_model)), n = 20)` reveals that observation 4 has a Cook's Distance value of 0.070 and observation 84 has a value of 0.269, both of which fall well under the Cook's Distance cutoff of 1. As a result, these observations (alongside the rest as none exceed the Cook's Distance cutoff of 1) will be kept as they are plausible and do not have much impact on the formation of the model.

After accounting for collinearity and potential outliers, further model diagnostics relating to multiple linear regression assumptions can be assessed. As stated in the Exploratory Data Analysis section, the data that the model is built on appears to abide by the Existence assumption when examining its considerable non-zero variance and the model's non-zero β parameters found with `summary(final_model)` reaffirm this idea. Additionally, the model does not appear to violate the Independence assumption as a Jackknife Residuals vs Observation Order plot created by `plot(studres(final_model), type = "p")` reveals that there is no clear trend in the concentration of the points over time.

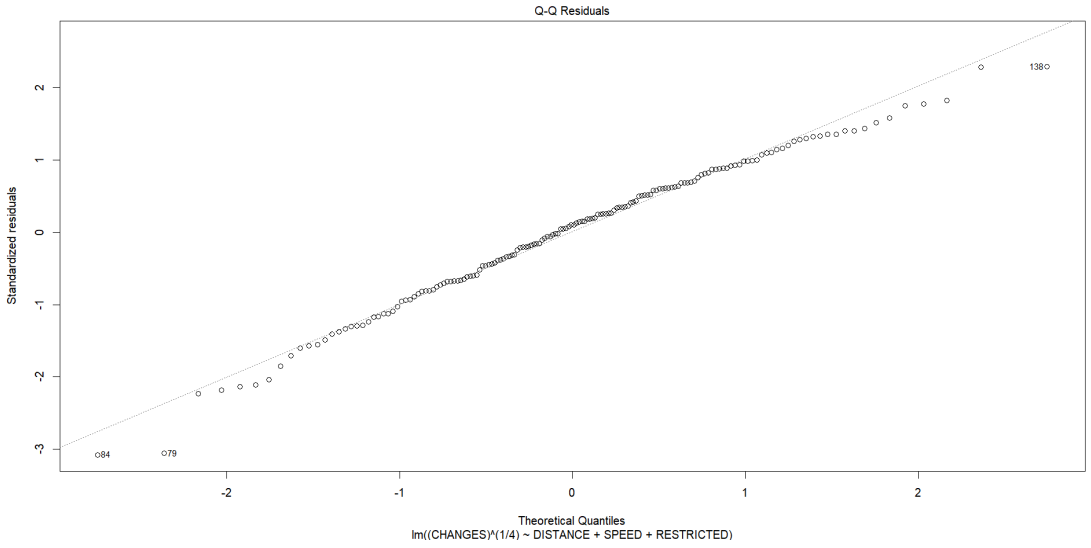



Likewise, Linearity does not appear to be violated when examining the Jackknife Residuals vs Fitted Values plot shown below with `plot(final_model$fitted.values, studres(final_model))` as the residuals remain centered around 0 with no clear curve across the predicted values. Using the same graph, Homoscedasticity also does not appear to be violated as the vertical width of the residuals remains roughly constant across the chart, with no clear conic shape.




Checking a Q-Q Residuals plot with `plot(final_model)` reveals the Normality assumption to not be violated either as the residuals closely hug the ideal normal distribution line, albeit with some deviation at more extreme z-values. A Shapiro-Wilk normality hypothesis test using `shapiro.test(final_model$residuals)` results in a p-value of 0.180, thus failing to indicate that the residuals are not normally distributed at the $\alpha = 0.05$ significance level. Therefore, after successfully completing diagnostics, the final sample model takes the form

$$\widehat{\sqrt[4]{CHANGES}} = 1.5573 + 0.0021 \cdot DISTANCE - 0.0027 \cdot SPEED + 0.3445 \cdot RESTRICTED.$$



**Evaluation of Model Reliability**

This model has an $r^2$ value of 0.668, indicating that about 66.8% of the variation in the fourth root of the number of lead changes is explained by the model. Because the standard deviation $s = 0.178$, approximately 95% of the observed values for the fourth root of lead changes fall within $2s = 0.356$ of their predicted values. The mean absolute error (MAE) of the model is 0.141, meaning that, on average, the model's predictions for the fourth root of lead changes differ from their observed values by 0.141. Considering these statistics and the range of observed response values being $[\sqrt[4]{2},\ \sqrt[4]{70}] = [1.189, 2.893]$, the model appears to fit the training data well. However, because a training-testing split was not performed, the potential for overfitting cannot be ruled out.

**Results, Summary, and Interpretations**

In summary, $\sqrt[4]{CHANGES} = \beta_0 + \beta_1 DISTANCE + \beta_2 SPEED + \beta_3 RESTRICTED + E$ ends up as the final population model with a fitted sample model of

$$\widehat{\sqrt[4]{CHANGES}} = 1.5573 + 0.0021 \cdot DISTANCE - 0.0027 \cdot SPEED + 0.3445 \cdot RESTRICTED.$$

This model provides useful interpretations which can be applied to future race design. As distance increases, the number of lead changes is predicted to increase slightly. As the average speed of the cars in a race increases, the number of lead changes decreases slightly. If the top speed is limited via a restrictor plate in the engine, the number of lead changes is predicted to significantly increase. All of these interpretations require assuming that only one variable changes at a time, and that the rest are held constant.

**Conclusion and Limitations**

The final model shows that race distance, average speed, and the usage of restrictor plates are all correlated with the number of lead changes in a race. The relationships between each of

these predictors and the response demonstrate routes NASCAR, as well as other motorsports sanctioning bodies, can take in generating more lead changes. Because race distance and restrictor plates are positively correlated with lead changes, more long-distance superspeedway types of tracks can be introduced into the schedule to take advantage of the pack-style racing and consistent position swapping upfront. To prevent the oversaturation of the race calendar with superspeedways while still not sacrificing lead changes, NASCAR can replace some of the medium-sized tracks with shorter and slower ones that provide more opportunities for drivers to use their skills to take the lead in the tighter corners as average race speed has been shown to be negatively correlated with lead changes. Interestingly, the implementation of the Next Gen car was shown to not be correlated with lead changes (after accounting for other variables), indicating that NASCAR's designers and engineers may need to further modify the design to improve overtaking opportunities. NASCAR's organizing body may also have some work to do in changing the playoff rules to increase the stakes for winning, as playoff status was also not correlated with more lead changes. Likewise, the caution system may need to be overhauled to generate more lead changes, albeit while still keeping safety in mind.

While certain obstacles were overcome and provided valuable learning opportunities, such as parsing, cleaning, and extracting data into a usable format, and carrying out a complete regression analysis from start to finish utilizing a statistics-focused programming language, other challenges proved to be much more inherently difficult to deal with. For example, a typical pitfall of most statistical analyses is the possibility of false negatives (Type II Errors) and false positives (Type I Errors). For instance, playoff status may actually correlate with lead changes, but this relationship may have gotten masked or ignored somewhere along the dataset and model selection processes. On the other hand, while race distance was shown to be "statistically"

significant in predicting lead changes, in practice it may have a minimal impact on affecting it and this avenue may not be worth pursuing. Regardless, correlation (or the lack thereof) does not necessarily indicate causation (or the lack thereof), so the previously mentioned actionable recommendations should be taken with a grain of salt. Another key limitation of this study was the small sample size of races. For example, two years' worth of racing with the Next Gen car may not be enough to assess whether the updated design had an impact on the number of lead changes. Additionally, the small sample size prevented the use of a training-testing data split, preventing a cross-validation assessment of whether the final model was overfitting the training data. Another limitation concerns the relatively few number of predictors as other variables that were not included in the dataset may provide a stronger correlation with lead changes.

Future research can expand and improve on this study by collecting race data from more years, using additional sources or personal observations of races to gather more predictors, exploring other interactions, transformations, and responses, focusing on different racing series, and utilizing training-testing data splits to better evaluate the model's predictive power and employ another regression method if necessary. The world of motorsports is an exciting and profitable one, offering rich opportunities for further research endeavors to maximize its entertainment value.

***We, the project team members, certify that below is an accurate account of the percentage of effort contributed by each team member in the project and report.***

| Project Team Member | Percentage of Total Effort |
|---|---|
| Farrukh Anwar | 33.33 |
| Connor Horan | 33.33 |
| Elishel Sason | 33.33 |

Appendix

```
import requests
import csv
# Fetch race data
races = []
race_seasons = [2020, 2021, 2022, 2023]
for race_season in race_seasons:
    races +=
requests.get(f"https://cf.nascar.com/cacher/{race_season}/1/race
_list_basic.json").json()
# Extract relevant race data and write into file
with open("races.csv", "w", newline = "") as races_csv:
    writer = csv.writer(races_csv)
    headers = ["CHANGES", "DISTANCE", "CAUTIONS", "SPEED",
"RESTRICTED", "PLAYOFF", "NEWGEN"]
    writer.writerow(headers)
    for race in races:
        CHANGES = race["number_of_lead_changes"]
        DISTANCE = float(race["actual_distance"])
        CAUTIONS = int(race["number_of_cautions"])
        SPEED = float(race["average_speed"])
        RESTRICTED = int(race["restrictor_plate"])'
        PLAYOFF = 0
        NEWGEN = 0
        if race["playoff_round"] > 0:
            PLAYOFF = 1
        if race["race_season"] == 2023 or race["race_season"] ==
2022:
            NEWGEN = 1
        row = [CHANGES, DISTANCE, CAUTIONS, SPEED, RESTRICTED,
PLAYOFF, NEWGEN]
        writer.writerow(row)
```

# Works Cited

Taranto, Steven. “NASCAR Next Gen car: Explaining the ins and outs of the NASCAR Cup Series' new racecar for 2022.” CBS Sports, 27 Feb. 2022, www.cbssports.com/nascar/news/nascar-next-gen-car-explaining-the-ins-and-outs-of-the-nascar-cup-series-new-racecar-for-2022/.