Regression Models - Course Project

adrianvs

Executive Summary

This report tries to answer two questions: which transmission type is better for MPG and how strong is its imapact on fuel efficiency. Although there is a significant difference in MPG between transmission types in the data set, with manual transition beeing more efficient, once the data is adjusted for other variables no significant difference remains. The estimated impact lies between -2.61 and 2.86 MPG for manual transmission (95% confidence interval for the manual transmission regression coefficient). The true impact can not be quantified from this data set.

Exploratory Analysis

The mtcars data set describes 10 variables of 32 cars (1973 - 1974 models) published in the 1974 Motor Trend US magazine. The observed varaibles are: miles per gallon, number of Cylinders, displacement, horsepower , rear axle ratio, weight, 1/4 mile time, v-shaped or straight engine, transmisison type, number of gears and number of carburetors. The following table gives a summary of the mileage with respect to transmission type.

Transmission	n	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
automatic	19	10.40	14.95	17.30	17.15	19.20	24.40	3.83
manual	13	15.00	21.00	22.80	24.39	30.40	33.90	6.17

Table 1: Summary: Miles per gallon by Transmission type

Figure 1 (appendix) shows, that if no other variables are accounted for there seems to be a distinct advantage for manual transmission in miles per gallon.

A closer look at the distribution of other variables that have an impact on miles per gallon with respect to transmission type point to a possible bias. Manual transmission cars in this data set seem to weigh less, have lower engine displacement, lower horsepower and lower acceleration (measured by quarter mile time). See **figure 2** of the appendix. The unpaired, two sample t-test results are shown in table 2.

```r
t1 <- t.test(df$mpg  ~ df$am); t2 <- t.test(df$wt   ~ df$am); t3 <- t.test(df$hp   ~ df$am)
t4 <- t.test(df$disp ~ df$am); t5 <- t.test(df$qsec ~ df$am)
```

	CI lower	CI upper	p-value
Mpg	-11.28	-3.21	0.001
Weight	0.85	1.86	0.000
Horsepower	-21.88	88.71	0.221
Displacement	75.33	218.37	0.000
QuarterMile	-0.49	2.14	0.209

Table 2: T-Test: Manual vs Automatic Transmission

Model Selection

To quantify the effect the mode of transmission has on fuel efficiency and account in part for the observed bias in other variables correlated with it, we fit a linear model that has to includes transmission type as a predictor. Since this model is not intended for prediction, all quadratic and interaction terms including the transmission variable will be omitted to allow for easy interpretability of this coefficient. In a model using all varibles as regressors none remains a significant predictor of fuel efficiency, displaying the effect of collinearity between the regressors. Coefficients p-values lie between 0.08 and 0.87 in this model.

Cyl.	Displ.	HP	Rear axle ratio	Weight	1/4 Mile time	Engine layout	Gears	Carburetors
-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	0.79	0.06

Table 3: Correlation of Varibles to Transmission Type

To adjust the effect of transmission type, variables, that are unevenly distributed between manual and automatic transmission and have a high correlation to fuel efficiency, are included stepwise into the model.

```
lm1 <- lm(mpg ~ am, data = df)
lm2 <- lm(mpg ~ am + wt, data = df)
lm3 <- lm(mpg ~ am + wt + hp, data = df)
lm4 <- lm(mpg ~ am + wt + hp + qsec, data = df)
lm5 <- lm(mpg ~ am + wt + hp + qsec + disp, data = df)
table3 <- anova(lm1,lm2,lm3,lm4,lm5)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.90				
29	278.32	1	442.58	74.99	0.0000
28	180.29	1	98.03	16.61	0.0004
27	160.07	1	20.22	3.43	0.0755
26	153.44	1	6.63	1.12	0.2990

Table 4: Analysis of Variance - Nested Likelihood Ratio

As can be seen by the rise in p-values, the inclusion of quarter mile time and engine disposition does not add significantly to the model but increases its variance. The third model including weight and horsepower ist the strongest so far. Since weight and horsepower are relatively strongly correlated, the inclusion of an interaction term in the linear model may better the fit.

```
lmInt <- lm(mpg ~ am + wt + hp + wt:hp, data = df)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
28	180.29				
27	129.72	1	50.57	10.53	0.0031

Table 5: Analysis of Variance(2) - Nested Likelihood Ratio

The inclusion of the interaction term adds significantly to the model. The residual sum of squares fell to below 130 and all coefficients are significant with pvalues from the F statistic below 0.005 in the ANOVA table. The adjusted R-squared of the model is 0.8677. The **manual transmission coefficient is 0.125** MPG more for manual transmission cars with all other model regressors held constant and a **95% Confidence Interval of -2.61, 2.86; no significant difference.**

Residuals and Diagnostics

A. Residuals The residual plot (**figure 3**, appendix) shows no obvious pattern in the residuals and confirms that a linear model fits the data. The Quantile-Quantile plot demonstrates that the residuals are nearly normally distributed.

B. Homoscedasticity The Scale Location plot tests whether the variance of the residuals change as a function of the fitted values. Since no clear trend in the red trend line is visible the residuals are assumed to be homoscedastic.

C. Influential Outliers No outlier has a Cook's Distance bigger than 0.2 and no high leverage, high standardized residual is present (see figure 3 (bottom right), appendix).

Appendix

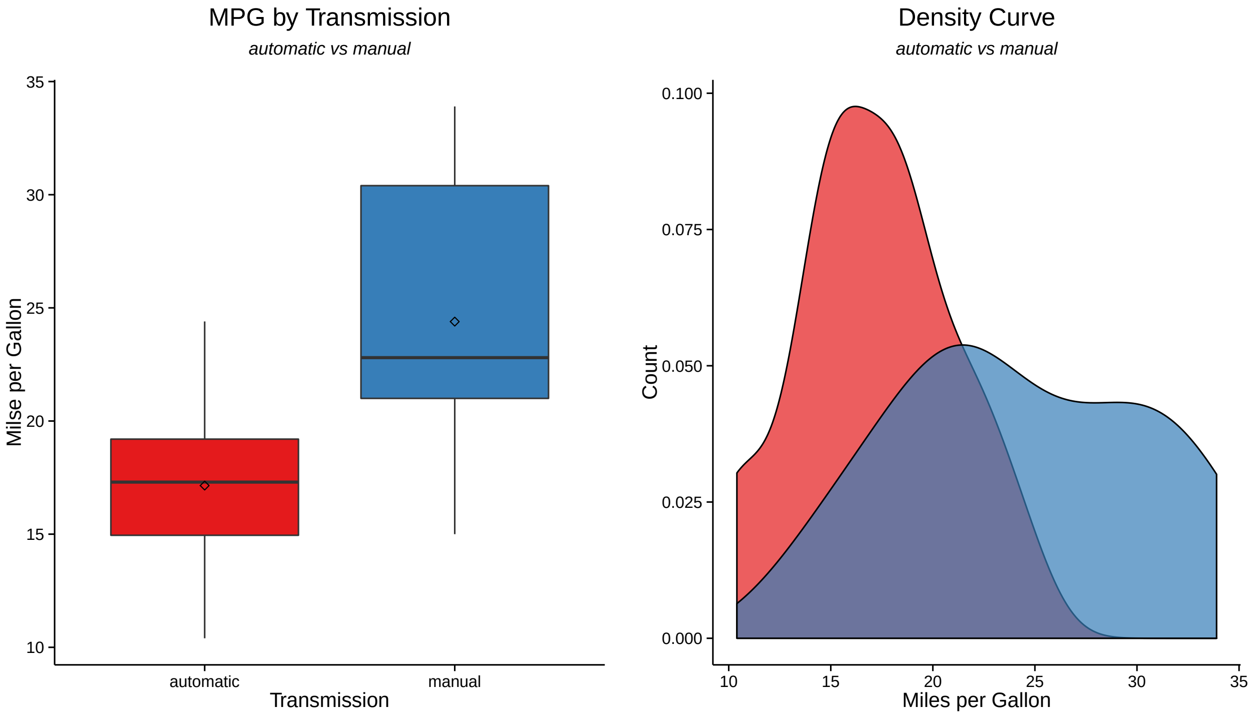

Figure 1: Miles per Gallon by Transmission Type

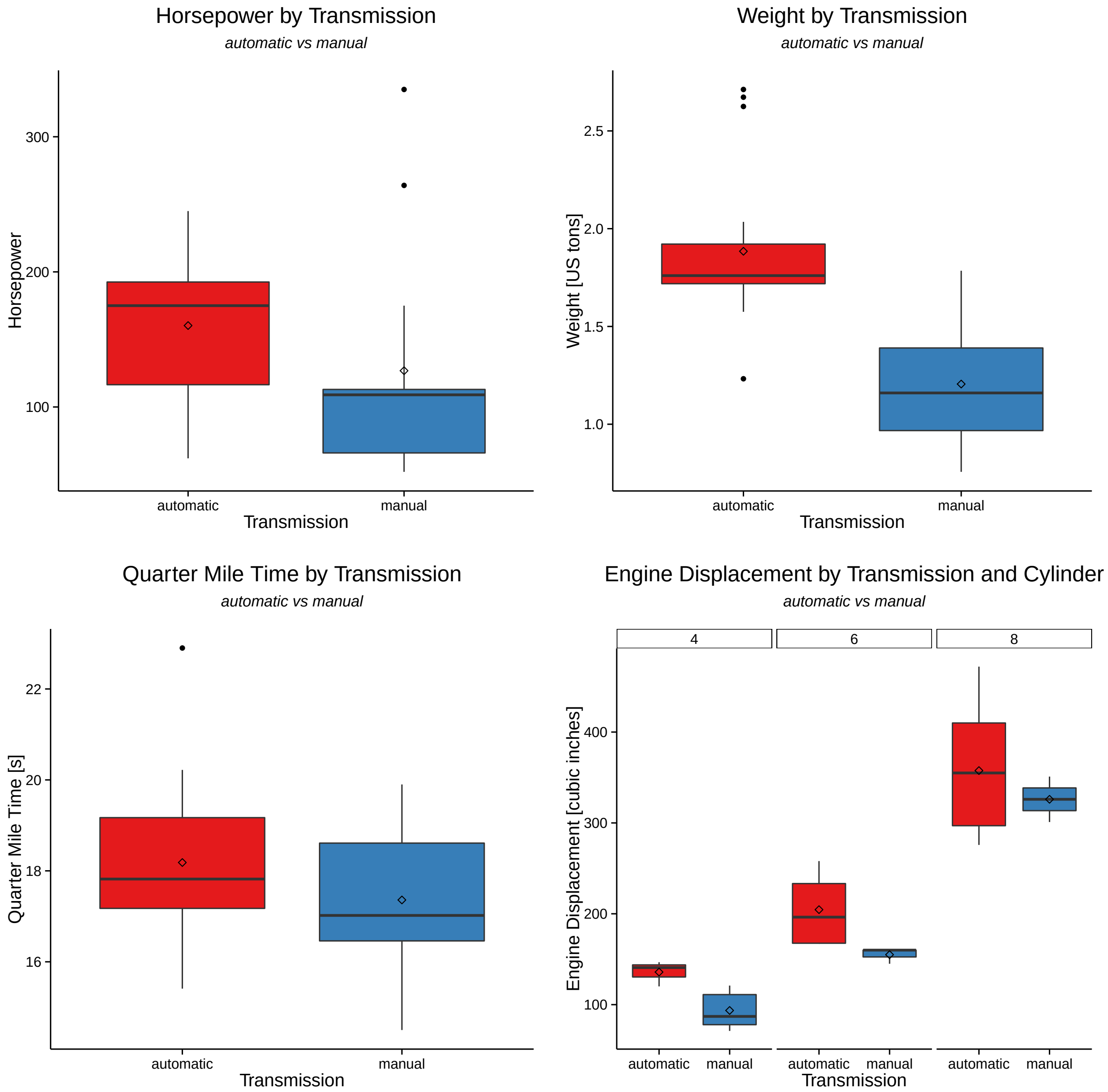

Figure 2: Distribution of Variables by Transmission Type

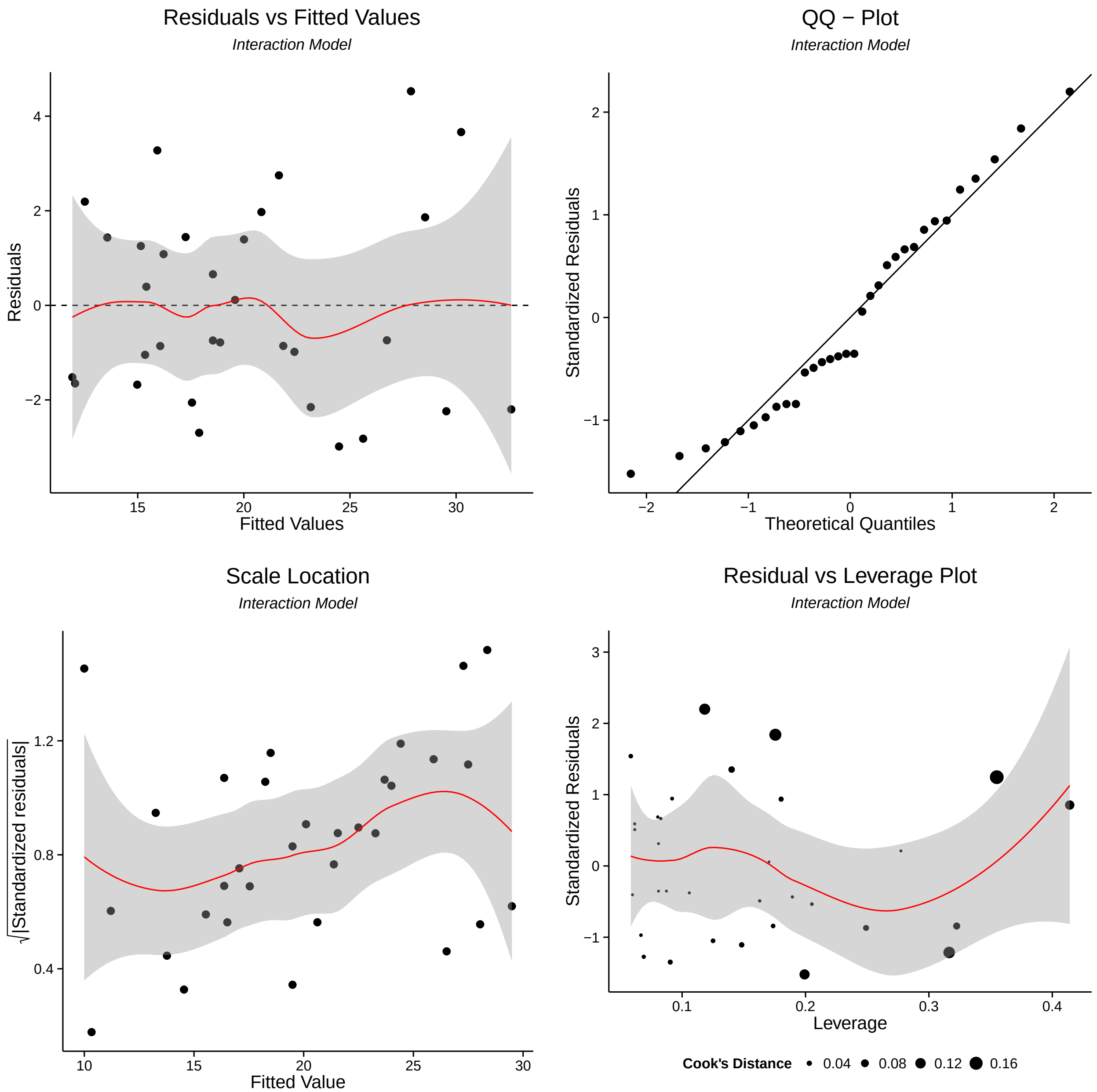

Figure 3: Residuals and Diagnostic Plots

The diagnostic plots used adapted ggplot2 code from **rpubs.com/therimalaya/43190**