Regression

Ally de Vera

2023-02-18

Combining consecutive years of data (needed to meet 10000+ row req) Seperates make and model for easier analysis

```
library(readxl)
library(stringr)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

EPA_2016 <- read_excel("EPA_2016.xlsx")
EPA_2017 <- read_excel("EPA_2017.xlsx")
EPA_2018 <- read_excel("EPA_2018.xlsx")
EPA_2019 <- read_excel("EPA_2019.xlsx")
EPA_2020 <- read_excel("EPA_2020.xlsx")


EPA_allYears <- rbind(EPA_2016, EPA_2017, EPA_2018, EPA_2019, EPA_2020)

MakeAndModel <- str_split_fixed(EPA_allYears$Model, " ", 2)

EPA_allYears$Model <- NULL

EPA_allYears <- cbind(MakeAndModel, EPA_allYears)

colnames(EPA_allYears)[1] <- 'Make'
colnames(EPA_allYears)[2] <- 'Model'

EPA_allYears$Make <- as.factor(EPA_allYears$Make)

EPA_allYears$Displ <- as.numeric(EPA_allYears$Displ)

## Warning: NAs introduced by coercion

EPA_allYears$Cyl <- as.numeric(EPA_allYears$Cyl)

## Warning: NAs introduced by coercion

EPA_allYears$Trans <- as.factor(EPA_allYears$Trans)

EPA_allYears$Drive <- as.factor(EPA_allYears$Drive)

EPA_allYears$Fuel <- as.factor(EPA_allYears$Fuel)
```

```
EPA_allYears$`Cert Region` <- as.factor(EPA_allYears$`Cert Region`)

EPA_allYears$`Veh Class` <- as.factor(EPA_allYears$`Veh Class`)

EPA_allYears$`Air Pollution Score`<- as.numeric(EPA_allYears$`Air Pollution
Score`)

## Warning: NAs introduced by coercion

EPA_allYears$`City MPG` <- as.numeric(EPA_allYears$`City MPG`)

## Warning: NAs introduced by coercion

EPA_allYears$`Hwy MPG` <- as.numeric(EPA_allYears$`Hwy MPG`)

## Warning: NAs introduced by coercion

EPA_allYears$`Cmb MPG` <- as.numeric(EPA_allYears$`Cmb MPG`)

## Warning: NAs introduced by coercion

EPA_allYears$`Greenhouse Gas Score` <- as.numeric(EPA_allYears$`Greenhouse
Gas Score`)

## Warning: NAs introduced by coercion

EPA_allYears$SmartWay <- as.factor(EPA_allYears$SmartWay)

EPA_allYears$`Comb CO2` <- as.numeric(EPA_allYears$`Comb CO2`)

## Warning: NAs introduced by coercion

str(EPA_allYears)

## 'data.frame':    13126 obs. of  19 variables:
##  $ Make               : Factor w/ 51 levels "ACURA","ALFA-ROMEO",..: 1 1
1 1 1 1 1 1 1 1 ...
##  $ Model              : chr  "ILX" "ILX" "MDX" "MDX" ...
##  $ Displ              : num  2.4 2.4 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 ...
##  $ Cyl                : num  4 4 6 6 6 6 6 6 6 6 ...
##  $ Trans              : Factor w/ 30 levels "AMS-6","AMS-7",..: 16 16 30
30 30 30 30 30 30 30 ...
##  $ Drive              : Factor w/ 2 levels "2WD","4WD": 1 1 1 1 1 1 2 2 2
2 ...
##  $ Fuel               : Factor w/ 7 levels "Diesel","Electricity",..: 5 5
5 5 5 5 5 5 5 5 ...
##  $ Cert Region        : Factor w/ 2 levels "CA","FA": 2 1 2 2 1 1 2 2 1 1
...
##  $ Stnd               : chr  "B5" "L3ULEV125" "B5" "B5" ...
##  $ Stnd Description   : chr  "Federal Tier 2 Bin 5" "California LEV-III
ULEV125" "Federal Tier 2 Bin 5" "Federal Tier 2 Bin 5" ...
```

```
##  $ Underhood ID        : chr  "GHNXV02.4XH3" "GHNXV02.4XH3" "GHNXV03.5VA3"
"GHNXV03.5VA3" ...
##  $ Veh Class           : Factor w/ 10 levels "large car","midsize car",..:
5 5 6 6 6 6 6 6 6 ...
##  $ Air Pollution Score : num  5 6 5 5 6 6 5 5 6 6 ...
##  $ City MPG            : num  25 25 19 20 19 20 18 19 18 19 ...
##  $ Hwy MPG             : num  36 36 27 27 27 27 26 26 26 26 ...
##  $ Cmb MPG             : num  29 29 22 23 22 23 21 22 21 22 ...
##  $ Greenhouse Gas Score: num  7 7 5 5 5 5 5 5 5 5 ...
##  $ SmartWay            : Factor w/ 3 levels "Elite","No","Yes": 2 3 2 2 2
2 2 2 2 2 ...
##  $ Comb CO2            : num  305 305 403 390 403 390 412 409 412 409 ...
```

training

```
set.seed(620)
i <- sample(1:nrow(EPA_allYears), 0.8*nrow(EPA_allYears), replace=FALSE)
train <- EPA_allYears[i,]
test <- EPA_allYears[-i,]
```

Gaussian GLM model

```
glm_gaus <- glm(`Greenhouse Gas Score` ~ `Cmb MPG`, data=train,
family=gaussian)
summary(glm_gaus)
```

```
##
## Call:
## glm(formula = `Greenhouse Gas Score` ~ `Cmb MPG`, family = gaussian,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.2272  -0.5772   0.1291   0.8355   3.2691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.4233617  0.0254716   95.14   <2e-16 ***
## `Cmb MPG`   0.0978996  0.0008648  113.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.519849)
##
##     Null deviance: 34479  on 9873  degrees of freedom
## Residual deviance: 15004  on 9872  degrees of freedom
##   (626 observations deleted due to missingness)
## AIC: 32159
##
## Number of Fisher Scoring iterations: 2
```

Quasi GLM model

```
glm_quasi <- glm(`Greenhouse Gas Score` ~ `Cmb MPG`, data=train,
family=quasi)
summary(glm_quasi)
```

```
##
## Call:
## glm(formula = `Greenhouse Gas Score` ~ `Cmb MPG`, family = quasi,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.2272  -0.5772   0.1291   0.8355   3.2691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.4233617  0.0254716   95.14   <2e-16 ***
## `Cmb MPG`   0.0978996  0.0008648  113.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 1.519849)
##
##     Null deviance: 34479  on 9873  degrees of freedom
## Residual deviance: 15004  on 9872  degrees of freedom
##   (626 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 2
```

Poission GLM model

```
glm_poiss <- glm(`Greenhouse Gas Score` ~ `Cmb MPG`, data=train,
family=poisson)
summary(glm_poiss)
```

```
##
## Call:
## glm(formula = `Greenhouse Gas Score` ~ `Cmb MPG`, family = poisson,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5254  -0.2940   0.0941   0.4452   1.4585
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.2565901  0.0075723  165.95   <2e-16 ***
## `Cmb MPG`   0.0124187  0.0002037   60.97   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7009.0  on 9873  degrees of freedom
## Residual deviance: 4181.2  on 9872  degrees of freedom
##   (626 observations deleted due to missingness)
## AIC: 37720
##
## Number of Fisher Scoring iterations: 4
```

Binomial GLM model

```
glm_binom <- glm(`Make` ~ `Cmb MPG`, data=train, family=binomial)
summary(glm_binom)
```

```
##
## Call:
## glm(formula = Make ~ `Cmb MPG`, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9783   0.1570   0.1605   0.1631   0.1722
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 4.093541   0.232489  17.607   <2e-16 ***
## `Cmb MPG`   0.010985   0.008876   1.238    0.216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1332.4  on 9889  degrees of freedom
## Residual deviance: 1330.5  on 9888  degrees of freedom
##   (610 observations deleted due to missingness)
## AIC: 1334.5
##
## Number of Fisher Scoring iterations: 7
```

Quasi Binomial GLM model

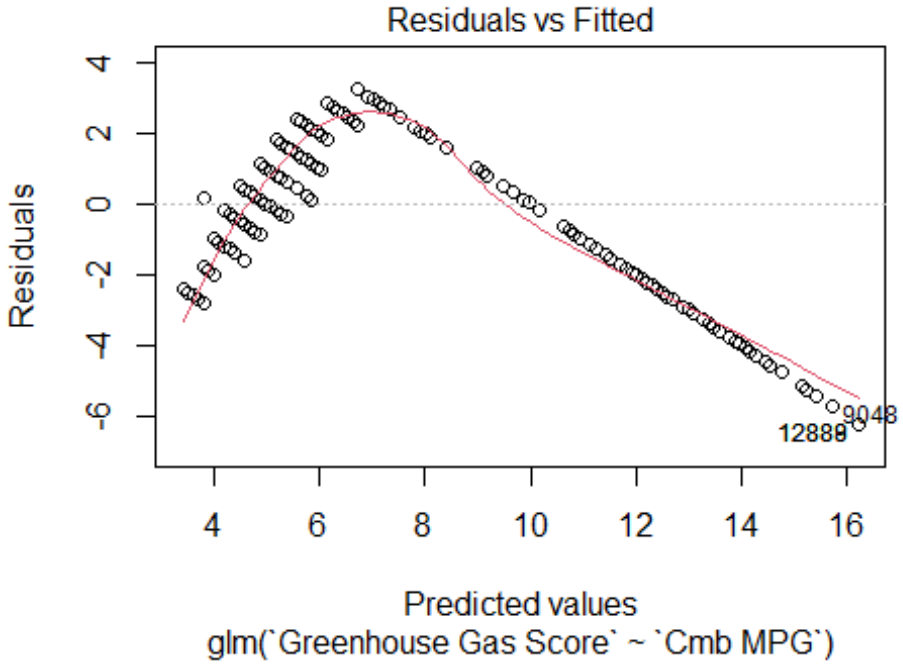
```
glm_qbi <- glm(`Make` ~ `Cmb MPG`, data=train, family=quasibinomial)
summary(glm_qbi)
```

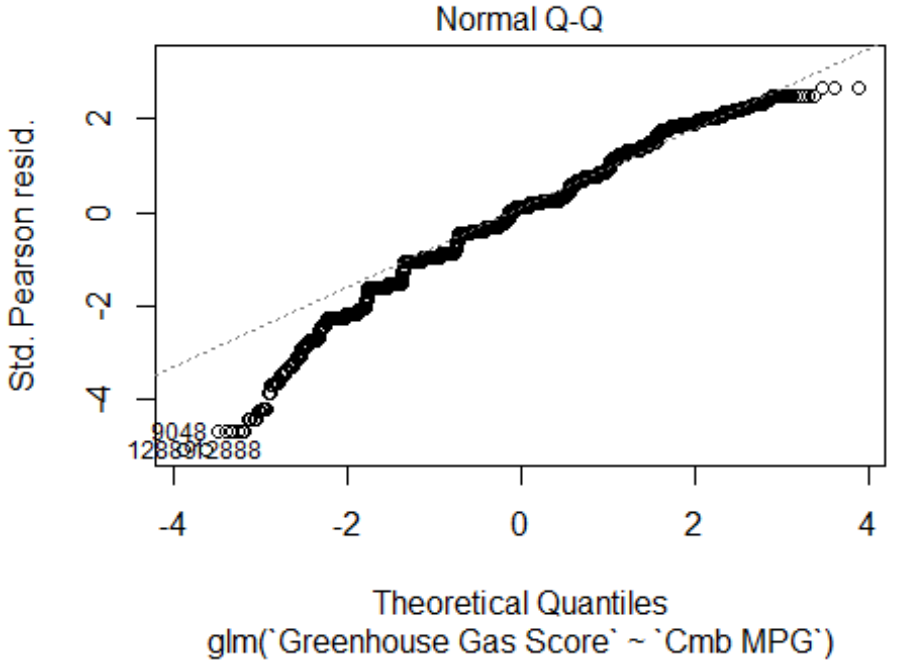
```
##
## Call:
## glm(formula = Make ~ `Cmb MPG`, family = quasibinomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

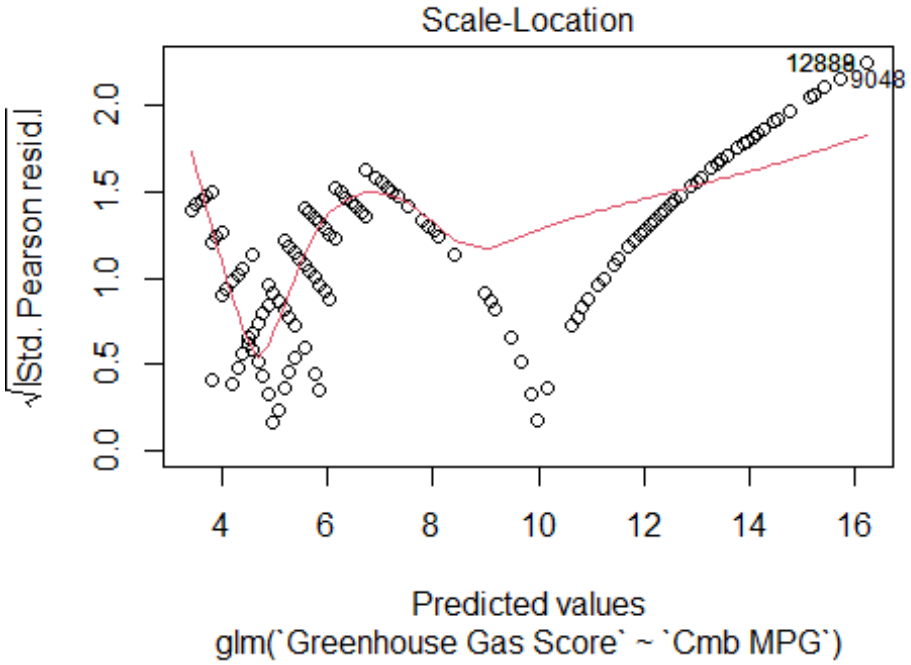
```
## -2.9783   0.1570   0.1605   0.1631   0.1722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.093541   0.231656  17.671   <2e-16 ***
## `Cmb MPG`   0.010985   0.008844   1.242    0.214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9928485)
##
##     Null deviance: 1332.4  on 9889  degrees of freedom
## Residual deviance: 1330.5  on 9888  degrees of freedom
##   (610 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```

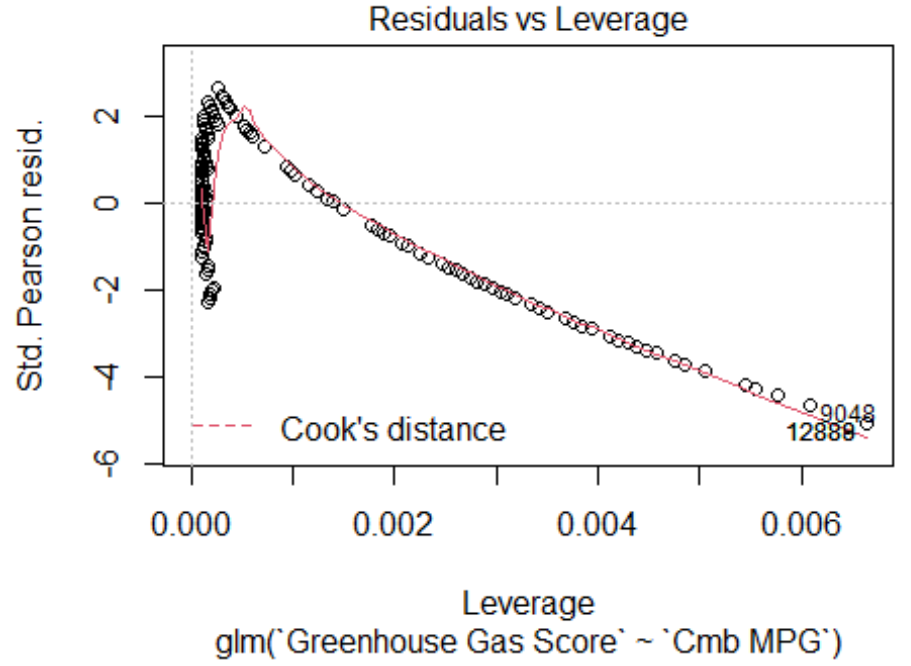
Gaussian GLM Graph

```
plot(glm_gaus)
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
-6
4
6
8
10
12
14
16
9048
12888
Predicted values
glm(`Greenhouse Gas Score` ~ `Cmb MPG`)

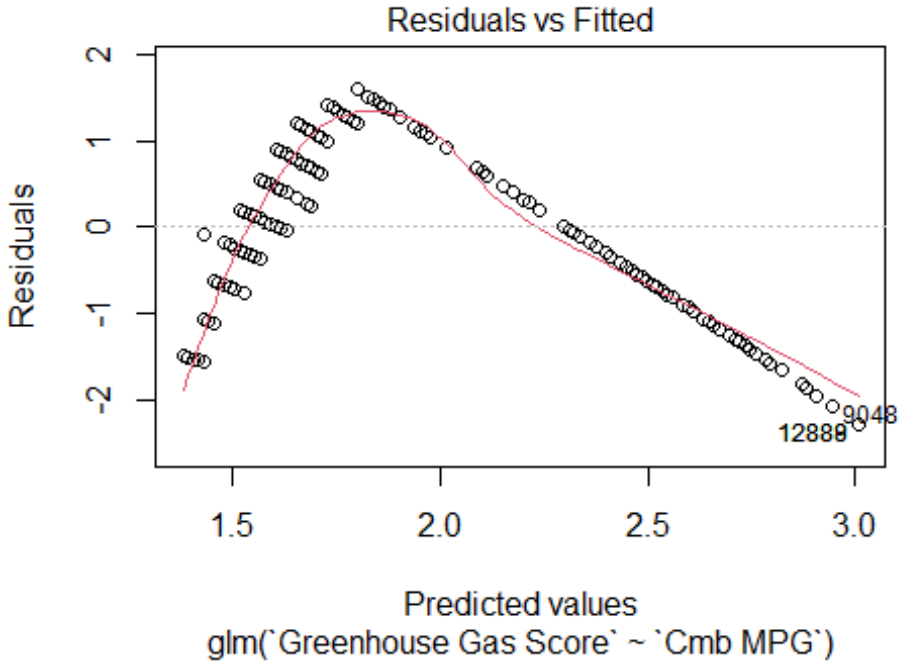

Normal Q-Q
Std. Pearson resid.
2
0
-2
-4
9048
12888
12888
-4
-2
0
2
4
Theoretical Quantiles
glm(`Greenhouse Gas Score` ~ `Cmb MPG`)

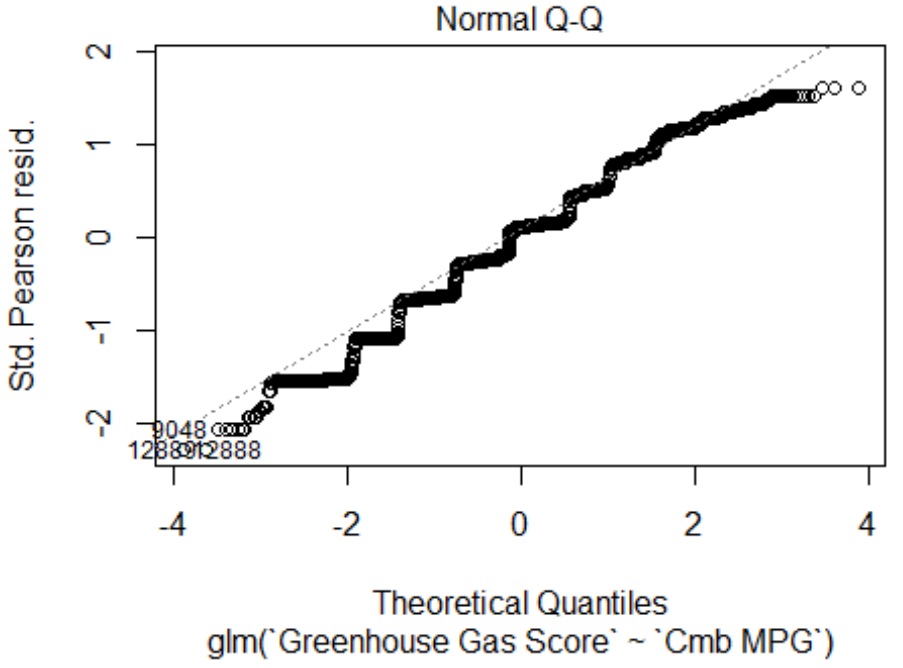


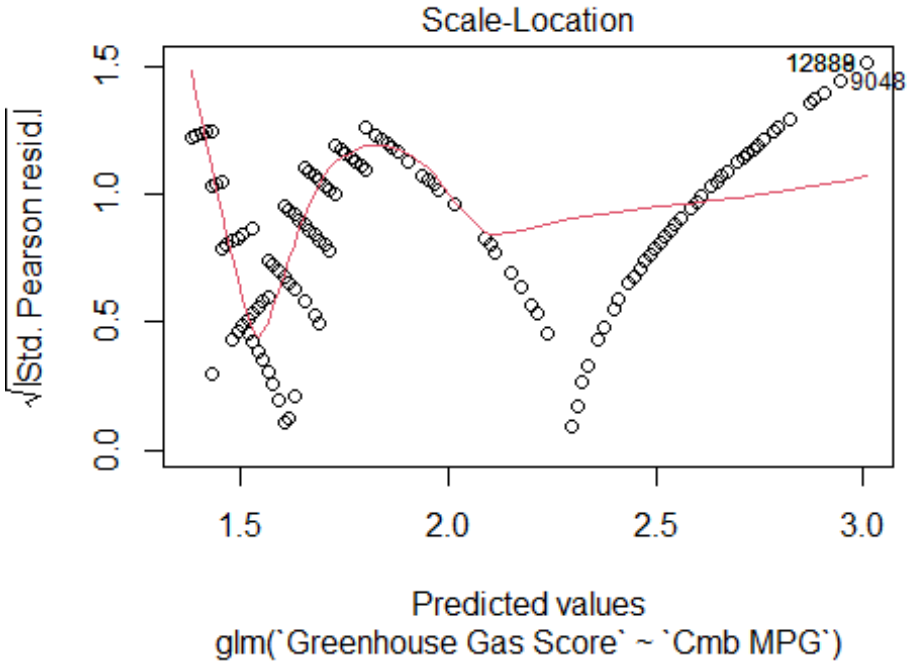


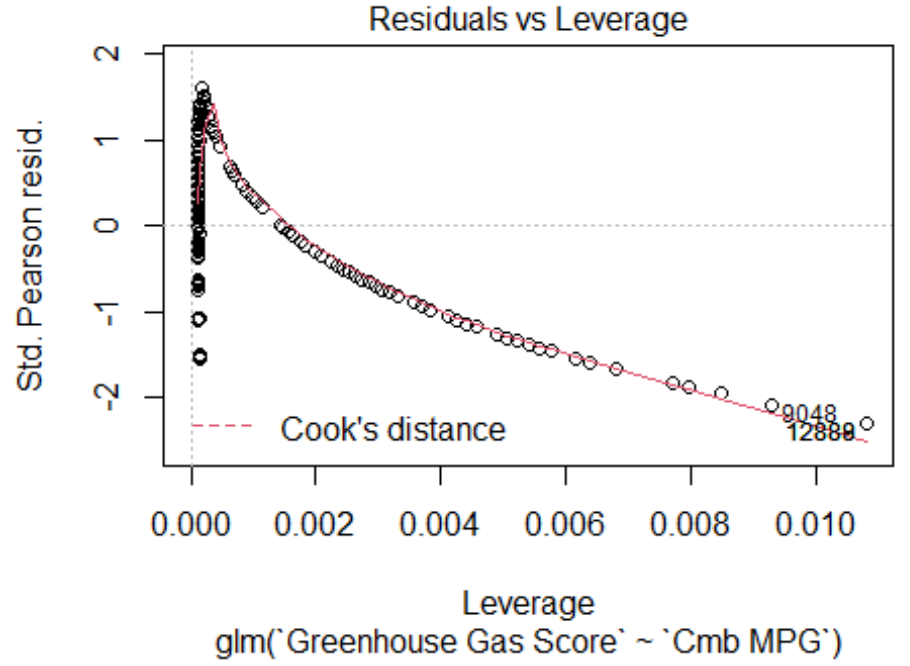
Poisson GLM Graph

```
plot(glm_poiss)
```


Residuals vs Fitted
Residuals
Predicted values
glm(`Greenhouse Gas Score` ~ `Cmb MPG`)
9048
12888


Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(`Greenhouse Gas Score` ~ `Cmb MPG`)
9048
12888


Scale-Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
12889
9048
1.5
2.0
2.5
3.0
Predicted values
glm(`Greenhouse Gas Score` ~ `Cmb MPG`)


Residuals vs Leverage
Std. Pearson resid.
-2
-1
0
1
2
Cook's distance
9048
12889
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
glm(`Greenhouse Gas Score` ~ `Cmb MPG`)

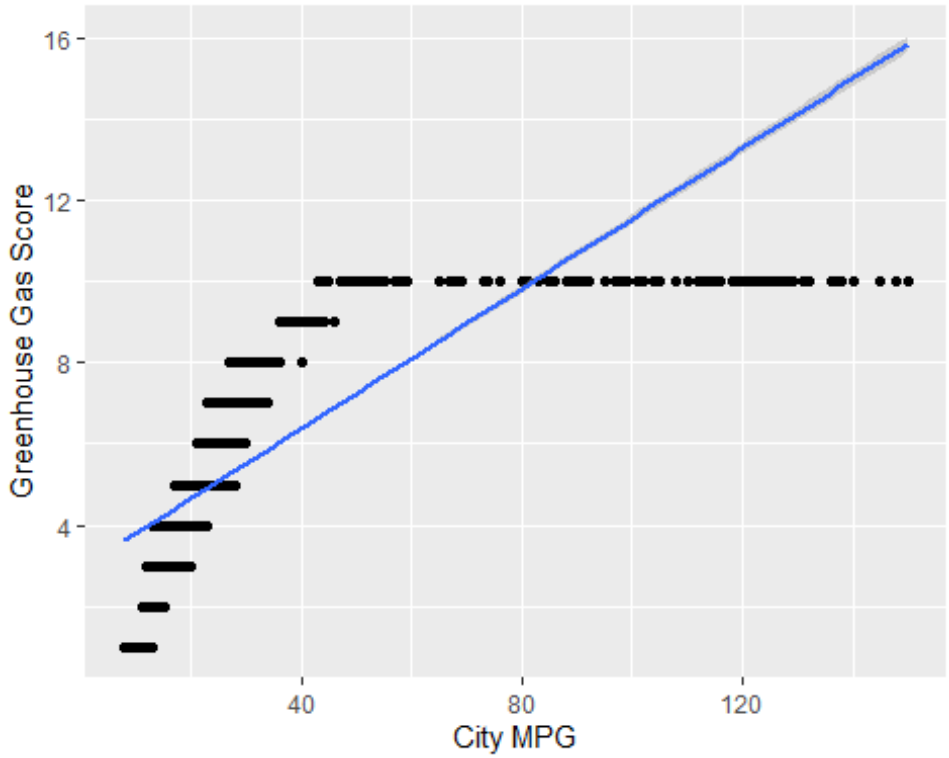
LM1 Graph

```
ggplot(EPA_allYears, aes(`City MPG`, `Greenhouse Gas Score`)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 781 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 781 rows containing missing values (geom_point).
```



LM1 Basics

```
lm1 <- lm(`City MPG`~ `Greenhouse Gas Score`, data = EPA_allYears)
lm1
```

```
##
## Call:
## lm(formula = `City MPG` ~ `Greenhouse Gas Score`, data = EPA_allYears)
##
## Coefficients:
##            (Intercept)  `Greenhouse Gas Score`
##                 -6.345                   5.991
```

LM1 Summary

```
summary(lm1)
```

```
##
## Call:
```

```
## lm(formula = `City MPG` ~ `Greenhouse Gas Score`, data = EPA_allYears)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.580  -4.608  -1.608   2.373  96.439
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.34478    0.27594  -22.99   <2e-16 ***
## `Greenhouse Gas Score`  5.99058    0.05226  114.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.82 on 12343 degrees of freedom
##   (781 observations deleted due to missingness)
## Multiple R-squared:  0.5156, Adjusted R-squared:  0.5156
## F-statistic: 1.314e+04 on 1 and 12343 DF,  p-value: < 2.2e-16
```

LM2 Basics

```
lm2 <- lm(`City MPG`~ `Greenhouse Gas Score`+`Air Pollution Score`, data =
EPA_allYears)
lm2

##
## Call:
## lm(formula = `City MPG` ~ `Greenhouse Gas Score` + `Air Pollution Score`,
##     data = EPA_allYears)
##
## Coefficients:
##            (Intercept)  `Greenhouse Gas Score`   `Air Pollution Score`
##                 -9.350                   5.303                   1.247
```

LM2 Summary

```
summary(lm2)

##
## Call:
## lm(formula = `City MPG` ~ `Greenhouse Gas Score` + `Air Pollution Score`,
##     data = EPA_allYears)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.296  -4.950  -1.513   2.397  93.851
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -9.34953    0.30447  -30.71   <2e-16 ***
## `Greenhouse Gas Score`  5.30315    0.06036   87.86   <2e-16 ***
## `Air Pollution Score`   1.24670    0.05769   21.61   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.63 on 12342 degrees of freedom
##   (781 observations deleted due to missingness)
## Multiple R-squared:  0.5333, Adjusted R-squared:  0.5332
## F-statistic:  7052 on 2 and 12342 DF,  p-value: < 2.2e-16
```

LM3 Basics

```
lm3 <- lm(`Cyl`~ `Cmb MPG` + `Air Pollution Score`, data = EPA_allYears)
lm3

##
## Call:
## lm(formula = Cyl ~ `Cmb MPG` + `Air Pollution Score`, data = EPA_allYears)
##
## Coefficients:
##           (Intercept)              `Cmb MPG`  `Air Pollution Score`
##               10.9315                -0.2077                -0.1025
```

LM3 Summary

```
summary(lm3)

##
## Call:
## lm(formula = Cyl ~ `Cmb MPG` + `Air Pollution Score`, data = EPA_allYears)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0899 -0.8131 -0.1399  0.5720  7.4560
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.931459   0.051498  212.27   <2e-16 ***
## `Cmb MPG`             -0.207728   0.002129  -97.58   <2e-16 ***
## `Air Pollution Score` -0.102492   0.007048  -14.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.299 on 12007 degrees of freedom
##   (1116 observations deleted due to missingness)
## Multiple R-squared:  0.5055, Adjusted R-squared:  0.5054
## F-statistic:  6136 on 2 and 12007 DF,  p-value: < 2.2e-16
```

of my three linear regression models, lm2 using combined miles per gallon with greenhouse gas score, and air pollution score had the highest r^2 value, implying that it had the strongest correlation at .5333.