Exploring the relationship between a car's transmission type and miles per gallon

Executive summary

This study examines the relationship between a set of variables and miles per gallon (MPG) for a collection of cars. The main purpose of the investigation was to establish whether automatic or manual transmission is better for MPG, and to quantify the relationship between transmission type and MPG, if it indeed exists. When ignoring other variables, data seemed to suggest that manual transmission is better for MPG. However, basic exploratory analysis revealed that a car's weight and horse power are important confounding variables. Consequently, proper regression analysis that accounted for confounding variables suggested that the difference in MPG by transmission type was not statistically significant (based on 5% confidence level). Therefore, a quantitative relationship between MPG and transmission type could not be established. Several regression models were explored and discussed in the context of the analysis.

Analysis

In order to address the question of whether transmission type is related to MPG, we first proceed by displaying a 2d plot of the MPG vs. transmission type data. Figure 1 of Appendix shows that manual transmission cars (blue dots) have, on average, higher MGP compared to automatic transmission cars (red dots). A regression model that further considers the sole relationship between MPG and transmission type (as a factor variable) further confirms the significance of the relationship. Such model suggests that manual transmission cars have 3.6-10.9 higher Miles/(US) gallon than automatic transmission cars (with p-value<0.0003) as shown below:

```
fit1 <- lm(mpg ~ factor(am), data=mtcars)
summary(fit1)$coef
```

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

```
confint(fit1,level = 0.95)
```

```
##              2.5 % 97.5 %
## (Intercept) 14.851  19.44
## factor(am)1  3.642  10.85
```

However, this single variable model doesn't seem to explain much of the variation in the MPG data as evident by the r.squared value:

```
summary(fit1)$r.squared
```

```
## [1] 0.3598
```

More importantly, one can question the validity of the relationship established above, in light of exclusion of other variables from the analysis. For instance, it is rather intuitive that other variables such as horse power and weight affect fuel consumption. Could exclusion of these variables have led us to the wrong conclusion about the relationship between MPG and transmission type? Clues to this question are found in the following exploratory analysis. In Figure 2 of Appendix we look at the relationship between MPG and selected variables while highlighting whether the data points correspond to manual or automatic transmission. It is clear from the data displayed in Figure 2 of Appendix that cars with manual transmission happen to have, on average, lower horse power and lower weight. These correlations cannot be ignored, and should've therefore guided our selections of variables in the regression model.

Next, we attempt to account for confounding variables by considering the following two alternative multi-variable regression models: (1) a model that considers all variables in the dataset, and (2) a model that considers those variables that are expected to be important, guided by our knowledge of factors that impact fuel consumption in a car and by the exploratory analysis performed before.

We find that when including all variables in the regression model, no statistically significant relations are established:

```
fit2 <- lm(mpg ~ factor(am) + disp + drat + qsec + wt + hp + factor(cyl) + gear + factor(vs) + factor(carb),
data=mtcars)
summary(fit2)$coef
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.44392   18.14546  1.1267  0.27650
## factor(am)1    1.12376    2.61635  0.4295  0.67328
## disp           0.03627    0.02766  1.3110  0.20836
## drat           1.15841    2.35934  0.4910  0.63010
## qsec           0.34902    0.83164  0.4197  0.68030
## wt            -4.56753    2.35059 -1.9431  0.06981
## hp            -0.06977    0.03552 -1.9644  0.06710
## factor(cyl)6  -2.70547    2.73843 -0.9880  0.33788
## factor(cyl)8  -0.54243    5.70947 -0.0950  0.92549
## gear           1.28349    1.77122  0.7246  0.47914
## factor(vs)1    1.87682    2.58241  0.7268  0.47787
## factor(carb)2 -1.02048    2.10315 -0.4852  0.63410
## factor(carb)3  3.01294    4.14982  0.7260  0.47830
## factor(carb)4  0.98284    3.77875  0.2601  0.79811
## factor(carb)6  4.42059    6.08568  0.7264  0.47809
## factor(carb)8  7.12434    7.73072  0.9216  0.37044
```

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

```
## [1] 0.8931
```

This is expected since the sample size is small and the standard error on the regression coefficients increases as the number of regressors are increased. Residual plots of the above fit are shown in Figure 3 of Appendix. Even though r.squared is large, p-values for all coefficients are greater than 5%. This is clearly a problem with the model itself since we know a priori that some relationships should be strong (e.g between MGP and weight).

Next, we attempt to explain the MPG data using the least number of necessary variables .The strategy here is to include two confounding variables weight and horse power (in addition to transmission type). The former is a physical property of the car which we know has influence on fuel consumption; with the expectation that the heavier the car, the lower the MPG. The latter also has influence on fuel consumption, where it is expected that higher horse power implies lower MPG. We elect not to select number of cylinders since it is highly correlated with horse power as shown here:

```
cor(mtcars$cyl, mtcars$hp)
```

```
## [1] 0.8324
```

We find that this alternative regression model yields some interesting results:

```
fit3 <- lm(mpg ~ factor(am) + wt + hp, data=mtcars)
summary(fit3)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 34.00288   2.642659  12.867 2.824e-13
## factor(am)1  2.08371   1.376420   1.514 1.413e-01
## wt          -2.87858   0.904971  -3.181 3.574e-03
## hp          -0.03748   0.009605  -3.902 5.464e-04
```

```
summary(fit3)$r.squared
```

```
## [1] 0.8399
```

```
confint(fit3,level = 0.95)
```

```
##                2.5 %  97.5 %
## (Intercept) 28.58963 39.4161
## factor(am)1 -0.73576  4.9032
## wt          -4.73232 -1.0248
## hp          -0.05715 -0.0178
```

Residual plots of the above fit are shown in Figure 4 of Appendix. There are significant relationships between MPG and both horse power and weight, as one would expect. However, when horse power and weight are held constant, switching from automatic to manual transmission does not increase MPG in a statistically significant amount. In specific, switching from automatic transmission to manual transmission while holding horse power and weight constant, changes MPG by an amount from -0.74 to +4.90 Miles/(US) gallon. Because zero lies within this range, we fail to reject the null hypothesis that the coefficient relating to transmission type is zero (this is also evident from the p-value of 0.14, which is larger than 0.05). Therefore, a quantitative relationship between MPG and transmission type cannot be established.

Appendix

Figure 1

```
plot(mtcars$am, mtcars$mpg, ylab="MPG", pch = 19, xlab="Transmission", col=ifelse(mtcars$am==0, "red", "blue"),
main="red=automatic, blue=manual")
```

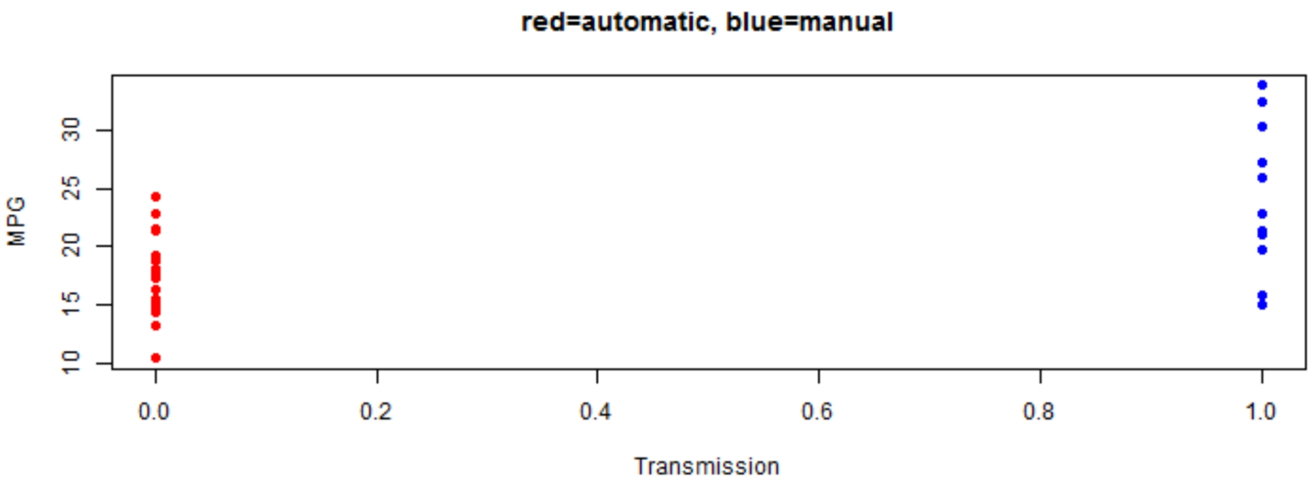

Figure 1: Scatter plot showing MGP vs. transmission type. Data suggests that manual transmission cars (blue) have higher MPG compared to automatic transmission cars (red dots).

Figure 2

```
par(mfrow=c(1,2))
plot(mtcars$hp, mtcars$mpg, ylab="MPG", xlab= "horse power", pch = 19, col=ifelse(mtcars$am==0, "red", "blue"),
main="red=automatic, blue=manual")
plot(mtcars$wt, mtcars$mpg, ylab="MPG", xlab= "weight", pch = 19, col=ifelse(mtcars$am==0, "red", "blue"),
main="red=automatic, blue=manual")
```

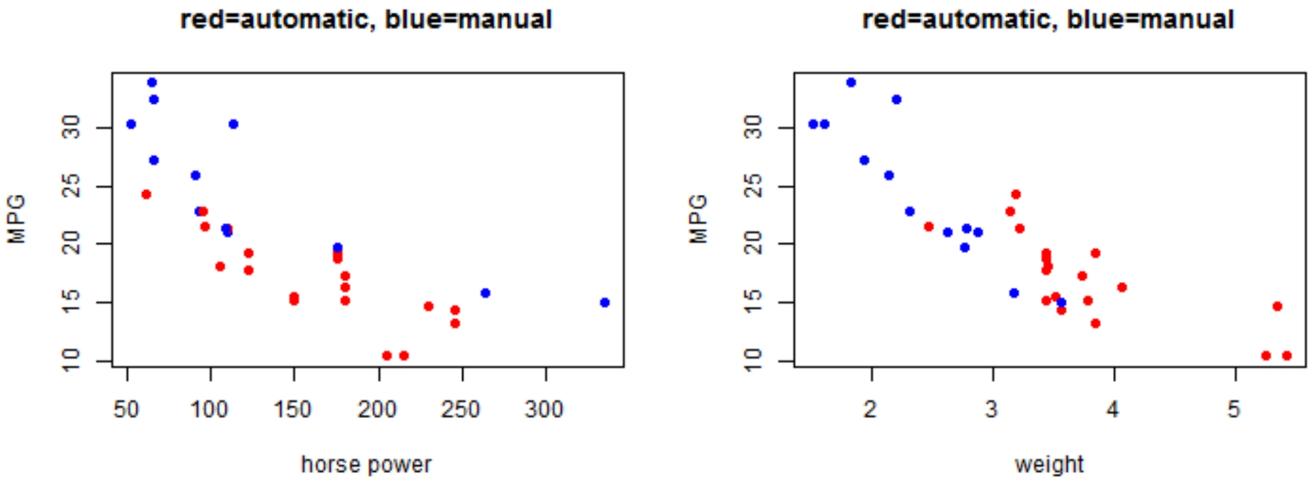

Figure 2: Scatter plots showing MGP vs. horse power (left) and MPG vs. weight (right). The two plots suggest that manual transmission cars (blue) are over represented in the lower horse power and lower weight ranges of the data.

Figure 3

```
par(mfrow=c(2,2))
plot(fit2)
```

```
## Warning: not plotting observations with leverage one:
##   30, 31
## Warning: not plotting observations with leverage one:
##   30, 31
```

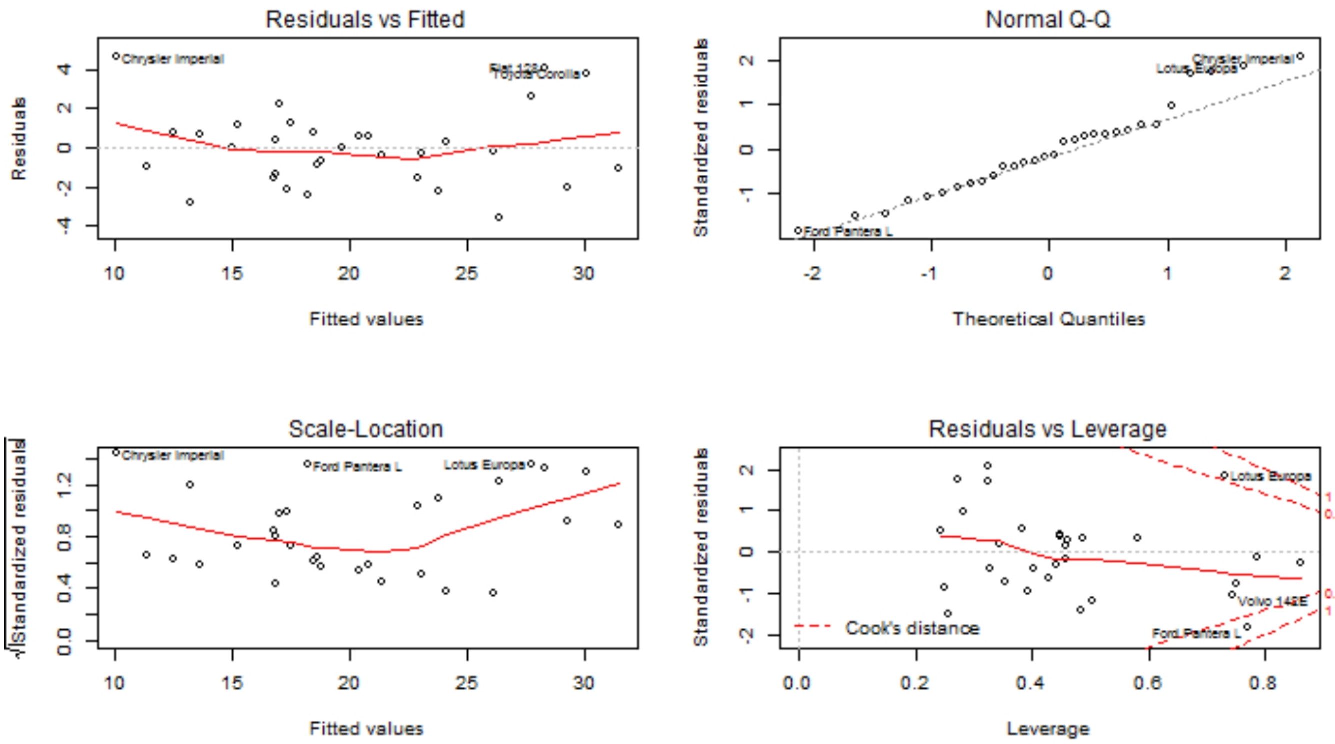

Figure 3: Various residual plots relating to the regression model fit2. Aside from few outliers, residuals seem to be randomly distributed around the zero baseline.

Figure 4

```
par(mfrow=c(2,2))
plot(fit3)
```

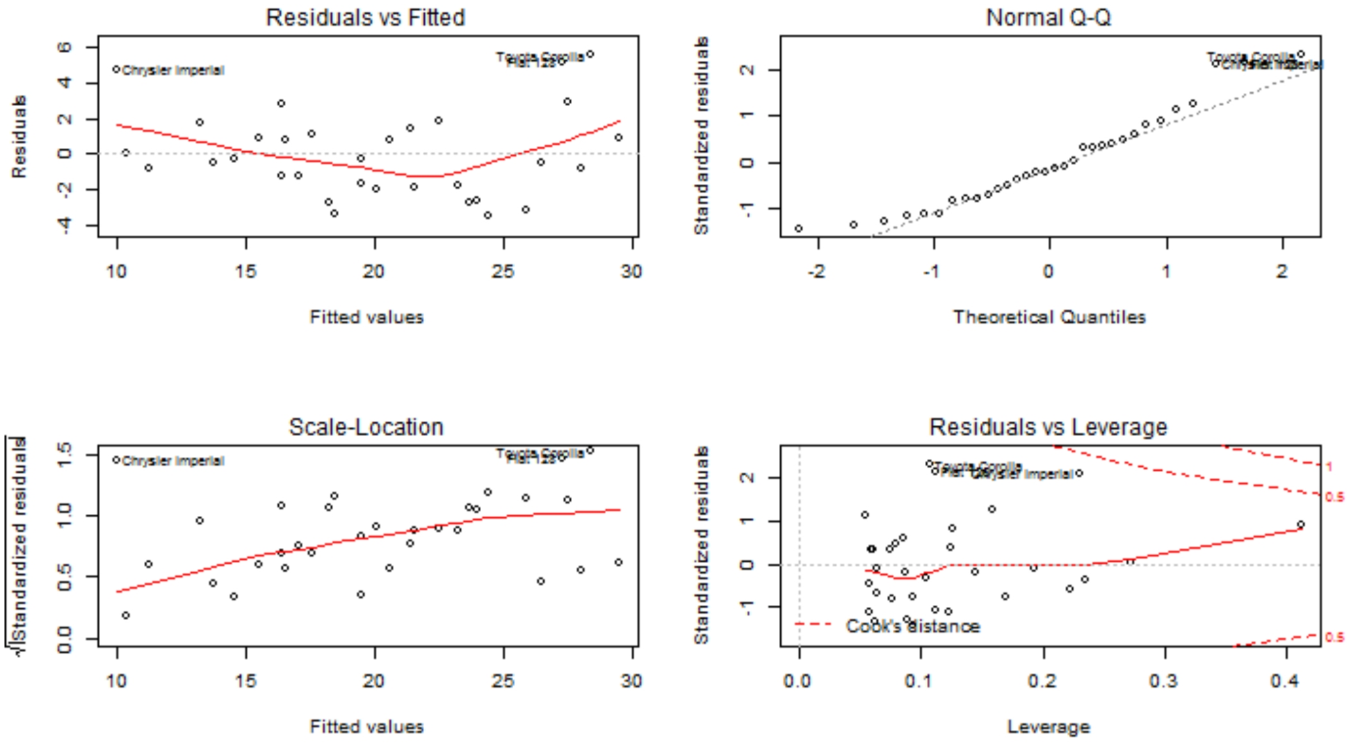

Figure 4: Various residual plots relating to the regression model fit3. Aside from few outliers, residuals seem to be randomly distributed around the zero baseline.