Dissecting the Relationship between Fuel Efficiency (MPG) and Transmission Mode (AT/MT)

Shivendra Sharma

28/02/2017

Synopsis

The `mtcars` data is an inbuilt dataset that comes with the `datasets` package and can be easily loaded up with the `mtcars` command. Its a fairly simple dataset that lists down 11 characteristics of 32 automobiles used by Henderson and Velleman in their book *Biometrics (1981)*. The data was extracted from the 1974 edition of the *Motor Trend* magazine and contains automobiles manufactured from 1973-74. This exercise will involve the same dataset to look into the relationship between an automobile's performance (MPG) with the mode of transmission (auto or manual) and conclude on which transmission is seemingly the better option. **Plots have been shown under appendix only.**

Selecting the Best Regression Model

Strategizing on selecting the best regression model is an essential task for anyone to conduct an efficient analysis. However, there is no specific regression model in general except for such situations and data where the researcher has either binomial observations or data different from general observations (dummy variables, categorical, etc.). In such cases, most researchers go ahead with just simple linear models modified to contain interactions and multiple variables to explain relationships; it all depends on the type of data and hence, an exploratory analysis is one of the first few steps.

Exploratory Analysis

A summary of our automobiles dataset should reveal the necessary details for each variable:

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
```

```
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Our areas of importance would be the `mpg` variable and the `am` variable that might seem confusing at first but basically comprises of zeroes and ones that relate to automatic or manual transmission respectively. In fact, the following code will suggest that the dataset is somewhat biased towards automatic automobiles.

```
table(mtcars$am)
```

```
##
##  0  1
## 19 13
```

There are 19 vehicles with automatic transmission and 13 for manual; although the difference ain't that much, regression for automatic vehicles will fit more precisely as compared to manual cars.

Analysing Model Fits

With the essentials of exploratory analysis done, a brief regression analysis will follow. We'll use the simple linear model with `am` as a factorised predictor. The predictor here is a binomial variable but it should be noted that the outcome is not. So, although binomial regression too can be used here, it will require `mpg` to be divided by 100 since the latter model requires y or the outcome to have values between 0 and 1.

```
fit <- lm(mpg ~ factor(am) - 1, data = mtcars)
fit2 <-glm(mpg ~ factor(am), data = mtcars)
summary(fit)$coeff
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## factor(am)0 17.14737   1.124603 15.24749 1.133983e-15
## factor(am)1 24.39231   1.359578 17.94109 1.376283e-17
```

```
summary(fit2)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

The first fit gives us the exact coefficients that we want. Although the intercept for the second model is the same as the coefficient for automatic transmission `(am)0`, adding the coefficient of (am)1 will give us the actual value for manual transmission (for the latter fit). And so, there we have it, efficiency of vehicles according to transmission mode. Manual cars seem to be more comparably efficient than manual ones simply because automatic cars consume more fuel per mileage. Of course, there are are other statistics of importance here; minute p-values for both the fits suggest proper coefficient values and low chances of comitting errors and t-values that suggest that the estimated values are significant and close to the studentized values.

Appendix

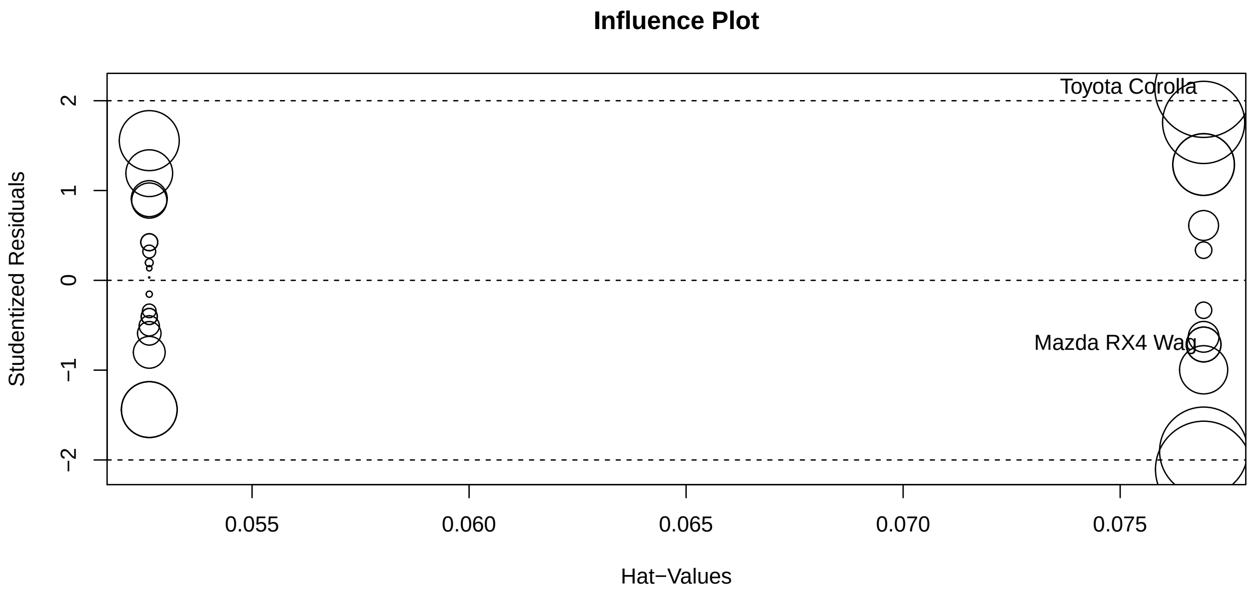
Several methods exist to diagnose linear models; some can be advanced and require knowledge regarding their use and interpretation while others are more standardised and globally used. The `gvlma` package comes really handy here since it carries out a number of tests with just one command, like so:

```
gv <- gvlma(fit)
gv
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) - 1, data = mtcars)
##
## Coefficients:
## factor(am)0  factor(am)1
##       17.15        24.39
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = fit)
##
##                          Value p-value                Decision
## Global Stat          2.667e+00  0.6150 Assumptions acceptable.
## Skewness             1.398e-02  0.9059 Assumptions acceptable.
## Kurtosis             5.752e-01  0.4482 Assumptions acceptable.
## Link Function       -1.298e-31  1.0000 Assumptions acceptable.
## Heteroscedasticity  2.078e+00  0.1495 Assumptions acceptable.
```

The `gvlma()` conducts a global assumptions test that lists down the necessary assumptions and whether they can be accepted or not, as can be seen above. Our fit is thus acceptable in almost all of the tests carried out. Similar, more advanced tests can be carried out too like Bonferroni's test of influential observations, influence plots, Durbin-Watson test, etc. For example, here's a simple influence plot for the fitted model, although I would doubt if everything is showing up correctly.

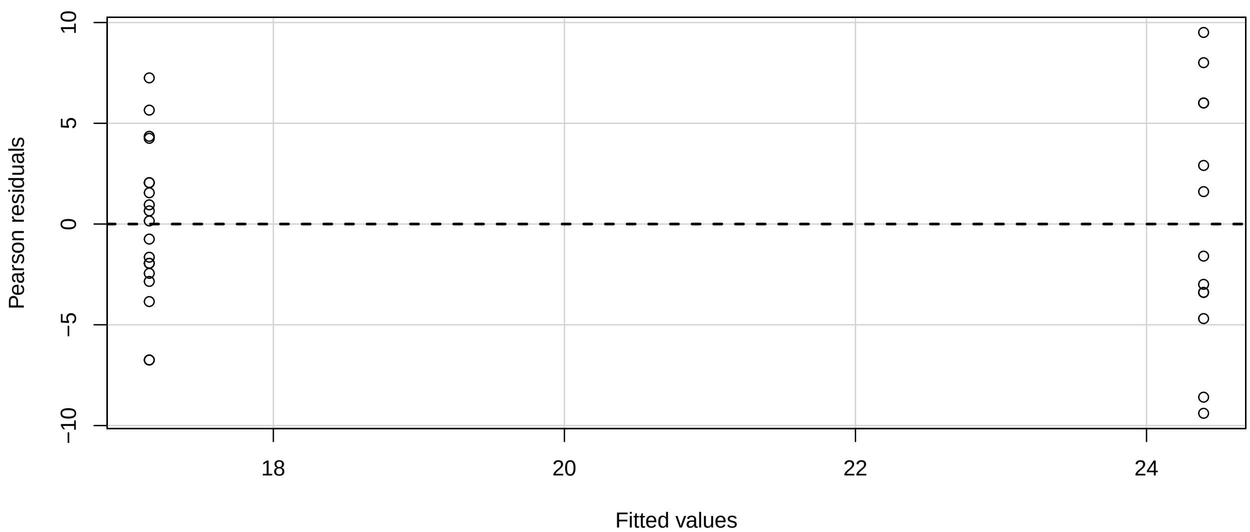
```
influencePlot(fit, main = 'Influence Plot')
```




```
##                    StudRes        Hat      CookD
## Mazda RX4 Wag  -0.714376 0.07692308 0.02161671
## Toyota Corolla  2.135121 0.07692308 0.16980457
```

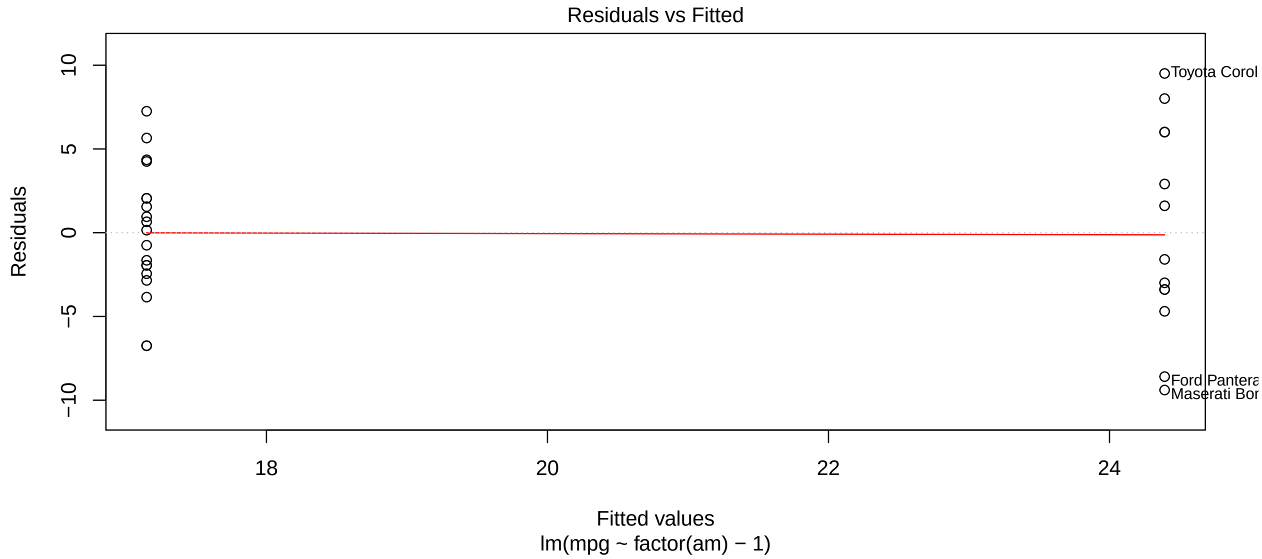
Similarly, a simple residual plot like below:

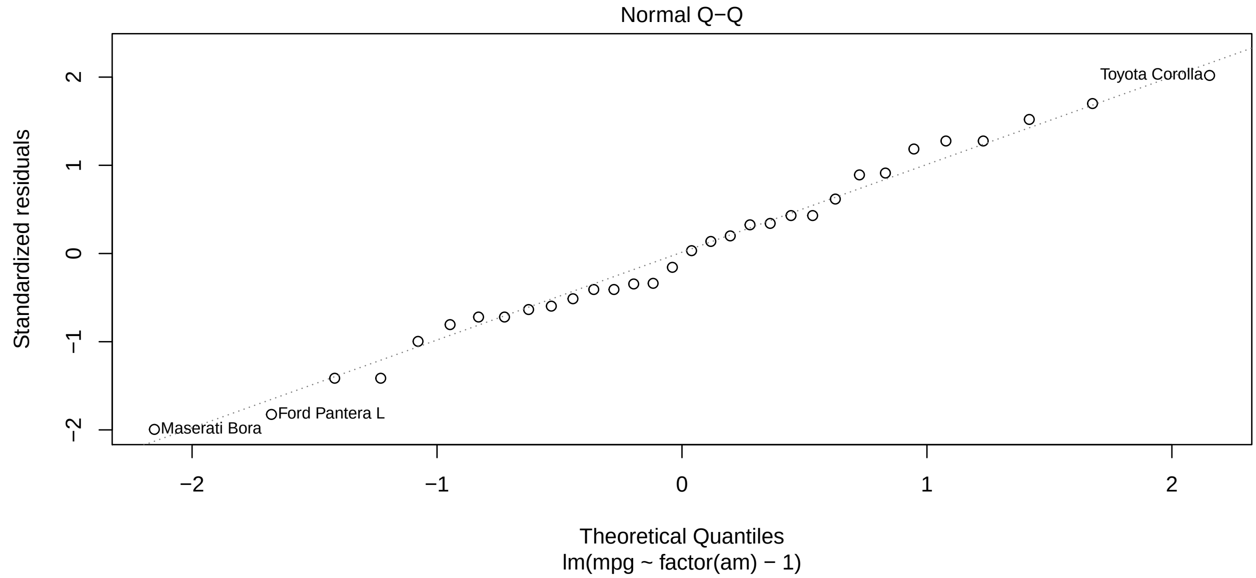
```r
residualPlot(fit)
```

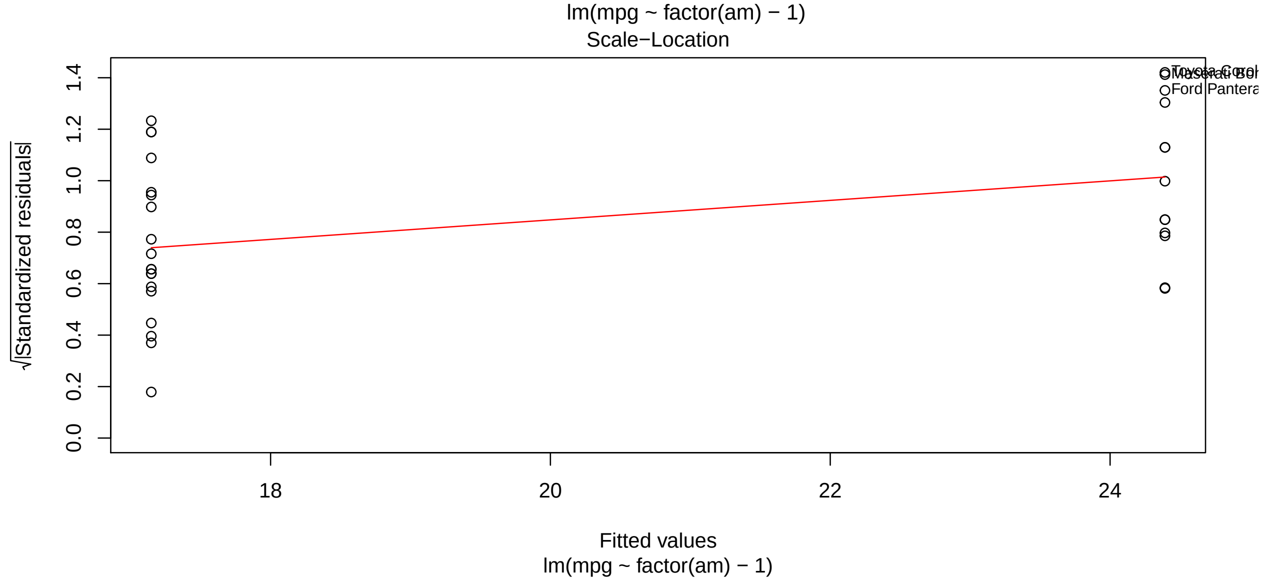


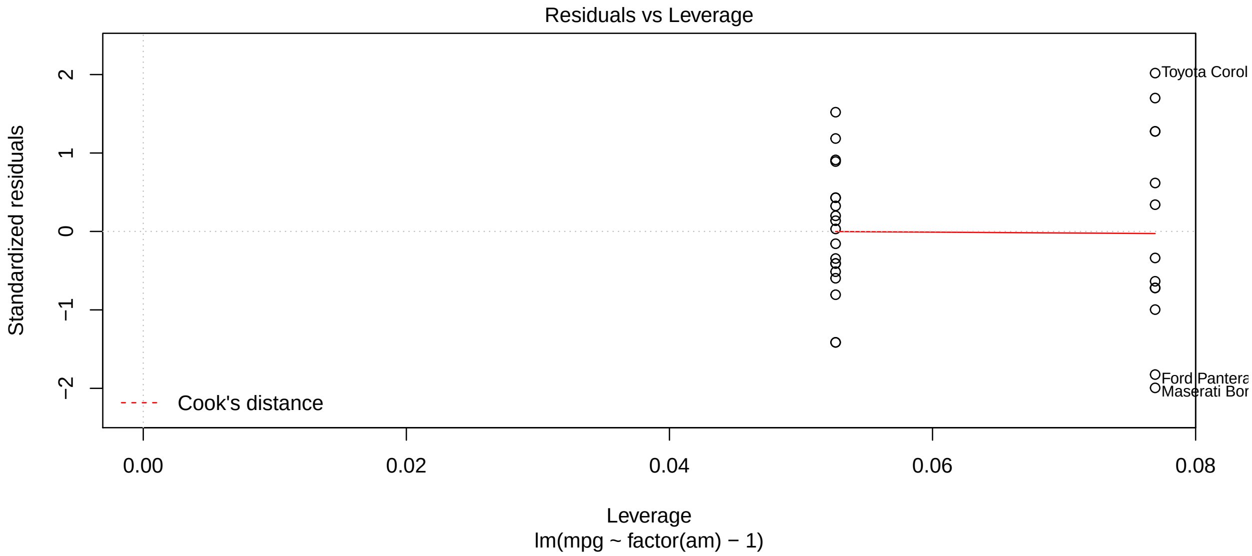
Finally, we can have lots of diagnostic plots that cannot be plotted with conventional commands.

```r
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ factor(am) − 1)
Toyota Corol
Ford Pantera
Maserati Bor


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ factor(am) − 1)
Toyota Corolla
Ford Pantera L
Maserati Bora


Scale−Location
√|Standardized residuals|
Fitted values
lm(mpg ~ factor(am) − 1)
Ford Pantera



Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
Toyota Corol
Ford Pantera
Maserati Bor
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
lm(mpg ~ factor(am) − 1)