Transmission Effects upon Fuel Economy in Various Automobiles: a *Motor Trend* analysis

Executive Summary

As part of *Motor Trend*'s ongoing role as an authority in making actionable recommendations about fuel economy, we explored the relationship between a set of variables and miles per gallon (MPG) for a sample of cars. We asked the following questions: (a) Is an automatic or manual transmission better for MPG? And (b) In quantitative terms, what is the MPG difference between automatic and manual transmissions? Question (a) turned out to be preliminary at best. Indeed, the cars in our sample with manual transmissions were found to have better fuel economy. However, MPG figures also vary indirectly with a bundle of other, correlated attributes including weight and horsepower. Therefore, manual transmissions should not be considered "better" without this complicating factor. The remainder of the report fleshes this out and answers (b) in detail.

Preparation and Exploratory Analysis

Note: In order to conserve report space, we have removed the ```{r} and ``` labels which designate runnable R code. Reproducibility is still a priority. The R code is offset by solid grey boxes. This diverges from knitr style slightly but we have done it concertedly for space reasons.

We began with the following:

```
library(datasets);data(mtcars);library(dplyr);library(ggplot2)
# output included warning messages which are omitted here
```

We ran **str()** to get our bearings for the various columns and understand the datatypes and the range of values for a given field.

```
str(mtcars)
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The following manipulations are simply to be able to represent the transmission type as "automatic" or "manual" on the subsequent boxplot. We ran **mutate, rbind** and so forth, to produce **transmission**.

```
mtcars_automatic <- subset(mtcars,am=="0"); mtcars_manual <- subset(mtcars,am=="1"); mtcars_automatic_2 <-
mutate(mtcars_automatic,transmission="automatic"); mtcars_manual_2 <- mutate(mtcars_manual,transmission="manual"); mtcars2 <-
rbind(mtcars_automatic_2,mtcars_manual_2)
```

Next, we ran a boxplot which shows that the manual-transmission data has somewhat higher MPG figures, yet also raises questions about what else could be in play simultaneously. **Plots are in the report appendix. For the boxplot, please refer to Figure 1 in the appendix.**

While the boxplot might suggest that the transmission tells the story, we then fit a model based on **mpg** and **transmission**. The summary output has been hand-reformatted to conserve report space.

```
fit1 <- lm(mpg ~ factor(transmission),data=mtcars2)
summary(fit1)
## Residuals: Min -9.3923 / 1Q -3.0923 / Med  -0.2974 / 3Q 3.2439 / Max 9.5077
## Coefficients:
##                                Estimate    Std. Error    t value    Pr(>|t|)
## (Intercept)                    17.147      1.125         15.247     1.13e-15 ***
## factor(transmission)manual     7.245       1.764         4.106      0.000285 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The R-Squared value is only about 36%. This suggests that a multivariate approach is warranted. We will not dwell on the inference work which is inline in the summary of this model, since it is quickly superceded by the multivariate section.

Multivariate regression: What else is going on?

At this point, we explore the data a little more. This might also be attained with a **ggpairs** plot, but that might be a barrage of information. We wanted to dig in

and understand the possibilities, since the number of fields does not make it impractical to go one by one. The goal at this point is the comprehensive and thorny problem of fitting a new, multivariate model that will include what needs to be included while avoiding underfitting and overfitting.

To begin, we discovered that several variables vary indirectly with **mpg**, and are strongly correlated with each other. For example: the correlation of **mpg** and **wt** (weight) is -0.87. The correlation of **mpg** and **hp** (horsepower) is -0.78. And **wt** and **hp** themselves have a correlation of 0.66. In the interest of pursuing this discernible inverse relationship, we disqualified the remaining fields from going into the new model. The fields **cyl**, **disp**, **hp**, **wt** and **carb** exhibit the relationship of interest. **For scatterplots showing this trend, please refer to Figure 2 in the appendix.**

This set of five may still be excessive for the model. At this point, we considered parsimony. We de-emphasized 'displacement' and the 'number of carburetors' because it departs from what is the most easily explainable in layman's terms. We also felt that this discarding was safe since the cluster of five are highly correlated with each other. We were going to want to throw out some of these by some means, to minimize variance inflation issues. Therefore, **disp** and **carb** are eliminated.

We assembled the remaining variables in order of decreasing ease and parsimony: **wt**, then **hp**, then the discrete number of cylinders, represented as a factor. We created these models and compared them in a nested fashion using ANOVA.

```
fit_anova1 <- lm(mpg ~ factor(am),data = mtcars)
fit_anova2 <- lm(mpg ~ factor(am) + wt ,data = mtcars)
fit_anova3 <- lm(mpg ~ factor(am) + wt + hp ,data = mtcars)
fit_anova4 <- lm(mpg ~ factor(am) + wt + hp + factor(cyl) ,data = mtcars)
anova(fit_anova1,fit_anova2,fit_anova3,fit_anova4)
## Analysis of Variance Table
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + hp
## Model 4: mpg ~ factor(am) + wt + hp + factor(cyl)
##   Res.Df RSS    Df Sum of Sq F      Pr(>F)
## 1 30     720.90
## 2 29     278.32 1  442.58    76.1924 3.32e-09 ***
## 3 28     180.29 1  98.03     16.8762 0.0003525 ***
## 4 26     151.03 2  29.27     2.5191  0.0999982 .
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using standard threshholds, models through **fit_anova3** are statistically significant (p ~= 0.0003). We discarded **cyl** at this point. The data therefore suggests that the added explanatory value of **fit_anova3** is not due to mere chance, and we are ready to draw conclusions based upon it.

Conclusions

With the model finalized, we ran the summary. Please see below. The Conclusions and Residuals sections should be read in tandem. We accept the following conclusions in conjunction with scrutinizing the residuals, which follow. The summary output has been hand-reformatted to conserve report space.

```
summary(fit_anova3)
## Residuals: Min  -3.4221 / 1Q  -1.7924 / Med  -0.3788 / 3Q 1.2249 / Max 5.5317
## Coefficients:
##                Estimate    Std. Error  t value  Pr(>|t|)
## (Intercept)    34.002875   2.642659    12.867   2.82e-13 ***
## factor(am)1    2.083710    1.376420    1.514    0.141268
## wt             -2.878575   0.904971    -3.181   0.003574 **
## hp             -0.037479   0.009605    -3.902   0.000546 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

The R-Squared value is substantially higher with weight and horsepower included: around 84% of the variation is explained by the multivariate model, instead of only 36%. Now that the confounds have been accounted for, we can answer both questions. Question (a) turned out to be preliminary at best. Indeed, **the cars in our sample with manual transmissions were found to have better fuel economy. However, MPG figures also vary indirectly with a bundle of other, intra-correlated attributes including weight and horsepower. Therefore, manual transmissions should not be considered "better" without this complicating factor. Our conclusion to (b) is as follows: we expect a 2.08 mpg increase given a manual-transmission car (a discrete question, with only two possibilities.) There is a simultaneous decrease of -2.88 mpg for every 1000 lbs of weight added, as well as a nominal decrease as gross horsepower increases.**

Residuals

Plots are in the report appendix. For the residuals of fit_anova3, please refer to Figure 3 in the appendix. The technique used for the residuals is to take **predict(fit_anova3)** by **resid(fit_anova3)**. Following the residuals are the four diagnostic plots. **For the diagnostic plots of fit_anova3, please refer to Figure 4 in the appendix.** The residual plot and the first of the diagnostic plots appears to show some "stripey" sets of points, which may suggest that there is additional regularity above and beyond the 84% we have explained. In the interest of concise work, we conclude this phase of the analysis here.

Appendix

Figure 1. Boxplot of miles-per-gallon by transmission type

```
plot(mpg ~ factor(transmission),data=mtcars2)
```

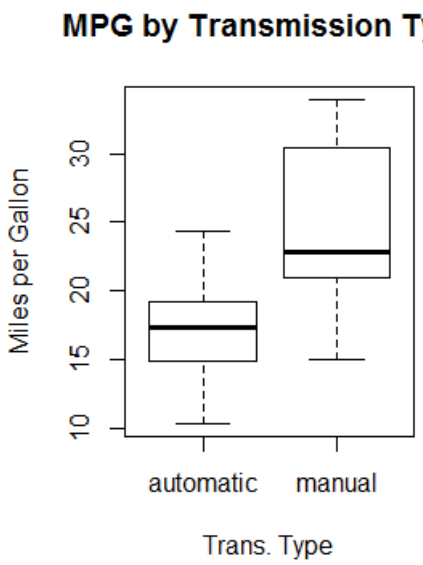
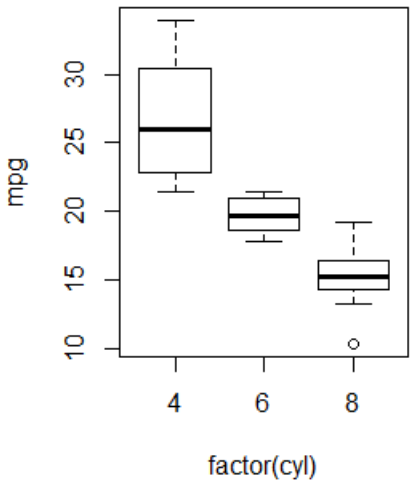
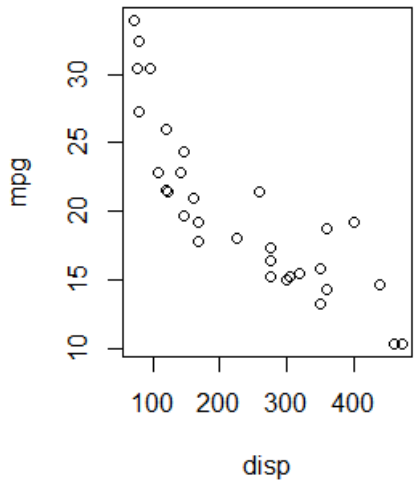


Figure 2. The following plots show that cyl, disp, wt, hp and carb vary indirectly with mpg:

```
plot(mpg ~ factor(cyl),data=mtcars2)
plot(mpg ~ disp,data=mtcars2)
plot(mpg ~ wt ,data=mtcars2)
```





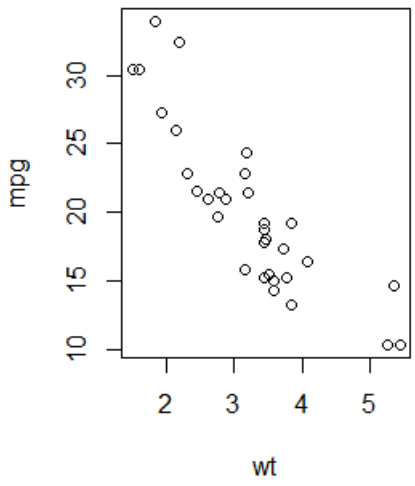


Figure 2. The following plots show that cyl, disp, wt, hp and carb vary indirectly with mpg (continued):

```
plot(mpg ~ hp ,data=mtcars2)
plot(mpg ~ factor(carb) ,data=mtcars2)
```

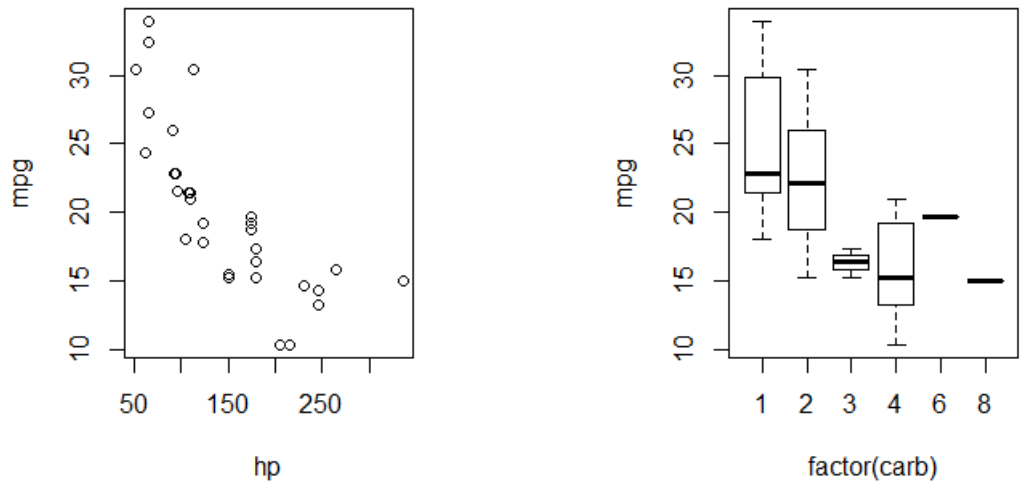


Figure 3. Residuals of our multivariate model: there may be additional regularity in the remaining ~14% variation

```
plot(predict(fit_anova3),resid(fit_anova3))
```

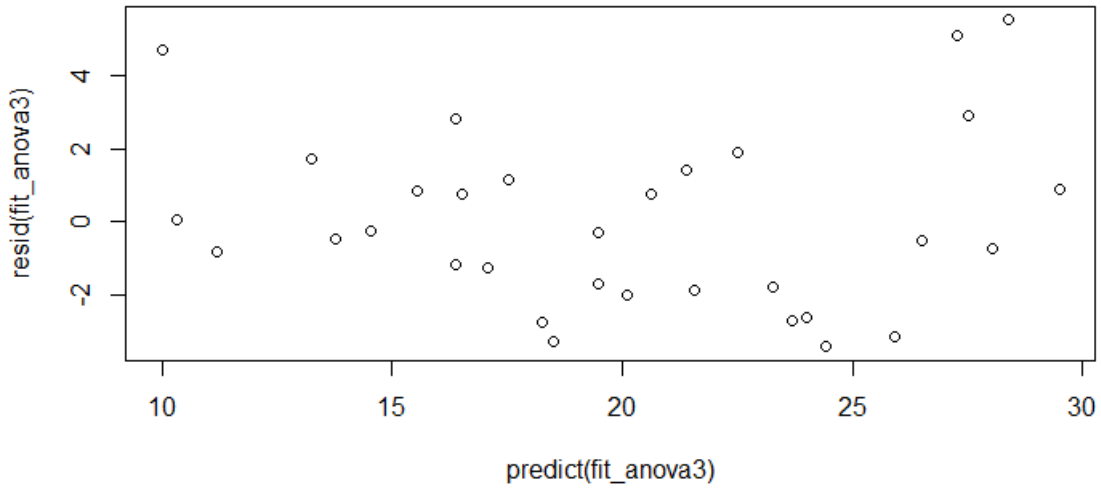


Figure 4. Diagnostic plots of our multivariate model: there may be additional regularity in the remaining ~14% variation

```
par(mfrow=c(2,2))
plot(fit_anova3)
```