# Report on auto-mpg Data Set

Saikat Bera, Presidency University, KOLKATA

December 2, 2022

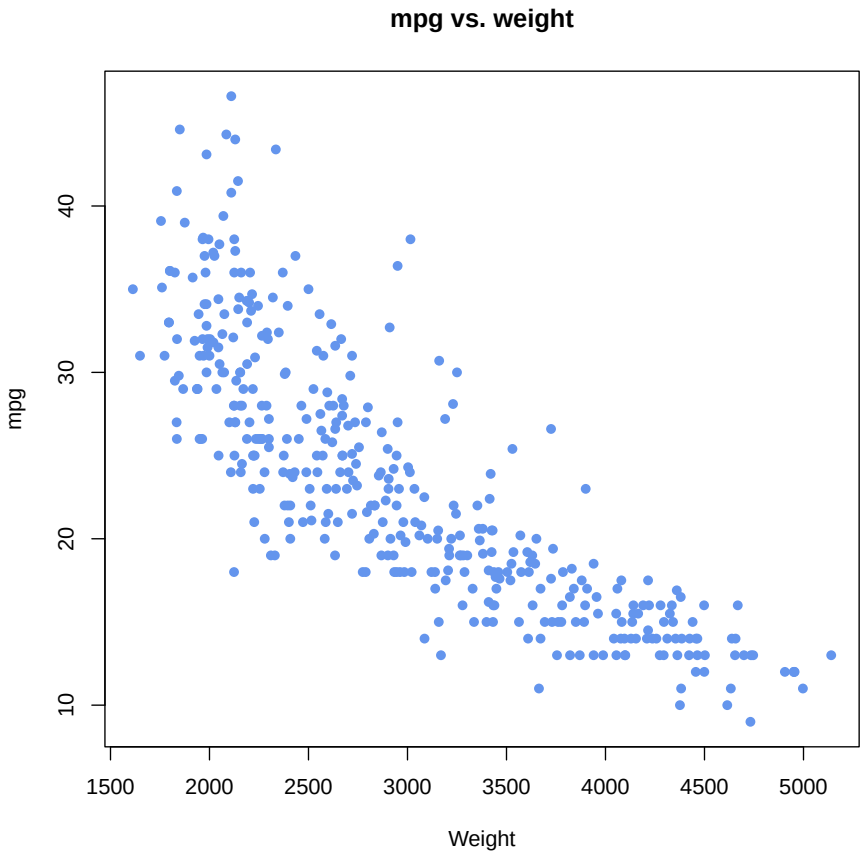
**Abstract**

The dataset auto-mpg.csv, comes from the 1983 American Statistical Association Exposition. The response variable of interest is fuel consumption, measured in miles per gallon. Other attributes of cars like the weight, horsepower, number cylinders, and acceleration time were also recorded for each car. We will study the relationship between mpg and weight (in lbs).

We first load the dataset and view the first few entries as well as run a summary:

```
##                   car.name mpg weight
## 1 chevrolet chevelle malibu  18   3504
## 2         buick skylark 320  15   3693
## 3        plymouth satellite  18   3436
## 4             amc rebel sst  16   3433
## 5               ford torino  17   3449
## 6          ford galaxie 500  15   4341
##    car.name              mpg            weight
##  Length:398         Min.   : 9.00   Min.   :1613
##  Class :character   1st Qu.:17.50   1st Qu.:2224
##  Mode  :character   Median :23.00   Median :2804
##                     Mean   :23.51   Mean   :2970
##                     3rd Qu.:29.00   3rd Qu.:3608
##                     Max.   :46.60   Max.   :5140
```

We can plot a mpg vs. weight scatterplot and find out the relationship between the two variables.

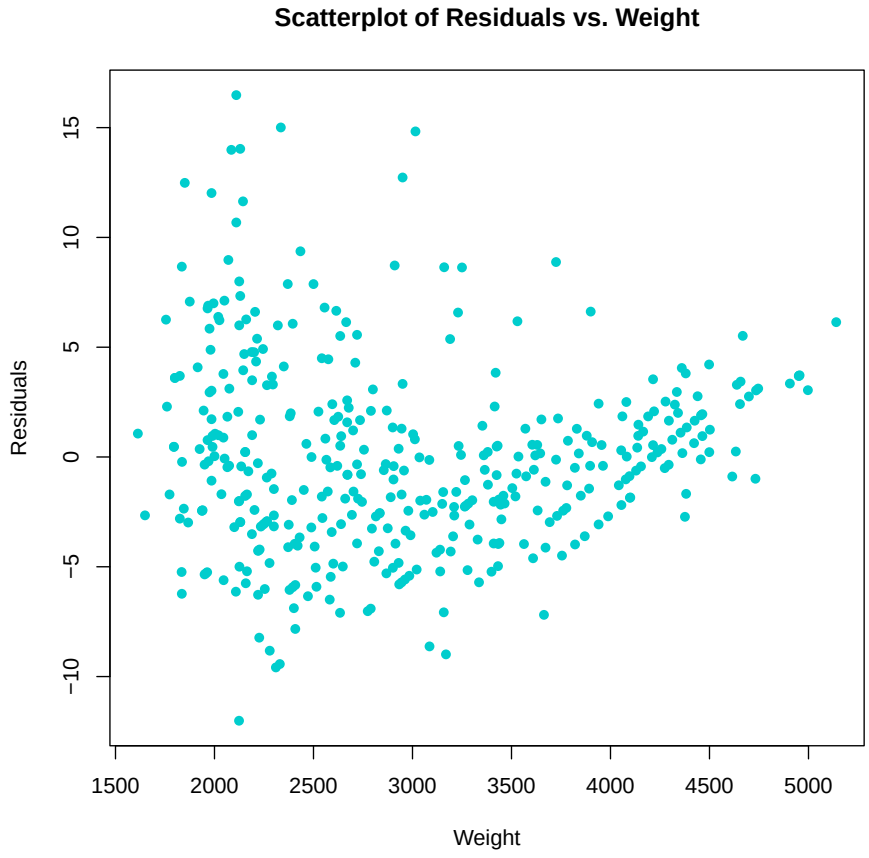



Looking at the plot, we find that a linear model would be a good fit for the dataset. We thus fit a linear model to the data:

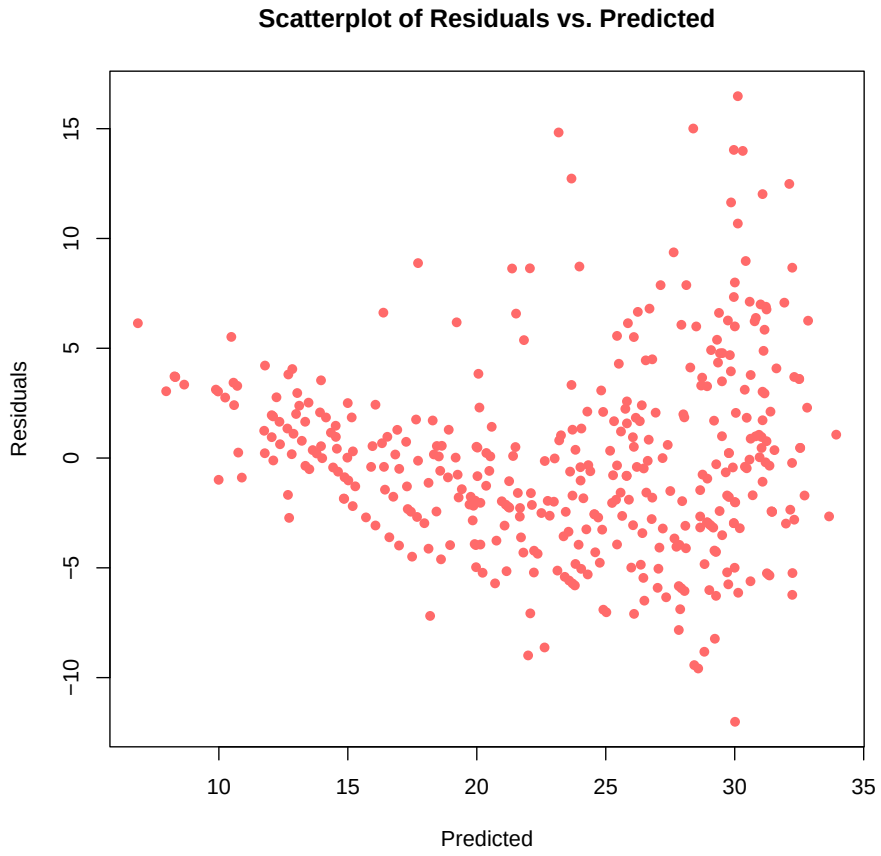
```
##
## Call:
## lm(formula = df$mpg ~ df$weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.012  -2.801  -0.351   2.114  16.480
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.3173644  0.7952452   58.24   <2e-16 ***
## df$weight   -0.0076766  0.0002575  -29.81   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.345 on 396 degrees of freedom
## Multiple R-squared:  0.6918,Adjusted R-squared:  0.691
## F-statistic: 888.9 on 1 and 396 DF,  p-value: < 2.2e-16
```

To look at whether our simple linear model assumption is right or not, we might plot the residuals on the y-axis and the predictor variables on the x-axis. The resultant plot is:

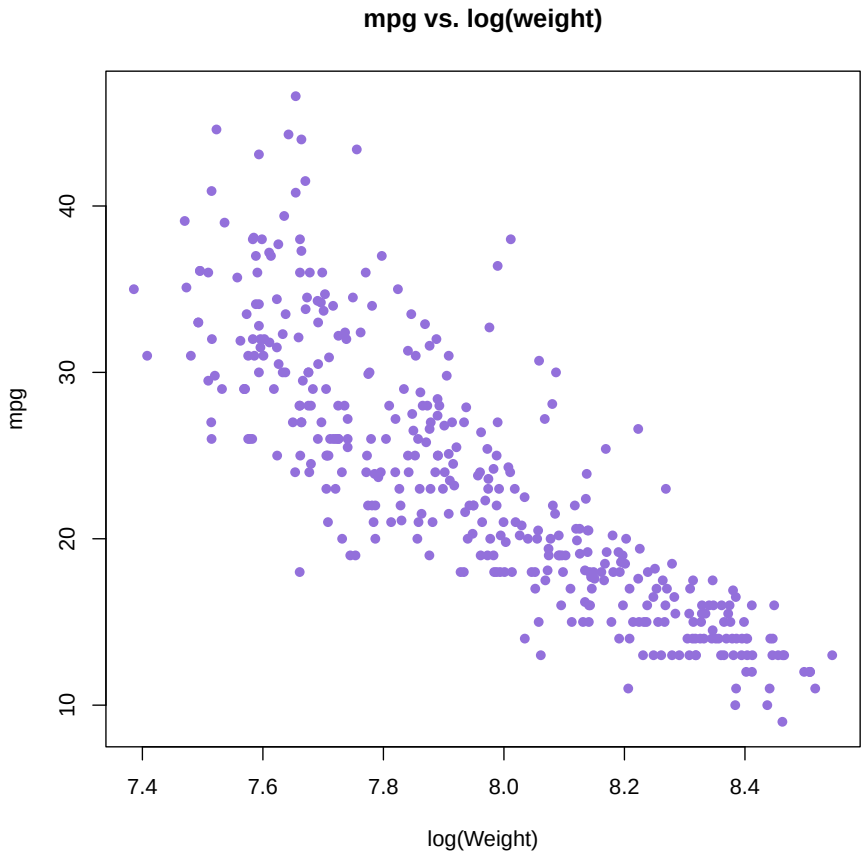



We observe that the plot shows a stepped pattern which gives evidence that the simple-linear part in the simple linear model is wrong. Also, the there is a changing width which also points to the fact that the model is mis-specified. Thus, we either need more predictor variables, or a different model, or both. We can derive a similar result using the residual vs predicted value plot:




We observe that this plot is more similar to the previous one. The reason being the predicted values are a function of the covariates and thus, the graphs are similar.

Taking the log transformation of the covariate, we obtaing the following graph:

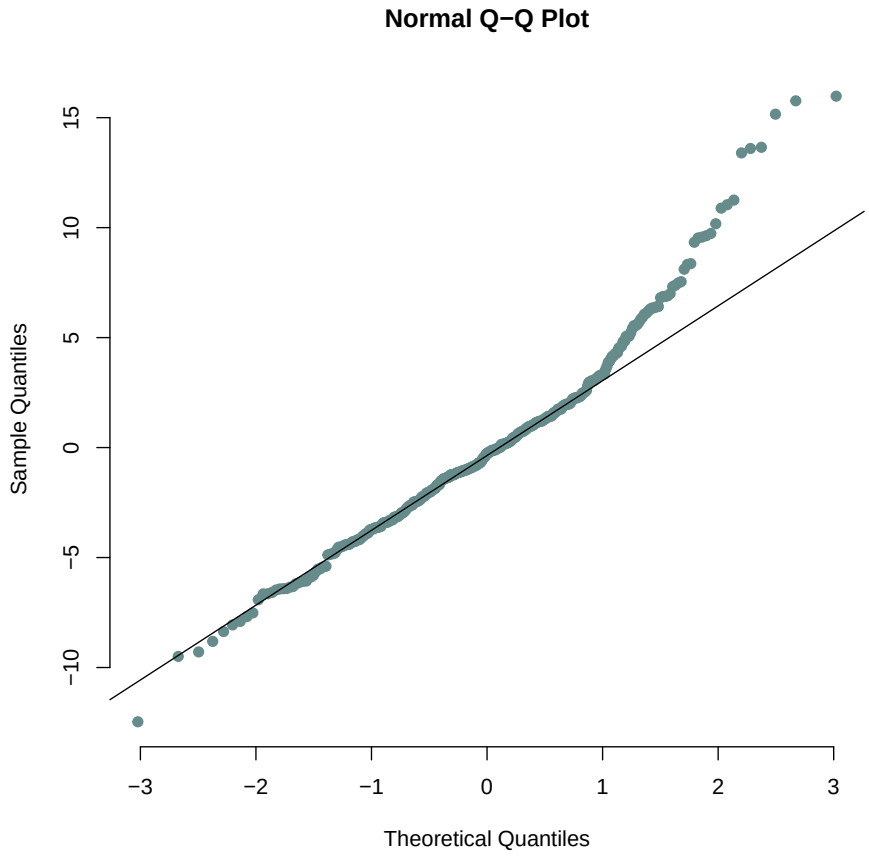



This is a straight line once again, which suggests that the regression is not linear, but rather exponential. It also suggests that the normality assumption of the error terms was false previously.

We once again fit a linear regression on this transformed data and see if the residuals satisfy the normality assumption:

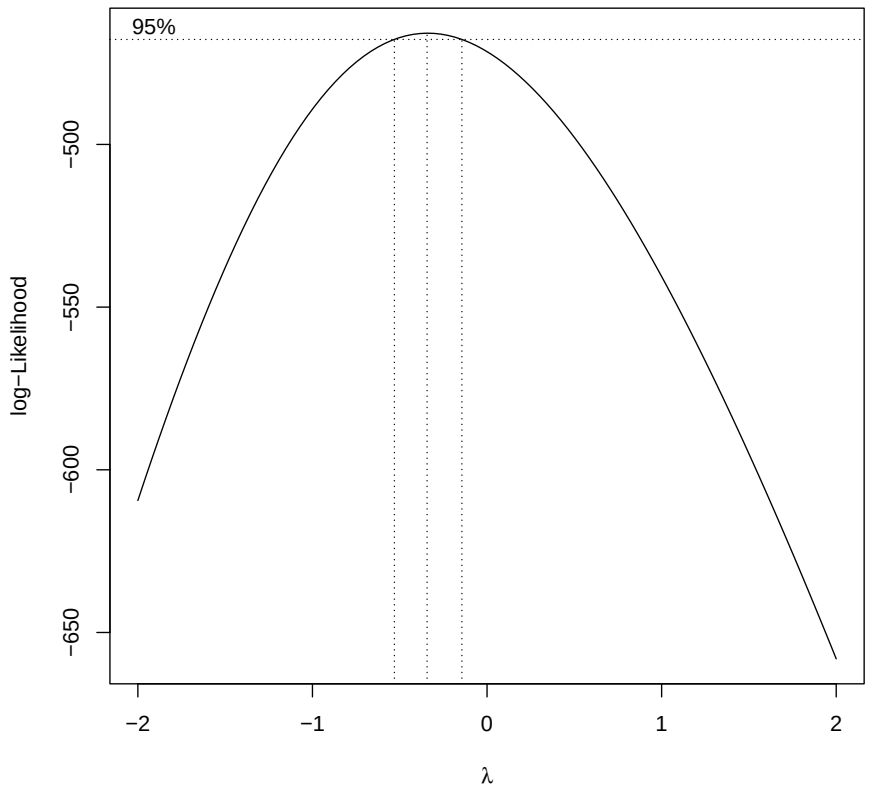
```
##
## Call:
## lm(formula = df$mpg ~ log(df$weight))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4676  -2.6558  -0.2669   1.9333  15.9770
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    210.5270     5.9763   35.23   <2e-16 ***
## log(df$weight) -23.5032     0.7506  -31.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.198 on 396 degrees of freedom
## Multiple R-squared:  0.7123,Adjusted R-squared:  0.7116
## F-statistic: 980.4 on 1 and 396 DF,  p-value: < 2.2e-16
```



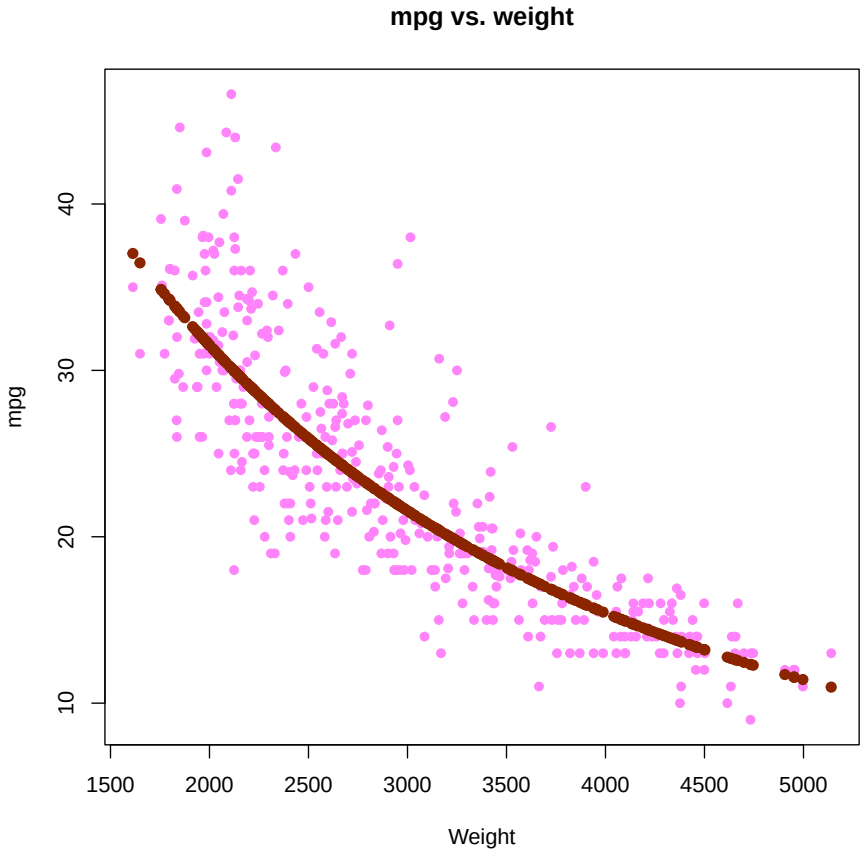

As anticipated, the residuals do not satisfy the normality assumption.

Finally, we might use the Box-Cox transformation and refit our model using the appropriate transformation:




```
## [1] -0.3434343
```

We finally plot our Box-Cox transformed model with our estimated lambda:




Among the three models, the Box-Cox transformed model looks the best fit. The slope of the model infers that if the weight is changed by one unit measure, on an average the mpg would change by (in units):

```
##     df$weight
## -0.0001239117
```

We can verify if the association between weight and mpg is linear or not. The null hypothesis being that the correlation is 0 with the alternative being that the correlation is not 0. Under the null hypothesis, the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

where r: sample correlation coefficient
and n: number of observations.

We accept the null hypothesis if our p-value of the test $> 0.05$ else we reject the null hypothesis. In case the null hypothesis is rejected then we might assume that the association between our variables is linear. We apply a correlation test on the variables and the result we obtained are as follows:

```
##
##  Pearson's product-moment correlation
##
## data:  df$weight and df$bc.mpg
## t = -37.449, df = 396, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9029842 -0.8593671
## sample estimates:
##        cor
## -0.8830687
```

We observe that the p-value for this test is:

```
## [1] 3.390755e-132
```

which is less than our minimum threshold for accepting the null hypothesis. Thus, with 0.05 level of significance, we reject the null hypothesis that the correlation is 0. Thus, we may safely assume that the there is some level of linear association between our variables.

Further, we might want to find a 90% confidence interval for $\beta\hat{}$ of our fitted model. We find that the required confidence interval is:

```
##                       5 %          95 %
## df$weight -0.0001293669 -0.0001184565
```

This tells us that if we find the estimated beta 100 times, in 90 of those, our beta would fall within this range.