Regression Models: Course Project

December 21, 2014

Executive Summary

The goal of this assignment is to find a relationship between a set of variables and miles per gallon (MPG) from the `mtcars` dataset. I particular, I am seeking to answer the following two questions: (1) *Is an automatic or manual transmission better for MPG?* and (2) *What is the MPG difference between automatic and manual transmissions?* The analysis finds that on average, standard transmissions beat automatic transmission in vehicle performance. This difference is 7.24 miles to the gallon in a simple linear regression, and a smaller (but nonetheless significant) 2.94 miles to the gallon when confounding variables are accounted for.

Exploratory Data Analysis

I have pre-processed the data so that `am` is a logical variable. An exploratory chart of all variables is in *Appendix A* (Chart #1). From initial inspection of this visualization, there appears to be strong correlation between `mpg` and the `cyl`, `disp`, `wt` variables (the lines seem to fit the data well). A more detailed visualization (Chart #2) of these variables faceted by `am` suggests these relationships exist regardless of transmission type (the green and blue lines are similar).

Model Fitting

Model 1

I begin the analysis by fitting a simple linear regression model of mpg to one of the hypothesized covariates from the exploratory stage. In this case, I have chosen to model automatic transmission, `am`, as it is the major variable of interest to this study. The coefficients of this model are presented here:

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

The `amTRUE` variable above is a dummy variable for standard transmission. The interpretation of the coefficients is that standard transmission improves car performance by an average 7.24 miles per gallon compared to automatic transmissions. A review of the model summary (see appendices) reveals that the t-test for $H_0 : \beta_{am} = 0$ versus $H_a : \beta_{am} \neq 0$ is statistically signficant ($p < .01$). The `summary(mod1)` printout (not shown) further reveals that the `am` variable captures an estimated 36% of the variance in `mpg`.

However, there may be confounding variables that explain this covariance. For this we turn to a more complex model that explores all variables.

Model 2

A more complex model may shed light on the true nature of the relationship between `mpg` and `am`. Many other linear models could be fit using combinations of the 10 available predictors. My strategy for model selection will be to use `R`'s built-in backward stepwise selection using the Akaike Information Criterion (AIC). The summary printout of this model follows:

```
## 
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4811 -1.5555 -0.7257  1.4110  4.6610 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.6178     6.9596   1.382 0.177915    
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amTRUE        2.9358     1.4109   2.081 0.046716 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336 
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The step function has trimmed the 10 independent variables down to a list of 3 predictors that best explain the variation in `mpg`. These are weight (`wt`), quarter-mile time (`qsec`), and transmission type (`am`). The interpretation of the coefficients is that holding all other variables fixed, standard transmissions increase MPG performance by 2.93 units. Fortunately, although this covariation is not as strong as originally thought, it is still significant (p < .05) and the magnitude is the same direction. The residual sum of squares is also much lower in the multivariate model (85% compared to 36%). The uncertainty is also lower in this model (Standard Error = 1.41 vs 1.76).

The diagnostic plots (*Appendix B*) reveal that there is no apparent heteroskedasticity or covariance in the standarized residuals. The Q-Q plot nicely fits the experimental quantiles along the identity line, suggesting normality. Accordingly, I am confident that parametic tests are appropriate and conclusions are robust.

Findings

An analysis of variance reveals that the improvements the second model are highly significant ($p < .01$). The printout of the anova is presented:

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ wt + qsec + am
## Model 2: mpg ~ am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)    
## 1     28 169.29                                 
## 2     30 720.90 -2   -551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

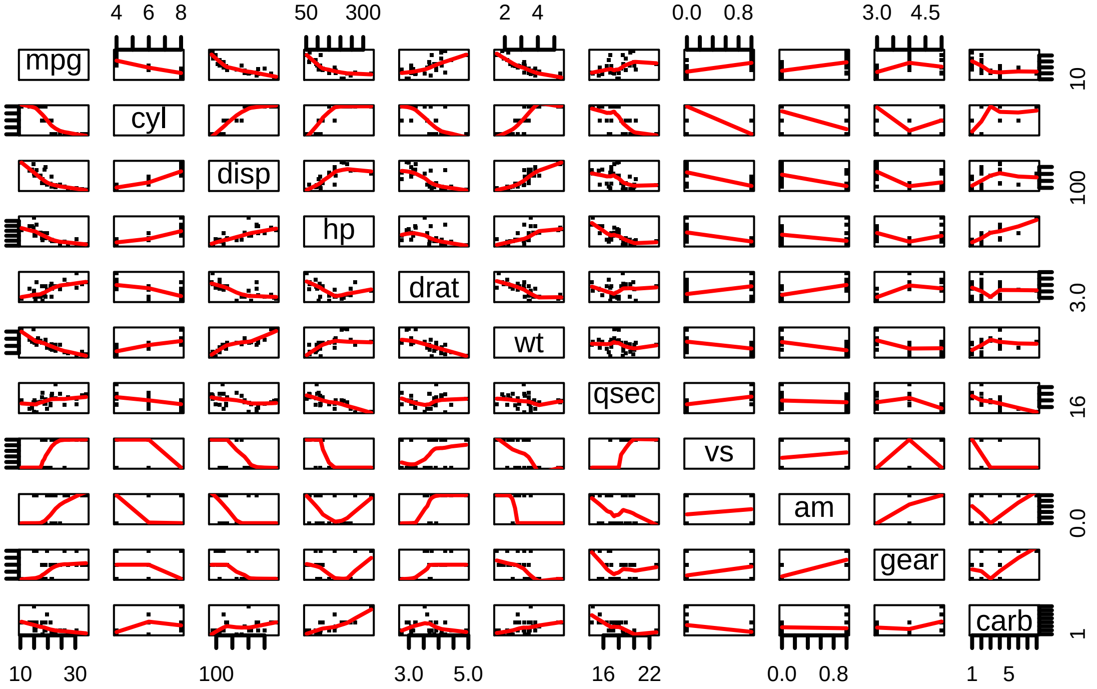
From the information above, we see that there are substantial variables that confound the relationship between MPG and the transmission type. Once these aspects are accounted for, we conclude that:

- a manual transmission is better for MPG; and
- Having a standard transmission improves car performance by an average 2.94 mpg compared to an automatic transmission.

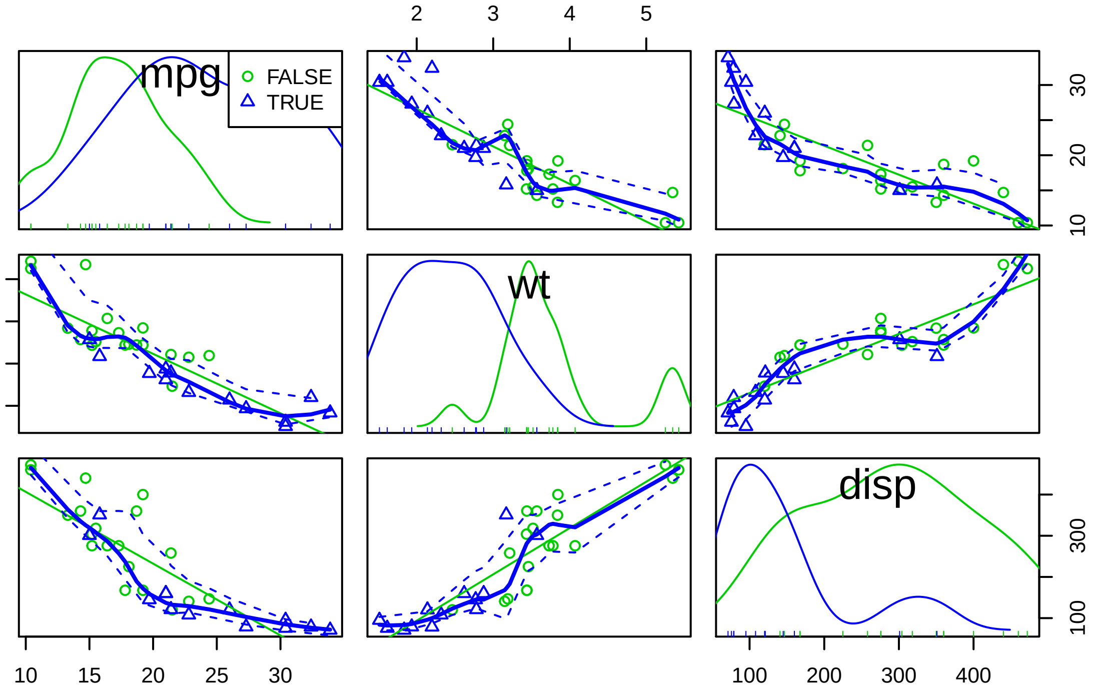
Appendices

Appendix A: Exploratory Charts

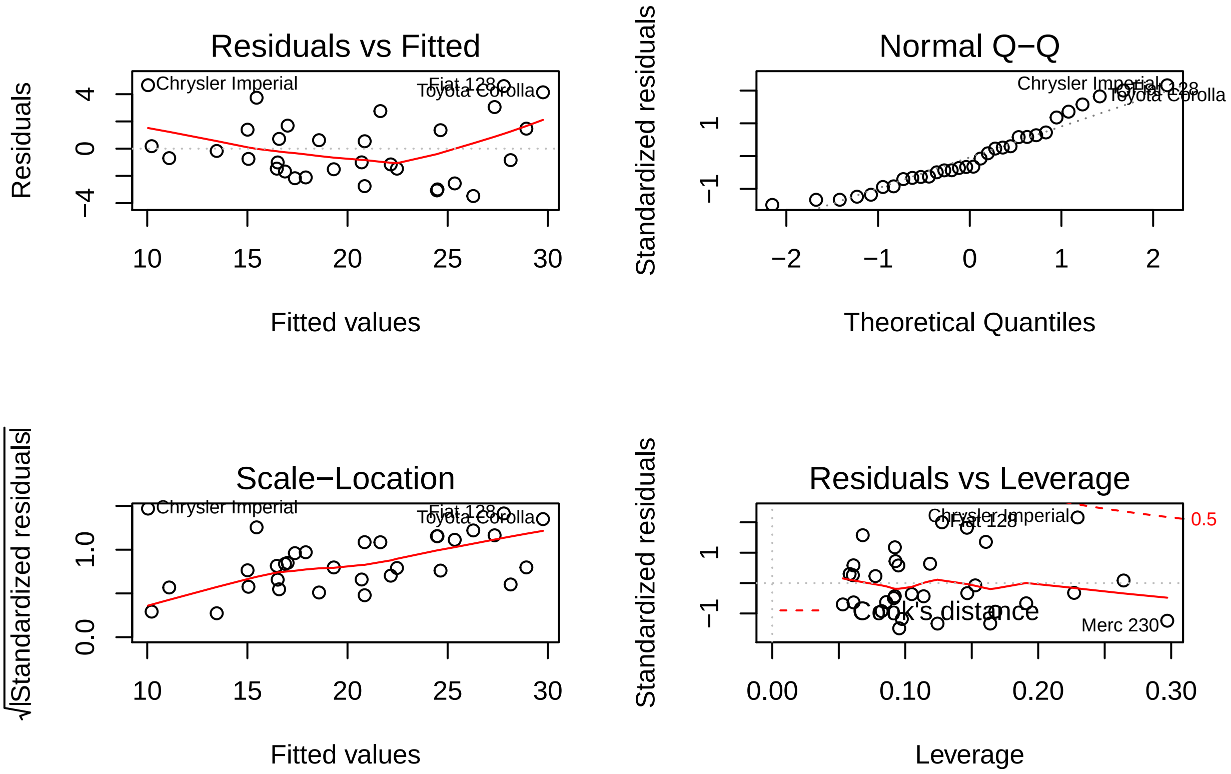
Exploratory Chart #1



Exploratory Chart #2



Appendix B: Diagnostic Plots (Multivariate Model)


Residuals vs Fitted
Residuals
−4
0
4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
−1
1
−2
−1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
0.0
1.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
−1
1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
Merc 230
0.5
Cook's distance