Data606 Project

Vic Chan

Introduction

In this project I will be analyzing a data set called "Fuel Economy" from the United States Department of Energy with the oversight of the Environmental Protection Agency. This dataset contains every single make and model of a car from 1984 - 2021 and is constantly being updated with new makes and models. With this dataset we will be answer if "Fuel Economy(MPG) of gas engine cars have peaked throughout the year." With the increase in emission regulations, it has been increasingly more difficult for car companies to create more efficient cars in order to meet regulations. Auto makers have been trying to create lighter, more aerodynamic, and more efficient gas engines in order to meet the ever increasing stringent regulations. I would like to see if auto makers are at its apex in creating more efficient gas powered cars.

Data

In the dataset there is 43,541 different vehicles with 83 different attributes. This dataset includes gas, diesel, electric, hybrid, and hydrogen cars. For this project we will only be focusing on gas powered cars and seeing if the MPG has truly been increasing over the years and if auto makers are at its limits for the efficiency of gas powered cars. This dataset does not include any information on how many of the make and model are sold each year. We will also be only looking at specific columns that I choose as I believe they may be related to MPG

- city08 - City MPG
- highway08 - Highway MPG
- cylinders - Number of cylinders in the engine
- displ - Engine displacement in liters
- drive - Drive Axle Type (eg. 2-Wheel Drive, 4-Wheel Drive, All-Wheel Drive)
- fuelType - Type of Fuel (eg. Regular, Premium)
- make - Make of the car
- model - Model of the car
- year - Model year
- trany - Transmission type

This is a observational study as we are looking at the collected data and basing our hypothesis off the dataset.

Statistics

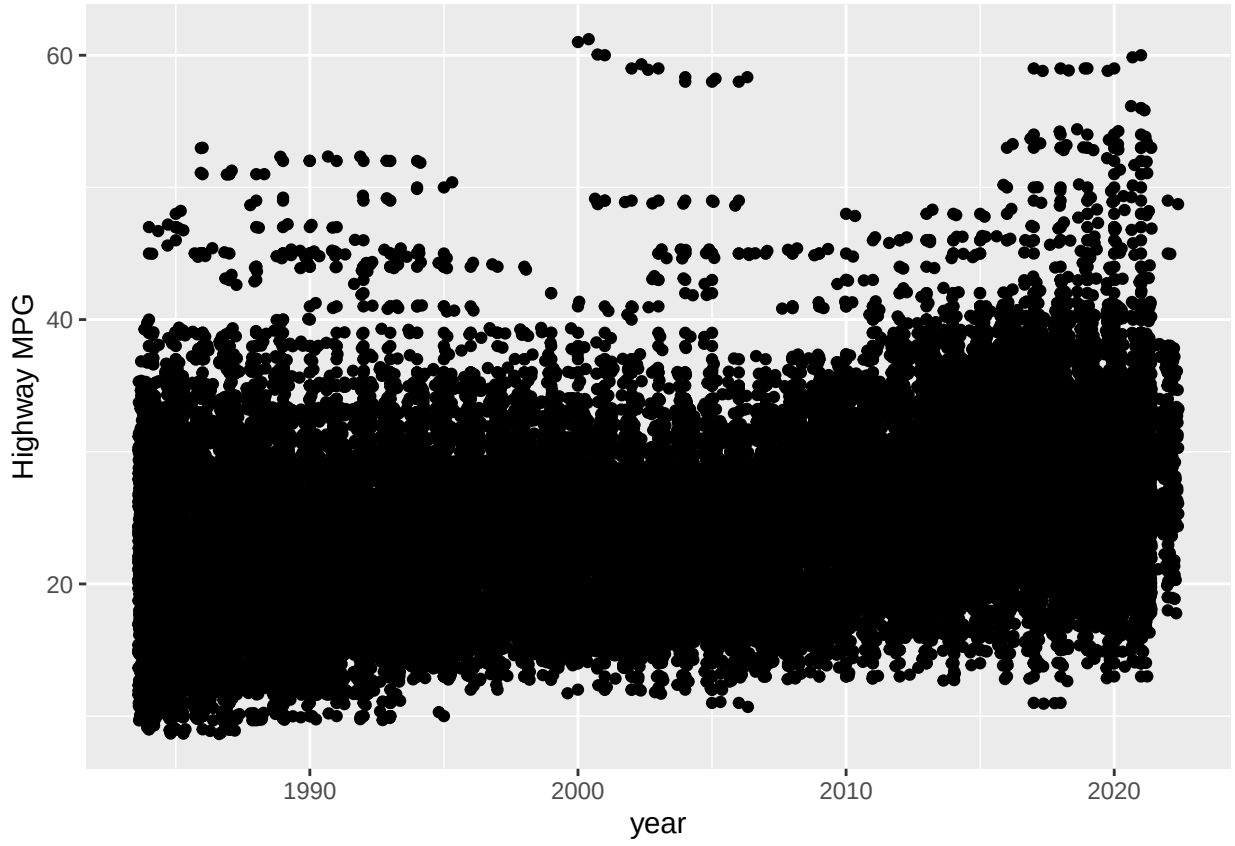
```
# load data
dataset = read.csv('https://www.fueleconomy.gov/feg/epadata/vehicles.csv')
```

```
# Filtering out all options that are unnecessary

dataset = dataset %>%
  filter(fuelType == 'Regular' |  fuelType == 'Premium' | fuelType == 'Midgrade') %>%
  select(city08, highway08, cylinders, displ, drive, fuelType, make, model, year, trany)
```

Just looking at the simple year and mpg chart we can see that there is a lot of data points and they are all clumped together. From a quick look you can see that there is a slight positive trend line, but we will be needing to fit a linear model to verify that. We can also see that in the 1980's the minimum fuel economy is lower compared to what we have today. This is most likely due to regulations created by the EPA.

```
dataset %>%
  ggplot(aes(x=year, y=highway08)) + geom_point() + geom_jitter() + ylab('Highway MPG')
```



Loess Regression

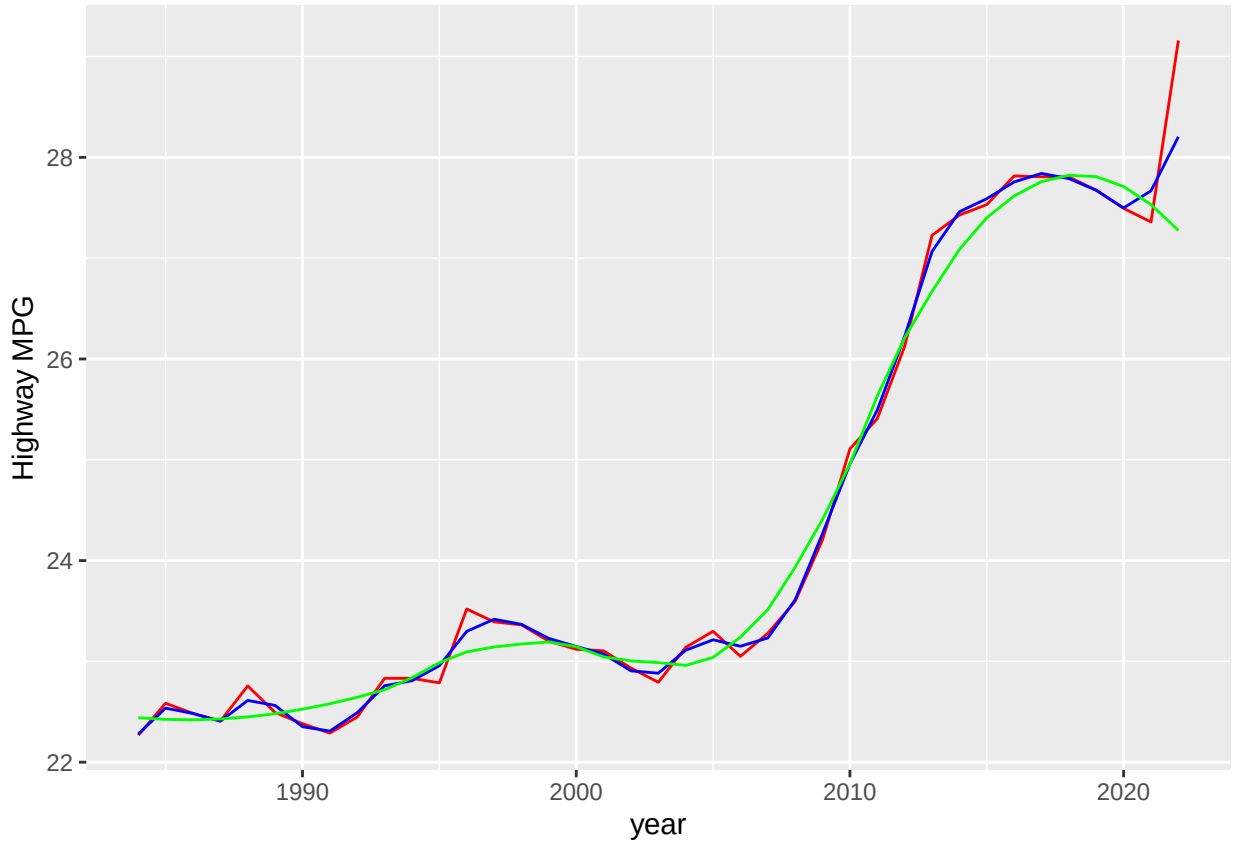
The Loess Regression is "a nonparametric technique that uses local weighted regression to fit a smooth curve through points in a scatter plot." This means that we are trying to fit smaller best fit lines together based on a specific span size, which are then put together to create the regression. Based on the different span sizes there will be different results. With a larger span size the Loess regression will use more data to create a estimate which will result in a less overfitting compared to use of a small span size. This is the reason why one needs to play around with the different parameters in order to get the best fit line for the Loess Regression while trying to minimize the Sum of Squared Errors(SSE).

Highway Fuel Efficency

We can see in the graph below the Loess Regression with different span parameters. We can see that the Green fitted line is the smoothest has as it has the highest span compared to the red line which is very jagged because of the low span. While each of the line are different we can see a general trend which is showing a increase in MPG from 2005 to 2016. This is due to to the financial crisis of 2008 when gas prices were very high. This meant that consumers were looking to buy more fuel efficient cars and car manufactures followed suit. But now we see that in the recent year the fuel economy has decreased due to the fact that consumers now want SUV. This is the reason why many car manufacturers are now no longer making sedans and only making SUV now.

```
m1 = loess(highway08 ~ year, data=dataset, span=0.05)
m2 = loess(highway08 ~ year, data=dataset, span=0.15)
m3 = loess(highway08 ~ year, data=dataset, span=0.50)
```

```
ggplot(data=dataset, aes(x=year, y=m1$fitted)) +
  geom_line(color='RED') + geom_line(aes(x=year, y=m2$fitted), color='BLUE') +
  geom_line(aes(x=year, y=m3$fitted), color='GREEN') + ylab('Highway MPG')
```

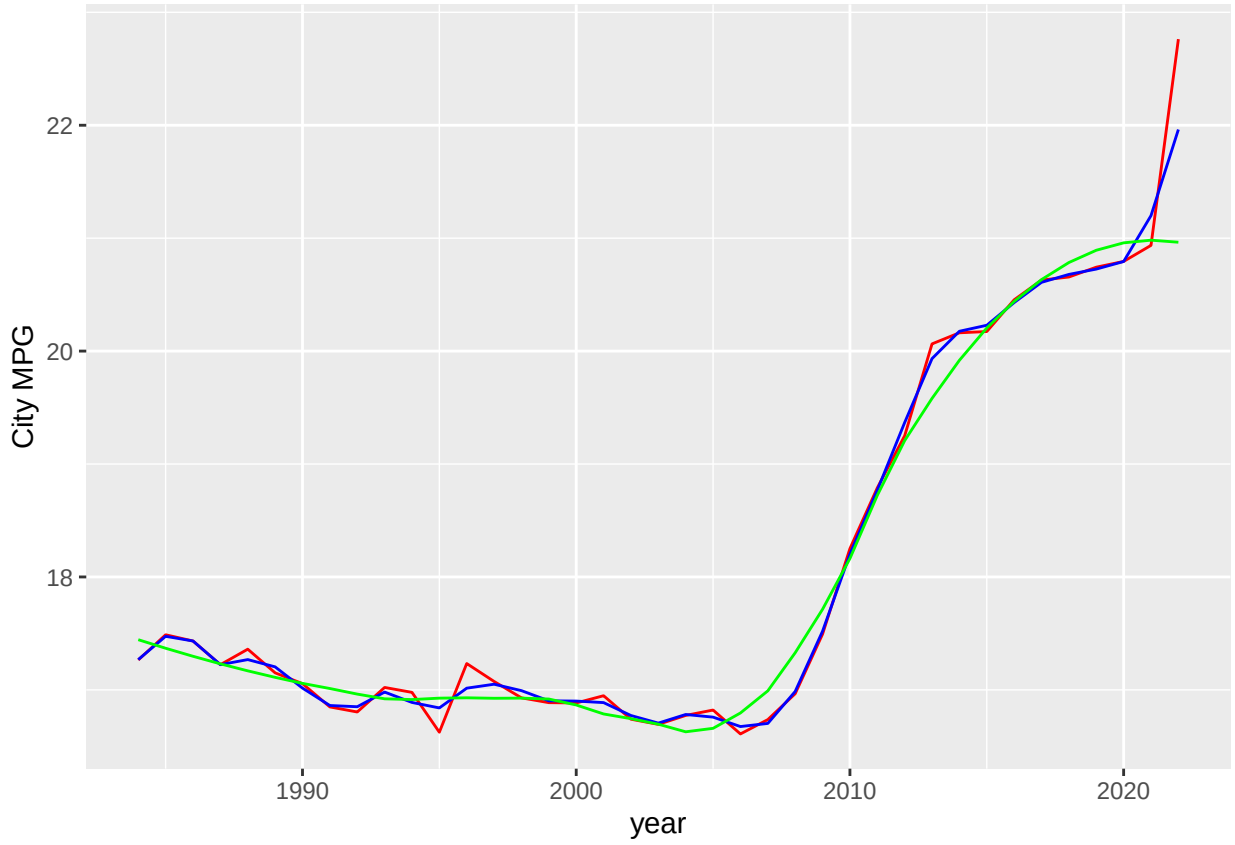


City Fuel Efficency

With the City Fuel Efficiency we can also see that there is a drastic increase from 2005 to 2016 because of the recession. We can also see that the city fuel efficiency follows the highway fuel efficiency curve very closely, especially the Loess Regression with a span of 0.50

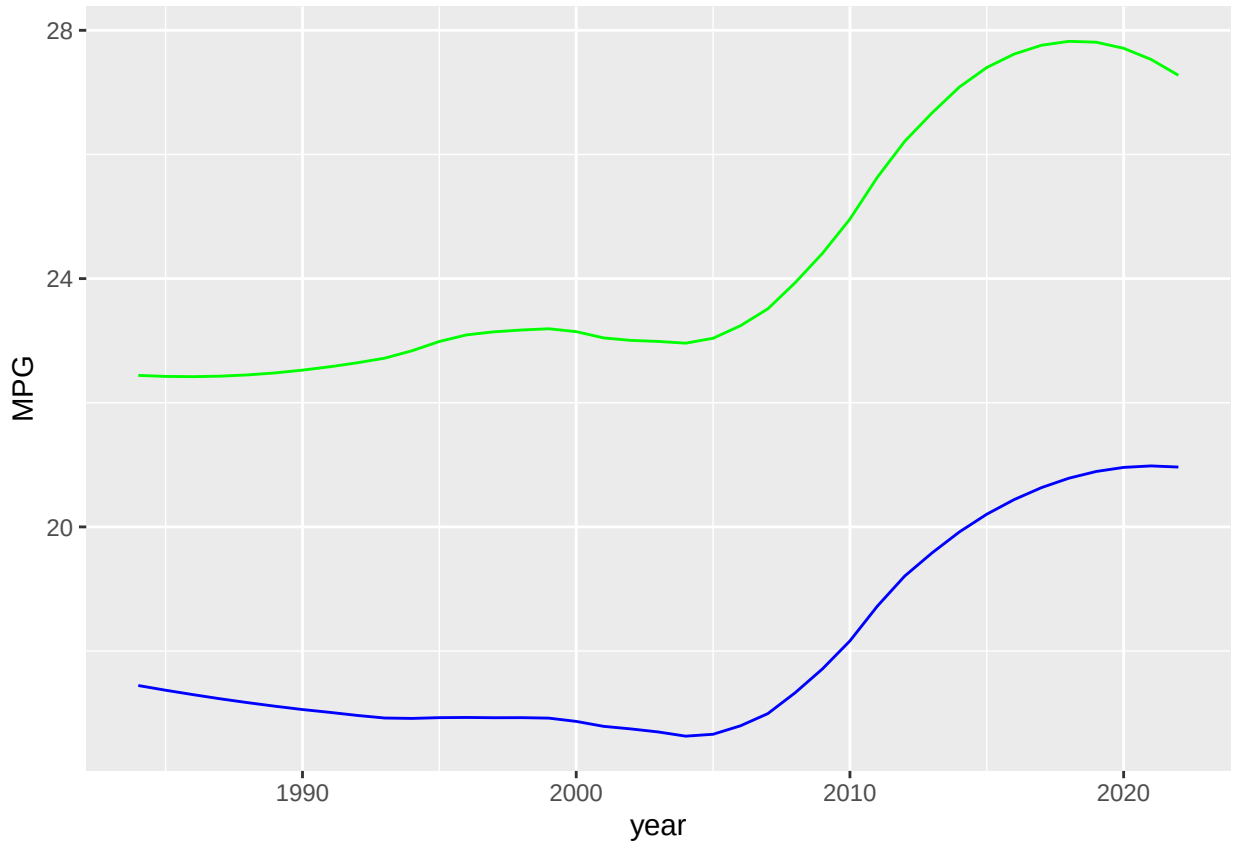
```
m4 = loess(city08 ~ year, data=dataset, span=0.05)
m5 = loess(city08 ~ year, data=dataset, span=0.15)
m6 = loess(city08 ~ year, data=dataset, span=0.50)
```

```
ggplot(data=dataset, aes(x=year, y=m4$fitted)) + geom_line(color='RED') +
  geom_line(aes(x=year, y=m5$fitted), color='BLUE') +
  geom_line(aes(x=year, y=m6$fitted), color='GREEN') + ylab('City MPG')
```



As you can see the city and highway fuel efficiency pattern follows very closely with each other. Both have a increase in fuel efficiency during the recession with the fuel economy peaking in recent years.

```
ggplot(data=dataset, aes(x=year)) +
  geom_line(aes(y=m3$fitted), color='GREEN') +
  geom_line(aes(y=m6$fitted), color='BLUE') + ylab('MPG')
```



The can definitely see that the fuel efficiency has peaked in the recent years, but this may not be due the increase difficulty in developing more efficient cars. Instead this is most likely due to the recent change in consumer demand for SUV instead of the more fuel efficient sedans. With this new demand by the customers, car manufactures are now no longer making sedans and is instead making more SUV. This will be interesting to see if this trend will go up and down with the COVID recession.

Finding The Optimal Span Which Minimizes Sum of Squared Errors (SSE)

One can just randomly keep creating different span numbers until they get the lowest SSE or we can write a function which can calculate all the different SSE for all the different span. This will allow us to find the most accurate fitting of the curve. As the span changes the curves smoothing the error also changes which is why we would like to minimize the error as much as possible. Based on intuition though a lower span number will have a lower SSE because of the closer fit to the curve.

```
calcSSE = function(x)
{
  m = loess(highway08 ~ year, data=dataset, span = x)
  return(sum(m$residuals^2))
}
```

```
optimal_span = function(span, lower, upper, n)
{
  message(span, ":", lower, ":", upper, ":", n)

  lower_bound = (span - lower) / 2
```

```
  upper_bound = (upper - span) / 2

  lower_SSE = calcSSE(lower_bound)
  upper_SSE = calcSSE(upper_bound)

  if (n == 0)
  {
    if(lower_SSE <= upper_SSE)
    {
      return(lower_bound)
    }
    else
    {
      return(upper_bound)
    }
  }
  else
  {
    if(lower_SSE <= upper_SSE)
    {
      return(optimal_span(lower_bound, lower, span, n-1))
    }
    else
    {
      return(optimal_span(upper_bound, span, upper, n-1))
    }
  }
}
```

```
optimal_span(0.5, 0, 1, 2)
```

```
## 0.5:0:1:2
```

```
## 0.25:0:0.5:1
```

```
## 0.125:0:0.25:0
```

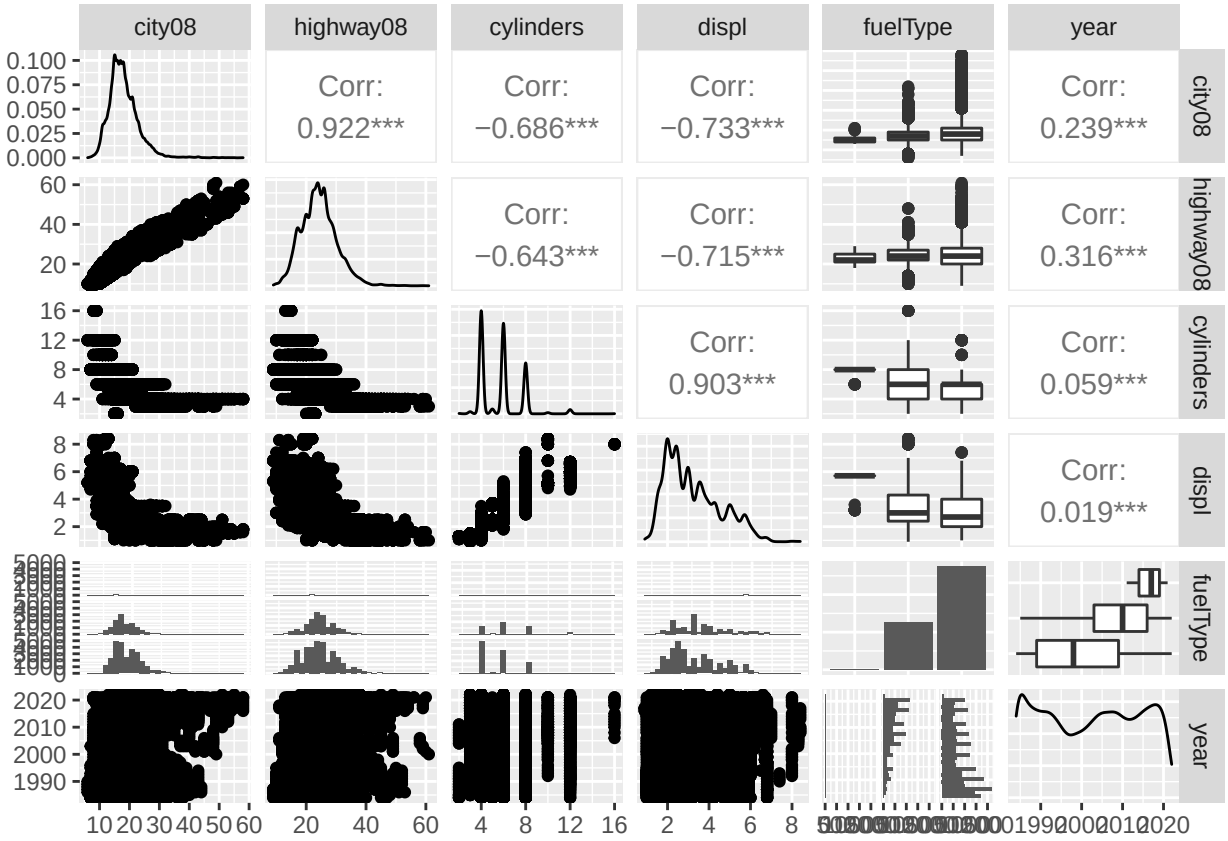
```
## [1] 0.0625
```

We can see that based on the code above the best SSE is when the span is the smallest. The issue with having the smallest span though is that we are going to be overfitting the regression which is not what we want. This is the reason why I selected the span of 0.5 instead of 0.05. The regression with span of 0.5 fits better on the data and the variance of the residual standard error is not that much greater compared to using a span of 0.05

GGPairs

I will now be using GGPairs to see if there is any coorelation with the other columns in the dataset

```
dataset %>%
  select('city08', 'highway08', 'cylinders', 'displ', 'fuelType', 'year') %>%
  ggpairs()
```



Linear Regression: City

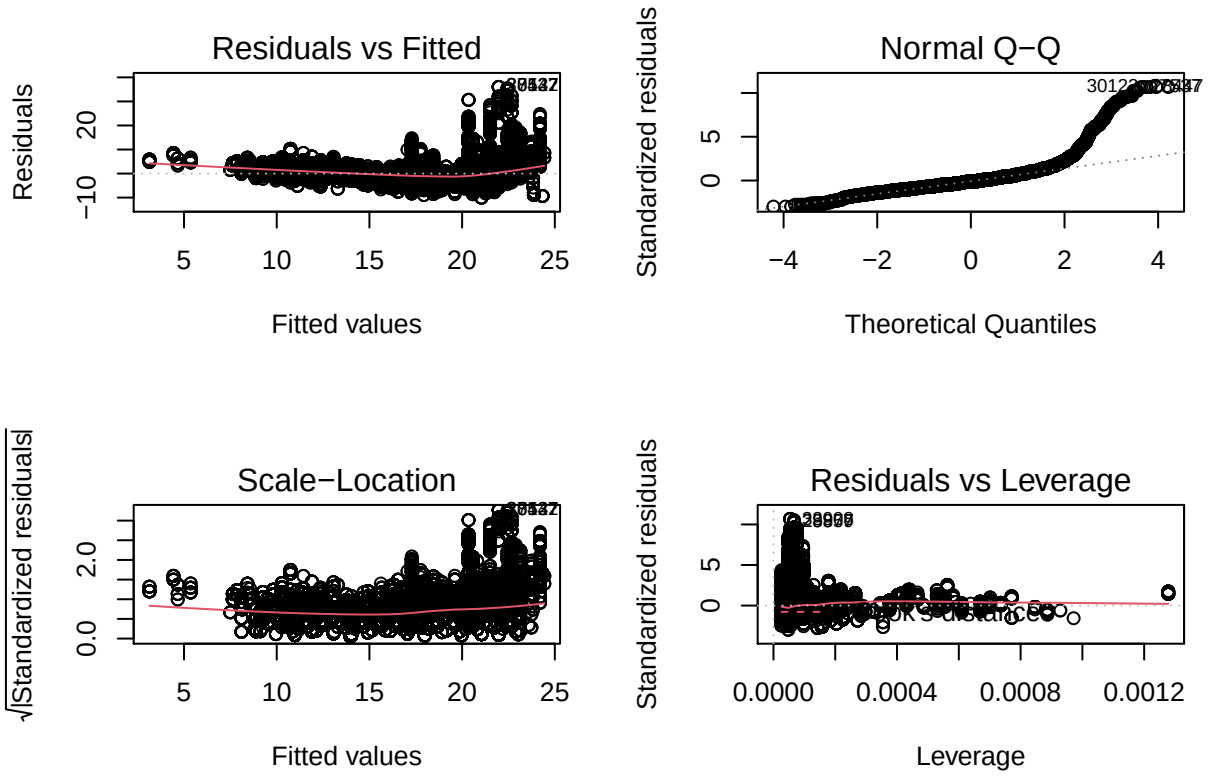
We can see that the number of cylinders and displacement has a positive correlation with fuel economy. We are going to see which of the variables contributes to the fuel efficiency

```
m7 = lm(city08 ~ cylinders + displ, data = dataset)
summary(m7)
```

```
##
## Call:
## lm(formula = city08 ~ cylinders + displ, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.043  -1.910  -0.446   1.394  36.028
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.63155    0.05962  463.45   <2e-16 ***
## cylinders   -0.37033    0.02235  -16.57   <2e-16 ***
```

```
## displ       -2.32128    0.02976  -78.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.374 on 40249 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5403, Adjusted R-squared:  0.5402
## F-statistic: 2.365e+04 on 2 and 40249 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(m7)
```



We can see from the diagnostic that a linear regression may not be the best fit for dataset. Looking at the Residual vs Fitted it has a cone like shape going further out on the x axis. Also the QQ Plot is not linear and shows a huge jump around the 2nd to 4th quantiles. We can also see that the adjusted R^2 value of 0.5402 which is pretty good.

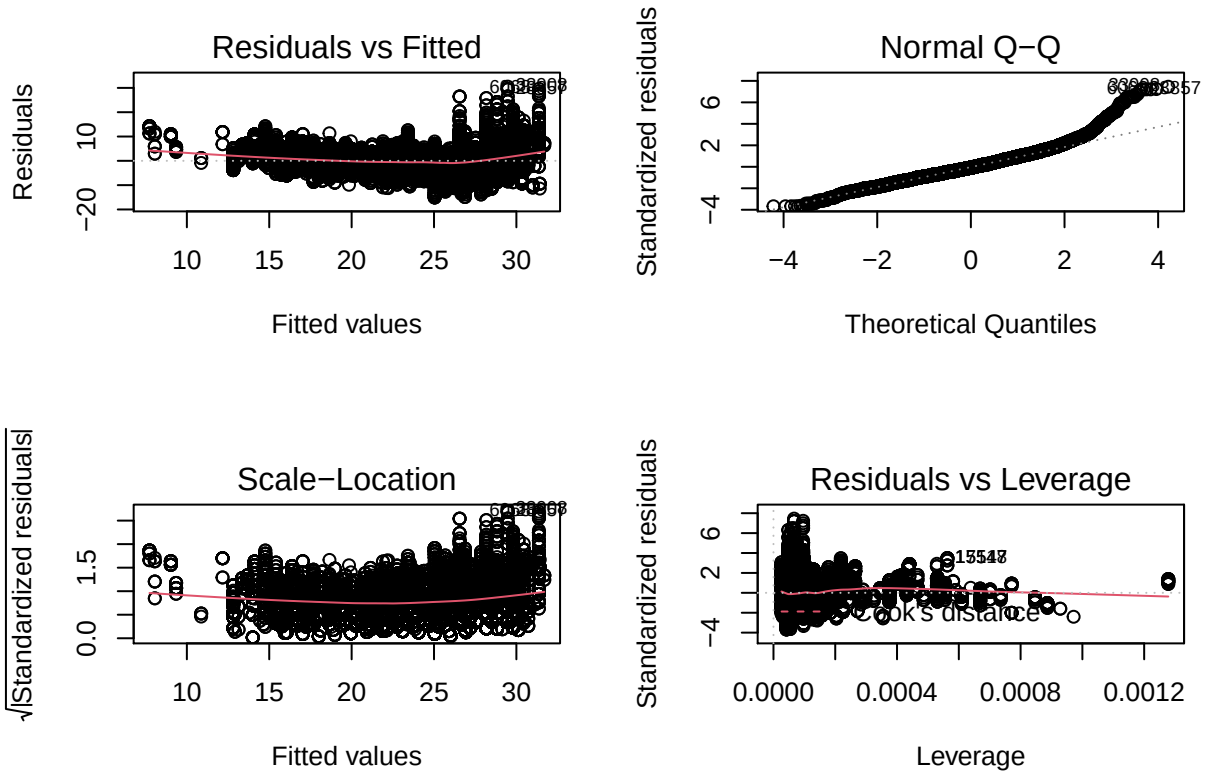
Linear Regression: Highway

```
m8 = lm(highway08 ~ cylinders + displ, data = dataset)
summary(m8)
```

```
##
```

```
## Call:
## lm(formula = highway08 ~ cylinders + displ, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0384  -2.6873  -0.2566   2.4722  30.5255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.45326    0.07236  476.17   <2e-16 ***
## cylinders    0.05315    0.02712    1.96     0.05 .
## displ       -3.24459    0.03612  -89.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.094 on 40249 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5118, Adjusted R-squared:  0.5118
## F-statistic: 2.11e+04 on 2 and 40249 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(m8)
```



We can also see that the highway MPG linear regression has very similar diagnostic plots and the adjusted R^2 value compared to the city MPG linear regression.

Conclusion

We can see that fuel economy for gas engine cars has indeed plateaued in the recent years. This is due to cheap gas prices and consumers buying more SUV instead of more fuel efficient sedans. This is a stark contrast to the 2008 recession when gas prices was expensive and consumers wanted more fuel efficient cars. We can see that the fuel economy trend is impacted by the economy which also effects gas prices. The reason why this analysis is important is because we might be seeing a new trend now due to the recession caused by COVID-19. This recession may create a wave of consumers wanting more fuel efficient vehicles instead of SUV. Somethings to improve on in this analysis is seeing how gas prices or the economy GDP correlates with the fuel efficiency.