STAT410 Final Project Code

Group 1: Miranda Gutierrez, Ishmael Guadarrama, Zachary Horner, Stephanie Meza

2022-12-01

First Question

What is the probability of Fuel Type and Engine Size being a factor for the Fuel Consumption in the City and the Highway?

- Simple Linear Regression Model with the Engine Size and Fuel Type as the predictor (in separate models) and the percent cover of Fuel Consumption in both City and Highway as the response
- Residuals QQplot of both models
- 95% confidence intervals for the effect of the predictors
- Plot of residuals as a function of percent cover of Fuel Consumption to find any evidence of heteroskedasticity.

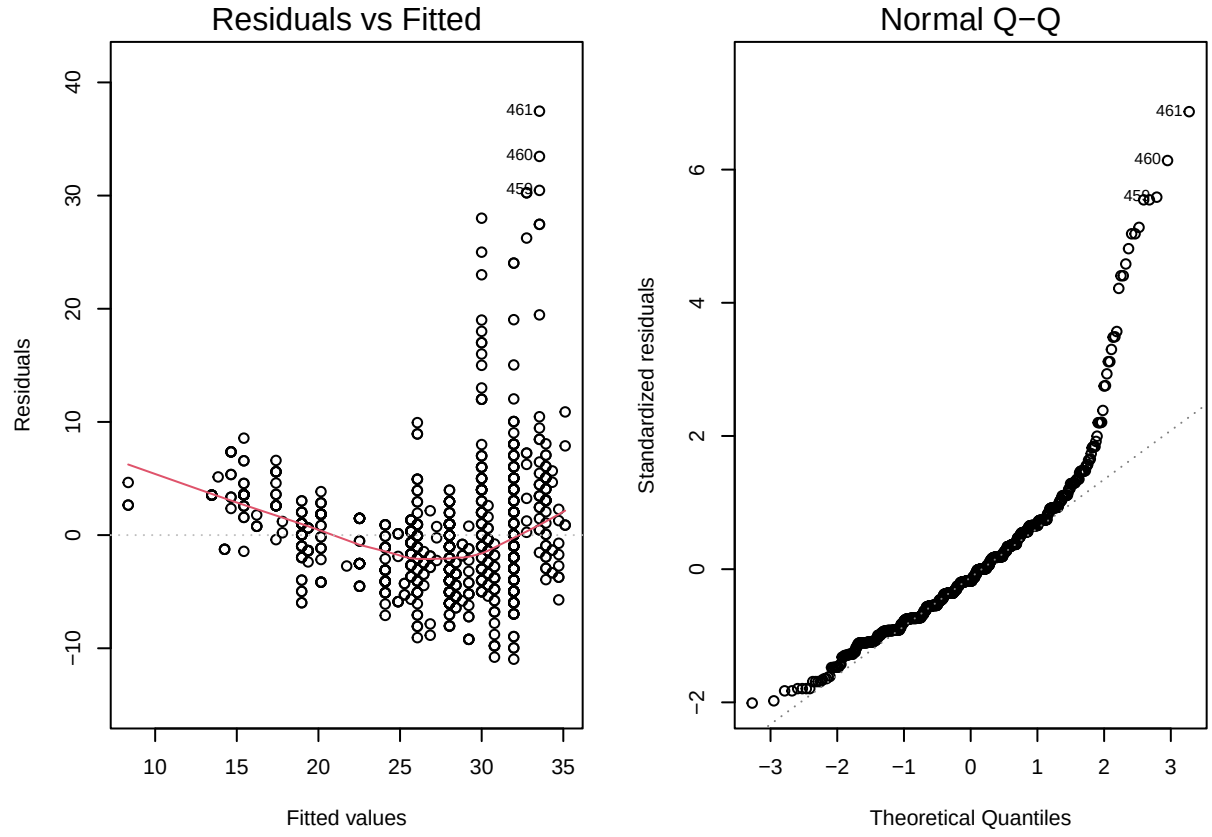
```
#first we turn Fuel.Type into a factor for better analysis
fuel$Fuel.Type <- as.factor(fuel$Fuel.Type)

#lm with engine size as solo predictor
enginesize_fit <- lm(Fuel.Consumption.Comb..mpg.. ~ Engine.Size.L., fuel)
#fit summary
summary(enginesize_fit)
```

```
##
## Call:
## lm(formula = Fuel.Consumption.Comb..mpg.. ~ Engine.Size.L., data = fuel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.966  -3.375  -0.966   2.034  37.460
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     39.8384     0.4497   88.58   <2e-16 ***
## Engine.Size.L.  -3.9363     0.1292  -30.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.46 on 944 degrees of freedom
## Multiple R-squared:  0.4958, Adjusted R-squared:  0.4953
## F-statistic: 928.4 on 1 and 944 DF,  p-value: < 2.2e-16
```

```r
#confidence interval for coefficients
confint(enginesize_fit)
```

```
##                    2.5 %   97.5 %
## (Intercept)    38.955829 40.72103
## Engine.Size.L. -4.189791 -3.68275
```

```r
#plots for assumption checks
layout(matrix(1:2, 3, 2, byrow = T))
par(mar = c(5,5,2,1))
plot(enginesize_fit, c(1:3))
```




```r
#lm with fuel type as solo predictor
fueltype_fit <- lm(Fuel.Consumption.Comb..mpg.. ~ Fuel.Type, fuel)
#fit summary
summary(fueltype_fit)
```

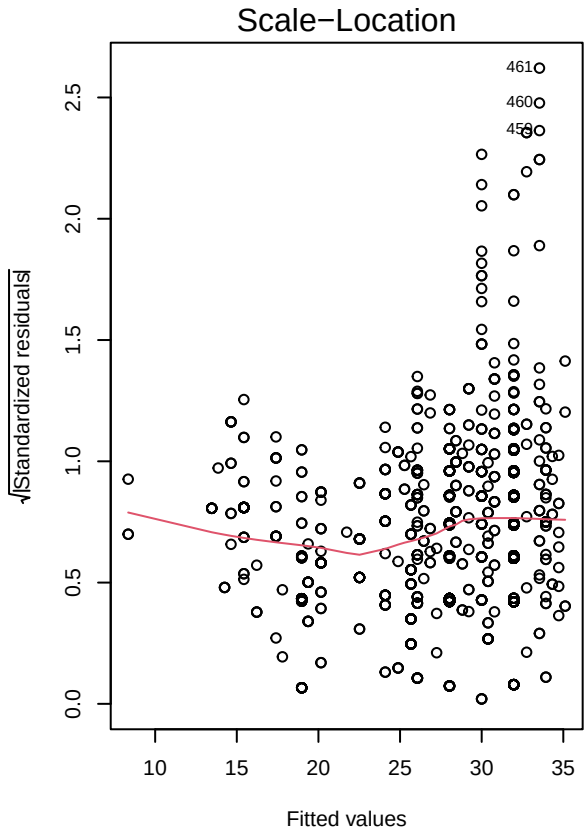
```
##
## Call:
## lm(formula = Fuel.Consumption.Comb..mpg.. ~ Fuel.Type, data = fuel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.812  -5.038  -0.812   4.188  40.962
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.107      1.348  20.852  < 2e-16 ***
## Fuel.TypeE   -11.821      2.335  -5.063 4.95e-07 ***
## Fuel.TypeX     1.931      1.390   1.390   0.1650
## Fuel.TypeZ    -3.295      1.389  -2.373   0.0178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.132 on 942 degrees of freedom
## Multiple R-squared:  0.1414, Adjusted R-squared:  0.1387
## F-statistic: 51.71 on 3 and 942 DF,  p-value: < 2.2e-16
```

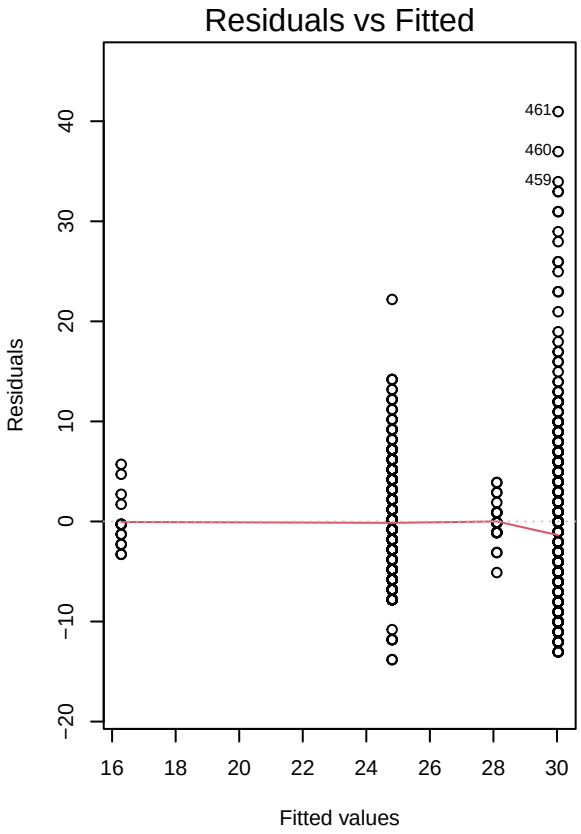
```
#confidence interval for coefficients
confint(fueltype_fit)
```

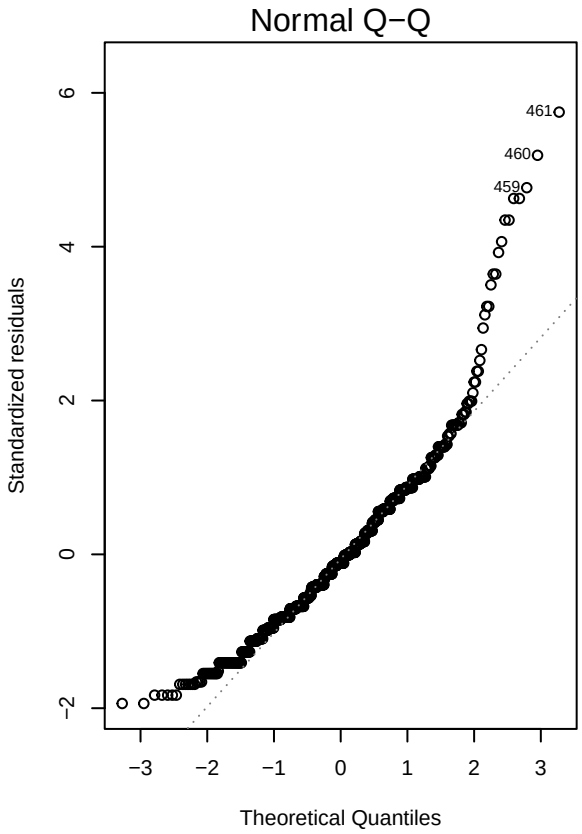
```
##                   2.5 %     97.5 %
## (Intercept)  25.4618834 30.7524023
## Fuel.TypeE  -16.4031524 -7.2397048
## Fuel.TypeX   -0.7960569  4.6580044
## Fuel.TypeZ   -6.0198353 -0.5699963
```

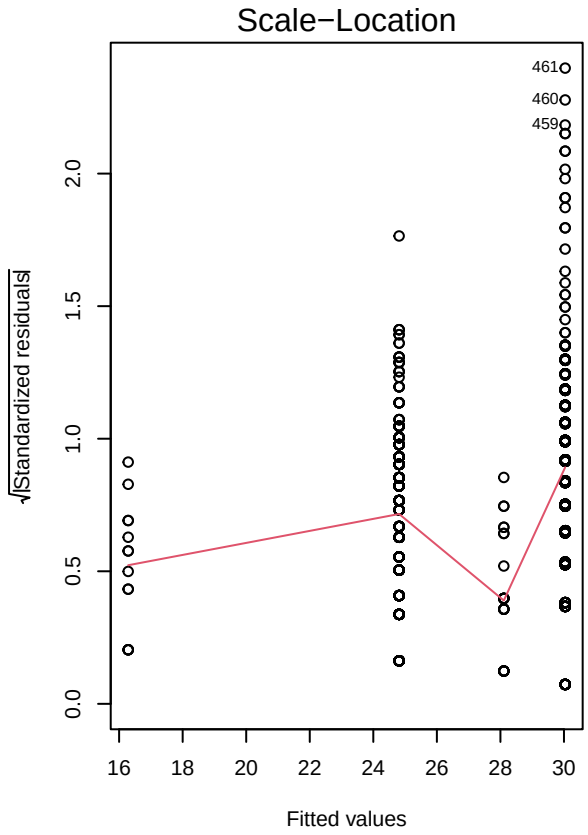
```
#plots for assumption checks
layout(matrix(1:2, 3, 2, byrow = T))
```



```
par(mar = c(5,5,2,1))
plot(fueltype_fit, c(1:3))
```







- The linear model with engine size as the predictor tells us there is a strong (-3.9) and significant (p-val $< .05$) relationship between engine size and fuel consumption. However, the residual normality and homoskedasticity assumptions are violated as seen in the plots
- The linear model with fuel type as the predictor tells us that Fuel Type E and Z have strong (-11.8 and -3.2, respectively) and significant (p-val $< .05$) relationships with Fuel Consumption. Again, we see violated linear model assumptions which will have to be addressed before fitting a final model
- Overall we can infer that Engine Size and Fuel Type will have a significant impact on Fuel Consumption and should be included in the final model

Second Question

Compare cylinder and engine size, how will that affect the fuel consumption?

- Do Cylinders and Engine Size have any relationship? If so, how will it affect the combined fuel consumption?
- Correlation Test to show the strongest correlation and if Cylinder and Engine Size plays a huge part
- linear model between these 2 to determine relationship and p-value
- Statistical test to evaluate the two variables and whether there is a significant association between the categories between the two variables.

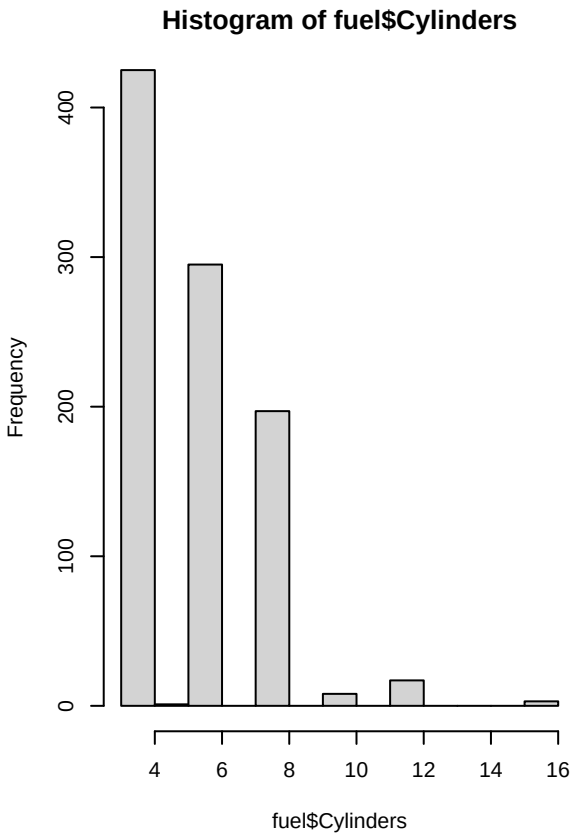
```
#first we check the correlation between Engine Size and Cylinders to see if there is a relationship
relation <- cor(fuel$Engine.Size.L., fuel$Cylinders)
relation
```

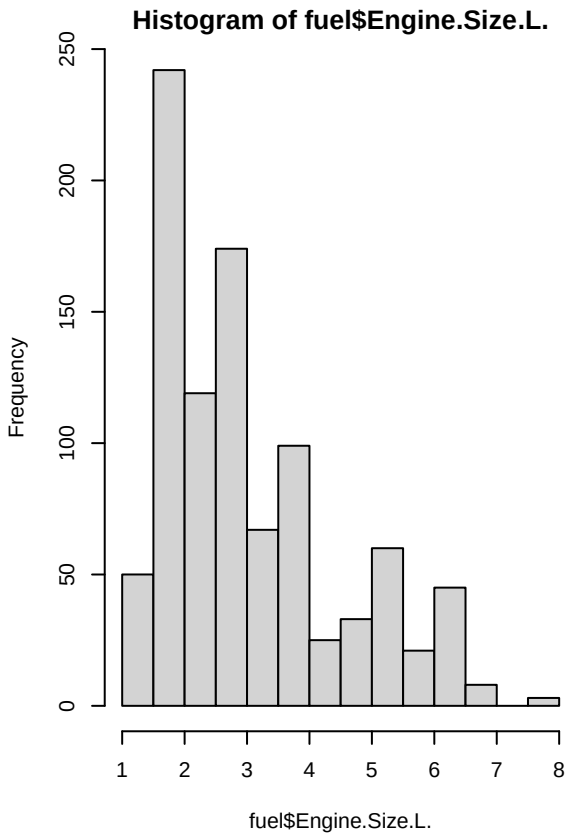
```
## [1] 0.9206976
```

```
#now we fit a model to better assess the relationship with a summary
fit <- lm(fuel$Cylinders ~ fuel$Engine.Size.L.)
summary(fit)
```

```
##
## Call:
## lm(formula = fuel$Cylinders ~ fuel$Engine.Size.L.)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8114 -0.3877 -0.1166  0.5891  4.1177
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.52800    0.06217   24.58   <2e-16 ***
## fuel$Engine.Size.L.   1.29429    0.01786   72.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7547 on 944 degrees of freedom
## Multiple R-squared:  0.8477, Adjusted R-squared:  0.8475
## F-statistic:  5254 on 1 and 944 DF,  p-value: < 2.2e-16
```

```
#We can see that these 2 variables are highly correlated
#The next step is to perform a statistical test, but first we need to figure out which test is best for
layout(matrix(1:2, 3, 2, byrow = T))
par(mar = c(5,5,2,1))
hist(fuel$Cylinders)
hist(fuel$Engine.Size.L.)
```






```
#non-normal distributions with similar variances (right skewed) lead us to the wilcox test
wilcox.test(fuel$Engine.Size.L., fuel$Cylinders)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  fuel$Engine.Size.L. and fuel$Cylinders
## W = 113407, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

- Since the correlation value is very high we can infer that the relationship between Engine Size and Cylinders is very strong
- This can lead to a confounding effect that can damage the model
- We'll make sure that our final model only includes one of these variables to avoid any issues
- The summary of our model confirms that the relationship between Engine Size and Cylinders is strong (1.2) and significant (p-val $< .05$)
- Since the p-value for the Wilcox test was smaller than .05 we can also infer that the true location shift is not equal to 0, meaning engine size and cylinders have different means (this doesn't change the fact that they are highly correlated)
- So, while the variables are different they are still highly correlated and should be watched for confounding effects

Third Question

Which variables are the most effective in determining Combined Fuel Consumption in a linear model?

- Use score functions, like stepwise model selection using AIC, to determine the most effective model.
- This method is appropriate for our data since we have various continuous and coded/binary numerical predictors. Using too many variables could over fit the model, so these score functions will provide the most effective combination of predictors.

```
#first, we turn any categorical variables into factors
fuel$Transmission <- as.factor(fuel$Transmission)
fuel$Vehicle.Class <- as.factor(fuel$Vehicle.Class)

#Since our predictor is is in mpg, we also need to convert CO2.Emissions.g.km to miles
fuel$CO2.Emissions.g.miles <- 1.60934 * fuel$CO2.Emissions.g.km

#The next step is to fit our model.
#Predictors include all variables not related to Fuel Consumption, with a few exceptions

#we removed Model.Year since all the cars in the dataset are 2022

#we removed Model and Make since there is an extensive number of unique values for both (39 for Make an
#Determining the relationship with the response for every combination of Make and Model will be computa
#We're more interested in the factors that impact fuel consumption, not exact vehicles

#we left the Transmission and Vehicle Class variables in the model since they provide more generalized

#now we can start fitting models with our selected predictors
formula <- (Fuel.Consumption.Comb..mpg..) ~
  Vehicle.Class + Engine.Size.L. + Cylinders +
  Transmission + Fuel.Type + CO2.Emissions.g.miles +
  CO2.Rating + Smog.Rating

raw_bestfit <- lm(formula, data = fuel )

#We'd usually verify model assumptions at this point, but
#Since we're using score functions we'll let them choose the best model for us and check/adjust afterwa

#backward
stepwise_bw <- step(raw_bestfit)
```

```
## Start:  AIC=1034
## (Fuel.Consumption.Comb..mpg..) ~ Vehicle.Class + Engine.Size.L. +
##     Cylinders + Transmission + Fuel.Type + CO2.Emissions.g.miles +
##     CO2.Rating + Smog.Rating
##
##                         Df Sum of Sq    RSS    AIC
## - Engine.Size.L.         1      3.66 2575.1 1033.3
## <none>                               2571.5 1034.0
## - Cylinders              1     19.39 2590.9 1039.1
## - Smog.Rating            1     24.81 2596.3 1041.1
## - Vehicle.Class         13     92.94 2664.4 1041.6
## - CO2.Emissions.g.miles  1    303.45 2874.9 1137.5
## - Fuel.Type              3    862.34 3433.8 1301.6
## - CO2.Rating             1   1094.82 3666.3 1367.5
## - Transmission          22   1444.25 4015.7 1411.7
##
## Step:  AIC=1033.34
```

```
## (Fuel.Consumption.Comb..mpg..) ~ Vehicle.Class + Cylinders +
##     Transmission + Fuel.Type + CO2.Emissions.g.miles + CO2.Rating +
##     Smog.Rating
##
##                          Df Sum of Sq    RSS    AIC
## <none>                                  2575.1 1033.3
## - Smog.Rating             1     23.58 2598.7 1040.0
## - Vehicle.Class          13     92.64 2667.8 1040.8
## - Cylinders               1     73.92 2649.1 1058.1
## - CO2.Emissions.g.miles   1    301.96 2877.1 1136.2
## - Fuel.Type               3    871.58 3446.7 1303.1
## - CO2.Rating              1   1091.57 3666.7 1365.7
## - Transmission           22   1443.97 4019.1 1410.5
```

```
#forward
stepwise_fw <- step(lm(Fuel.Consumption.Comb..mpg.. ~ 1, data = fuel),
scope = formula, direction = "forward")
```

```
## Start:  AIC=3859.35
## Fuel.Consumption.Comb..mpg.. ~ 1
##
##                          Df Sum of Sq   RSS    AIC
## + CO2.Rating              1     50326  5488 1667.2
## + CO2.Emissions.g.miles   1     46527  9287 2164.8
## + Transmission           22     29961 25853 3175.3
## + Engine.Size.L.          1     27675 28139 3213.5
## + Cylinders               1     26851 28964 3240.8
## + Vehicle.Class          13     15467 40347 3578.4
## + Smog.Rating             1     12540 43275 3620.6
## + Fuel.Type               3      7892 47922 3721.1
## <none>                                55814 3859.4
##
## Step:  AIC=1667.21
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating
##
##                          Df Sum of Sq    RSS    AIC
## + Transmission           22   1612.11 3876.3 1382.2
## + Fuel.Type               3    903.30 4585.1 1503.1
## + Cylinders               1    120.82 5367.6 1648.2
## + Vehicle.Class          13    228.53 5259.9 1653.0
## + Engine.Size.L.          1     74.80 5413.6 1656.2
## + CO2.Emissions.g.miles   1     27.18 5461.2 1664.5
## <none>                                5488.4 1667.2
## + Smog.Rating             1      0.81 5487.6 1669.1
##
## Step:  AIC=1382.23
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating + Transmission
##
##                          Df Sum of Sq    RSS    AIC
## + Fuel.Type               3    838.51 3037.8 1157.6
## + CO2.Emissions.g.miles   1    283.51 3592.8 1312.4
## + Engine.Size.L.          1     25.42 3850.9 1378.0
## + Vehicle.Class          13    111.13 3765.2 1380.7
## + Cylinders               1      9.62 3866.7 1381.9
```

```
## <none>                              3876.3 1382.2
## + Smog.Rating            1      0.11 3876.2 1384.2
##
## Step:  AIC=1157.64
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating + Transmission + Fuel.Type
##
##                         Df Sum of Sq    RSS    AIC
## + CO2.Emissions.g.miles  1   240.080 2797.7 1081.8
## + Vehicle.Class         13   118.420 2919.4 1146.0
## + Smog.Rating            1    21.033 3016.8 1153.1
## <none>                              3037.8 1157.6
## + Engine.Size.L.         1     2.411 3035.4 1158.9
## + Cylinders              1     1.333 3036.4 1159.2
##
## Step:  AIC=1081.76
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating + Transmission + Fuel.Type +
##     CO2.Emissions.g.miles
##
##                   Df Sum of Sq    RSS    AIC
## + Cylinders        1   115.390 2682.3 1043.9
## + Engine.Size.L.   1    86.541 2711.2 1054.0
## + Vehicle.Class   13   134.148 2663.6 1061.3
## <none>                         2797.7 1081.8
## + Smog.Rating      1     2.984 2794.7 1082.8
##
## Step:  AIC=1043.91
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating + Transmission + Fuel.Type +
##     CO2.Emissions.g.miles + Cylinders
##
##                   Df Sum of Sq    RSS    AIC
## + Vehicle.Class   13    83.593 2598.7 1040.0
## + Smog.Rating      1    14.530 2667.8 1040.8
## <none>                         2682.3 1043.9
## + Engine.Size.L.   1     2.479 2679.8 1045.0
##
## Step:  AIC=1039.96
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating + Transmission + Fuel.Type +
##     CO2.Emissions.g.miles + Cylinders + Vehicle.Class
##
##                   Df Sum of Sq    RSS    AIC
## + Smog.Rating      1   23.5760 2575.1 1033.3
## <none>                         2598.7 1040.0
## + Engine.Size.L.   1    2.4236 2596.3 1041.1
##
## Step:  AIC=1033.34
## Fuel.Consumption.Comb..mpg.. ~ CO2.Rating + Transmission + Fuel.Type +
##     CO2.Emissions.g.miles + Cylinders + Vehicle.Class + Smog.Rating
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         2575.1 1033.3
## + Engine.Size.L.   1    3.6557 2571.5 1034.0
```

```
#both ways
stepwise_both <- step(lm(Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class, data = fuel),
```

```
scope = list(upper = formula,
lower = log(Fuel.Consumption.Comb..mpg..) ~ 1), direction = "both")
```

```
## Start:  AIC=3578.36
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class
##
##                          Df Sum of Sq   RSS    AIC
## + CO2.Rating              1     35087  5260 1653.0
## + CO2.Emissions.g.miles   1     31626  8720 2131.2
## + Transmission           22     19301 21045 3006.7
## + Cylinders               1     16349 23998 3088.9
## + Engine.Size.L.          1     16152 24194 3096.6
## + Smog.Rating             1      9222 31124 3334.9
## + Fuel.Type               3      7909 32437 3377.9
## <none>                                 40347 3578.4
## - Vehicle.Class          13     15467 55814 3859.4
##
## Step:  AIC=1652.98
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class + CO2.Rating
##
##                          Df Sum of Sq   RSS    AIC
## + Transmission           22      1495  3765 1380.7
## + Fuel.Type               3       885  4375 1484.8
## + Cylinders               1        84  5176 1639.8
## + CO2.Emissions.g.miles   1        44  5216 1647.1
## + Engine.Size.L.          1        43  5216 1647.1
## <none>                                  5260 1653.0
## + Smog.Rating             1         0  5260 1655.0
## - Vehicle.Class          13       229  5488 1667.2
## - CO2.Rating              1     35087 40347 3578.4
##
## Step:  AIC=1380.71
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class + CO2.Rating + Transmission
##
##                          Df Sum of Sq     RSS    AIC
## + Fuel.Type               3     845.8  2919.4 1146.0
## + CO2.Emissions.g.miles   1     296.4  3468.8 1305.2
## + Engine.Size.L.          1      43.1  3722.0 1371.8
## + Cylinders               1      23.5  3741.6 1376.8
## <none>                                 3765.2 1380.7
## - Vehicle.Class          13     111.1  3876.3 1382.2
## + Smog.Rating             1       1.2  3764.0 1382.4
## - Transmission           22    1494.7  5259.9 1653.0
## - CO2.Rating              1   17280.2 21045.4 3006.7
##
## Step:  AIC=1146.03
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class + CO2.Rating + Transmission +
##     Fuel.Type
##
##                          Df Sum of Sq     RSS    AIC
## + CO2.Emissions.g.miles   1     255.8  2663.6 1061.3
## + Smog.Rating             1      42.2  2877.1 1134.2
## <none>                                 2919.4 1146.0
```

```
## + Cylinders              1       2.2  2917.2 1147.3
## + Engine.Size.L.         1       0.3  2919.0 1147.9
## - Vehicle.Class         13     118.4  3037.8 1157.6
## - Fuel.Type              3     845.8  3765.2 1380.7
## - Transmission          22    1456.0  4375.3 1484.8
## - CO2.Rating             1   14558.8 17478.2 2837.0
##
## Step:  AIC=1061.28
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class + CO2.Rating + Transmission +
##     Fuel.Type + CO2.Emissions.g.miles
##
##                         Df Sum of Sq    RSS    AIC
## + Cylinders              1     64.84 2598.7 1040.0
## + Engine.Size.L.         1     49.20 2614.4 1045.6
## + Smog.Rating            1     14.49 2649.1 1058.1
## <none>                               2663.6 1061.3
## - Vehicle.Class         13    134.15 2797.7 1081.8
## - CO2.Emissions.g.miles  1    255.81 2919.4 1146.0
## - Fuel.Type              3    805.24 3468.8 1305.2
## - CO2.Rating             1   1166.01 3829.6 1402.8
## - Transmission          22   1672.57 4336.1 1478.3
##
## Step:  AIC=1039.96
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class + CO2.Rating + Transmission +
##     Fuel.Type + CO2.Emissions.g.miles + Cylinders
##
##                         Df Sum of Sq    RSS    AIC
## + Smog.Rating            1     23.58 2575.1 1033.3
## <none>                               2598.7 1040.0
## + Engine.Size.L.         1      2.42 2596.3 1041.1
## - Vehicle.Class         13     83.59 2682.3 1043.9
## - Cylinders              1     64.84 2663.6 1061.3
## - CO2.Emissions.g.miles  1    318.46 2917.2 1147.3
## - Fuel.Type              3    848.21 3446.9 1301.2
## - CO2.Rating             1   1128.04 3726.8 1379.0
## - Transmission          22   1430.54 4029.3 1410.8
##
## Step:  AIC=1033.34
## Fuel.Consumption.Comb..mpg.. ~ Vehicle.Class + CO2.Rating + Transmission +
##     Fuel.Type + CO2.Emissions.g.miles + Cylinders + Smog.Rating
##
##                         Df Sum of Sq    RSS    AIC
## <none>                               2575.1 1033.3
## + Engine.Size.L.         1      3.66 2571.5 1034.0
## - Smog.Rating            1     23.58 2598.7 1040.0
## - Vehicle.Class         13     92.64 2667.8 1040.8
## - Cylinders              1     73.92 2649.1 1058.1
## - CO2.Emissions.g.miles  1    301.96 2877.1 1136.2
## - Fuel.Type              3    871.58 3446.7 1303.1
## - CO2.Rating             1   1091.57 3666.7 1365.7
## - Transmission          22   1443.97 4019.1 1410.5
```

```
#comparing the predictor variables in each score method
sort(names(stepwise_bw$coefficients))
```

```
##  [1] "(Intercept)"
##  [2] "CO2.Emissions.g.miles"
##  [3] "CO2.Rating"
##  [4] "Cylinders"
##  [5] "Fuel.TypeE"
##  [6] "Fuel.TypeX"
##  [7] "Fuel.TypeZ"
##  [8] "Smog.Rating"
##  [9] "TransmissionA6"
## [10] "TransmissionA7"
## [11] "TransmissionA8"
## [12] "TransmissionA9"
## [13] "TransmissionAM6"
## [14] "TransmissionAM7"
## [15] "TransmissionAM8"
## [16] "TransmissionAS10"
## [17] "TransmissionAS5"
## [18] "TransmissionAS6"
## [19] "TransmissionAS7"
## [20] "TransmissionAS8"
## [21] "TransmissionAS9"
## [22] "TransmissionAV"
## [23] "TransmissionAV1"
## [24] "TransmissionAV10"
## [25] "TransmissionAV6"
## [26] "TransmissionAV7"
## [27] "TransmissionAV8"
## [28] "TransmissionM5"
## [29] "TransmissionM6"
## [30] "TransmissionM7"
## [31] "Vehicle.ClassFull-size"
## [32] "Vehicle.ClassMid-size"
## [33] "Vehicle.ClassMinicompact"
## [34] "Vehicle.ClassMinivan"
## [35] "Vehicle.ClassPickup truck: Small"
## [36] "Vehicle.ClassPickup truck: Standard"
## [37] "Vehicle.ClassSpecial purpose vehicle"
## [38] "Vehicle.ClassStation wagon: Mid-size"
## [39] "Vehicle.ClassStation wagon: Small"
## [40] "Vehicle.ClassSubcompact"
## [41] "Vehicle.ClassSUV: Small"
## [42] "Vehicle.ClassSUV: Standard"
## [43] "Vehicle.ClassTwo-seater"
```

```
sort(names(stepwise_fw$coefficients))
```

```
##  [1] "(Intercept)"
##  [2] "CO2.Emissions.g.miles"
##  [3] "CO2.Rating"
##  [4] "Cylinders"
##  [5] "Fuel.TypeE"
##  [6] "Fuel.TypeX"
##  [7] "Fuel.TypeZ"
##  [8] "Smog.Rating"
```

```
##  [9] "TransmissionA6"
## [10] "TransmissionA7"
## [11] "TransmissionA8"
## [12] "TransmissionA9"
## [13] "TransmissionAM6"
## [14] "TransmissionAM7"
## [15] "TransmissionAM8"
## [16] "TransmissionAS10"
## [17] "TransmissionAS5"
## [18] "TransmissionAS6"
## [19] "TransmissionAS7"
## [20] "TransmissionAS8"
## [21] "TransmissionAS9"
## [22] "TransmissionAV"
## [23] "TransmissionAV1"
## [24] "TransmissionAV10"
## [25] "TransmissionAV6"
## [26] "TransmissionAV7"
## [27] "TransmissionAV8"
## [28] "TransmissionM5"
## [29] "TransmissionM6"
## [30] "TransmissionM7"
## [31] "Vehicle.ClassFull-size"
## [32] "Vehicle.ClassMid-size"
## [33] "Vehicle.ClassMinicompact"
## [34] "Vehicle.ClassMinivan"
## [35] "Vehicle.ClassPickup truck: Small"
## [36] "Vehicle.ClassPickup truck: Standard"
## [37] "Vehicle.ClassSpecial purpose vehicle"
## [38] "Vehicle.ClassStation wagon: Mid-size"
## [39] "Vehicle.ClassStation wagon: Small"
## [40] "Vehicle.ClassSubcompact"
## [41] "Vehicle.ClassSUV: Small"
## [42] "Vehicle.ClassSUV: Standard"
## [43] "Vehicle.ClassTwo-seater"
```

```
sort(names(stepwise_both$coefficients))
```

```
##  [1] "(Intercept)"
##  [2] "CO2.Emissions.g.miles"
##  [3] "CO2.Rating"
##  [4] "Cylinders"
##  [5] "Fuel.TypeE"
##  [6] "Fuel.TypeX"
##  [7] "Fuel.TypeZ"
##  [8] "Smog.Rating"
##  [9] "TransmissionA6"
## [10] "TransmissionA7"
## [11] "TransmissionA8"
## [12] "TransmissionA9"
## [13] "TransmissionAM6"
## [14] "TransmissionAM7"
## [15] "TransmissionAM8"
## [16] "TransmissionAS10"
```

```
## [17] "TransmissionAS5"
## [18] "TransmissionAS6"
## [19] "TransmissionAS7"
## [20] "TransmissionAS8"
## [21] "TransmissionAS9"
## [22] "TransmissionAV"
## [23] "TransmissionAV1"
## [24] "TransmissionAV10"
## [25] "TransmissionAV6"
## [26] "TransmissionAV7"
## [27] "TransmissionAV8"
## [28] "TransmissionM5"
## [29] "TransmissionM6"
## [30] "TransmissionM7"
## [31] "Vehicle.ClassFull-size"
## [32] "Vehicle.ClassMid-size"
## [33] "Vehicle.ClassMinicompact"
## [34] "Vehicle.ClassMinivan"
## [35] "Vehicle.ClassPickup truck: Small"
## [36] "Vehicle.ClassPickup truck: Standard"
## [37] "Vehicle.ClassSpecial purpose vehicle"
## [38] "Vehicle.ClassStation wagon: Mid-size"
## [39] "Vehicle.ClassStation wagon: Small"
## [40] "Vehicle.ClassSubcompact"
## [41] "Vehicle.ClassSUV: Small"
## [42] "Vehicle.ClassSUV: Standard"
## [43] "Vehicle.ClassTwo-seater"
```

```
#from these lists we can tell that Transmission and Vehicle Class levels are included in every method
#So we we'll look at the top 10 variables to get a clearer picture of included predictors
sort(names(stepwise_bw$coefficients))[1:10]
```

```
##  [1] "(Intercept)"           "CO2.Emissions.g.miles" "CO2.Rating"
##  [4] "Cylinders"             "Fuel.TypeE"            "Fuel.TypeX"
##  [7] "Fuel.TypeZ"            "Smog.Rating"           "TransmissionA6"
## [10] "TransmissionA7"
```

```
sort(names(stepwise_fw$coefficients))[1:10]
```

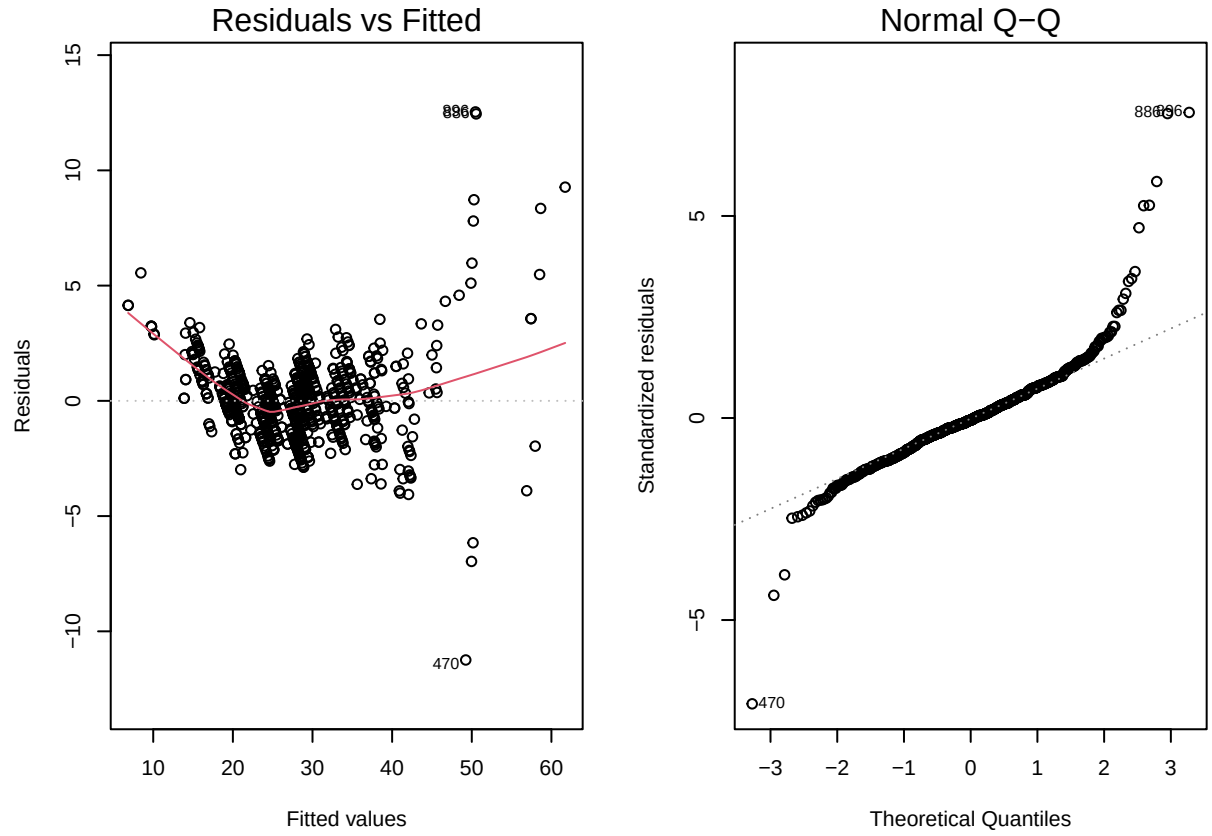
```
##  [1] "(Intercept)"           "CO2.Emissions.g.miles" "CO2.Rating"
##  [4] "Cylinders"             "Fuel.TypeE"            "Fuel.TypeX"
##  [7] "Fuel.TypeZ"            "Smog.Rating"           "TransmissionA6"
## [10] "TransmissionA7"
```

```
sort(names(stepwise_both$coefficients))[1:10]
```

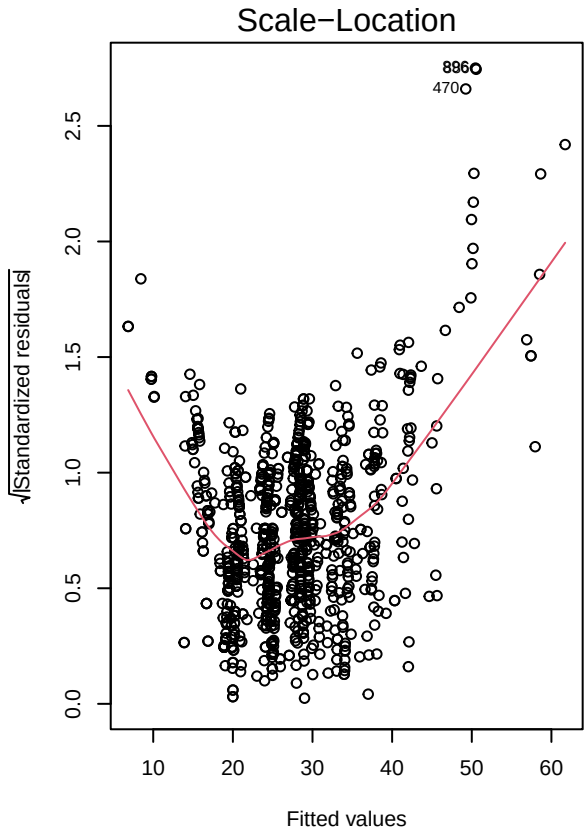
```
##  [1] "(Intercept)"           "CO2.Emissions.g.miles" "CO2.Rating"
##  [4] "Cylinders"             "Fuel.TypeE"            "Fuel.TypeX"
##  [7] "Fuel.TypeZ"            "Smog.Rating"           "TransmissionA6"
## [10] "TransmissionA7"
```

```
#Here we see that all the methods produce the same variables, with the exception of ___ that is missing
#We can be confident that the best fit includes CO2.Emissions.g.miles, CO2.Rating, Cylinders, Fuel.Type

#Now we can fit a model with the recommended predictors and check the assumptions
refined_model <- lm(Fuel.Consumption.Comb..mpg.. ~ CO2.Emissions.g.miles + CO2.Rating + Cylinders + Fuel
layout(matrix(1:2, 3, 2, byrow = T))
par(mar = c(5,5,2,1))
plot(refined_model, 1:3)
```

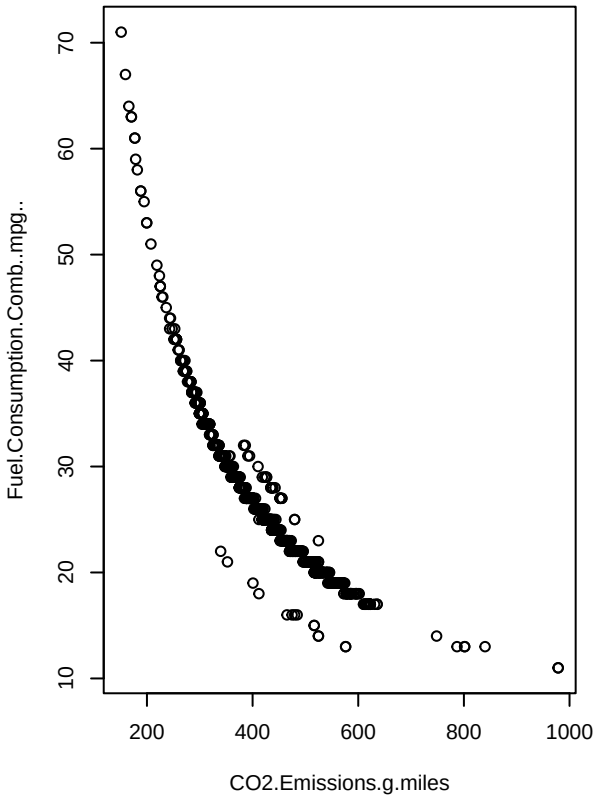
```
## Warning: not plotting observations with leverage one:
##   932
```

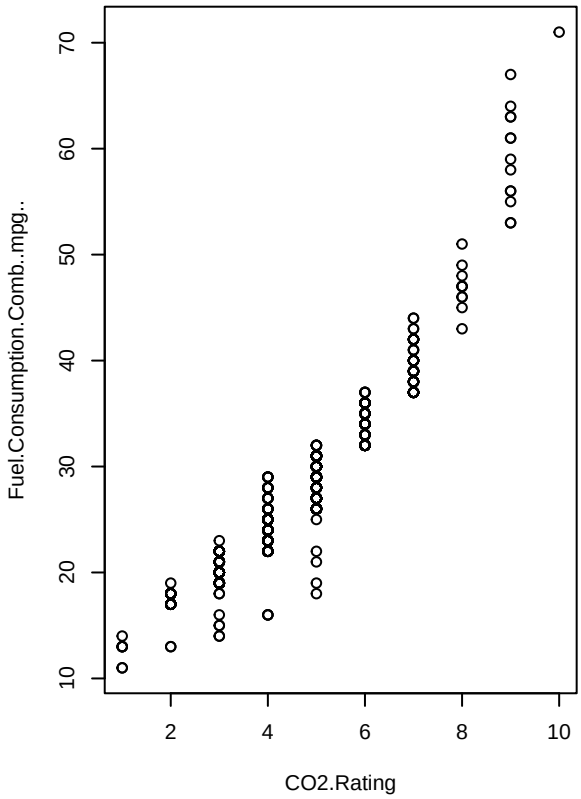



```
#we see heavy deviations from the normal line in the QQplot for residuals, and a parabolic relationship
#we'll fit a GAM to try and resolve these issues and find the final model

#first we need to get a rough estimate of each predictor's relationship with the response (so we know w
#dataframe with only variables of interest
newdata <- fuel[, c(4,6,7,8,12,14,15,16)]
layout(matrix(1:2, 3, 2, byrow = T))
```

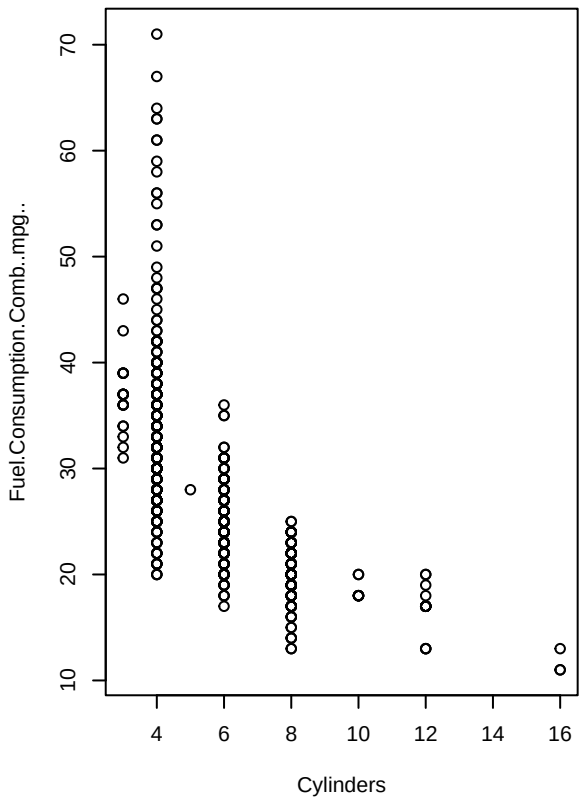


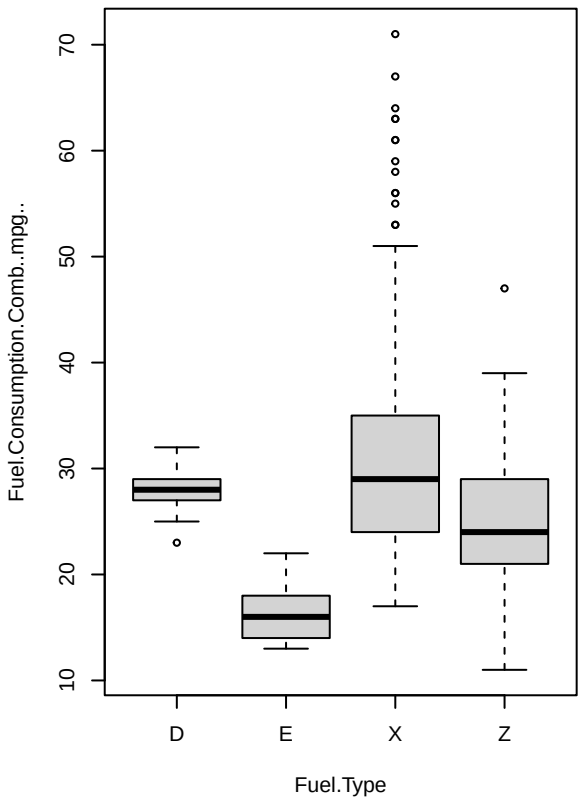

```
par(mar = c(5,5,2,1))
plot(Fuel.Consumption.Comb..mpg.. ~ CO2.Emissions.g.miles, fuel)
plot(Fuel.Consumption.Comb..mpg.. ~ CO2.Rating, fuel)
```






```
plot(Fuel.Consumption.Comb..mpg.. ~ Cylinders, fuel)
plot(Fuel.Consumption.Comb..mpg.. ~ Fuel.Type, fuel)
```





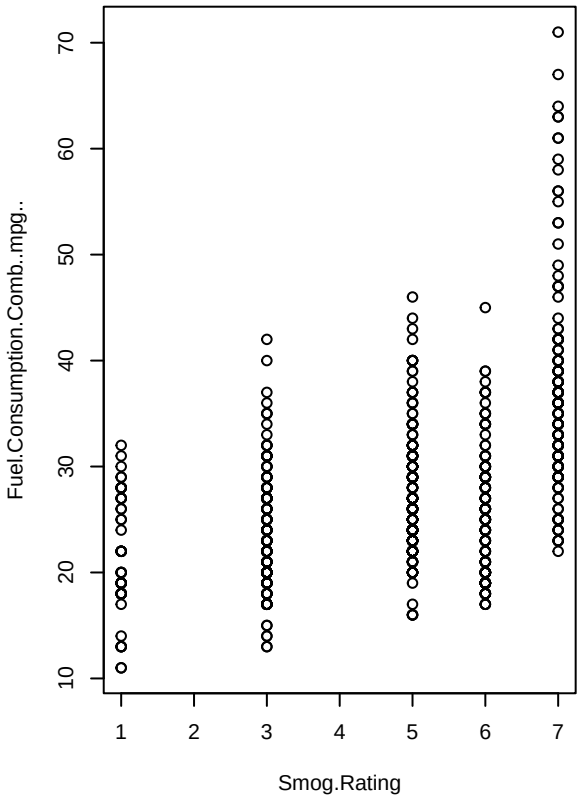

```
plot(Fuel.Consumption.Comb..mpg.. ~ Smog.Rating, fuel)
#Transmission and Vehicle Class can't be plotted but we'll assume a linear (or at the very least a rand


#From these we can see that CO2.Emissions, CO2.Rating, Cylinders, and Smog Rating have a non-linear rel

#Now we can fit an appropriate GAM model that targets the non-linear variables
library(mgcv)
```

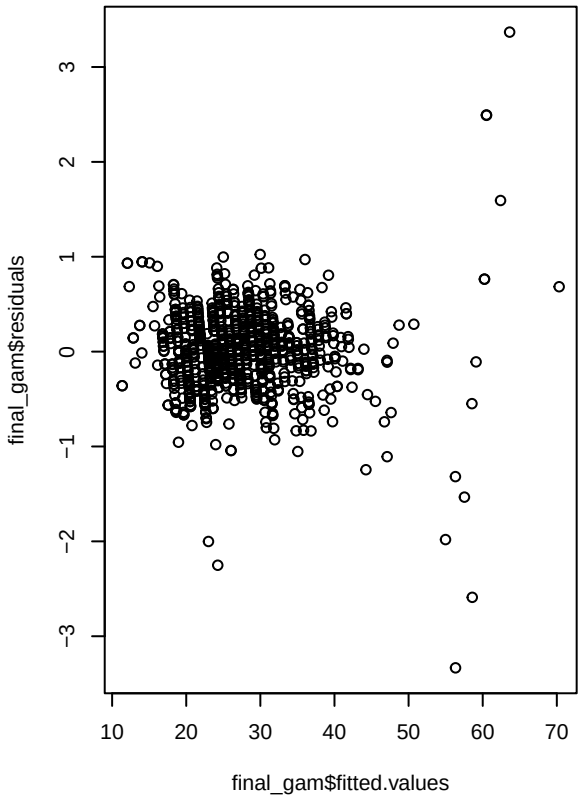
```
## Warning: package 'mgcv' was built under R version 4.2.2
```

```
## Loading required package: nlme
```

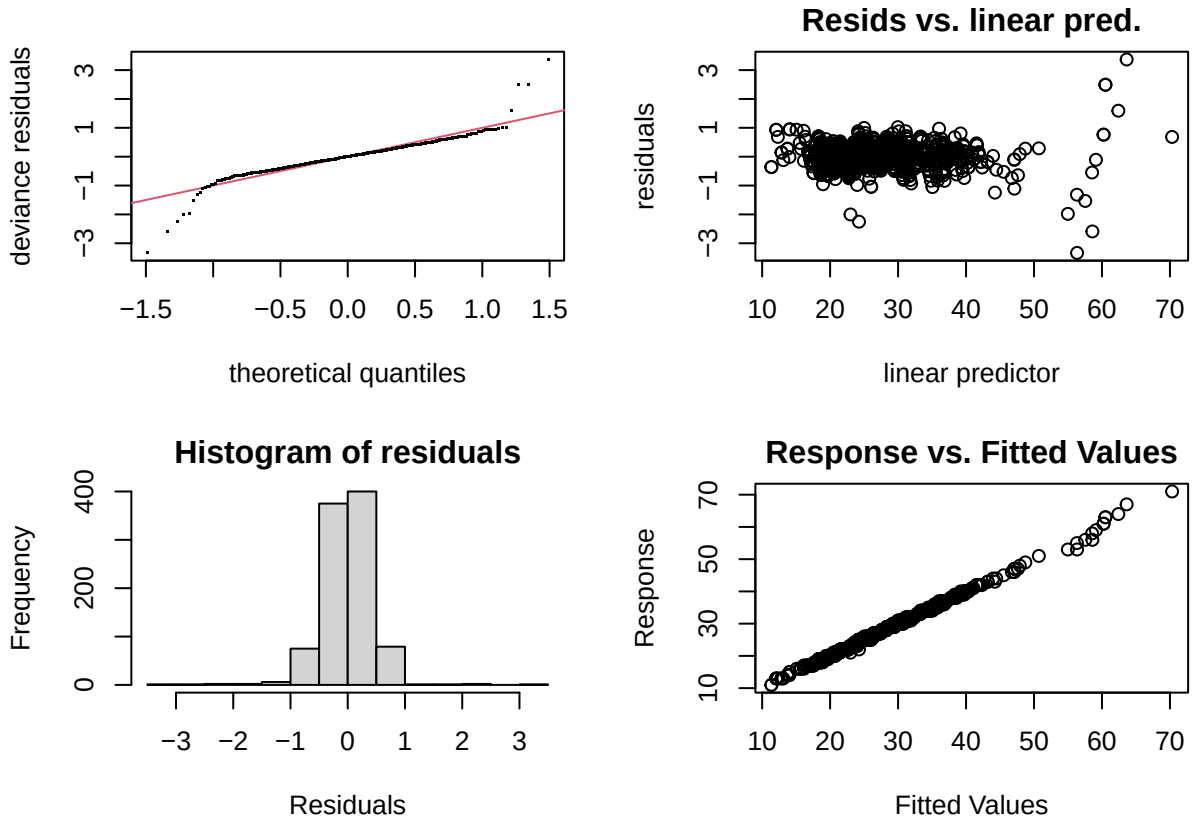
```
## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.
```

```
final_gam <- gam(Fuel.Consumption.Comb..mpg.. ~ s(CO2.Emissions.g.miles, k=5) + s(CO2.Rating, k=5) + s(
plot(final_gam$fitted.values, final_gam$residuals)
```






```
gam.check(final_gam)
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 18 iterations.
## The RMS GCV score gradient at convergence was 3.214518e-07 .
## The Hessian was positive definite.
## Model rank =  55 / 55
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                            k'  edf k-index p-value
## s(CO2.Emissions.g.miles) 4.00 4.00    0.52  <2e-16 ***
## s(CO2.Rating)            4.00 3.99    0.94   0.035 *
## s(Cylinders)             4.00 1.00    0.95   0.035 *
## s(Smog.Rating)           4.00 2.23    0.97   0.180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(final_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

```
## Fuel.Consumption.Comb..mpg.. ~ s(CO2.Emissions.g.miles, k = 5) +
##     s(CO2.Rating, k = 5) + s(Cylinders, k = 5) + Fuel.Type +
##     s(Smog.Rating, k = 5) + Transmission + Vehicle.Class
##
## Parametric coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3.086e+01  1.238e-01 249.311  < 2e-16 ***
## Fuel.TypeE                          -1.049e+01  1.733e-01 -60.509  < 2e-16 ***
## Fuel.TypeX                          -3.551e+00  1.160e-01 -30.624  < 2e-16 ***
## Fuel.TypeZ                          -3.597e+00  1.185e-01 -30.350  < 2e-16 ***
## TransmissionA6                       2.651e-01  1.306e-01   2.030  0.04263 *
## TransmissionA7                      -8.859e-02  4.640e-01  -0.191  0.84862
## TransmissionA8                      -1.343e-01  8.665e-02  -1.550  0.12146
## TransmissionA9                       8.784e-02  9.562e-02   0.919  0.35849
## TransmissionAM6                      9.954e-01  2.220e-01   4.484 8.28e-06 ***
## TransmissionAM7                     -4.612e-02  1.050e-01  -0.439  0.66053
## TransmissionAM8                      1.393e-01  1.099e-01   1.267  0.20540
## TransmissionAS10                    -2.386e-02  8.637e-02  -0.276  0.78240
## TransmissionAS5                     -6.387e-01  3.311e-01  -1.929  0.05407 .
## TransmissionAS6                      2.103e-02  1.099e-01   0.191  0.84829
## TransmissionAS7                     -9.843e-02  1.928e-01  -0.511  0.60978
## TransmissionAS8                     -1.143e-01  8.378e-02  -1.364  0.17284
## TransmissionAS9                     -3.173e-02  1.333e-01  -0.238  0.81191
## TransmissionAV                      -6.087e-02  1.283e-01  -0.474  0.63526
## TransmissionAV1                      3.489e-01  2.603e-01   1.341  0.18035
## TransmissionAV10                    -5.146e-02  1.887e-01  -0.273  0.78516
## TransmissionAV6                     -5.651e-01  1.874e-01  -3.016  0.00263 **
## TransmissionAV7                      3.471e-02  1.743e-01   0.199  0.84222
## TransmissionAV8                     -1.161e-01  1.308e-01  -0.888  0.37504
## TransmissionM5                      -4.558e-01  2.250e-01  -2.026  0.04309 *
## TransmissionM6                      -5.507e-02  1.007e-01  -0.547  0.58454
## TransmissionM7                       5.439e-04  1.584e-01   0.003  0.99726
## Vehicle.ClassFull-size               1.066e-01  8.609e-02   1.239  0.21576
## Vehicle.ClassMid-size                4.179e-03  7.219e-02   0.058  0.95385
## Vehicle.ClassMinicompact            -4.621e-02  1.008e-01  -0.458  0.64671
## Vehicle.ClassMinivan                 3.343e-01  1.875e-01   1.782  0.07503 .
## Vehicle.ClassPickup truck: Small    -1.917e-01  1.390e-01  -1.380  0.16808
## Vehicle.ClassPickup truck: Standard  1.281e-02  9.123e-02   0.140  0.88835
## Vehicle.ClassSpecial purpose vehicle -2.224e-01  1.537e-01  -1.447  0.14828
## Vehicle.ClassStation wagon: Mid-size  8.322e-02  1.764e-01   0.472  0.63720
## Vehicle.ClassStation wagon: Small    -2.043e-01  1.269e-01  -1.610  0.10764
## Vehicle.ClassSubcompact              -1.328e-01  7.823e-02  -1.698  0.08993 .
## Vehicle.ClassSUV: Small               2.706e-02  7.404e-02   0.365  0.71489
## Vehicle.ClassSUV: Standard            2.801e-02  8.106e-02   0.346  0.72978
## Vehicle.ClassTwo-seater              -1.441e-01  1.003e-01  -1.438  0.15088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                             edf Ref.df        F p-value
## s(CO2.Emissions.g.miles) 3.996   4.00 1878.593  <2e-16 ***
## s(CO2.Rating)            3.991   4.00  225.530  <2e-16 ***
## s(Cylinders)             1.000   1.00    0.136  0.7128
## s(Smog.Rating)           2.229   2.72    1.959  0.0849 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.997   Deviance explained = 99.7%
## GCV = 0.21828  Scale est. = 0.20669   n = 946
```

```
#assumptions have been cleared and we're left with a final model!
#QQplot normality is met, a linear trend is visible, and the residuals don't show any clear trend which
```

- Using the results from our three score functions we can infer that the best model will include predictor variables CO2.Emissions.g.miles, CO2.Rating, Cylinders, Fuel.Type, Smog.Rating, Transmission, and Vehicle.Class.
- We also note that only 1 of Engine Size or Cylinders is included to avoid confounding effects
- Adding more predictor variables will not produce improved results, so we can conclude that we have the most effective predictor variables
- More specifically, a GAM is the best fit with predictors CO2.Emissions.g.miles, CO2.Rating, Cylinders, and Smog.Rating being transformed with thin plate regression splines and a basis dimension of 5
- From the final GAM summary we see that 99.7% of the deviance in the response is explained by our model, and that Fuel.Type, TransmissionA6, TransmissionAM6, TransmissionAV6, CO2.Emissions, and CO2.Rating are particularly significant predictors
- Interpret some of the model findings to end the presentation?