Automatic or Manual: Effect of Transmission Type on mpg

Amogu Ukairo

July 27, 2014

Executive Summary

The number of miles a vehicle can go per gallon of gas/fuel is influenced by many contribtory factors. In order to examine the question we apply multiple linear regression analysis to the mtcars dataset to determine which transmision type(automatic or manual) offers better fuel efficience and by how much. Our model shows that of cars in this dataset, manual cars have higher mpg. Considering the interaction between weight and transmission type, this result is clearly influenced by the small sample size, and the relative number and weight of the automatic models in our dataset.

Data Processing

First we did some exploratory analysis of the dataset to get a sense of what the how the variables relate to each other. From Fig 1, it is clear that there variables such as weight (wt) and horsepower (hp) have defined relationship to mpg. Infact the boxplot of mpg vs. am suggests that manual cars having higher mpg than autamatic cars.

Indeed when we build a linear OLS model of these two, lm(mpg ~ am, data = mtcars1), we get what seems to be significant coeficients. However with R-squared at 0.3384589, it is clear that our model can account for only about 34% of the result so we obviously need a better fit.

Running a model with all the variables thrown in increases adjusted R to 0.779, but by using the stepAIC tool from the cars package and a knowledge of the correlation of wt and hp, we select the model with even higher adjusted R values.

```
fit2 <- lm(formula = mpg ~ am*wt, data = mtcars1) #
summary(fit2)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   31.416     3.0201  10.402 4.001e-11
## amManual      14.878     4.2640   3.489 1.621e-03
## wt            -3.786     0.7856  -4.819 4.551e-05
## amManual:wt   -5.298     1.4447  -3.667 1.017e-03
```

```
summary(fit2)$adj.r.squared
```

```
## [1] 0.8151
```

Interpreting the coeficients

Looking at the coeficients, clearly Automatic cars are the reference category. The intercept for automatics is 31.4161 with a slope of -3.7859. 14.8784 is the change in intercept for manuals so that 31.4161 plus

14.8784 is the intercept for manuals. The sum of -5.2984 and -3.7859 is our new slope for Manuals. The p-values of amManual and amManual:wt indeed confirm that we should expect different slopes and intercepts. Specifically, the significant interaction value -5.2984 tells us that the relationship betwwen mpg and am varies by car weight. In fact when we examine this some more using the effects plot, we see that at lower tonnage, automatic cars have higher mpg than manuals. Therefore, the model we have will be influenced by the number of automatics in the sample and their tonnage.

Diagnostics

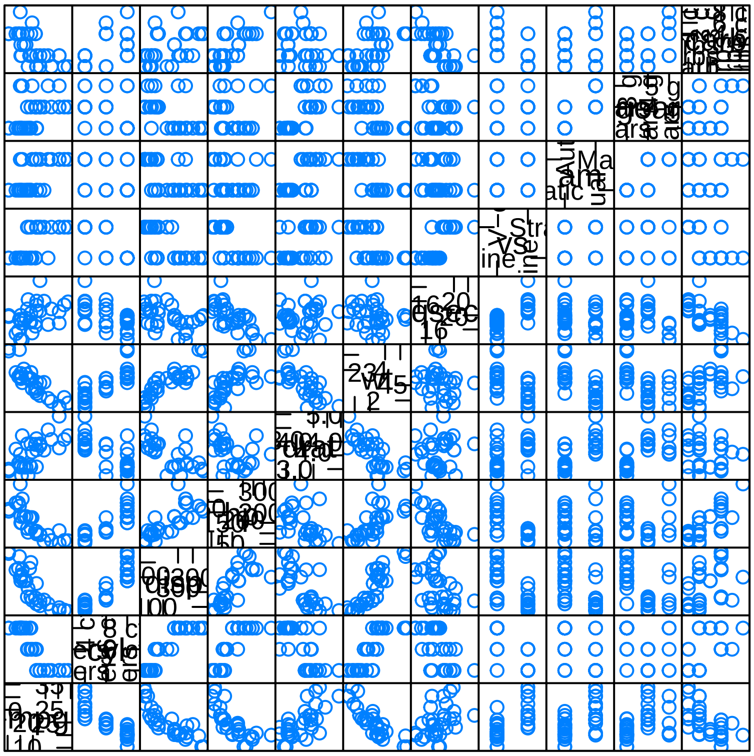
But how well does our model conform to the required assumptions that the dependent variable is normally distributed for a fixed set of predictor values? We can check that by ploting the studentized residuals against a t distribution with n-p-1 degrees of freedom, where n is the sample size and p is the number of regression parameters (including the intercept). The resulting grapgh shows that the points are close to the line and/or withing the confidence envelope.

To validate all of this, we use GVLMA, the Global Validation of Linear Model Assumptions. From the output, the general statistics and link function tell us that linear model fits our data and that linearity assumption of our model is acceptable. To check for homoscedasticity, also confirms that we met the constant variance assumption.

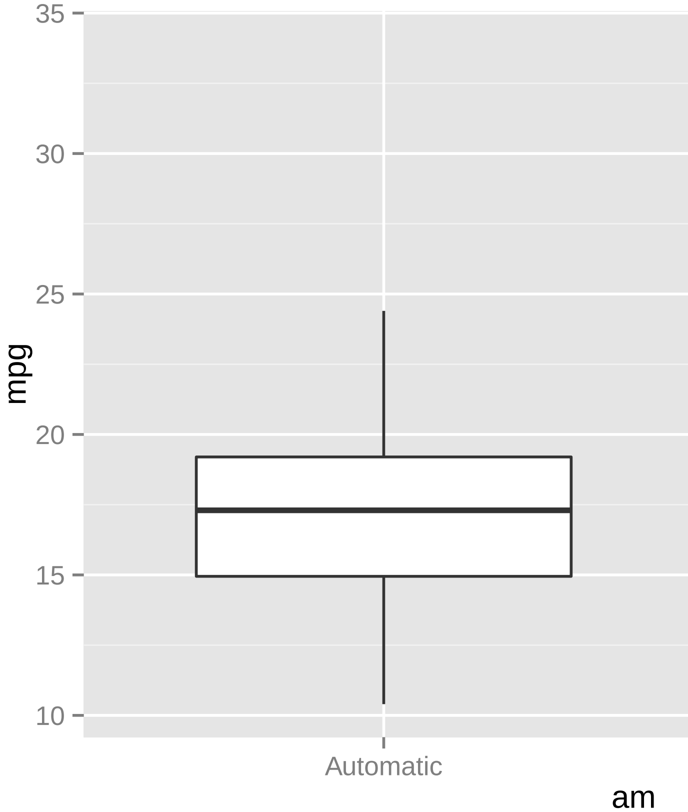
```r
gvlma(fit2)
```

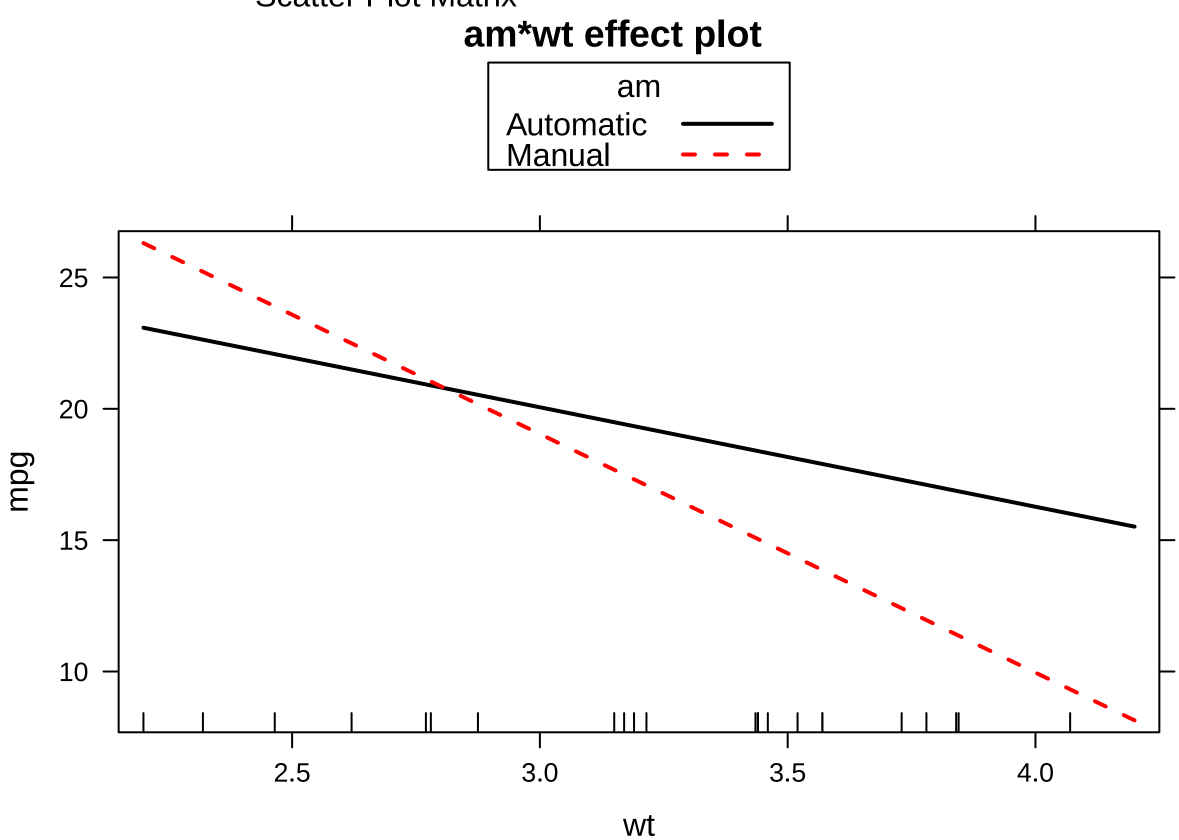
```
##
## Call:
## lm(formula = mpg ~ am * wt, data = mtcars1)
##
## Coefficients:
## (Intercept)     amManual           wt  amManual:wt
##       31.42        14.88        -3.79        -5.30
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = fit2)
##
##                      Value p-value                Decision
## Global Stat        3.14643  0.5336 Assumptions acceptable.
## Skewness           3.08649  0.0789 Assumptions acceptable.
## Kurtosis           0.00187  0.9655 Assumptions acceptable.
## Link Function      0.02220  0.8816 Assumptions acceptable.
## Heteroscedasticity 0.03587  0.8498 Assumptions acceptable.
```

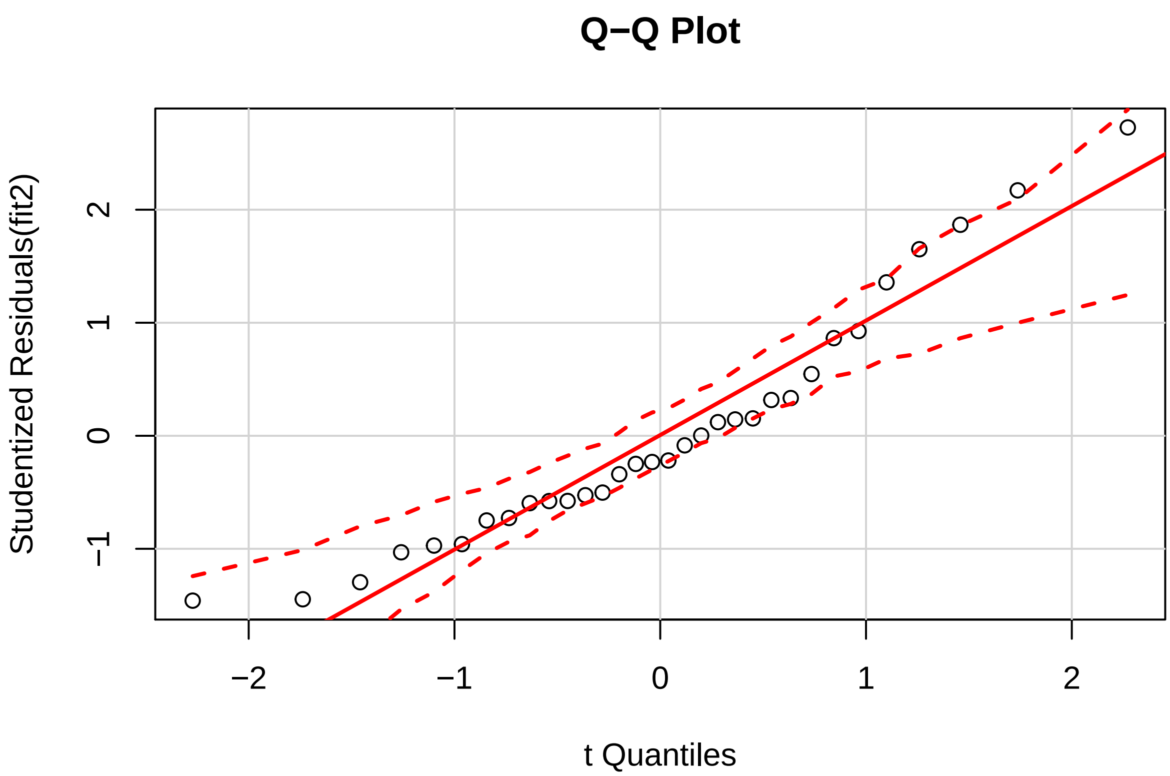
Appendix



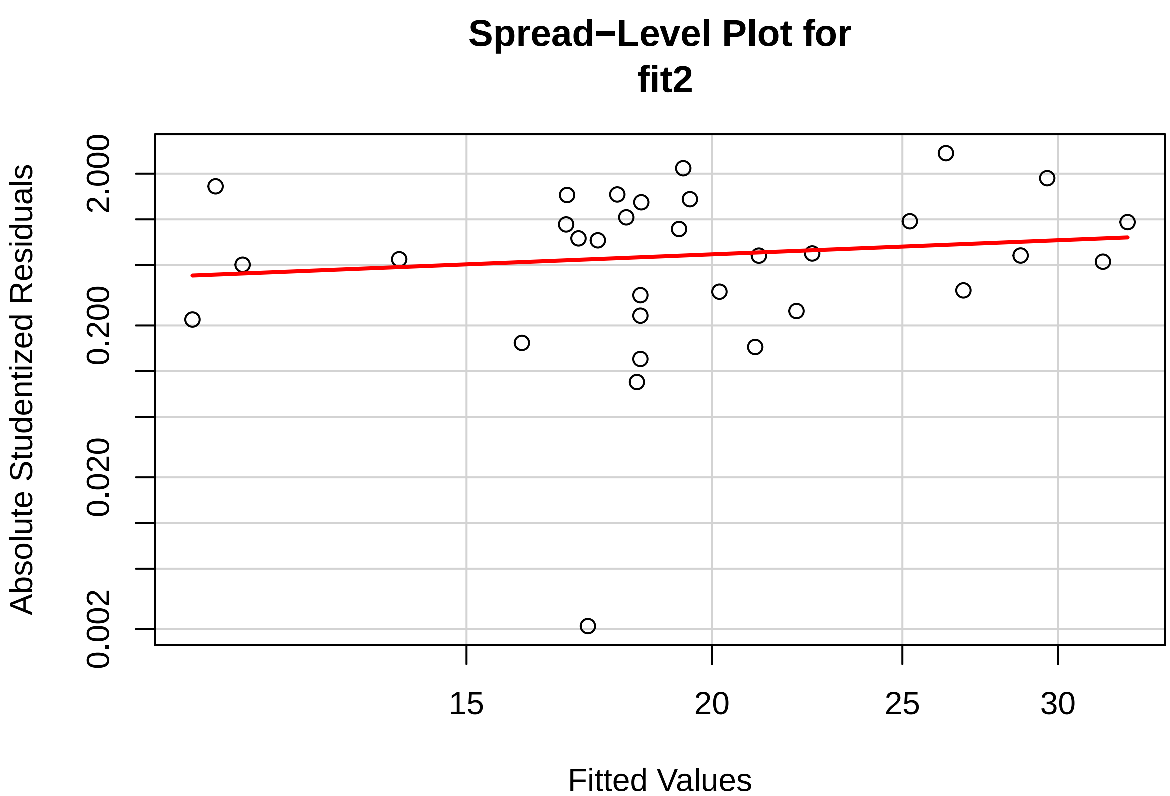
Scatter Plot Matrix








```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8905    Df = 1     p = 0.3453
```



```
##
## Suggested power transformation:  0.4726
```