Homwork-2

Econometrics of Data Science

Name: Mohammad Shaik

Student Id:16319967

For the available data, estimate a linear regression model with mpg (miles per gallon) as a function of selected variables, including (but not limited to) binary variables for model year and the variable "foreign" (Think about your model!).

1. State and justify your model.

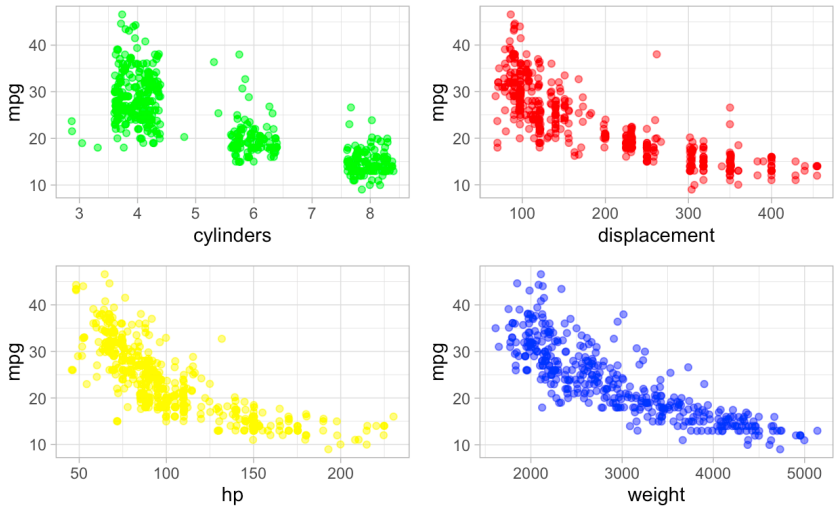
As given, we need to build a Linear Regression Model by considering mpg variable as a function. Since the dataset is having more than two independent variables, the model to be build is Multilinear Regression Model. Therefore, the Linear Regression Model build on the top of the car dataset with mpg as a function is as follows: mpg = β_0+$\beta_1$cylinders + β_2 displacement +$\beta_3$hp+$\beta_4$weight +$\beta_5$acceleration +β_6 modelyr+$\beta_7$origin +$\beta_8$foreign + $\beta_9$modyr80+ β_{10}modyr81+β_{11} modyr82+u From the above equation, mpg is a Dependent variable cylinders, displacement, hp, …….. , modyr82 are Independent variables.

β_0, β_1, β_2, β_3, β_4, ….., β_{11} are the coefficients. u

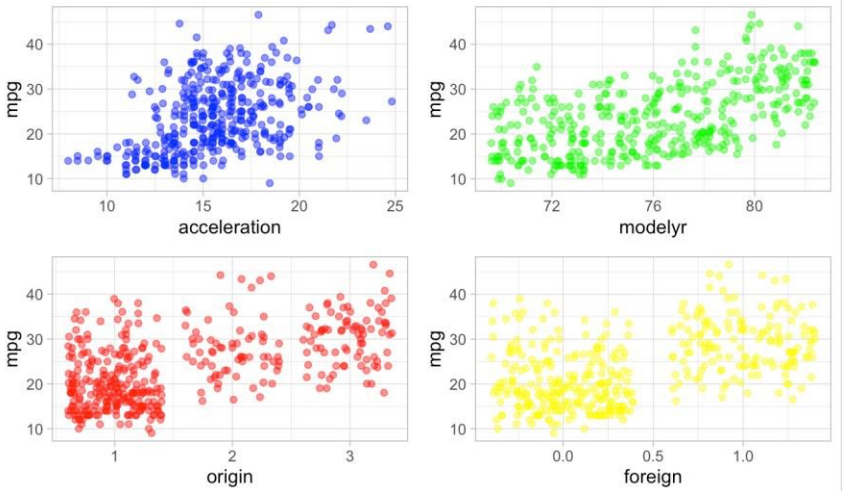
is a y-intercept.

When we built the model with all the independent variables from the dataset, Adjusted R-squared: 0.8479 and build multiple models with different sets of variables. Also by plotting different graphs, we found the variables which result in better R-squared value. Hence finally by plotting Correlation graph, we conclude the model with the above independent variables. From the below graphs, all the independent variables are directly proportional to target variable mpg and the 1980's are heavily related to mpg than 1970.

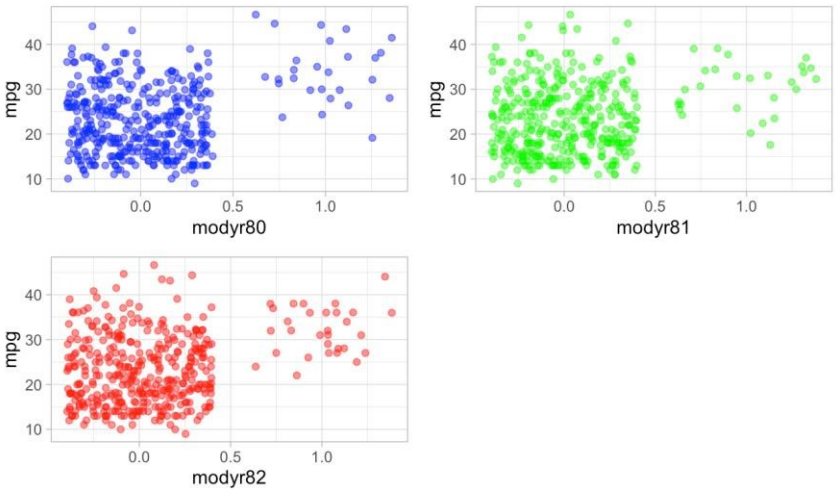
1. Correlation between mpg and cylinders / displacement / hp / weight



2. Correlation between mpg and acceleration/modelyr/origin/foreign



3. Correlation between mpg and modyr80/modyr81/modyr82



2. Present summary statistics (min, max, standard deviation, and selected quantiles) for the variables in your model and briefly comment.

Ans.

The summary statistics for the variables that we build for the model:

```
      mpg          cylinders      displacement         hp            weight      acceleration
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00
 1st Qu.:17.38   1st Qu.:4.000   1st Qu.:100.2   1st Qu.: 75.0   1st Qu.:2220   1st Qu.:13.80
 Median :23.00   Median :4.000   Median :146.0   Median : 92.0   Median :2792   Median :15.50
 Mean   :23.50   Mean   :5.457   Mean   :193.4   Mean   :104.2   Mean   :2969   Mean   :15.57
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:263.2   3rd Qu.:125.0   3rd Qu.:3610   3rd Qu.:17.12
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80
    modelyr          origin         foreign          modyr80           modyr81
 Min.   :70.00   Min.   :1.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:73.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :76.00   Median :1.000   Median :0.0000   Median :0.00000   Median :0.00000
 Mean   :75.95   Mean   :1.591   Mean   :0.3813   Mean   :0.06818   Mean   :0.07071
 3rd Qu.:79.00   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :82.00   Max.   :3.000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
    modyr82
 Min.   :0.00000
 1st Qu.:0.00000
 Median :0.00000
 Mean   :0.07576
 3rd Qu.:0.00000
 Max.   :1.00000
```

```
       mpg    cylinders displacement           hp       weight acceleration      modelyr
 7.7875490    1.7035110  104.6421834   38.4058785  850.0533252    2.7685813    3.6802891
    origin      foreign      modyr80      modyr81      modyr82
 0.8140029    0.4863236    0.2523765    0.2566592    0.2649446
```

Table representation:

	mpg	cylinders	displacement	hp	weight	acceleration	modelyr	origin	foreign	modyr80	modyr81	modyr82
	Min. : 9.00	Min. :3.000	Min. : 68.0	Min. : 46.0	Min. :1613	Min. : 8.00	Min. :70.00	Min. :1.000	Min. :0.0000	Min. :0.00000	Min. :0.00000	Min. :0.00000
	1st Qu.:17.38	1st Qu.:4.000	1st Qu.:100.2	1st Qu.: 75.0	1st Qu.:2220	1st Qu.:13.80	1st Qu.:73.00	1st Qu.:1.000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000
	Median :23.00	Median :4.000	Median :146.0	Median : 92.0	Median :2792	Median :15.50	Median :76.00	Median :1.000	Median :0.0000	Median :0.00000	Median :0.00000	Median :0.00000
	Mean : 23.50	Mean :5.457	Mean :193.4	Mean :104.2	Mean :2969	Mean :15.57	Mean :75.95	Mean :1.591	Mean :0.3813	Mean :0.06818	Mean :0.07071	Mean :0.07576
	3rd Qu.:29.00	3rd Qu.:8.000	3rd Qu.:263.2	3rd Qu.:125.0	3rd Qu.:3610	3rd Qu.:17.12	3rd Qu.:79.00	3rd Qu.:2.000	3rd Qu.:1.0000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000
	Max. :46.60	Max. :8.000	Max. :455.0	Max. :230.0	Max. :5140	Max. :24.80	Max. :82.00	Max. :3.000	Max. :1.0000	Max. :1.00000	Max. :1.00000	Max. :1.00000
	S.dev :7.793952	S.dev :1.7029297	S.dev :104.6902392	S.dev :38.4027282	S.dev :851.0898765	S.dev :2.7838445	S.dev :3.6840323	S.dev :0.8160446	S.dev :0.4867007	S.dev: 0.2520809	S.dev: 0.2563595	S.dev: 0.2646372

By using the summary() function, the statistics will be displayed as below and then furthermore, the mean is used to compute the standard deviation or by using apply() function we got the standard deviation. From the above stats, there is a large discrepancy between the maximum and minimum values of most independent variables. ‘mpg’, and variables i.e., ‘displacement’, 'weight', 'horsepower' are terms used interchangeably. The mean and median in ‘mpg’, ‘cylinders’, ‘acceleration’, ‘modelyr’, ‘origin’ are nearly identical. Furthermore, the stats for 1980’s are mostly similar to each other.

3. Using R, generate your model results and interpret them. Interpret the coefficients in terms of the sign, size (magnitude), and significance.

```
Call:
lm(formula = mpg ~ cylinders + displacement + hp + weight + acceleration +
    modelyr + origin + foreign + modyr80 + modyr81 + modyr82,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2865 -1.9114  0.0608  1.8308 11.4348

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5186714  5.5019066   0.094  0.92494
cylinders    -0.1279539  0.3071903  -0.417  0.67726
displacement  0.0181537  0.0072499   2.504  0.01269 *
hp           -0.0225027  0.0129489  -1.738  0.08305 .
weight       -0.0065177  0.0006185 -10.538  < 2e-16 ***
acceleration  0.0681151  0.0912195   0.747  0.45569
modelyr       0.5152224  0.0674397   7.640 1.75e-13 ***
origin        0.0499772  0.5277433   0.095  0.92460
foreign       2.2905813  0.9287746   2.466  0.01409 *
modyr80       5.0942997  0.7327259   6.953 1.55e-11 ***
modyr81       1.9260898  0.7438494   2.589  0.00998 **
modyr82       3.0950799  0.7828220   3.954 9.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.11 on 384 degrees of freedom
Multiple R-squared:  0.845,     Adjusted R-squared:  0.8406
F-statistic: 190.3 on 11 and 384 DF,  p-value: < 2.2e-16
```

From the model summary, we can see that the coefficients weight, modelyr, modyr80, modyr82 are highly significant at 0.1% level of significance level. Therfore, the Null-Hypothesis can be rejected as we can insights from these variables more accurately. Similarly, modyr81 is also highly significant at 1% level of significance level and displacement, foreign are significant at 5% level of significance level. But also the other independent variables contribute towards the target variable as per correlation graph. Moreover, these independent variables as per the alternative hypothesis is that they are jointly significant to each other

Further, all these explanatory variables account for 84% of the version in mpg even as the relaxation are accounted for with the aid of using random factors and are mostly having positively related with each other as we can see the sign and size. F-

statistic is 190 which means the explanatory variables coefficients are jointly significant.

4. Do American cars run less mileage per gallon than foreign cars? Use an appropriate statistical test

To find that American cars run less mpg than foreign cars, firstly, we are computing the sample statistics for foreign and non-foreign vehicles by filtering from the actual dataset where we get mean, standard deviation, and number of observations to use in the significance of the two-sample z-test of the hypothesis.

Later we need to form a hypothesis and run z-tests from the BSDA package which we installed and then draw conclusions from these filtered. From the below Two-sample z-Test, p value is 0.6054 which is very high

Hypothesis: The term Hypothesis means where we can insights or trends from the dataset

Null hypothesis: There is high significant difference in the average mpg of foreign and non-foreign cars.

Alternative hypothesis: The average mpg for American cars runs less than the foreign cars.

```
        Two-sample z-Test

data:  foreign$mpg and nonforeign$mpg
z = 0.26737, p-value = 0.6054
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       NA 65.05771
sample estimates:
mean of x mean of y
 29.12980  20.03347
```

Using the p-value, where it is more than 5%, we cannot reject the null hypothesis and conclude that American vehicles runs less than foreign ones.

5. Do the model year variables jointly have explanatory power? Use an appropriate statistical test.

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.4142 -1.8877 -0.0098  1.8040 11.8098

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.503e+01  6.615e+01  -1.437  0.15166
cylinders    -1.804e-01  3.021e-01  -0.597  0.55090
displacement  1.716e-02  7.170e-03   2.394  0.01717 *
hp           -2.270e-02  1.347e-02  -1.685  0.09277 .
weight       -6.175e-03  6.429e-04  -9.604  < 2e-16 ***
acceleration  6.796e-02  8.997e-02   0.755  0.45048
modelyr       1.712e+00  8.104e-01   2.112  0.03532 *
origin        4.661e-02  5.205e-01   0.090  0.92870
foreign       2.451e+00  9.144e-01   2.680  0.00768 **
modyr70       1.193e+01  9.376e+00   1.272  0.20404
modyr71       1.135e+01  8.550e+00   1.328  0.18513
modyr72       8.577e+00  7.744e+00   1.108  0.26874
modyr73       6.270e+00  6.933e+00   0.904  0.36635
modyr74       6.778e+00  6.121e+00   1.107  0.26881
modyr75       4.326e+00  5.317e+00   0.814  0.41633
modyr76       3.415e+00  4.508e+00   0.757  0.44925
modyr77       3.190e+00  3.724e+00   0.857  0.39208
modyr78       1.316e+00  2.918e+00   0.451  0.65240
modyr79       1.915e+00  2.152e+00   0.890  0.37423
modyr80       4.310e+00  1.415e+00   3.047  0.00248 **
modyr81              NA         NA      NA       NA
modyr82              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.037 on 376 degrees of freedom
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8479
F-statistic: 116.9 on 19 and 376 DF,  p-value: < 2.2e-16
```

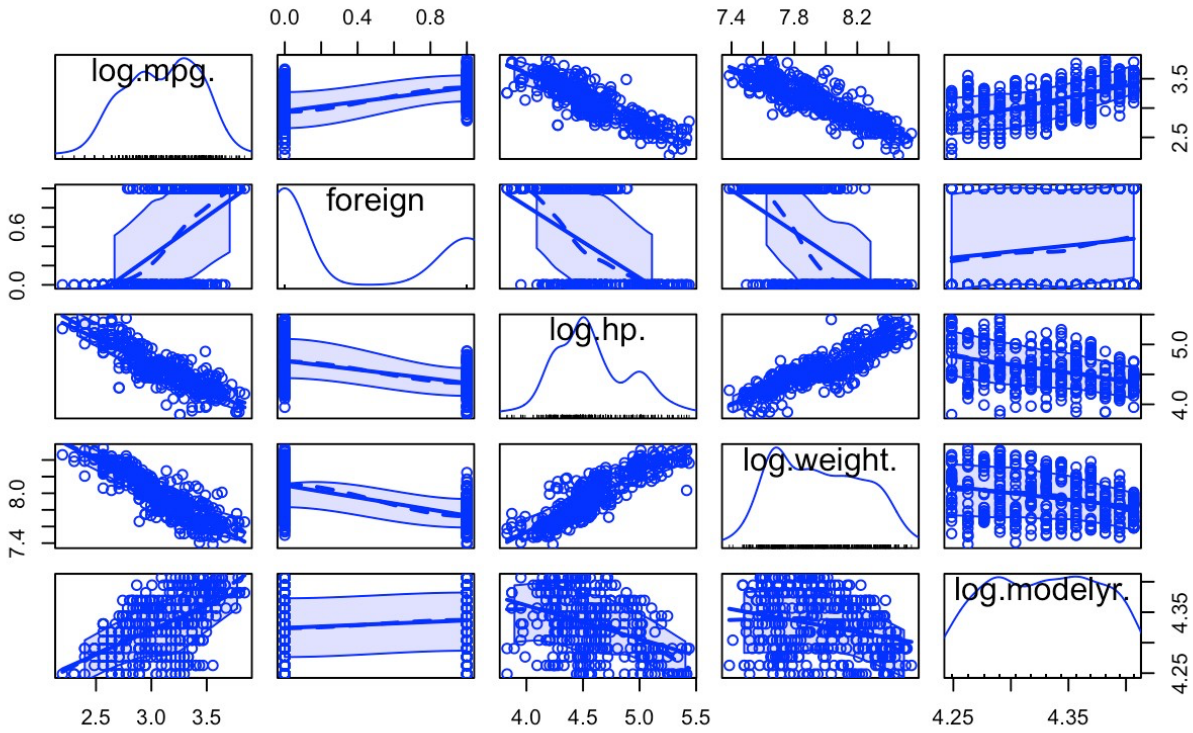
```
Call:
lm(formula = mpg ~ cylinders + displacement + hp + weight + acceleration +
    modelyr + origin + foreign + modyr80 + modyr81 + modyr82,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2865 -1.9114  0.0608  1.8308 11.4348

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5186714  5.5019066   0.094  0.92494
cylinders    -0.1279539  0.3071903  -0.417  0.67726
displacement  0.0181537  0.0072499   2.504  0.01269 *
hp           -0.0225027  0.0129489  -1.738  0.08305 .
weight       -0.0065177  0.0006185 -10.538  < 2e-16 ***
acceleration  0.0681151  0.0912195   0.747  0.45569
modelyr       0.5152224  0.0674397   7.640 1.75e-13 ***
origin        0.0499772  0.5277433   0.095  0.92460
foreign       2.2905813  0.9287746   2.466  0.01409 *
modyr80       5.0942997  0.7327259   6.953 1.55e-11 ***
modyr81       1.9260898  0.7438494   2.589  0.00998 **
modyr82       3.0950799  0.7828220   3.954 9.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.11 on 384 degrees of freedom
Multiple R-squared:  0.845,     Adjusted R-squared:  0.8406
F-statistic: 190.3 on 11 and 384 DF,  p-value: < 2.2e-16
```

When we considered all the independent variables i.e., the entire dataset and building the model as shown in fig1, we can see the p-values are all greater than 5% threshold of significance.

However, after playing with the dataset and building multiple models by considering different set of independent variables as shown in fig2, the variables for 1970’s doesn’t have any statistical explanatory power, but modyr80, modyr81, modyr82 have more statistical power.

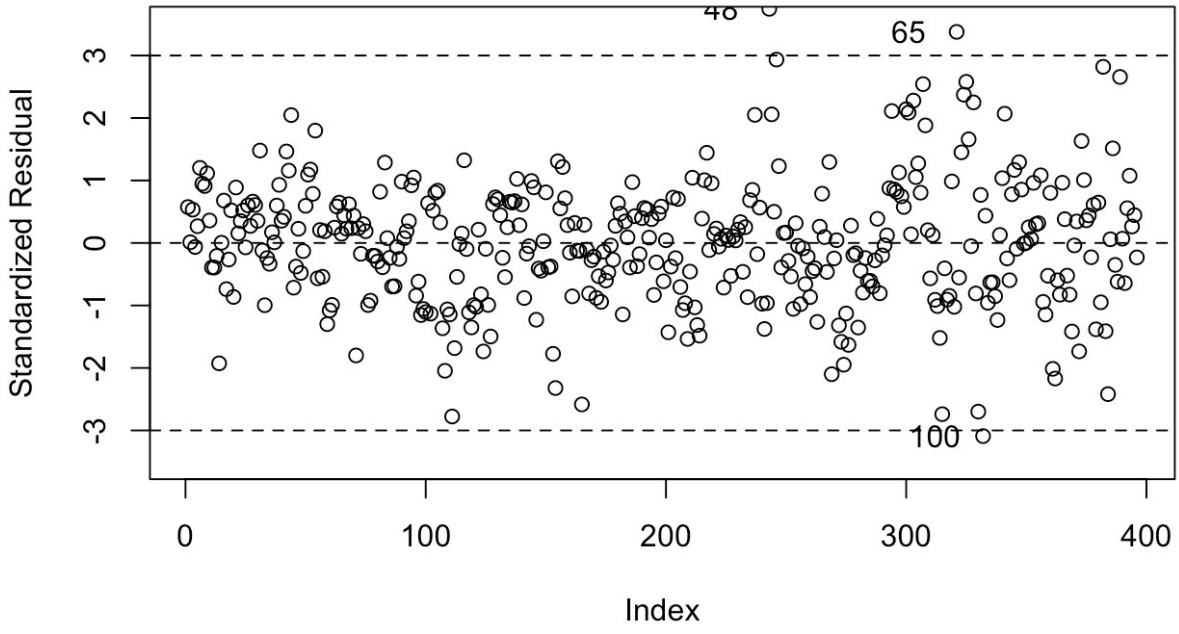
6. Check the assumptions of the linear regression model.

Scatterplot:



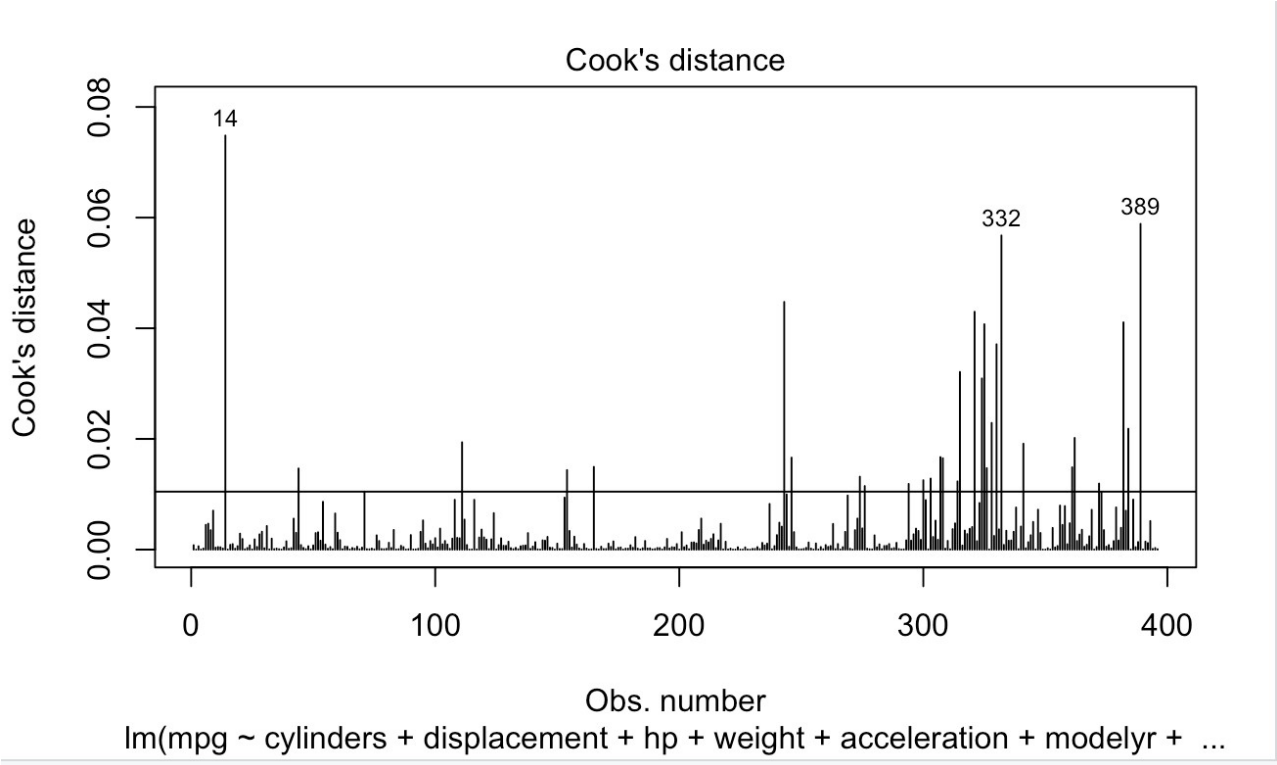
Plotting a Scatterplot graph by considering log as because to remove non-linearity. From the below plot each represents w.r.t two variables.

Outliers Identification:



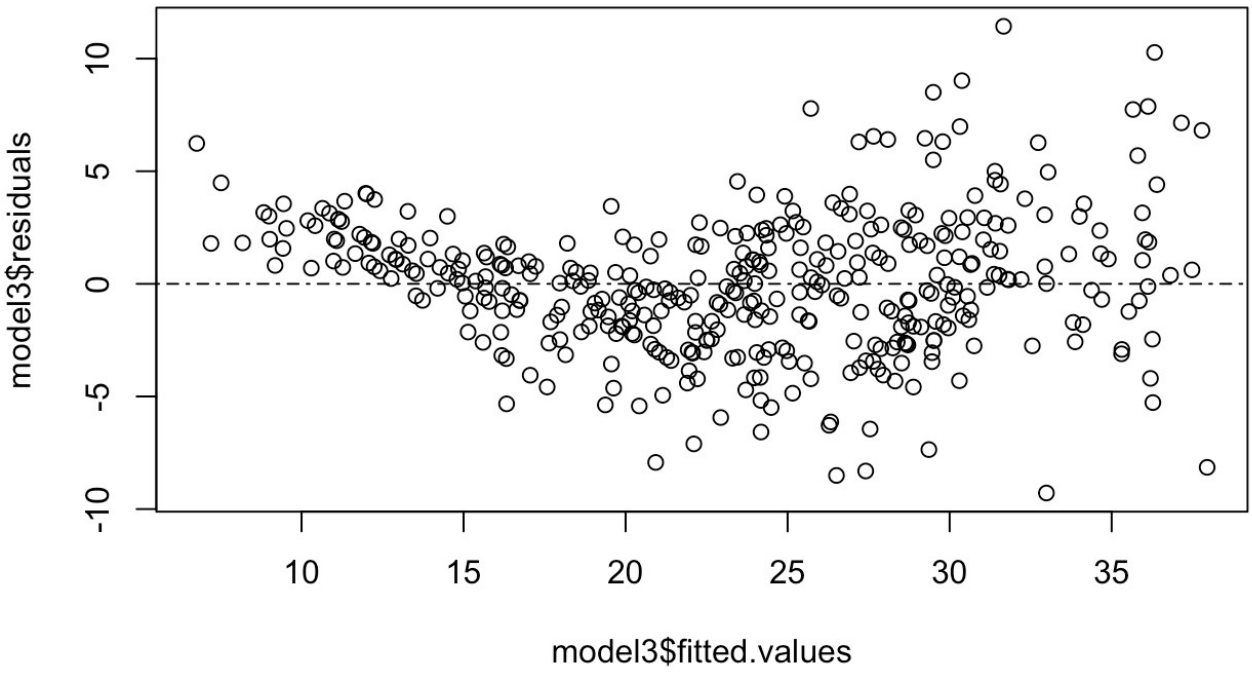
First we plot the graph without residual line then included. Finally, print the index number of respective observation in order to remove the outliers by considering Horse power(hp) variable.

Cooks distance:



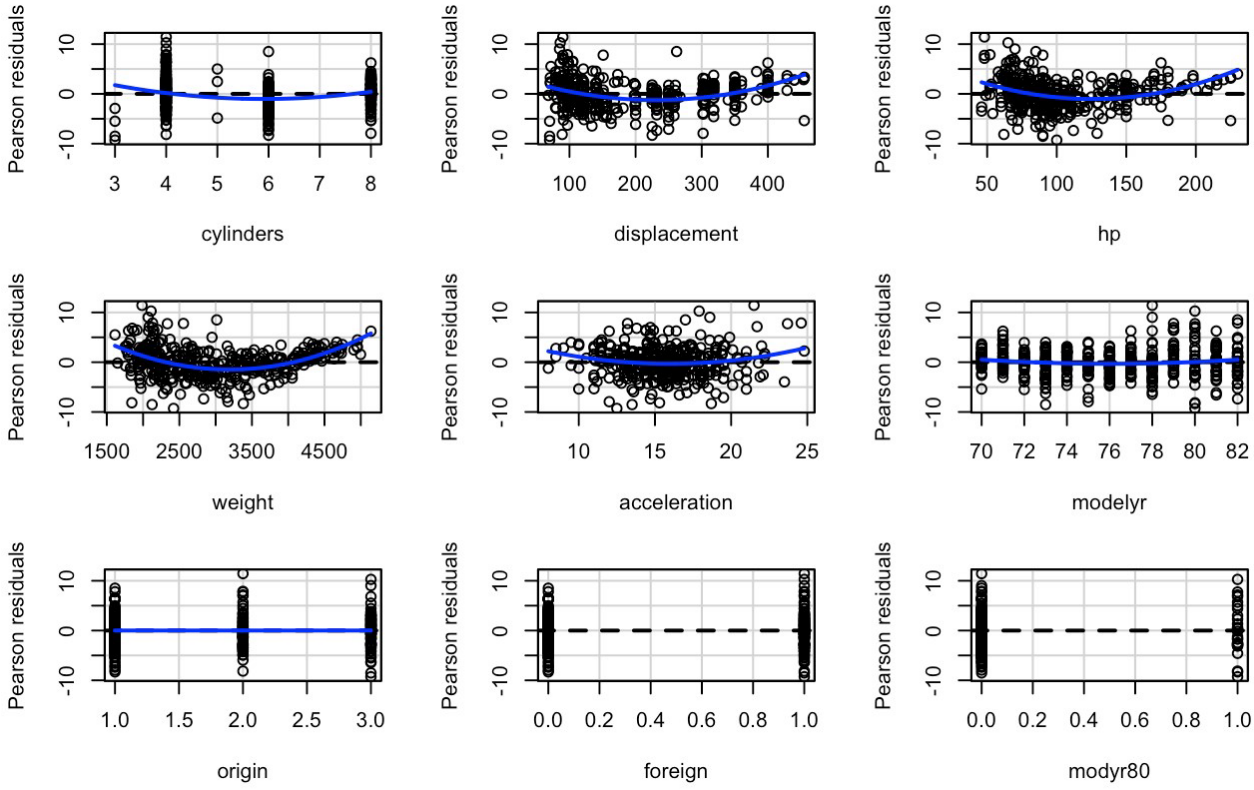
The plot above shows the influential observations having cooks distance w.r.t the model build.

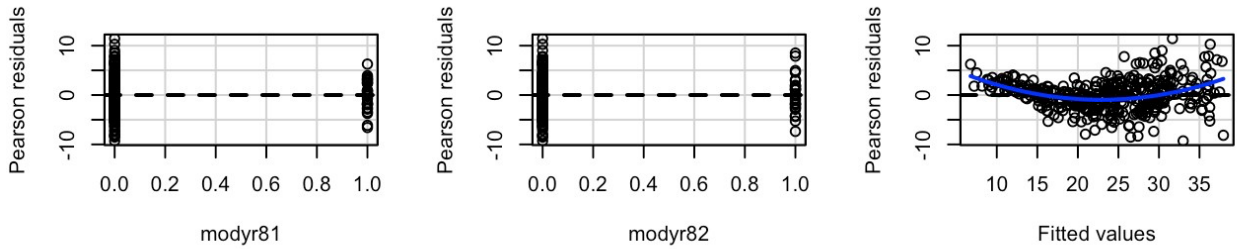
Checking homoscedasticity:



Since the residual plot from the figure is approximately horizontal at zero and hence linearity is assumed where a fitted pattern will not be shown.

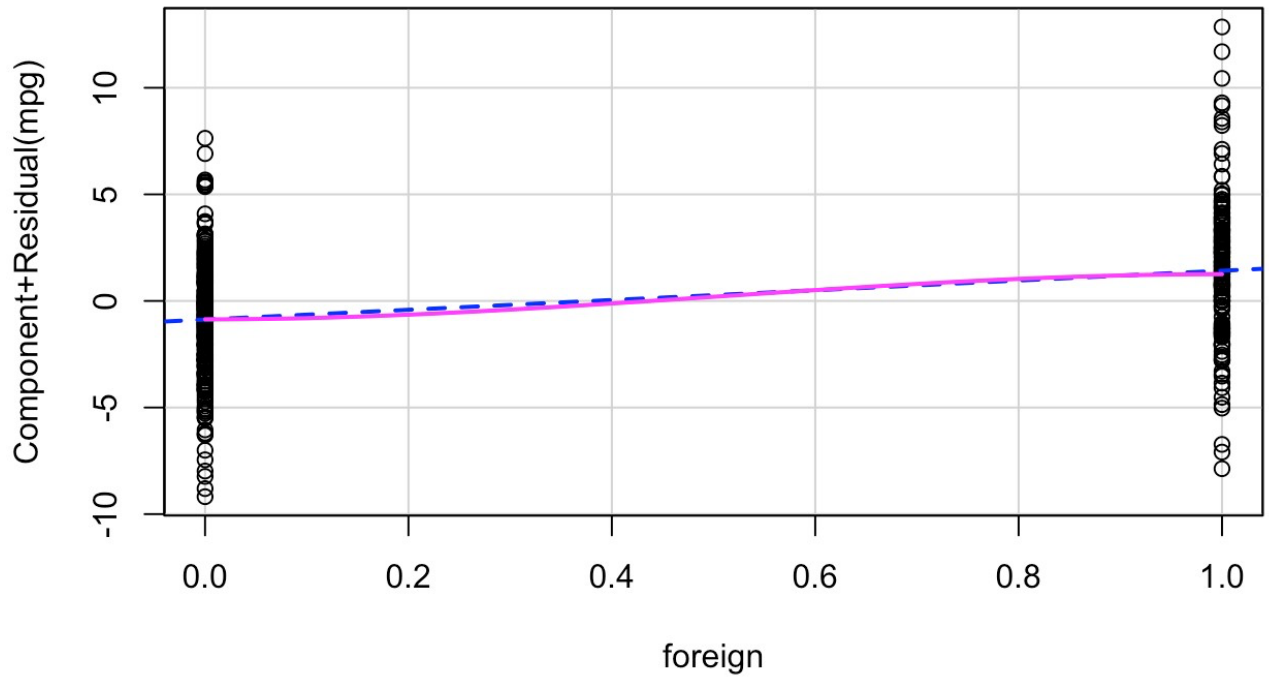
Checking linearity:



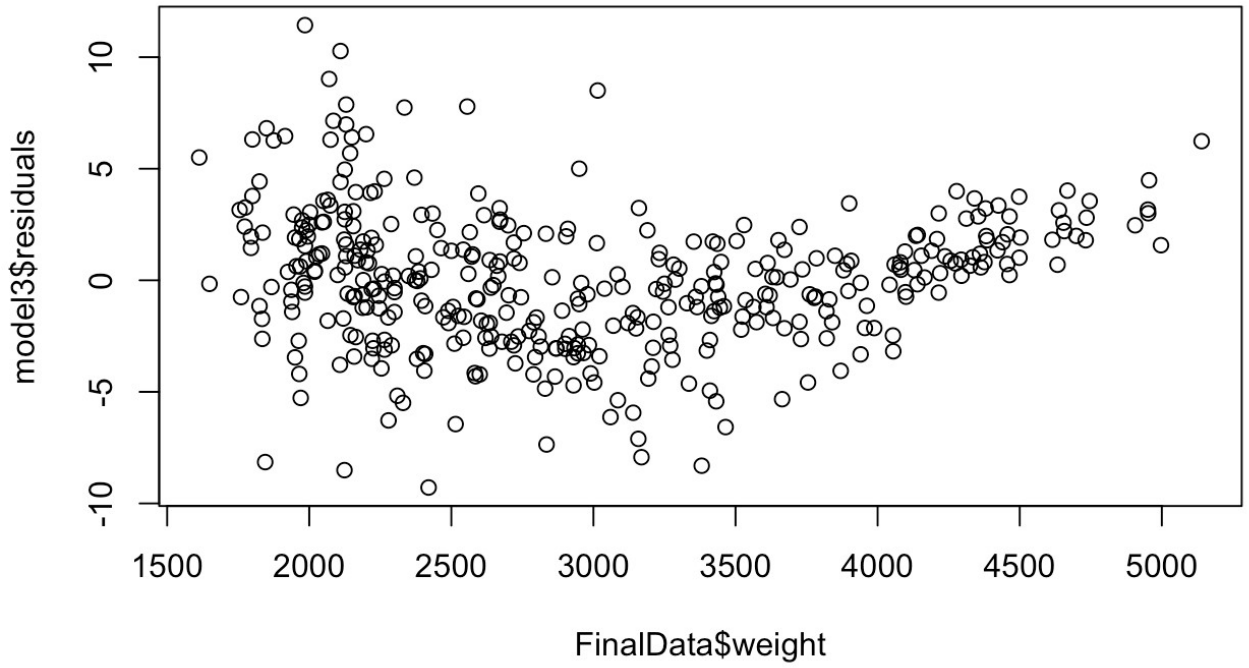


We had check linearity for all the independent variables with the pearson residuals. Instead of plotting each variable with residual, the entire linearity will be checked by using finalized data.

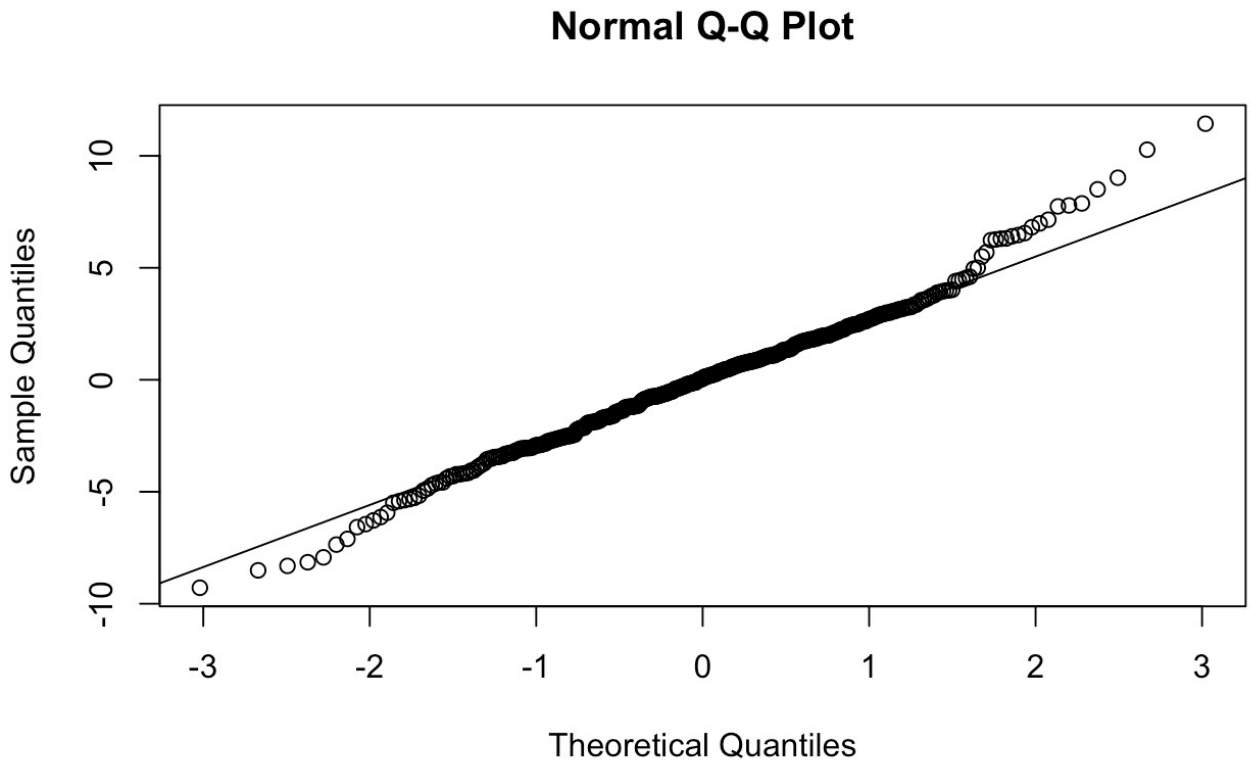
Linearity checking for 1 variable "foreign":



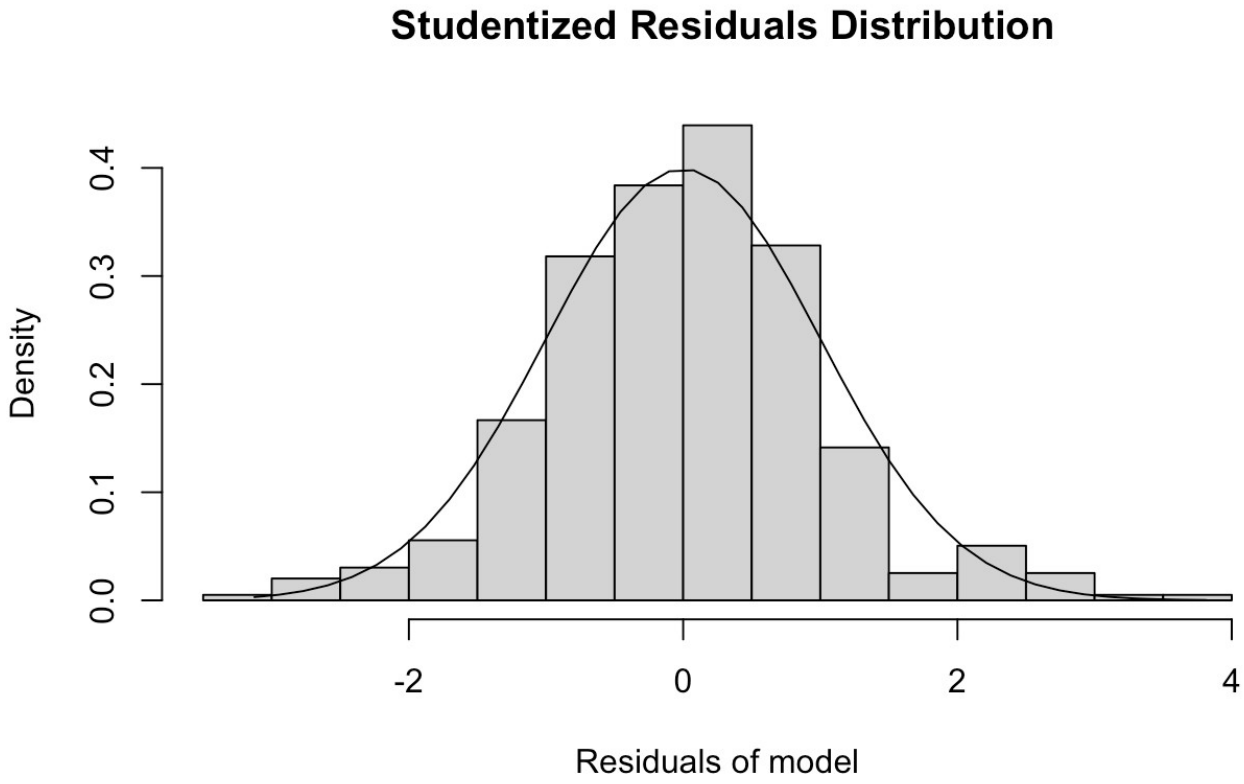
Independence:



Similarly, checking linearity for weight variable how the data is plotted **Normality:**



From the Normal Q-Q Plot, the data is linearly distributed resulting in the formation of linear line. Hence linearity is assumed.



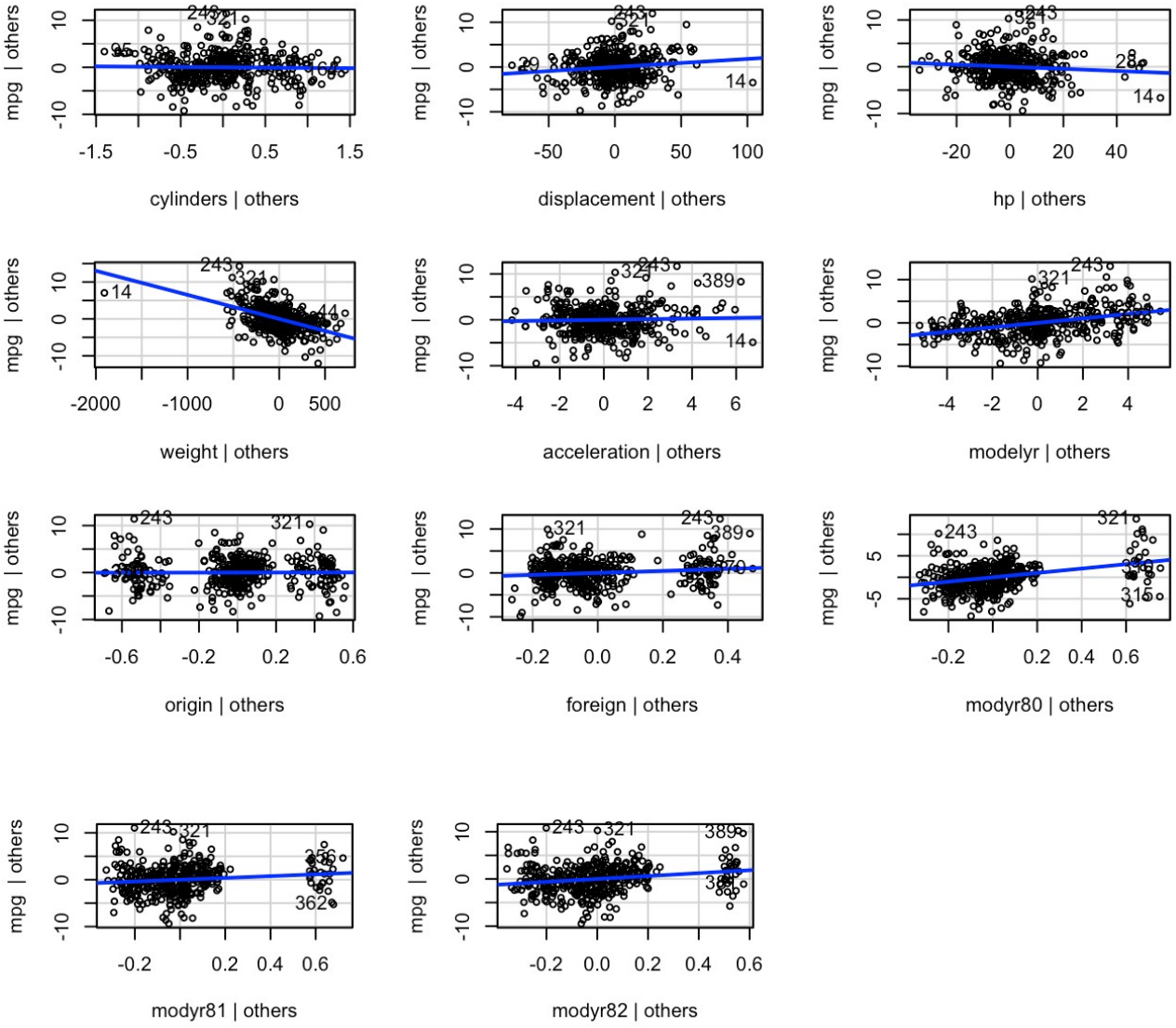
From the above, we are plotting Normal distribution where the data is normally distributed.

Multicollinearity Checking:

cylinders	displacement	hp	weight	acceleration	modelyr	origin
11.186298	23.510247	10.102817	11.290930	2.605381	2.516381	7.538395
foreign	modyr80	modyr81	modyr82			
8.333988	1.396891	1.488898	1.757186			

Checking the correlation between the independent variables.

avplot:



The above plots indicates added variable plot for the multilinear regression model in between the independent and dependent variable.

7. Comment on the model's goodness of fit.

Overall, for any p-value less than 5% level of significance level accounts for the model was more significant. Therefore, by building multiple models on different sets of data we finalize the where the explanatory variables accounted for 84% of the mpg variation which is significantly good as shown in question 3.

II. Summarize the main ideas in "Let's Take the Con Out of Econometrics," by Edward E. Leamer (source: The American Economic Review, Vol. 73, No. 1 (Mar., 1983), pp. 31-43). Ans.

Summary:

The author talks that the con out of Econometrics is depending upon what model you select for analyzing the data you get dramatically different estimates and conclusions and economists has not spend enough effort in drawing the conclusions or insights pretending that the datasets provide a clear information the kind of metric method to be used. From this paper, the clear idea is to develop the tools that will benefit individuals, researchers, statisticians in analyzing the data to sort out fragile conclusions and a method of communicating the insights you got from the dataset whether the dataset is small and assumptions are not really credible.

Also, the paper is about the dealing some of the issues regression discontinuity, instrumental variables natural experiments with econometrics during 1980. Medical research is evolved or dependent upon the results generated which will be studied upon the empirical world of economics called econometrics. The economists and medical scientist face randomized control trial. The con from econometric model is the author hidden the larger model and reduced method of reduction which means eliminated all the parameters by considering only few parameters. By considering the t values as well as p values the model build on the top of the data will give adequate results which is easier to draw the insights by considering only few variables or parameters. By doing minor changes in the model, which results in the sensitive analysis from building sensitive model.

The statement provided by the author "sex, is better demonstrated than discussed, though often better anticipated than experienced". Illustrating the statement with an example, there are 50 states in USA, during 1950's the murder rate is completely dependent upon the capital punishment. If low the capital, the low in murder rate. The author considered by taking Utah which is having less capital punishment, as by looking into the 50 states scenario, it can be considered as the 50 observations and by adding more parameters or variables to the existence one like culture, age etc. which results the model to generate better insights that may either decrease or increase the murder rate. Moreover, the author trying to explain, instead of considering less parameters where the dataset can speak about the questions asked, it's better to add or bring more variables which then generate adequate fragile meaning. The global sensitivity analysis for any of the dataset ever known until it is built upon the dataset when you consider 1 variable and build the model, the independent variable has more to say about that variable although you were 10 variables but not considering. But when you build the model by considering 10 variables, you will get bigger coefficient than with the one variable.

The paper also tells whether having pseudo randomization or real randomization control trails there's still serious persuasive findings to be known. As the experiments done in econometric laboratory may not apply to the real world, the economists working on the randomization may have large list of controls and these failure controls can result in the parameterized dimensionality problem. For example, when you have 100 randomized trails upon the 100 control variables but you won't get any estimated results.

After going through some of the additional papers written by authors Angrist and Pisgay stated after doing some kind of reboot on the same paper that extreme bound and the kind of these sensitive analysis were not the path to the truth, better research design will deal in issues. Most of the part the economists or statisticians work on the data analysis part rather on building ML models. Instead, they need to draw an insights from the data before building the model so more experimental analysis needed to be carried out which helps the other sectors or domains driven by these insights.