What is the key to win a Formula One Grand Prix

Caleb, Jac, Colby

Dataset

Kaggle Dataset

- Real Formula One data
- Length, Turns, Temperature, Birth Year, and Team Value variables added manually

225 Rows & 18 columns

- Each observation is a winner of a Formula One Grand Prix

Summary Statistics:

Fastest Lap Speed (km/h)

Minimum	151.388
Q1	198.046
Median	209.915
Mean	210.137
Q3	224.752
Max	252.794

Wins from 2014-2024

Lewis Hamilton	82
Max Verstappen	62
Nico Rosberg	20
Sebastian Vettel	14
Valtteri Bottas	10
Daniel Ricciardo	8

Time in Minutes

Minimum	73.686
Q1	87.776
Median	93.715
Mean	97.151
Q3	100.432
Max	181.733

First Research Question:

Can we predict the fastest speed when finishing their fastest lap?

1st Model – Extra Variables

avg_annual_temp – The average annual temperature for each country where a race is held

Description: df [6 × 2]

	circuit_name <chr>	avg_annual_temp <dbl>
1	70th Anniversary ...	48.63
2	Abu Dhabi Grand ...	82.71
3	Australian Grand ...	71.69
4	Austrian Grand Prix	45.39
5	Azerbaijan Grand ...	55.33
6	Bahrain Grand Prix	81.84

6 rows

birth_year - The year of birth for each driver who wins a grand prix

Description: df [6 × 2]

	driver_ref <chr>	birth_year <dbl>
1	bottas	1989
2	gasly	1996
3	hamilton	1985
4	leclerc	1997
5	max_verstappen	1997
6	norris	1999

6 rows

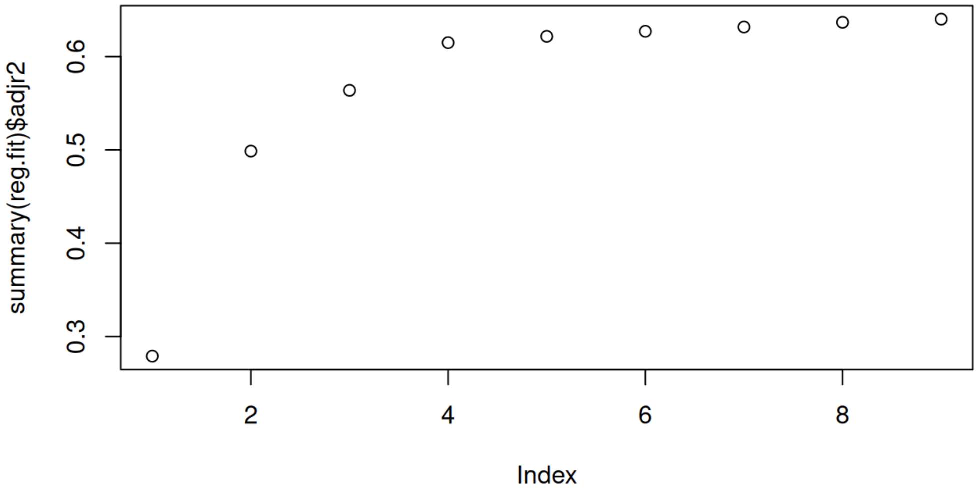
team_value - The value of each team that wins a grand prix in billions of dollars

Description: df [4 × 2]

	team_name <chr>	team_value <dbl>
1	Ferrari	6.50
2	McLaren	4.40
3	Mercedes	6.00
4	Red Bull	4.35

4 rows

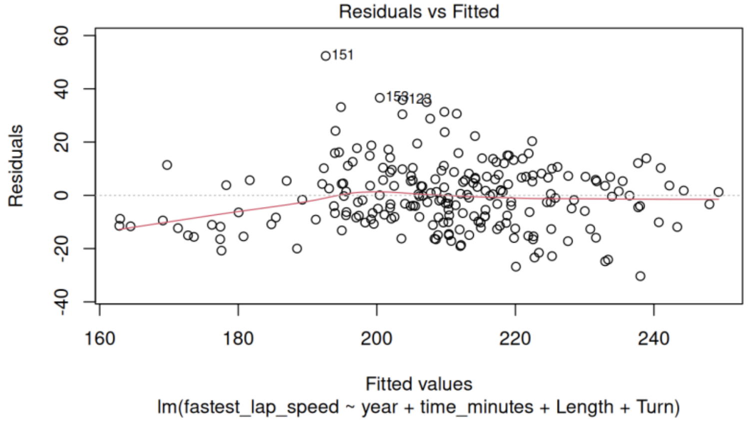
1st Model – Finding Significant Predictors

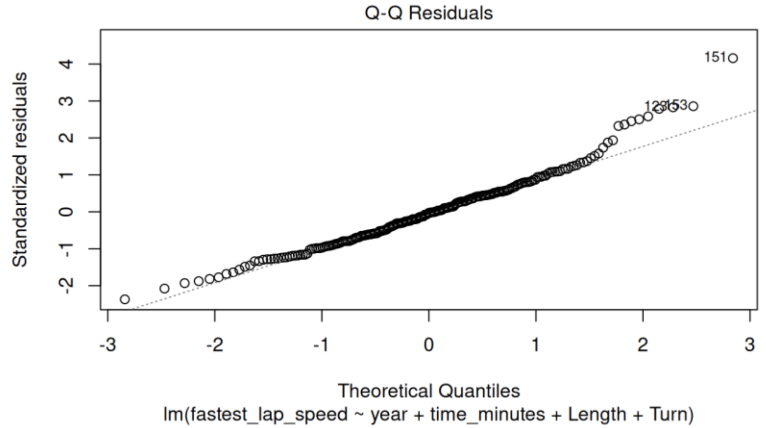



```
Call:
lm(formula = fastest_lap_speed ~ year + time_minutes + Length +
    Turn, data = f1_uid3)

Residuals:
    Min      1Q  Median      3Q     Max
-30.336  -8.947  -0.383   7.008  52.334

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.074e+03  5.508e+02  -5.581 7.10e-08 ***
year          1.623e+00  2.727e-01   5.949 1.07e-08 ***
time_minutes -3.084e-01  5.637e-02  -5.471 1.23e-07 ***
Length        1.572e+01  1.153e+00  13.639  < 2e-16 ***
Turn         -2.621e+00  2.809e-01  -9.329  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.95 on 217 degrees of freedom
Multiple R-squared:  0.6219,    Adjusted R-squared:  0.6149
F-statistic: 89.23 on 4 and 217 DF,  p-value: < 2.2e-16
```

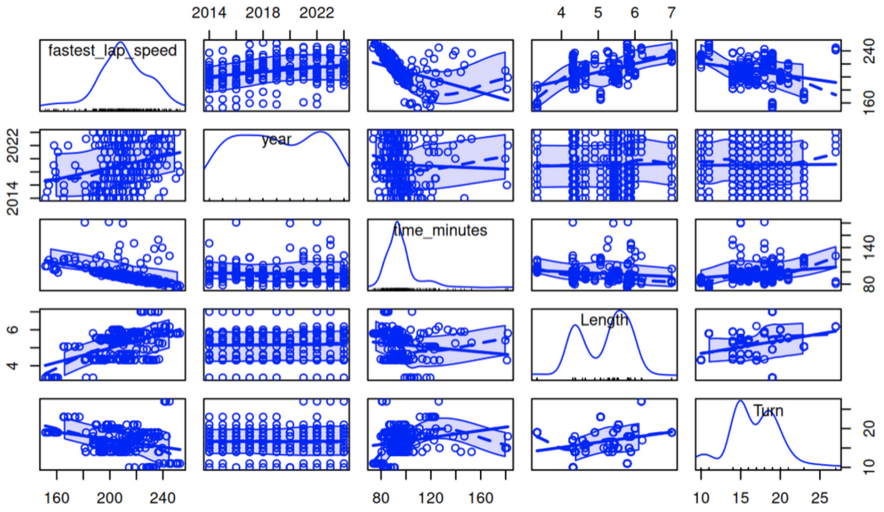
1st Model – Assumptions

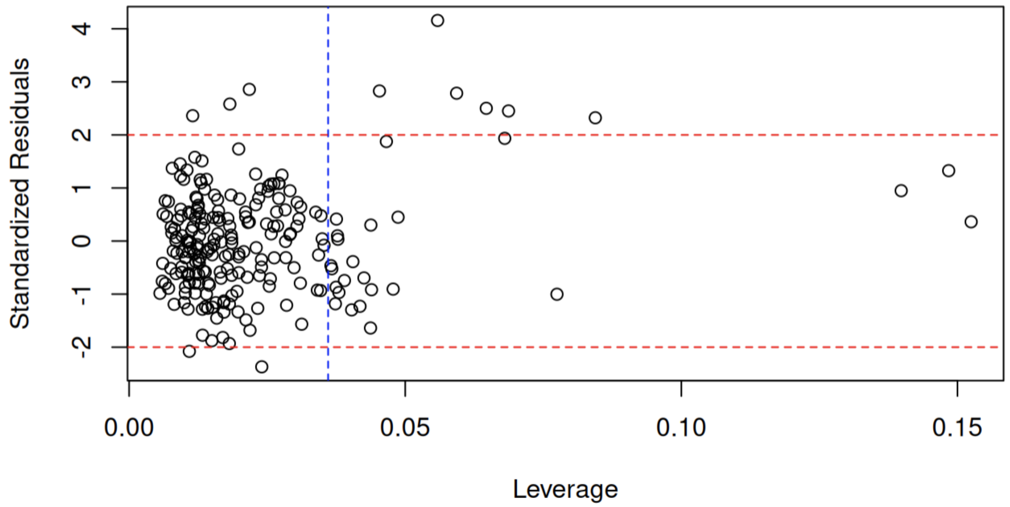
```
        Shapiro-Wilk normality test

data:  reduced.fit$residuals
W = 0.97401, p-value = 0.0004109

Non-constant Variance Score Test
Variance formula: ~fitted.values
Chisquare = 1.80369, Df = 1, p = 0.17927
```



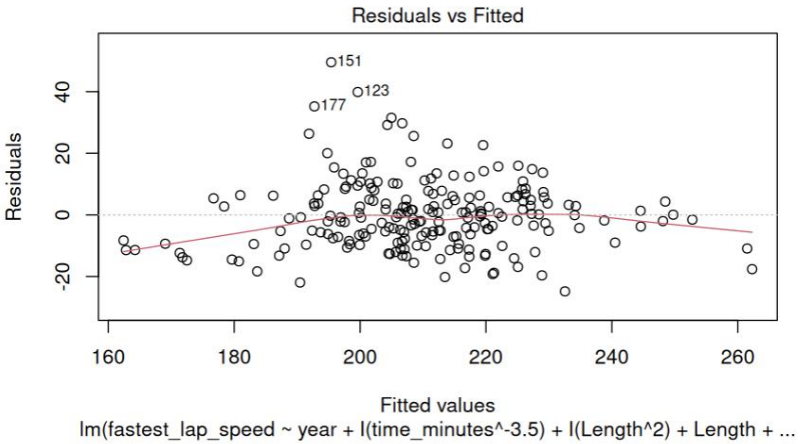


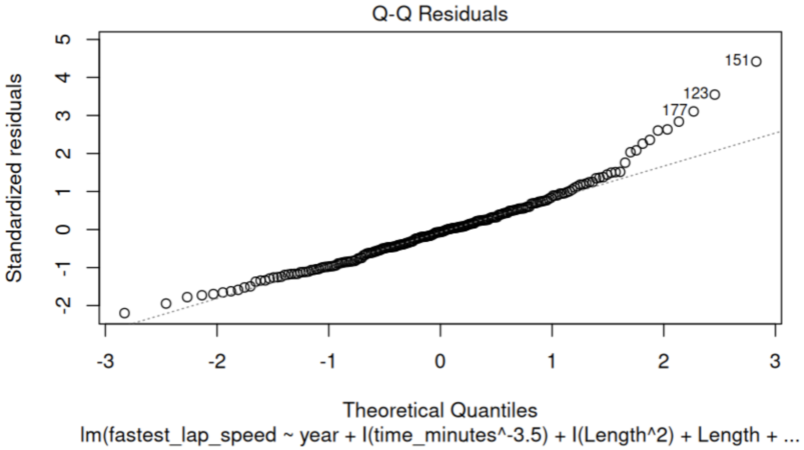
1st Model – Collinearity & Outliers





1st Model - Transformations



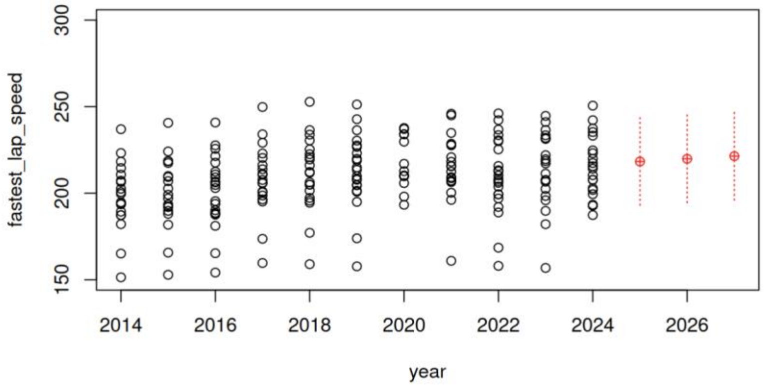



```
bcPower Transformations to Multinormality
             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
year            2.9992        1.00     -93.6200      99.6183
time_minutes   -3.4810       -3.48      -4.2142      -2.7478
Length          1.6993        2.00       1.0245       2.3741
Turn            0.3967        0.00      -0.0862       0.8796

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    1.0459           1       0.1439       1.9478
```

```
Call:
lm(formula = fastest_lap_speed ~ year + I(time_minutes^-3.5) +
    I(Length^2) + Length + log(Turn), data = f1_uid5)

Residuals:
    Min      1Q  Median      3Q     Max
-24.838  -7.488  -0.765   5.790  49.553

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.699e+03  5.024e+02  -5.371 2.08e-07 ***
year                  1.421e+00  2.491e-01   5.706 3.94e-08 ***
I(time_minutes^-3.5)  1.652e+08  1.876e+07   8.807 5.09e-16 ***
I(Length^2)          -1.471e+00  9.549e-01  -1.541  0.12494
Length                2.737e+01  9.703e+00   2.821  0.00525 **
log(Turn)            -3.001e+01  4.593e+00  -6.535 4.82e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.46 on 208 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.6801
F-statistic: 91.56 on 5 and 208 DF,  p-value: < 2.2e-16
```

1st Model - Prediction

```
# Prediction Plot

new_x <- data.frame(year=2027, time_minutes=97.290245, Length=5.166419, Turn=16.697674)
fastest_lap_speed_2027 <- predict(backtransformed.fit, newdata = new_x, interval =
"prediction")
plot(fastest_lap_speed ~ year, data = f1_uid6, xlim=c(2014, 2027), ylim=c(150, 300))
points(2025, fastest_lap_speed_2025[1,1], col="red", pch=10)
segments(x0 = 2025, y0 = fastest_lap_speed_2025[1,2], x1 = 2025, y1 =
fastest_lap_speed_2025[1,3], col = "red", lty = "dotted")
points(2026, fastest_lap_speed_2026[1,1], col="red", pch=10)
segments(x0 = 2026, y0 = fastest_lap_speed_2026[1,2], x1 = 2026, y1 =
fastest_lap_speed_2026[1,3], col = "red", lty = "dotted")
points(2027, fastest_lap_speed_2027[1,1], col="red", pch=10)
segments(x0 = 2027, y0 = fastest_lap_speed_2027[1,2], x1 = 2027, y1 =
fastest_lap_speed_2027[1,3], col = "red", lty = "dotted")

# Prediction Intervals
print(fastest_lap_speed_2025)
print(fastest_lap_speed_2026)
print(fastest_lap_speed_2027)
```



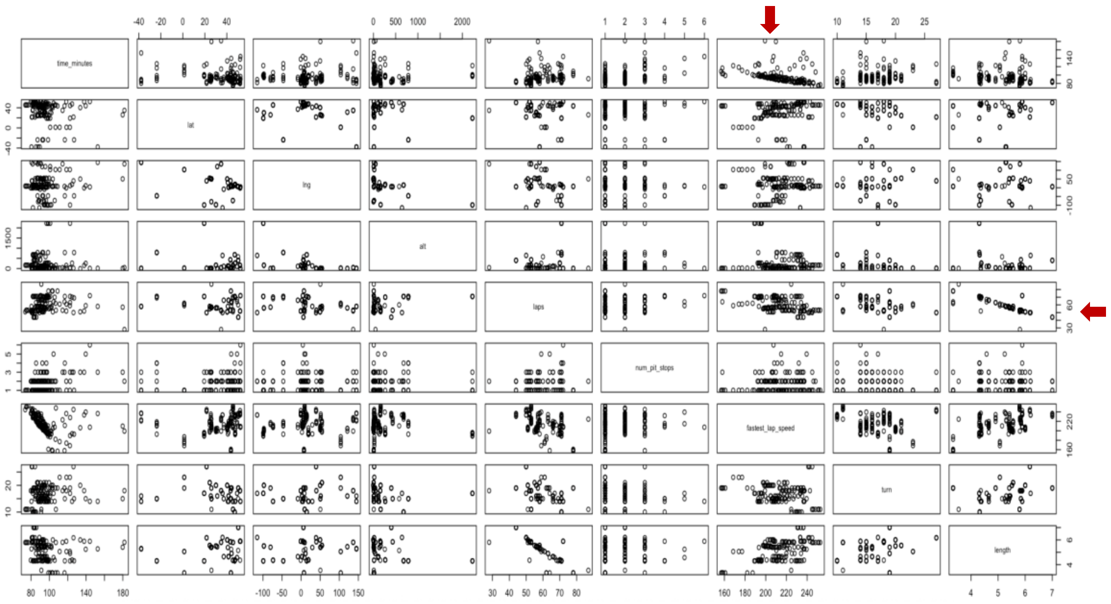

```
       fit      lwr      upr
1 218.3001 193.0065 243.5937
       fit      lwr      upr
1 219.8589 194.4931 245.2247
       fit      lwr      upr
1 221.4177 195.9687 246.8667
```

Second Research Question:

Can we predict the time it takes to win a Grand Prix

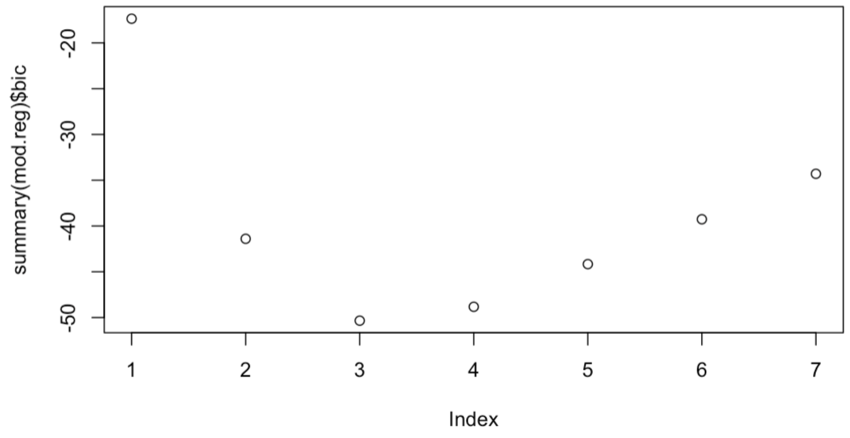
After the addition of the halo on Formula One racecars (2018-2024)

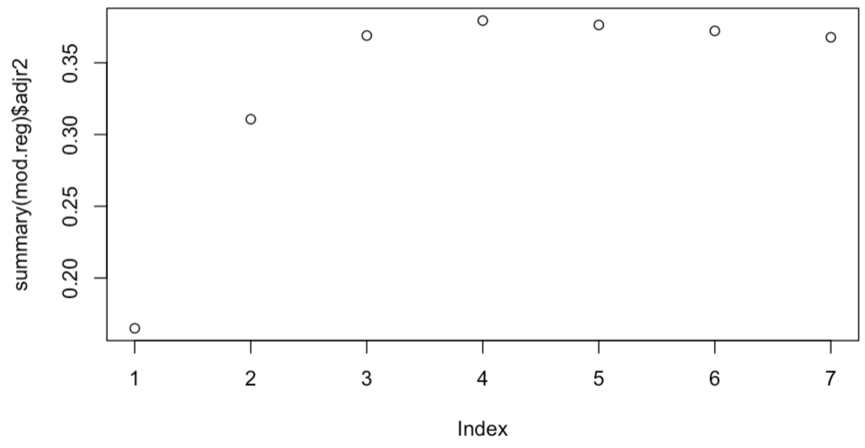
Visualizing data:



- Possible multicollinearity issues between `length` & `laps`
- Strong linear relationship between `time_minutes` & `fastest_lap_speed`

Variable Selection for Model with `regsubsets()`:





Significant predictors:

- `laps`: number of laps during grand prix
- `num_pit_stops`: Number of pit stops taken during the grand prix
- `fastest_lap_speed`: fastest speed during fastest lap (in kilometers per hour)

Testing for Multicollinearity with `vif()`:

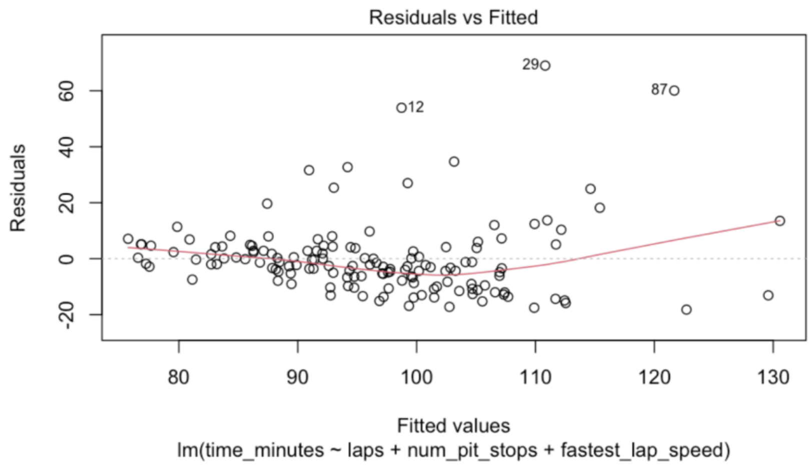
`laps`: 1.33

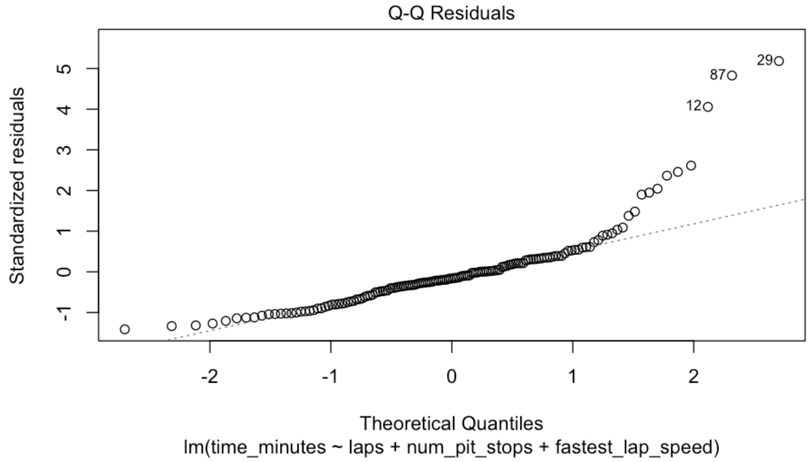
`num_pit_stops`: 1.06

`Fastest_lap_speed`: 1.30

Assumptions before Transformation:

- **L**inearity: No obvious pattern in fitted values v. residuals plot
- **I**ndependence: The observations are independent of each other
- **N**ormality: The residuals do not appear to be normally distributed (Shapiro-Wilk: 5.148e-12)
- **E**qual variance: The fitted values v. residuals plot shows datapoints are not randomly scattered around zero (ncvTest: 4.0825e-16)





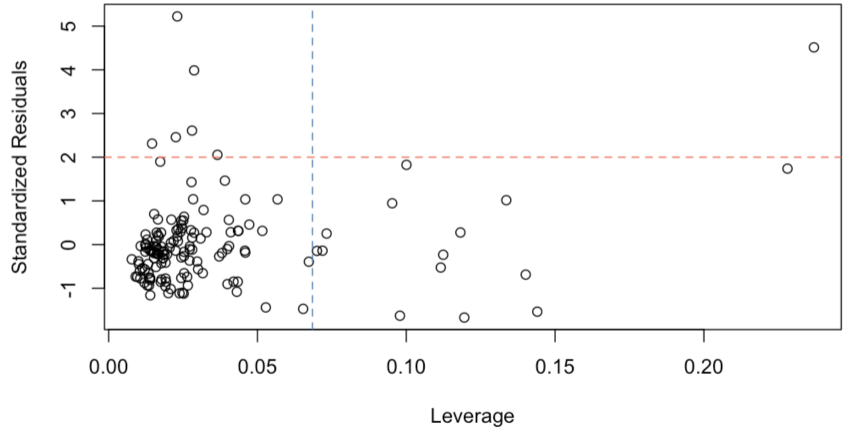
Outliers & bad leveraged datapoints

One bad leveraged datapoint:

- Max Verstappen: 2022 Japanese Grand Prix

Seven outliers:

- `time_minutes` over 120 minutes



Transformations using `powerTransform()`:

`time_minutes`

- Negative fourth power transformation

`laps`

- Log transformation

`num_pit_stops`

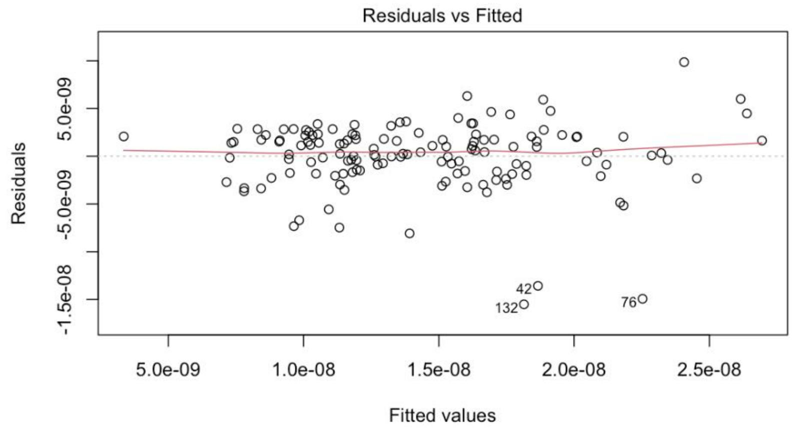
- Log transformation

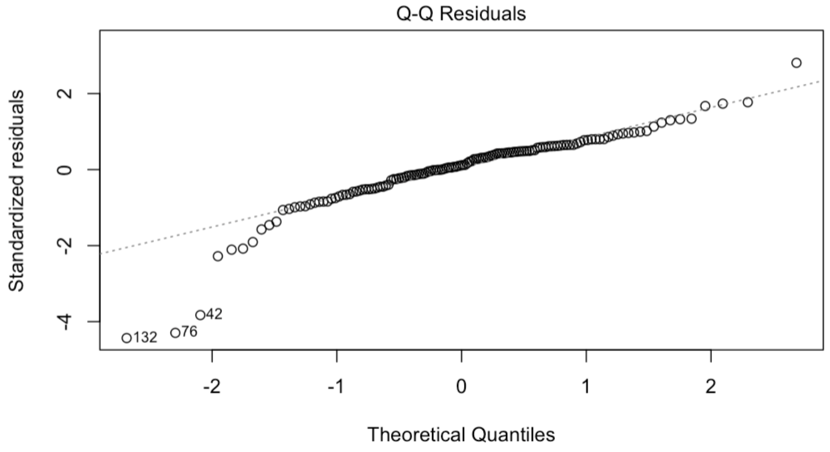
`fastest_lap_speed`

- Square transformation

Assumptions after Transformation:

- **L**inearity: No obvious pattern in fitted values v. residuals plot
- **I**ndependence: The observations are independent of each other
- **N**ormality: The residuals appear to be mostly normally distributed
- **E**qual variance: The fitted values v. residuals plot shows datapoints are mostly scattered randomly around zero





Final Model & Predictions

$$\widehat{\text{time\_minutes}} = 1.67e^{-9}\log(\text{laps}) + 4.78e^{-9}\log(\text{num\_pit\_stops}) + 1.82e^{-12}(\text{fastest\_lap\_speed})^2 - 5.30e^{-10}(\text{fastest\_lap\_speed})$$

R^2_{adj}:

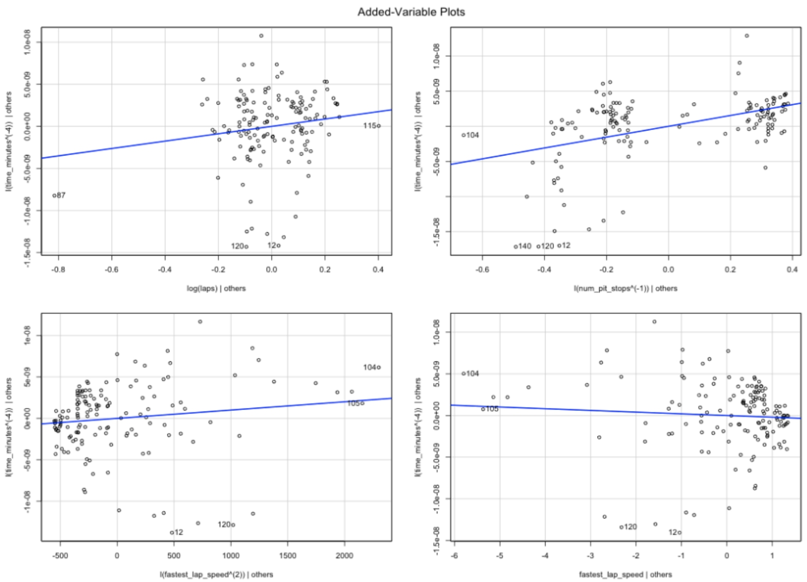
- 0.667 or 66.7%

Prediction:

- `laps`: 55
- `num_pit_stops`: 2
- `fastest_lap_speed`: 202

Result:

- 100.1285 minutes



QUESTIONS?