**Hursh Desai**

## Introduction

I was curious about what factors contributed to carbon emissions per state. So, I picked my numerical response variable as million metric tons of carbon dioxide in a state. For my six potential predicting variables I chose net generation of electricity, population, miles of public road in the state, GDP per capita, gasoline sold, and personal consumption expenditures per capita. I was contemplating choosing my response variable to be CO2 emissions per capita however, I felt that would be assuming too much because I still did not yet know if a higher population had to mean that the emissions from that state would be higher as well. I could imagine a state that had a low population but because it had a lot of power plants it still had high CO2 emissions. However, I do know that population will be highly correlated with GDP and consumption, so I made those two variables per capita to try and remove that correlation preemptively. I've heard that electricity generation and transportation are the biggest contributors to carbon emissions, so I tried to include those in the regression through the net generation of electricity for the former and miles of public road and gasoline sold for the latter. The GDP and personal consumption data came from the Bureau of Economic Analysis[1][2], the public road length and gasoline sold came from the US Department of Transportation Federal Highway Administration[3][4], the net generation of electricity and carbon emission data came from the US Energy Information Administration[5][6].

## Analysis

[1] https://apps.bea.gov/iTable/iTable.cfm?0=1200&1=1&2=200&3=sic&4=1&5=xx&6=-1&7=-1&8=-1&9=70&10=levels&isuri=1&reqid=70&step=10#reqid=70&step=10&isuri=1&7003=1000&7035=-1&7004=naics&7005=1&7006=xx&7036=-1&7001=11000&7002=1&7090=70&7007=-1&7093=levels

[2] https://apps.bea.gov/iTable/iTable.cfm?0=1200&1=1&2=200&3=sic&4=1&5=xx&6=-1&7=-1&8=-1&9=70&10=levels&isuri=1&reqid=70&step=10#

[3] https://www.fhwa.dot.gov/policyinformation/statistics/2017/hm20.cfm

[4] https://www.fhwa.dot.gov/policyinformation/statistics/2017/33ga.cfm

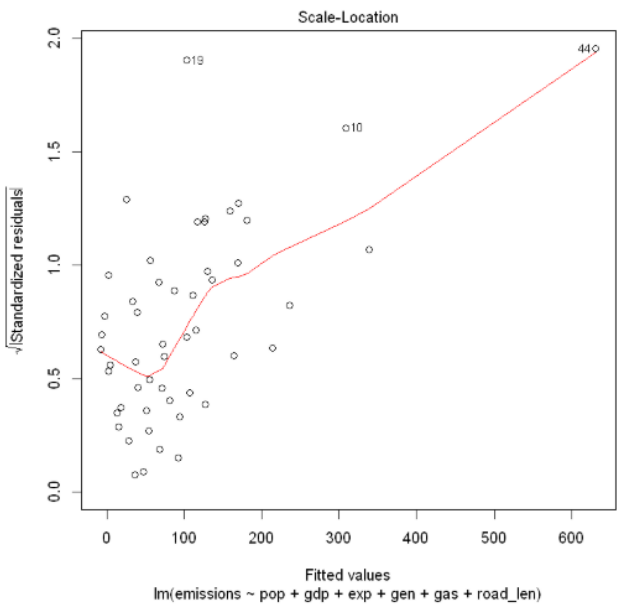
[5] https://www.eia.gov/electricity/annual/html/epa_03_07.html

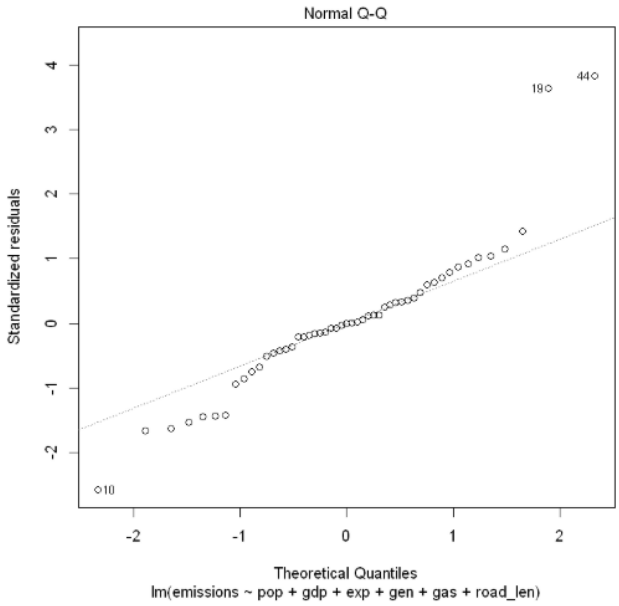
[6] https://www.eia.gov/environment/emissions/state/

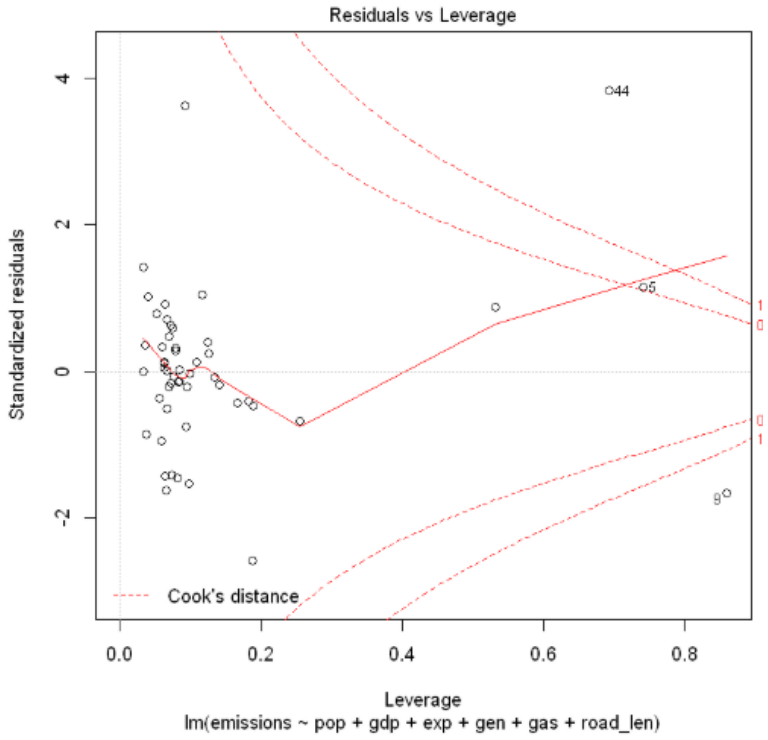
```
lm(formula = emissions ~ pop + gdp + exp + gen + gas + road_len,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-82.287 -14.306  -0.197  14.728 122.483

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.741e+01  4.725e+01  -1.003   0.3212
pop         -8.274e-06  4.150e-06  -1.994   0.0524 .
gdp          3.055e-04  3.787e-04   0.807   0.4241
exp          3.488e-04  1.341e-03   0.260   0.7959
gen          8.464e-04  1.667e-04   5.077 7.48e-06 ***
gas          2.955e-08  1.136e-08   2.602   0.0126 *
road_len     2.289e-04  1.772e-04   1.291   0.2033
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.4 on 44 degrees of freedom
Multiple R-squared:  0.9122,    Adjusted R-squared:  0.9002
F-statistic:  76.2 on 6 and 44 DF,  p-value: < 2.2e-16
```

As we can see this model does have a F-statistic with a very low p-value of 2.2e-16 which means that there is strong evidence of a relationship between these variables and carbon emissions. This model also has a R-squared of .9122 which means that there is evidence of a strong effect between these variables and carbon emissions. However, looking at the individual t-values and their related p-values there is not very strong evidence of a relationship between most of the variables. We are very confident though that net generation of electricity is related to carbon emissions while population and gasoline sold we are less so but still confident that there is a relationship between them and carbon emissions as well. GDP per capita, consumption per capita and road length however, we are not confident there is a relationship between them and carbon emissions. One part of this analysis that surprised me though is that population seems to be slightly negatively associated with carbon emissions.

However, we must be careful of taking this too seriously because the data could still hold strong outliers and leverage points that could be messing up the regression. And, so we check the quickly check the residuals vs fitted values plot, Normal Q-Q plot, and Residuals vs. Leverage plot.



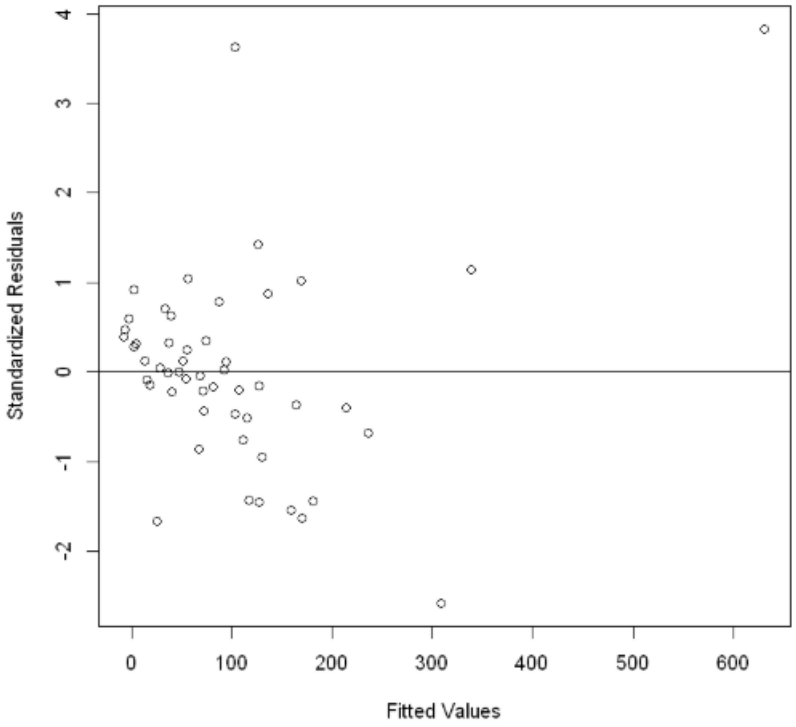




Using the these plots we can easily see that data point 44 (Texas) is both an outlier and a leverage point which is extremely dangerous to a regression. Texas being a leverage point makes sense since it is a big state so it will definitely have a large population along with the other variables but it also has the highest carbon emissions so I would have thought it wouldn't be an outlier as well. It also looks like 19 (Louisiana) has been pointed out as an outlier along with 10 (Florida). It seems as though these points could be caught up in a swapping effect so we will have refer to the residuals vs leverage plot that

superimposes the Cook's Distance onto the plot so that we can see which point cross the 1 threshold. From this plot it seems that Texas is definitely disrupting the regression the most since not only is at an outlier but it also has the largest leverage value with a hat value of 0.691934. However now we can see that point 9 (D.C) and 5 (California) have now also been identified as influential points D.C more so than California since California has not yet crossed the 1 threshold. Texas seems like it would do the

However, we must make sure that these results would be the same if we were comparing using standardized residuals. The three points 44, 19, 10 still stand out and we can also see that they are clearly outside the ±2.5 standard deviations the closest to that barrier with -2.579 being 10 (Florida).




Before we take any data points out, which would give us a new data set, I wanted to check what the best subsets plot gave me still using this data. Looking at the plot it lets us know that the best model will probably have 3 variables. This is because the lowest Mallows Cp is 4.111 but also this fits the other criteria of choosing a model which is that (p + 1) should equal the Mallows Cp. The best model that uses 3 variables includes population, net electricity generation, and amount of gasoline sold, which does correspond with the statistical significance of these variables since these were the only variables that were statistically significant. The VIF table shows that the population and gasoline sold variables are highly collinear with the other variables. So, it would make sense that the best model would use those variables.

A matrix: 11 × 6 of type lgl

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 1 | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE |
| 2 | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE |
| 2 | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 3 | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE |
| 3 | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE |
| 4 | TRUE | TRUE | FALSE | TRUE | TRUE | FALSE |
| 4 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| 5 | TRUE | TRUE | FALSE | TRUE | TRUE | TRUE |
| 5 | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE |
| 6 | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

$label

'(Intercept)' · '1' · '2' · '3' · '4' · '5' · '6'

$size

2 · 2 · 3 · 3 · 4 · 4 · 5 · 5 · 6 · 6 · 7

$Cp

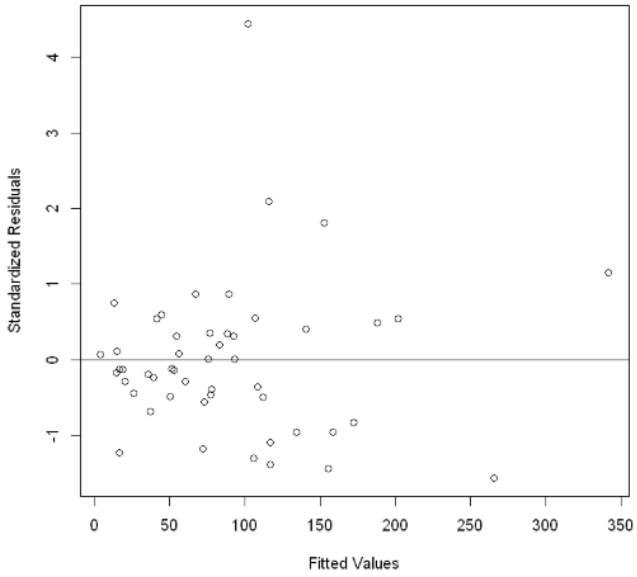
11.4669284026931 · 63.7931210191559 · 6.12700532287982 · 9.59846820305148 · 4.11115084398057 · 6.15143077765131 · 4.68487770315991 · 5.00533186813009 · 5.06771335184968 · 5.65095944058096 · 7

pop:
36.5164521892194
gdp:
2.1979034903063
exp:
2.84220702263115
gen:
6.62208509371001
gas:
49.4415316952638
road_len:
3.77235990657272

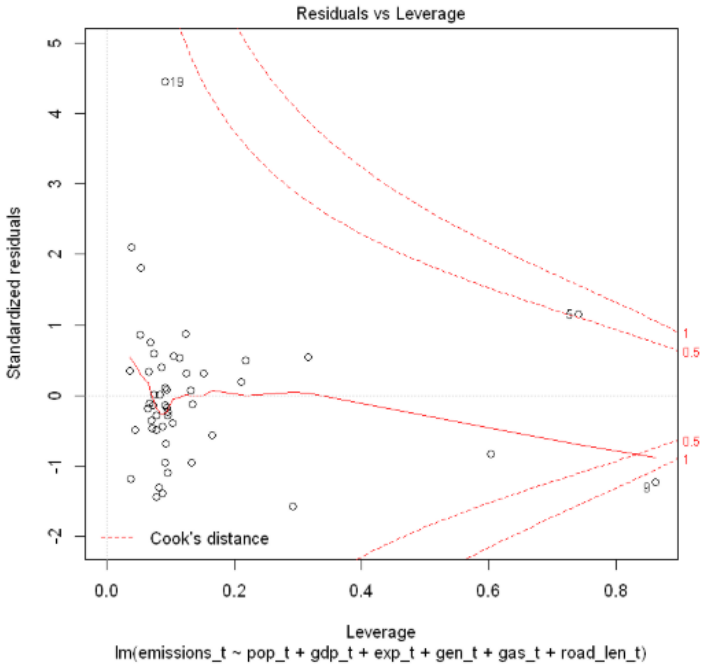
**Models without Unusual Observations**

Now let us compare the model for when we take out Texas.

```
Call:
lm(formula = emissions_t ~ pop_t + gdp_t + exp_t + gen_t + gas_t +
    road_len_t, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-40.514 -13.700  -3.611  10.917 123.840

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.832e+01  4.415e+01   1.094 0.279895
pop_t        2.594e-06  4.153e-06   0.625 0.535467
gdp_t        2.535e-04  3.130e-04   0.810 0.422381
exp_t       -1.305e-03  1.163e-03  -1.121 0.268332
gen_t        5.480e-04  1.520e-04   3.605 0.000806 ***
gas_t        6.486e-09  1.062e-08   0.611 0.544518
road_len_t   1.056e-04  1.488e-04   0.710 0.481518
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.24 on 43 degrees of freedom
Multiple R-squared:  0.8547,    Adjusted R-squared:  0.8344
F-statistic: 42.15 on 6 and 43 DF,  p-value: < 2.2e-16
```




Surprisingly taking out Texas has decreased our R-squared meaning that the relationship has become much weaker. And now there are two less variables that have evidence of a strong relationship.




The Residuals vs. Leverage plot still shows that Washington D.C as an influential point.

However, comparing this model to the model where I take out Louisiana and Florida, the model where I take out Louisiana and Florida gives us 3% more R-squared. And, it also makes the three variables that we previously found added the most information much more statistically significant.

```
Call:
lm(formula = emissions_t ~ pop_t + gdp_t + exp_t + gen_t + gas_t +
    road_len_t, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-53.605 -14.795   2.621  14.555  59.400

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.002e+01  3.769e+01  -2.123 0.039684 *
pop_t       -9.219e-06  3.226e-06  -2.858 0.006611 **
gdp_t        7.473e-05  2.968e-04   0.252 0.802463
exp_t        1.332e-03  1.061e-03   1.255 0.216280
gen_t        8.346e-04  1.335e-04   6.250 1.73e-07 ***
gas_t        3.345e-08  8.842e-09   3.783 0.000484 ***
road_len_t   2.311e-04  1.458e-04   1.585 0.120443
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.47 on 42 degrees of freedom
Multiple R-squared:  0.9467,    Adjusted R-squared:  0.9391
F-statistic: 124.4 on 6 and 42 DF,  p-value: < 2.2e-16
```

However, I found that no matter how many influential points I took out the R-squared remained the same and every time I plotted the standardized residuals vs fitted values as well as standardized residuals vs leverage points to see if there were any influential points according to Cook's Distance it always gave me a new state as another influential point. When I took out Louisiana and Florida the plots would tell me South Dakota was an influential point and if I took that out it would tell me South Carolina was an influential point, and then Rhode Island and so on. And, I could not figure out why but my best hypothesis is that there is a variable that I am not taking into account such as a subgroup that is causing there to be so many influential points. Nevertheless, you cannot just take out arbitrary points in the hopes of a marginally higher R-squared and 3% is not even that high of a leap in strength of relationship to rationalize taking out what seem to be influential outliers and leverage points.


```
Call:
lm(formula = emissions_t ~ pop_t + gdp_t + exp_t + gen_t + gas_t +
    road_len_t, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-53.605 -14.795   2.621  14.555  59.400

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.002e+01  3.769e+01  -2.123 0.039684 *
pop_t       -9.219e-06  3.226e-06  -2.858 0.006611 **
gdp_t        7.473e-05  2.968e-04   0.252 0.802463
exp_t        1.332e-03  1.061e-03   1.255 0.216280
gen_t        8.346e-04  1.335e-04   6.250 1.73e-07 ***
gas_t        3.345e-08  8.842e-09   3.783 0.000484 ***
road_len_t   2.311e-04  1.458e-04   1.585 0.120443
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.47 on 42 degrees of freedom
Multiple R-squared:  0.9467,    Adjusted R-squared:  0.9391
F-statistic: 124.4 on 6 and 42 DF,  p-value: < 2.2e-16
```

$which

A matrix: 11 × 6 of type lgl

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 1 | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE |
| 2 | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE |
| 2 | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 3 | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE |
| 3 | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE |
| 4 | TRUE | TRUE | TRUE | TRUE | FALSE | FALSE |
| 4 | FALSE | TRUE | TRUE | TRUE | TRUE | FALSE |
| 5 | TRUE | TRUE | TRUE | TRUE | FALSE | TRUE |
| 5 | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE |
| 6 | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

$label

'(Intercept)' · '1' · '2' · '3' · '4' · '5' · '6'

$size

2 · 2 · 3 · 3 · 4 · 4 · 5 · 5 · 6 · 6 · 7

$Cp

18.2400937297587 · 23.453089676594 · 1.61830415782997 · 2.80451781664973 · 2.36780174113674 · 2.65648745311854 · 3.92399125921464 · 3.92420134087848 · 5.37313116057319 · 5.39025336377593 · 7

Using best subsets on the model where I took out Texas would now yield a model with only 2 variables generation and gasoline to be the best model. However, this model would only give .84 as its R-squared which I would say is not worth losing in terms of how much the model has been simplified.

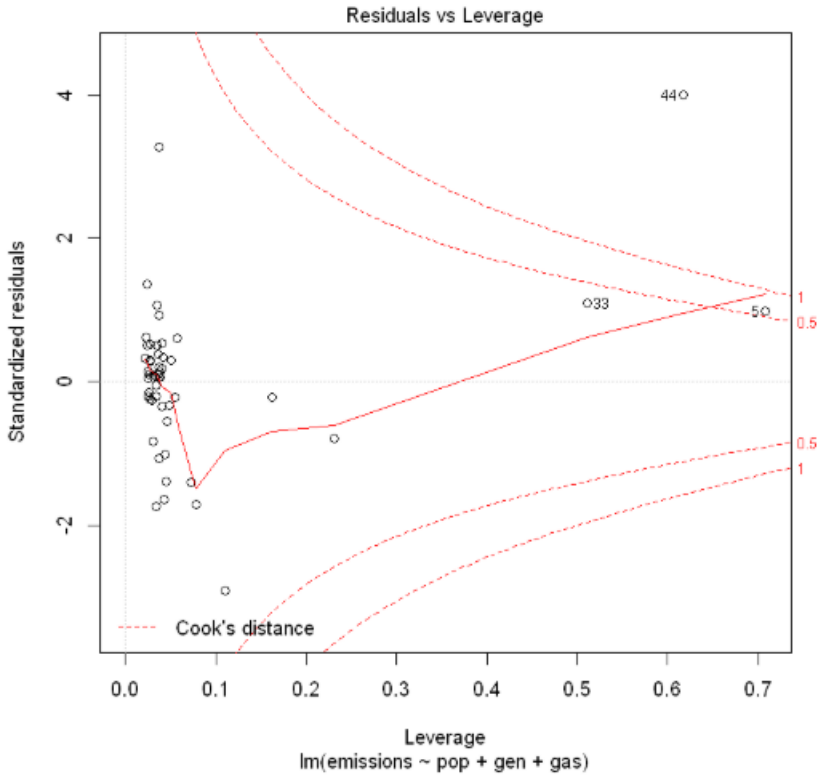
**Best Model**

Now to see what the regression looks like for the model that only uses the three variables we found that add the most to our model.

```
Call:
lm(formula = emissions ~ pop + gen + gas, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-97.155 -10.202   3.033  15.466 113.774

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.208e+00  7.146e+00  -1.009  0.31829
pop         -7.902e-06  3.948e-06  -2.002  0.05112 .
gen          9.162e-04  1.491e-04   6.145 1.62e-07 ***
gas          3.011e-08  1.102e-08   2.733  0.00882 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.44 on 47 degrees of freedom
Multiple R-squared:  0.906,     Adjusted R-squared:    0.9
F-statistic:   151 on 3 and 47 DF,  p-value: < 2.2e-16
```

As expected, we were able to simplify our model by three whole variables and the R-squared only decreased by .01. This means that the GDP, expenditure, and road length variables were only adding

unnecessary noise. The statistical significance of amount of gasoline sold has also risen in statistical significance while population has not. It is still surprising to see the population has a negative regression coefficient according to my assumptions at the start.




Texas again shows up as an influential point even in this model however again if I were to start taking out what seemed to be influential points based on the Cook's Distance rule and replotting this plot it would keep giving me new influential points.

**Conclusion**

From the analysis we can see that Net Generation of Electricity as expected predicted the most in terms of carbon emissions and amount of gasoline sold right after that. Which points are outliers or leverage points or what points have influenced the regression the most is still inconclusive?