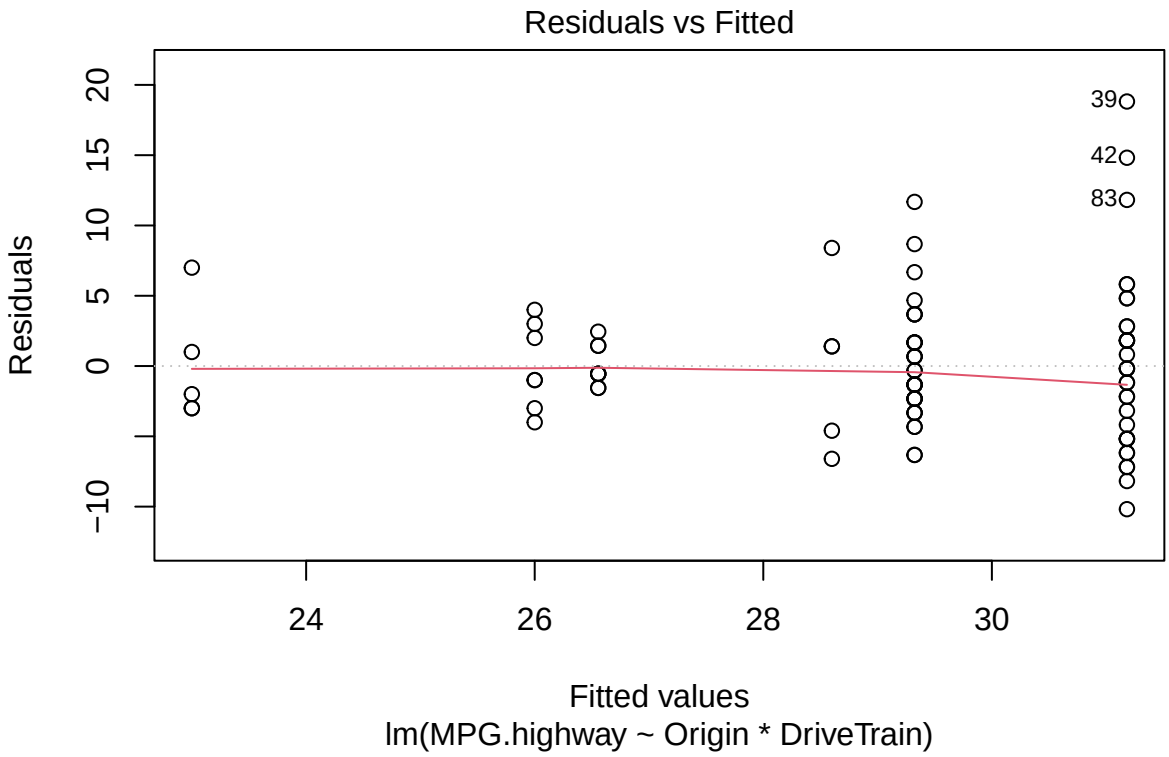
Michael_Ghattas_W3

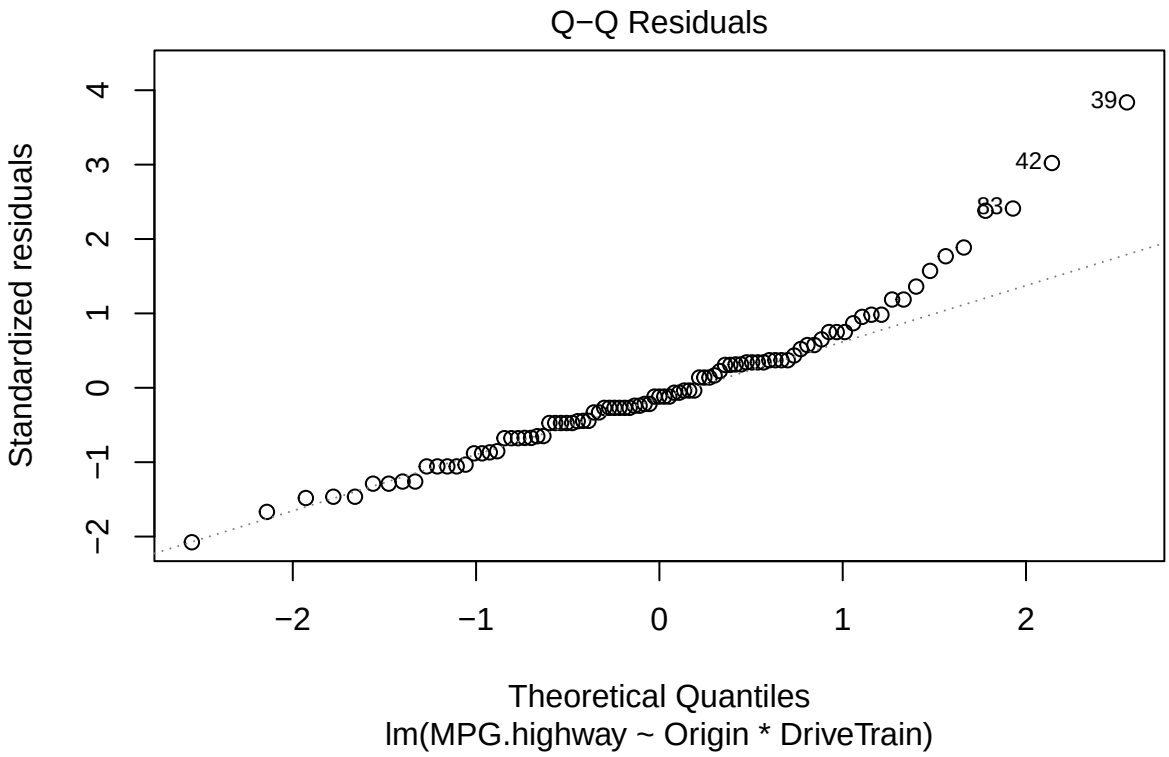
Michael Ghattas

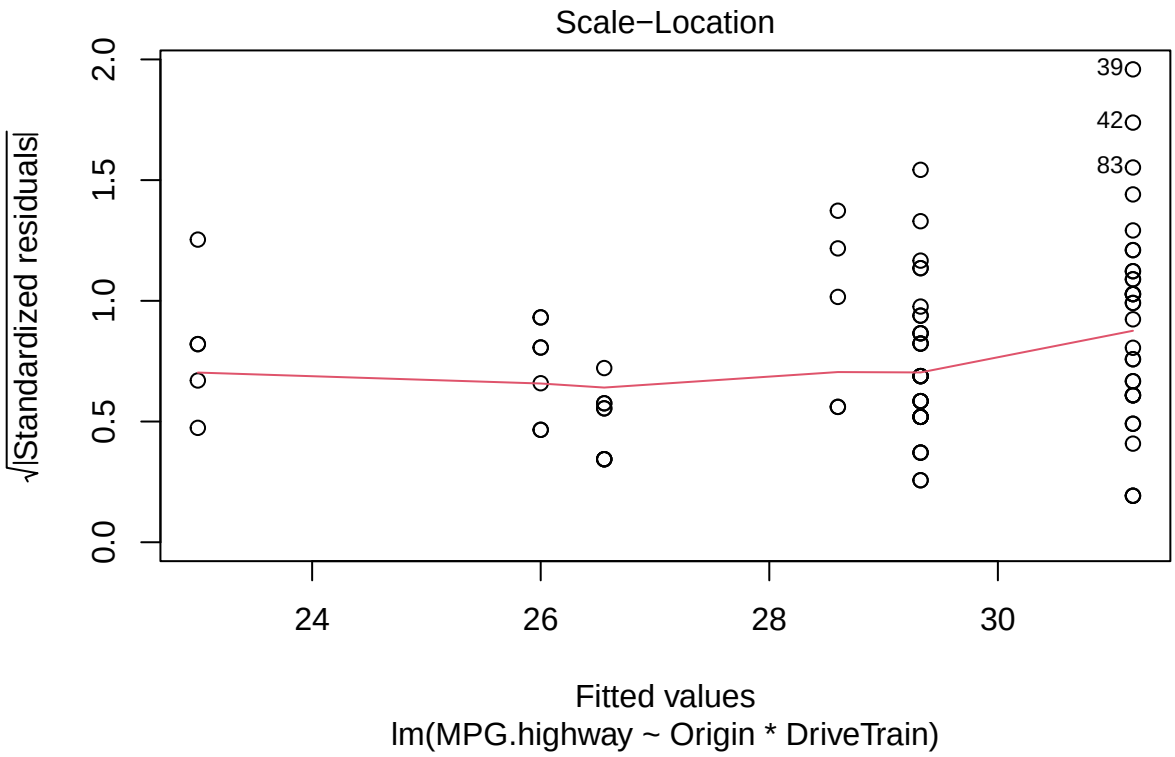
2024-09-28

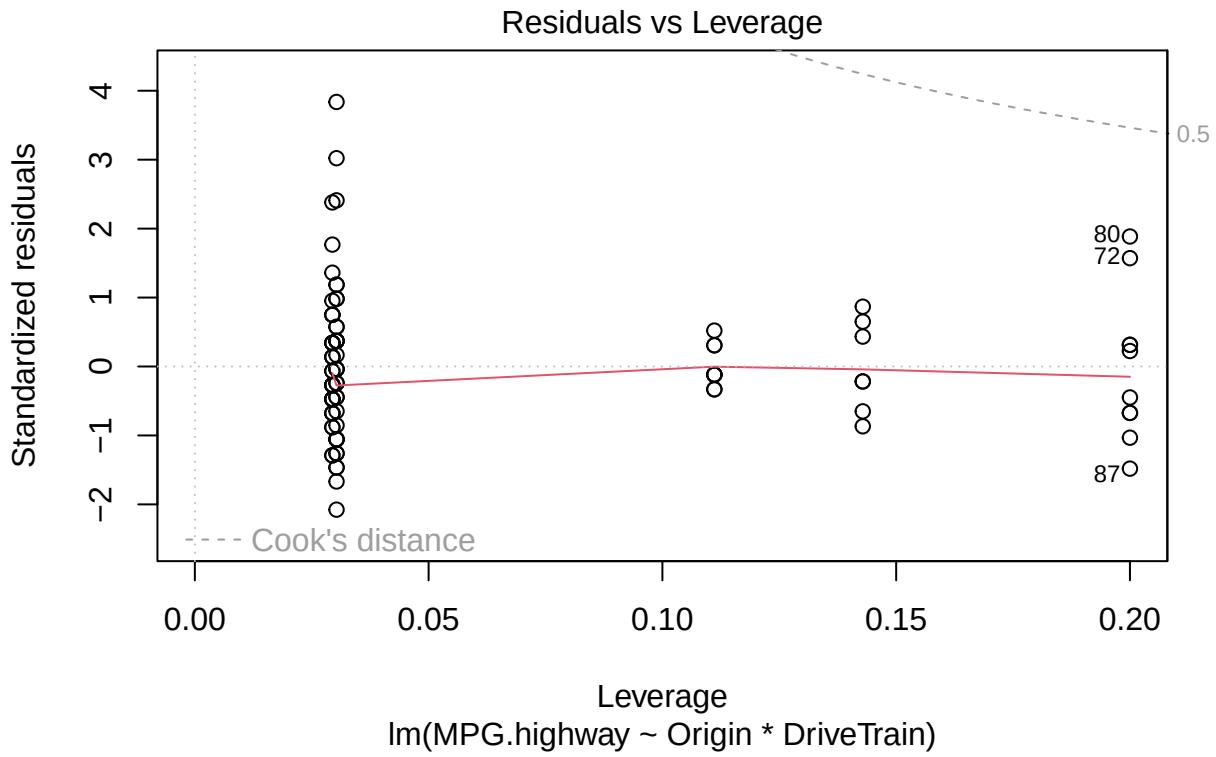
Problem 1: Does Origin and DriveTrain affect highway MPG?

```
# Linear model
hmodel1 <- lm(MPG.highway ~ Origin * DriveTrain, data = Cars93)
plot(hmodel1)
```




Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
3
4
39
42
83
−2
−1
0
1
2
Theoretical Quantiles
lm(MPG.highway ~ Origin * DriveTrain)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
24
26
28
30
39
42
83
Fitted values
lm(MPG.highway ~ Origin * DriveTrain)




```
summary(hmodel1)
```

```
##
## Call:
## lm(formula = MPG.highway ~ Origin * DriveTrain, data = Cars93)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1818  -3.0000  -0.5556   1.8182  18.8182
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     23.000      2.227  10.327  < 2e-16 ***
## Originnon-USA                    5.600      3.150   1.778  0.07891 .
## DriveTrainFront                  6.324      2.385   2.651  0.00953 **
## DriveTrainRear                   3.556      2.778   1.280  0.20395
## Originnon-USA:DriveTrainFront   -3.742      3.377  -1.108  0.27087
## Originnon-USA:DriveTrainRear    -6.156      4.027  -1.528  0.13003
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.98 on 87 degrees of freedom
## Multiple R-squared:  0.1749, Adjusted R-squared:  0.1275
## F-statistic:  3.69 on 5 and 87 DF,  p-value: 0.004476
```

```r
# ANOVA model
hmodel2 <- aov(MPG.highway ~ Origin * DriveTrain, data = Cars93)
summary(hmodel2)
```

```
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## Origin              1   87.7   87.69   3.536 0.06341 .
## DriveTrain          2  311.9  155.95   6.288 0.00281 **
## Origin:DriveTrain   2   57.9   28.97   1.168 0.31579
## Residuals          87 2157.8   24.80
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output Interpretation

Linear Model

- Intercept: The estimated intercept is 27, which represents the average MPG.highway for a car with the reference levels of the categorical variables (USA origin and 4-wheel drive).
- Originnon-USA: The estimated coefficient is 5.333, suggesting that non-USA cars tend to have approximately 5.33 higher MPG on highways than USA cars, although this is not statistically significant (p-value = 0.224).
- DriveTrainFront: The coefficient for front-wheel drive is 2.719, but it is also not statistically significant (p-value = 0.436).
- DriveTrainRear: The coefficient for rear-wheel drive is -0.25, but it is not statistically significant (p-value = 0.947).
- Interaction terms (Originnon-USA and Originnon-USA): These interaction terms are also not statistically significant (p-values > 0.05), meaning that the combined effect of origin and drivetrain doesn't significantly impact highway MPG.
- Residual Standard Error: 4.767, indicating the typical deviation of the observed highway MPG from the model's predicted values.
- R-squared (0.151): This suggests that the model explains only 15.1% of the variance in highway MPG, which is relatively low.

ANOVA Model

- Origin: The p-value for the effect of Origin is 0.0678, which is close to significant at the 5% level. This suggests that Origin might have a small effect on highway MPG, but it is not conclusively significant at the conventional threshold.
- DriveTrain: The p-value for DriveTrain is 0.0176, indicating that DriveTrain has a statistically significant effect on highway MPG.
- Interaction between Origin and DriveTrain: The interaction term has a p-value of 0.4617, which indicates that the interaction between Origin and DriveTrain is not significant. This means the combined effect of these variables does not meaningfully impact highway MPG.

Diagnostic Plots

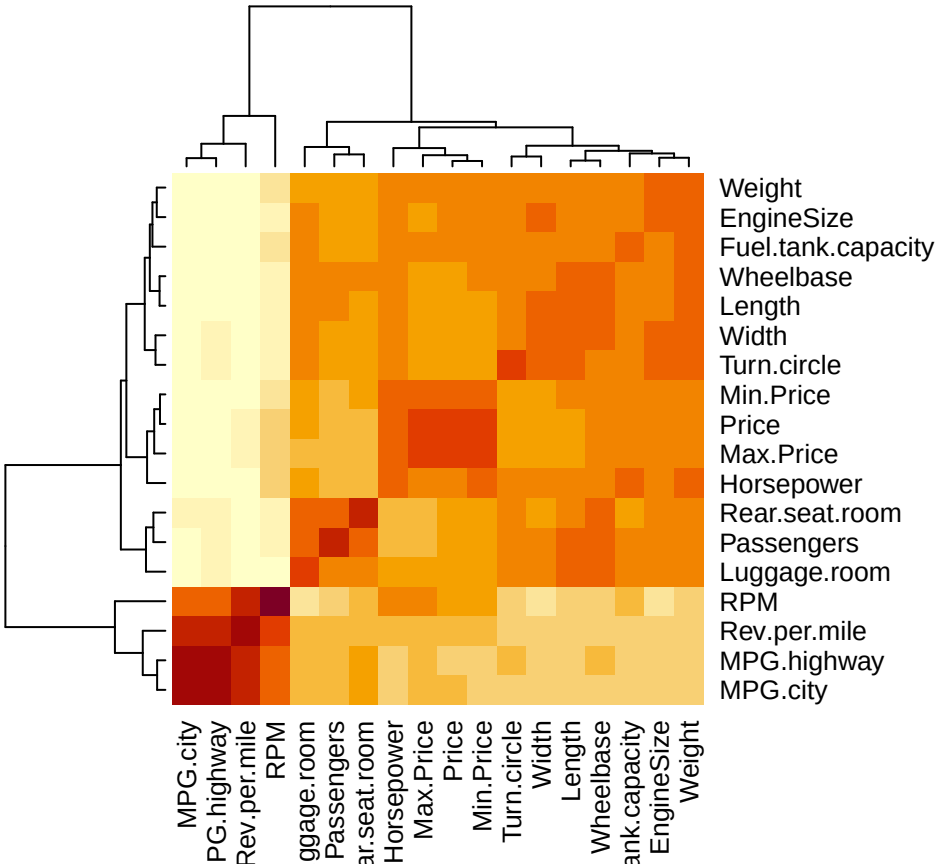
- Residuals vs Fitted: The plot shows a somewhat random spread around 0, which is a good sign. However, there may be some clusters or patterns indicating potential model improvement with other variables.
- Q-Q Plot: The points mostly follow the line, indicating that the residuals are approximately normally distributed, though there are some deviations at the ends (potential outliers).
- Scale-Location Plot: Shows relatively consistent variance across fitted values, though there is some slight increase in variance at higher fitted values.
- Residuals vs Leverage: No points have extremely high leverage, and there are no influential outliers based on Cook's distance.

Final Conclusion

- The DriveTrain variable has a significant effect on highway MPG, as shown by the ANOVA results (p-value = 0.0176). However, neither Origin nor the interaction between Origin and DriveTrain are significant at the 5% level.
- The R-squared value is relatively low (0.151), suggesting that the model does not explain much of the variability in highway MPG.
- Based on the linear model and ANOVA, the effect of Origin alone is marginally significant but not conclusive, and the interaction effect does not play a major role in determining highway MPG.

Problem 2: Use heat map of correlations to find input variables to model 'Price'

```r
# Heatmap visualization and correlation matrix
cor_data <- cor(Cars93[ , sapply(Cars93, is.numeric)], use = "complete.obs")
heatmap(cor_data)
```




```r
# Linear model using variables that are highly correlated with 'Price'
lm_model <- lm(Price ~ Horsepower + Fuel.tank.capacity + Weight, data = Cars93)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Price ~ Horsepower + Fuel.tank.capacity + Weight,
##     data = Cars93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.974  -2.977  -0.545   1.729  32.184
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.944658   3.538759  -1.680   0.0965 .
## Horsepower          0.125290   0.017888   7.004 4.54e-10 ***
## Fuel.tank.capacity  0.088717   0.429727   0.206   0.8369
## Weight              0.001938   0.002490   0.778   0.4384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.968 on 89 degrees of freedom
## Multiple R-squared:  0.6307, Adjusted R-squared:  0.6183
## F-statistic: 50.67 on 3 and 89 DF,  p-value: < 2.2e-16
```

Output Interpretation

Intercept: - The intercept is statistically significant (p-value < 0.05). This gives the estimated price when all other variables (Horsepower, Fuel tank capacity, and Weight) are zero, which is not directly interpretable but helps adjust the baseline of the model.

Horsepower: - Estimate: 0.1119. This means that for every 1 unit increase in horsepower, the price increases by approximately 0.1119 units, holding other variables constant. - p-value: 5.14e-05. Since this is less than 0.05, horsepower is statistically significant in predicting car price. This suggests that horsepower has a strong, positive effect on price.

Fuel.tank.capacity: - Estimate: 0.5388. For every 1 unit increase in fuel tank capacity, the price increases by approximately 0.5388 units. - p-value: 0.3027. This is greater than 0.05, so fuel tank capacity is not statistically significant in predicting price. This means its effect on price is not distinguishable from random noise at the 5% significance level.

Weight: - Estimate: 0.0017. For every 1 unit increase in weight, the price increases by approximately 0.0017 units. - p-value: 0.6001. This is also greater than 0.05, indicating that weight is not statistically significant in predicting price.

Final Conclusion: - The most significant predictor of Price in this model is Horsepower. - The variables Fuel.tank.capacity and Weight do not appear to be statistically significant in this model, as their p-values are greater than 0.05. - The model explains approximately 63.89% of the variability in the car prices (R-squared = 0.6389).

End.