Motor Trend Regression Report

Alex Petkovski

Sunday, June 21, 2015

1 Executive Summary

The following report analyses whether automatic or manual transmission improve MPG based on data extracted from 1974 Motor Trend US magazine and quantifies this difference. We found that manual performed better than automatic, when holding horse power and car weight fixed.

2 Data Exploration

After loading required packages (see Appendix), we construct a pairs plot in order to identify correlations by parameters shown in **mtcars**. Treating **Transmission(am)** as a factor variable shows boxplots and histograms for each possible parameter, which from initial inspection reveals that manual transmission generally has a higher mpg than automatic transmission when not factoring any other variables. The panel plot is shown in the Appendix. Next, we need to set factor variables for all factor variables (See Appendix).

3 Model Selection

As a consequence of the Panel Plot, we probably won't need qsec and gear due to low correlation with mpg. Also, from the model, we see highly correlated parameters likely potential to cause Variance Inflation.

3.1 All parameter model

To show this we fit a model with all parameters which showed high p-values and high variance inflation factors, thus our model requires a selection process. (See appendix for code and results)

Results in appendix show non-significant **p-values** for each beta.

4 Diagnostics

With each iteration we run diagnostics to determine if model has good R^2 and low **VIF**.

4.1 Initial Model using Transmission Type Only

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

From this model:
1. Estimate for mpg is 17.15 when transmission is automatic(am = 0) and 24.39 when transmission is manual(am = 1)
2. We also see that t values are sigmificant, both at less than 0.001

3. Without taking any other parameters other that transmission type, we see manual has better mpg than automatic, but we R-squared is at 0.36, so our model is biased and requires additional explanatory variables to make any reasoable conclusion

4.2 Add Horse Power (hp) to model

1. Coefficients have good p-values
2. This model with Horse Power added improves R^2 to 0.78
3. VIFs are low

```
##       am       hp
## 1.062867 1.062867
```

4.3 Add weight (wt) to model along with interaction on transmission

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 30.94733319 2.723410935 11.363446 8.546944e-12
## am1         11.55481296 4.023276579  2.871991 7.844579e-03
## hp          -0.02694935 0.009795903 -2.751084 1.047673e-02
## wt          -2.51558550 0.844496532 -2.978799 6.051842e-03
## am1:wt      -3.57790980 1.442795585 -2.479845 1.967639e-02
```

1. Coefficients continue to have good p-values
2. This model with Weight and interaction added improves R^2 to 0.87
3. VIFs are worse but still low

```
##        am        hp        wt     am:wt
## 22.972658  2.571184  3.891758 18.921765
```

4. Let's plot the residuals as this is our best model (See appendix for plot)
5. Residual plots don't show any obvious bias (See appendix for plot)

5 Conclusion

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp
## Model 3: mpg ~ am + hp + wt + am:wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 87.4211 5.885e-10 ***
## 3     27 146.85  2     98.59  9.0641 0.0009735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From ANOVA, we see that inclusion of all am, hp, and am:wt gives a model with high significance and low p-value. Best model from the three tested is fit3 and is significant in the alpha = 0.01 range. Finally, keeping hp and wt constant, manual transmission has higher mpg by 7.98. The table of standard errors are shown in 3.1 coefficients table.

6 APPENDIX

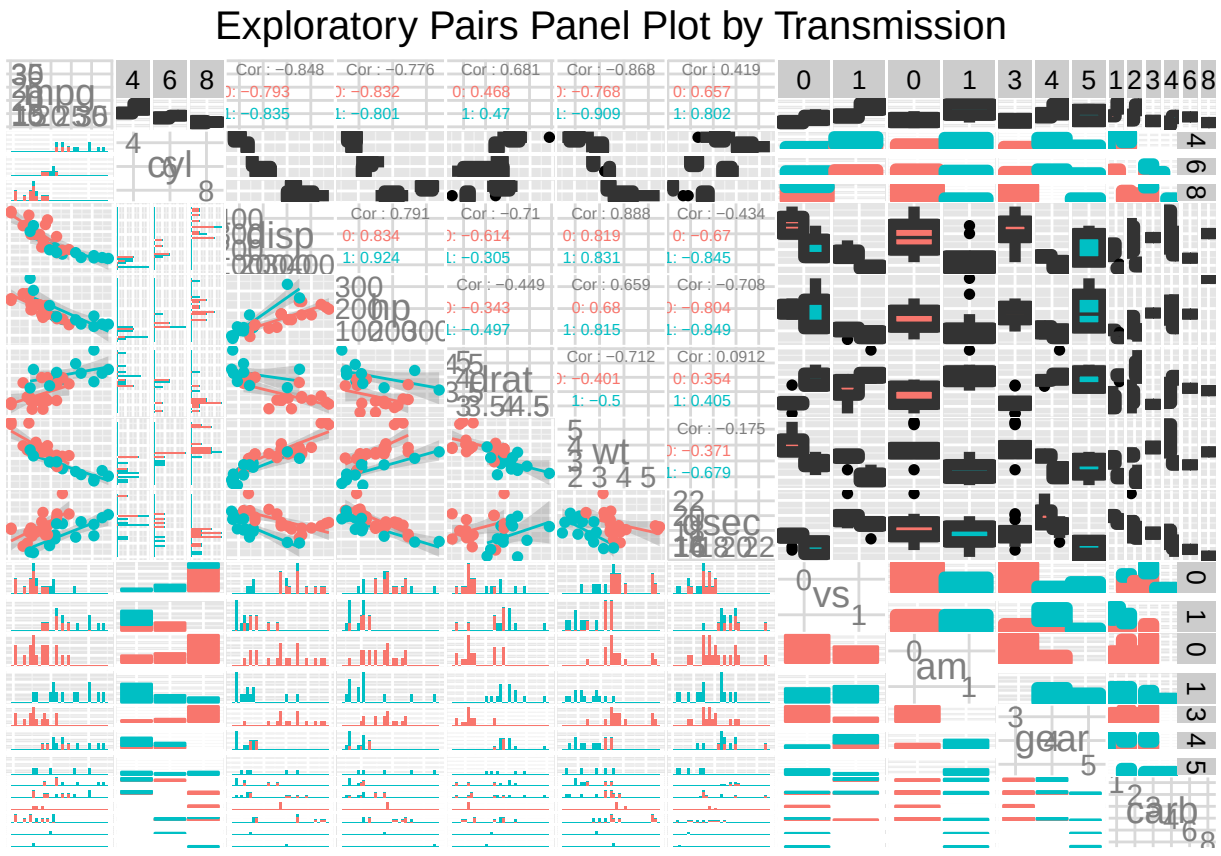
6.1 R Code used in the Study:

The following code was used to create and represent the data used in this analysis

```
library(datasets)
library(ggplot2)
library(GGally)
library(car)
```

The following code was used to create the panel plot described in Data Exploration section and the panel plot is shown

```
mtcars$am <- as.factor(mtcars$am)
g = ggpairs(mtcars,
            colour = "am",
            lower = list(continuous = "smooth"),
            upper = list(params=list(size = 2, fatten=0.3)),
            params = c(method = "loess"),
            axisLabels = "internal",
            title = "Exploratory Pairs Panel Plot by Transmission")
g
```



The following code was used to set factor variables

```r
mtcars$am <- as.factor(mtcars$am)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

The following shows code and coefs for the all parameter model

```r
par(mfrow = c(2, 2))
fit.all <- lm(mpg ~ ., mtcars)
summary(fit.all)$coef
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## am1          1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

The following shows code used to create the initial fit with mpg ~ am

```r
par(mfrow = c(2, 2))
fit <- lm(mpg ~ am, mtcars)
fit1 <- fit
summary(fit1)$coef
intercept <- summary(fit1)$coef[1]
am1 <- summary(fit1)$coef[2]
est_am1 <- intercept + am1
r.squared <- summary(fit1)$r.squared
```

The following shows code used to create the second fit adding hp

```r
par(mfrow = c(2, 2))
fit2 <- update(fit, mpg ~ am + hp)
coef2 <- summary(fit2)$coef
r.squared.2 <- summary(fit2)$r.squared
```

The following shows code used to create the second fit VIFs

```
vif(fit2)
```

The following shows code used to create the third fit adding wt and am:wt

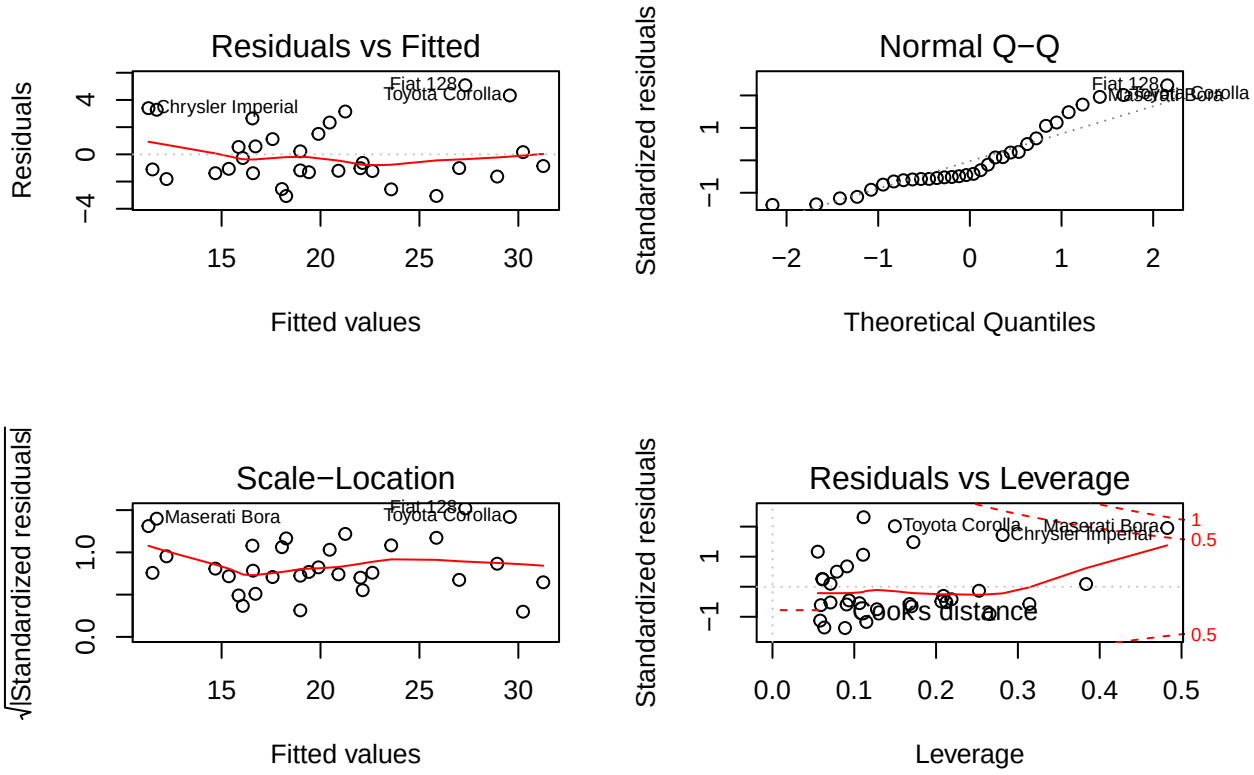
```
fit3 <- update(fit, mpg ~ am + hp + wt + wt:am)
summary(fit3)$coef
r.squared.3 <- summary(fit3)$r.squared
```

The following shows code used to create the third fit VIFs

```
vif(fit3)
```

The following shows code used to create the third fit residual plots

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



The following shows code used to create the ANOVA conclusion results

```
anova(fit1, fit2, fit3)
```