Car Fuel Efficiency

Srujan Routhu

Sunday, August 24, 2014

Executive Summary

This article is a report of an analysis done on a dataset containing details pertaining to various aspects of 32 different cars. The analysis was done to identify the parameters that significantly affect a car's fuel efficiency as measured in miles travelled per unit gallon. An interesting question that was expected to be answered was the effect of the transmission mode (automatic vs. manunal) on the fuel effeciency. Analysis was done by exploring the data using various statistical techniques including linear modelling and correlation analysis, and building a regression model to estimate fuel efficiency. The results were inconclusive as to the effect of transmission mode but 2 parameters explain the variation significantly up to a level of 82 percent.

Fuel Efficency Estimation

Description

To estimate the factors affecting the fuel efficiency in a car, we use the dataset "mtcars" which is available in the core R package "datasets". The data comprises fuel cosumption and 10 aspects of automobile design and performance for 32 automobiles (1973 - 74 models).

The 10 aspects are: number of cylinders, displacement (cu.in.), gross horsepower, rear axle ratio, weight (lb/1000), 1/4 mile time, V/S, transmission (0 = automatic, 1 = manual), number of forward gears, and number of carburetors.

Out of the above 10 aspects, we directly discard gross horsepower and 1/4 mile time, as there variable values are consequential to the other aspects rather than causal to the fuel consumption.

Analysis

A basic grasp on the process of fuel consumption points us to its dependence on Weight of the car. From figure 2 in the appendix we can see that the weight is correlated highly with the number of cylinders as well as the displacement. Hence, including all three of these aspects in the model could prove to be redundant.

Figure 3 shows us the scatter plots between MPG and the three variables mentioned in the above paragraph. As we mentioned earlier, consider all three variables in the model could prove to be redundant, and we test this by analysing the adjusted r-squared values for all combinations of these variables.

Figure 4 outlines the adjusted r - squared values for a linear model between MPG and all combinations of the three variables listed above. We can see that the most optimul inclusion is attained by discarding the displacement and choosing only the number of cylinders and weight, thereby reaching an adjusted r - squared of 82 percent. Our earlier concern that using all three variables could prove to be redundant is shown to be well founded.

Questions of Interest

We see that the other variables in the data are not well correlated with the fuel consumption figures. But our question of interest pertains to the effect of transmission on the fuel consumption. We add the variable "Transmission" to the model we selected in the previous section and analyse the changes in the results. Following is the summary of the results.

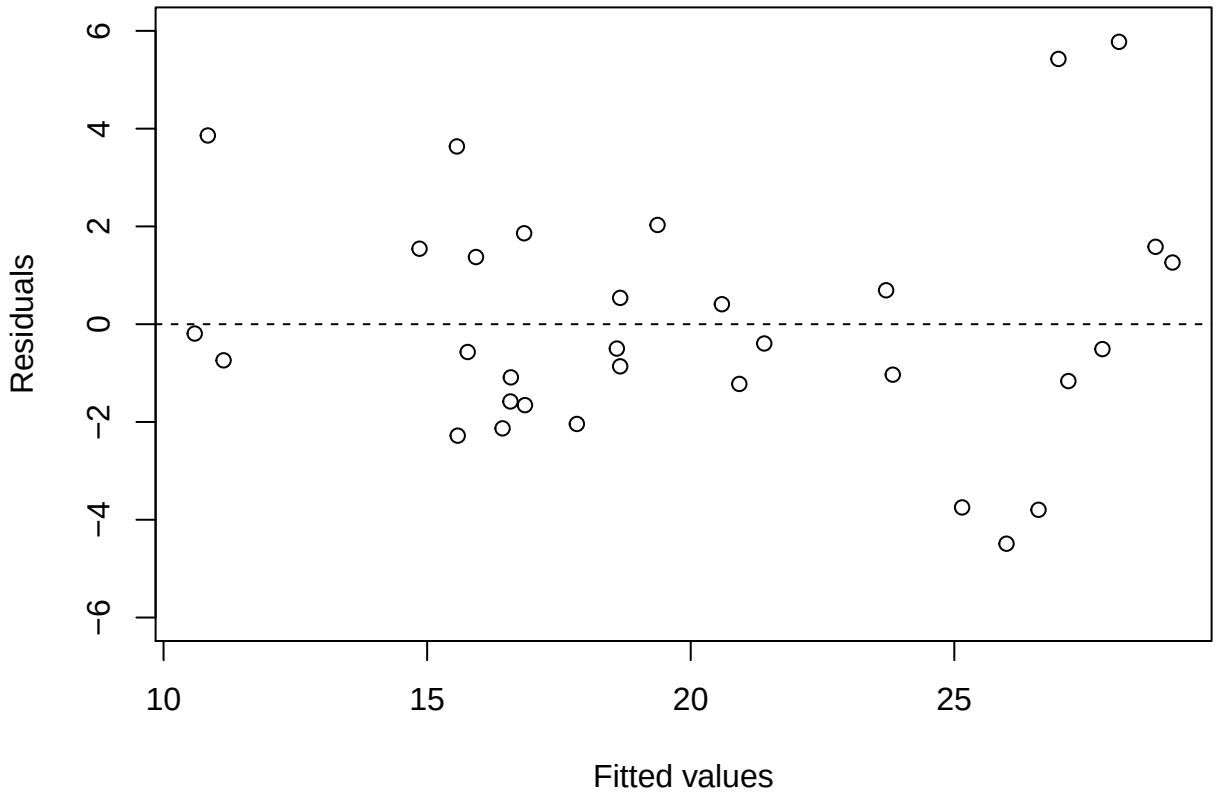
```
## lm(formula = mpg ~ as.factor(cyl) + wt + as.factor(am), data = z)

##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      33.7536      2.813 11.9971 2.496e-12
## as.factor(cyl)6  -4.2573      1.411 -3.0167 5.515e-03
## as.factor(cyl)8  -6.0791      1.684 -3.6105 1.228e-03
## wt               -3.1496      0.908 -3.4685 1.771e-03
## as.factor(am)1    0.1501      1.300  0.1154 9.089e-01
```

The inclusion of other variables was also analysed by adding the remaining variables one by one and calculating the linear model. All the remaining variables, Rear axle ratio, V/S, Number of forward gears, and Number of carburetors have been found to have insignificant effect on fuel consumption at reseonable levels of significance, while reducing the adjusted r - squared value by noticeable amount. The Number of carburetors variable can be ommitted directly based on the results of the univariate analysis shown in figure 1.

The results shown above suggest that there is a 0.15 miles increase per gallon for cars with manual transmission, though the result is insignificant at any reasonable level of significance. The p value for the acceptance of the alternative hypothesis is 0.9089, **taking the uncertainty levels in our estimate to a value of over 90 %**. Therefore, we fail to reject the null hypothesis that there is no difference between the two modes of transmission, and hence cannot reach any credible conclusion on the matter of which mode of transmission is better for fuel effeciency.

Residual Plot

Attached below is residual plot of the model specified in the previous section, showing fitted values on the x-axis and the residuals on the y-axis. Other diagnostic plots of the specified model can be found in the appendix (Figure 6).



Appendix

Figure 1: Summaries of Univariate Regressions

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       26.664     0.9718  27.437 2.688e-22
## as.factor(cyl)6   -6.921     1.5583  -4.441 1.195e-04
## as.factor(cyl)8  -11.564     1.2986  -8.905 8.568e-10
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 29.59985   1.229720  24.070 3.577e-21
## disp        -0.04122   0.004712  -8.747 9.380e-10
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   -7.525      5.477  -1.374 1.796e-01
## drat           7.678      1.507   5.096 1.776e-05
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   37.285     1.8776  19.858 8.242e-19
## wt            -5.344     0.5591  -9.559 1.294e-10
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       16.62      1.080  15.390 8.847e-16
## as.factor(vs)1     7.94      1.632   4.864 3.416e-05
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      17.147      1.125  15.247 1.134e-15
## as.factor(am)1    7.245      1.764   4.106 2.850e-04
```

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        16.107      1.216  13.250 7.867e-14
## as.factor(gear)4    8.427      1.823   4.621 7.257e-05
## as.factor(gear)5    5.273      2.431   2.169 3.842e-02
```

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        25.343      1.854  13.670 2.215e-13
## as.factor(carb)2   -2.943      2.417  -1.218 2.343e-01
## as.factor(carb)3   -9.043      3.385  -2.672 1.285e-02
## as.factor(carb)4   -9.553      2.417  -3.952 5.295e-04
## as.factor(carb)6   -5.643      5.243  -1.076 2.917e-01
## as.factor(carb)8  -10.343      5.243  -1.973 5.927e-02
```

Figure 2: Correlation Matrix

```
##        mpg   cyl  disp  drat    wt    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85  0.68 -0.87  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90 -0.70  0.78 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00 -0.71  0.89 -0.71 -0.59 -0.56  0.39
## drat  0.68 -0.70 -0.71  1.00 -0.71  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89 -0.71  1.00 -0.55 -0.69 -0.58  0.43
## vs    0.66 -0.81 -0.71  0.44 -0.55  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59  0.71 -0.69  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56  0.70 -0.58  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39 -0.09  0.43 -0.57  0.06  0.27  1.00
```

Figure 3: Scatter Plots for MPG Vs. Number of Cylinders, Weight and Displacement

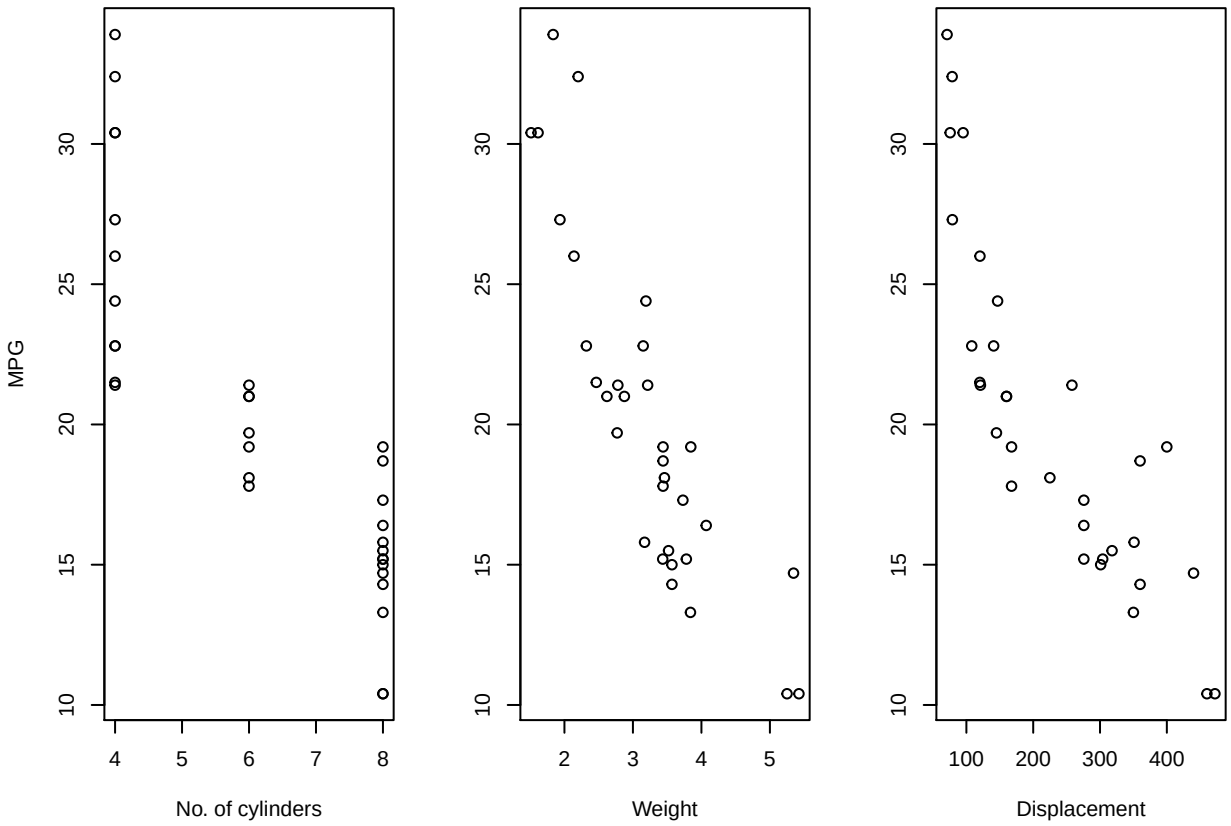


Figure 4: Adjusted R-Squared values for combinations of variables plotted in figure 3

1. mpg ~ as.factor(cyl) : adjusted r - squared = 0.714
2. mpg ~ wt : adjusted r - squared = 0.7446
3. mpg ~ disp : adjusted r - squared = 0.709
4. mpg ~ as.factor(cyl) + wt : adjusted r - squared = 0.82
5. mpg ~ wt + displacement : adjusted r - squared = 0.7658
6. mpg ~ as.factor(cyl) + displacement : adjusted r - squared = 0.7605
7. mpg ~ as.factor(cyl) + wt + displacement: adjusted r - squared = 0.8135

Figure 5: Summaries of Linear Models after including each of the remaining variables

```
## lm(formula = mpg ~ as.factor(cyl) + wt + as.factor(am) + drat,
##     data = z)

##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      34.7974     6.5736  5.2935 1.551e-05
## as.factor(cyl)6  -4.3374     1.5073 -2.8777 7.904e-03
## as.factor(cyl)8  -6.2054     1.8582 -3.3396 2.544e-03
## wt               -3.1656     0.9292 -3.4067 2.148e-03
## as.factor(am)1    0.2758     1.5036  0.1834 8.559e-01
## drat             -0.2699     1.5296 -0.1764 8.613e-01
```

```
## lm(formula = mpg ~ as.factor(cyl) + wt + as.factor(am) + as.factor(vs),
##     data = z)

##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      32.1811     3.7286  8.6309 4.140e-09
## as.factor(cyl)6  -3.7329     1.6376 -2.2795 3.108e-02
## as.factor(cyl)8  -4.7483     2.6572 -1.7869 8.561e-02
## wt               -3.1061     0.9203 -3.3751 2.326e-03
## as.factor(am)1    0.6232     1.5012  0.4151 6.814e-01
## as.factor(vs)1    1.2418     1.9042  0.6521 5.200e-01

## lm(formula = mpg ~ as.factor(cyl) + wt + as.factor(am) + as.factor(gear),
##     data = z)

##                  Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)       33.6102     2.9048 11.57075 1.562e-11
## as.factor(cyl)6   -4.0072     1.4609 -2.74288 1.110e-02
## as.factor(cyl)8   -5.4631     2.0946 -2.60815 1.514e-02
## wt                -3.2366     0.9395 -3.44503 2.026e-03
## as.factor(am)1     0.8814     1.8462  0.47744 6.372e-01
## as.factor(gear)4   0.1403     1.8915  0.07418 9.415e-01
## as.factor(gear)5  -1.6043     2.1989 -0.72961 4.724e-01
```

Figure 6: Other model diagnostics (QQ Plot)