Effects of Transmission Type on Miles per Gallon

Toby Popenfoose

June 20, 2015

Executive Summary:

This analysis is of the mtcars dataset described in https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html. There are 19 cars with manual transmissions and 13 with automatic transmissions in the dataset. This report is of the inference from modeling the dataset to determine and quantify the difference in miles per gallon for an automatic versus a manual transmission. The multivariable linear model shows there is a statistically significant difference at a 95% level between manual and automatic transmissions. Accounting for weight and quarter mile time, the predicted difference in miles per gallon is 2.94 miles per gallon better with an automatic transmission than with a manual transmission. The lower 95% confidence interval is 0.05 miles per gallon and the upper 95% confidence interval is 5.83 miles per gallon due to the difference in transmissions. The multivariable linear model using transmission, weight and quarter second time explains 85% of the model variation compared to only 36% explained in the simple linear model of just transmission to predict miles per gallon. When testing the two models, the multivariable model is statistically significantly better than the simple linear model ata the 95% confidence level. (See https://github.com/tpopenfoose/RegressionModels for knitr source and rough draft)

Exploratory Data Analysis:

There were no missing values found.

For EDA I used the str, summary, cor and table functions. See the roughDraft report on github for the particulars. The roughDraft also has the pairs plot of all variables. See Figure 1.

It appears that two variables, am and vs, do not make sense as numeric. I will transform am and vs to factors so the modeling functions will behave better.

```
mtcars$am <- factor(mtcars$am)
mtcars$vs <- factor(mtcars$vs)
```

Simple Linear Model:

A naive starting point would be to just do a simple linear regression of just the independent variable am to predict mpg. After doing a simple linear model with just 'am' for the predictor, only 36% of the variation is explained by the model suggesting there are other confounder variables to be accounted for.

```
fit1 <- lm(mpg ~ am, mtcars)
summary(fit1)$r.squared
```

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

Multivariate Linear Model:

With 10 possible predictor variables there are 2^10= 1,024 different models. To save time, I used stepFit and regsubsets from the leaps package for variable selection using AIC in stepFit and BIC in leaps. See Figure 2 for graphs of the results. Based on the multiple models both methods tested, I have chose to use additional variables of weight and quarter mile time. Using transmission type, weight and quarter mile time explains 85% of the variance.

```r
finalFit <- lm(mpg ~ am + wt + qsec, data=mtcars)
summary(finalFit)$r.squared
```

```
## [1] 0.8497
```

```r
tidy(finalFit, conf.int=TRUE)
```

```
##          term estimate std.error statistic     p.value conf.low conf.high
## 1 (Intercept)    9.618    6.9596     1.382 0.177915165 -4.63830    23.874
## 2         am1    2.936    1.4109     2.081 0.046715510  0.04573     5.826
## 3          wt   -3.917    0.7112    -5.507 0.000006953 -5.37333    -2.460
## 4        qsec    1.226    0.2887     4.247 0.000216174  0.63457     1.817
```

Diagnostics and Inference:

The residual plots in Figure 3 show there is little heteroscadicity, they are mostly normal, there does not appear to be any significant outliers and there does not appear to be any data points with high influence or high leverage.

```r
anova(fit1, finalFit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
##   Res.Df RSS Df Sum of Sq    F       Pr(>F)
## 1     30 721
## 2     28 169  2       552 45.6 0.0000000016
```

The results of the anova are statistically significant at the 95% level and suggests that the multivariable model is better by looking at the residual sum of squares.

```r
vif(finalFit)
```

```
##    am    wt  qsec
## 2.541 2.483 1.364
```

The variance inflation factor suggests there is only moderate correlation between the variables selected which suggests there is not a large amount of collinearity between the variables.

```r
PRESS(fit1)
```

```
## [1] 830.3
```

```r
PRESS(finalFit)
```

```
## [1] 231.3
```

The PRESS statistic is less for the multivariable linear model which suggest is has higher predictive ability due to less predictive error than the simple linear model.

Appendix:

Figure 1. Ggpairs Plot of Selected mtcars Variables.

```
ggpairs(mtcars[ ,c(1,6,7,9)], lower=list(continuous="smooth"), colour="am", alpha=0.4,
        params = c(binwidth=.4), title="Figure 1.  Ggpairs Plot of Selected mtcars Variables.")
```

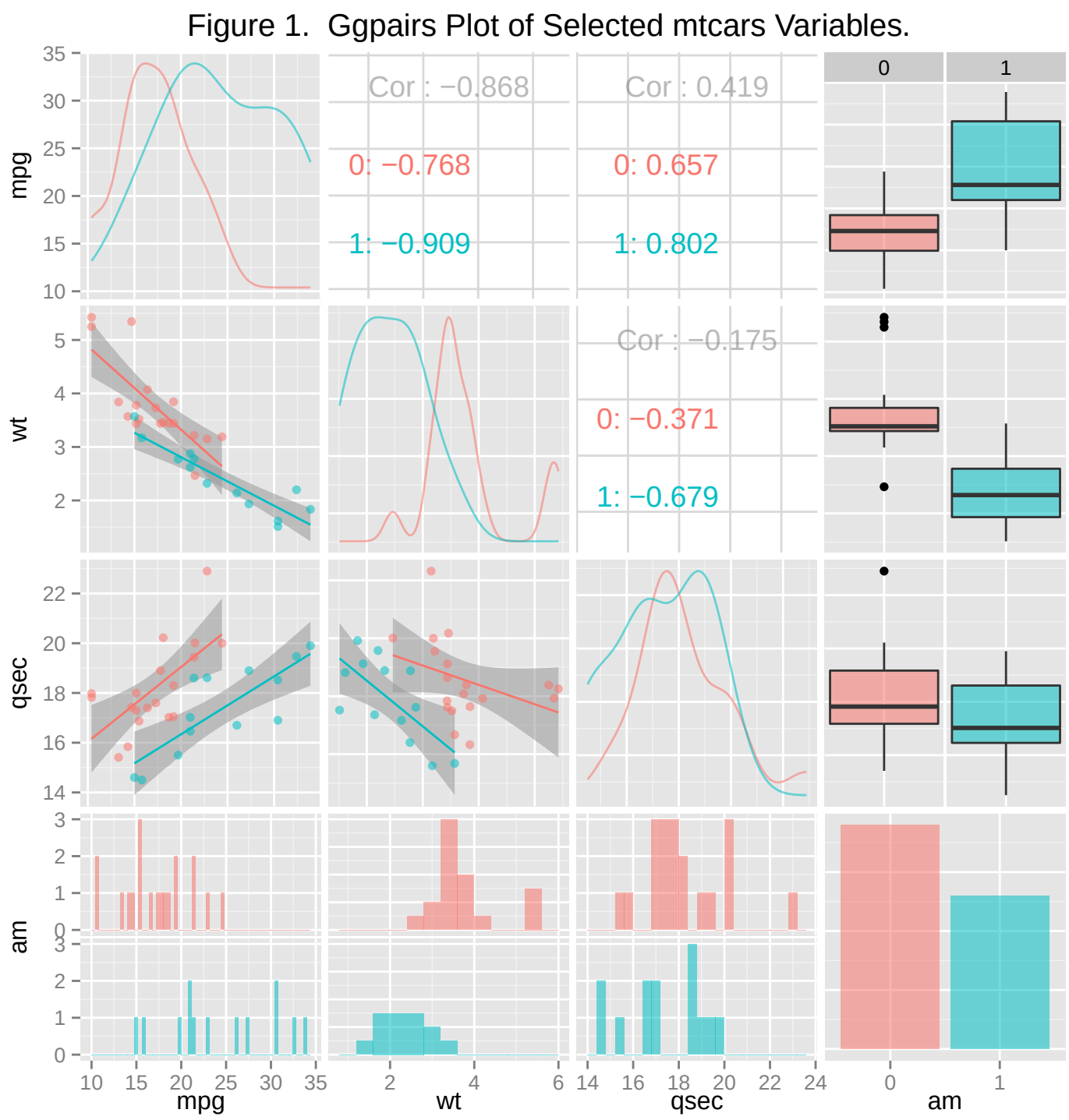

Figure 2. Multivariate Variable Selection.

```
par(mfrow=c(2,2))
plot(leapsSummary$rss, xlab="Number of Variables", ylab="RSS", type="l")
plot(leapsSummary$adjr2, xlab="Number of Variables", ylab="Adjusted RSq", type="l")
points(which.max(leapsSummary$adjr2),
       leapsSummary$adjr2[which.max(leapsSummary$adjr2)], col="red", cex=2, pch=20)
plot(leapsSummary$cp,xlab="Number of Variables", ylab="Cp", type="l")
points(which.min(leapsSummary$cp), leapsSummary$cp[which.min(leapsSummary$cp)],
       col="red", cex=2, pch=20)
plot(leapsSummary$bic, xlab="Number of Variables", ylab="BIC", type="l")
points(which.min(leapsSummary$bic), leapsSummary$bic[which.min(leapsSummary$bic)],
       col="red", cex=2, pch=20)
```

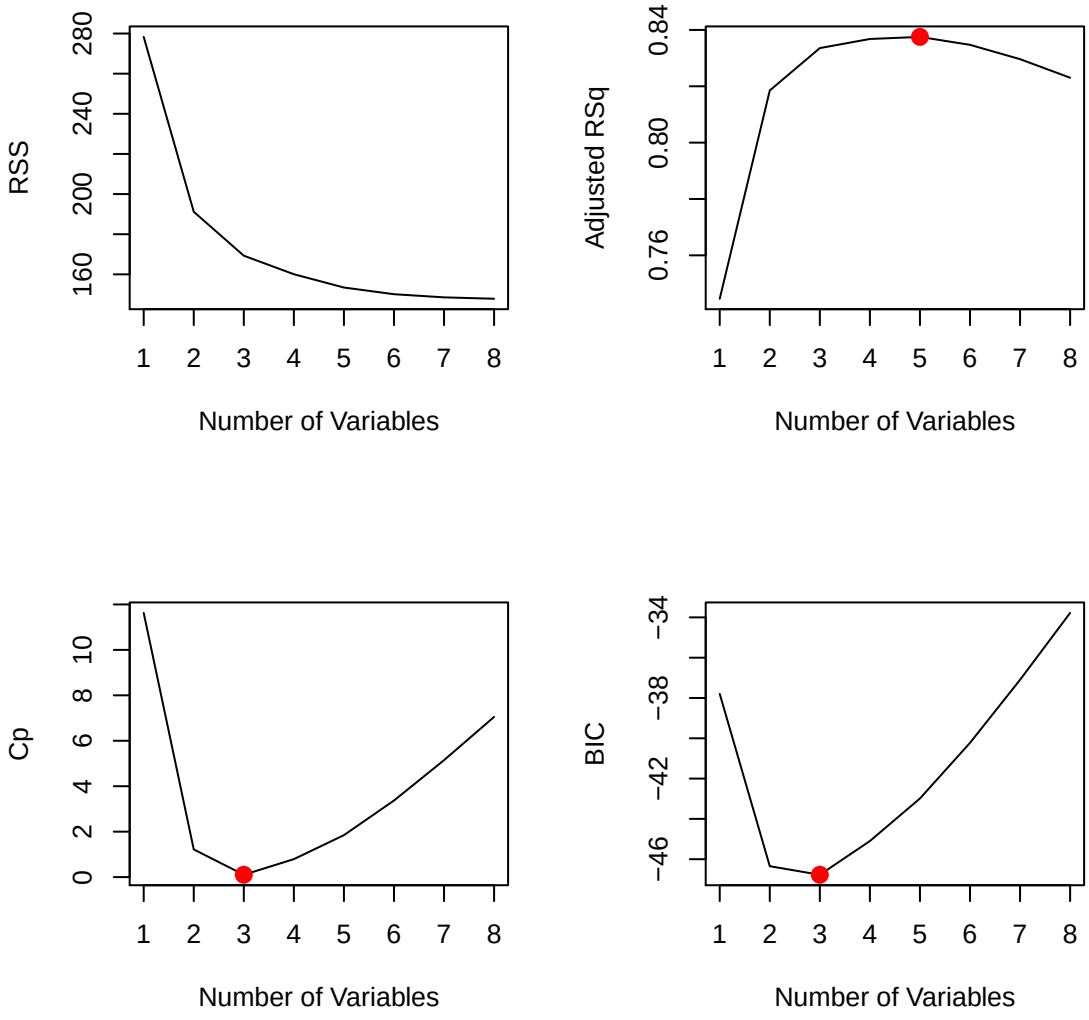

Figure 3. Residuals and Diagnostics Plots for Final Multivariable Linear Model.

```
par(mfrow=c(3,2)); plot(finalFit, which=1:6)
```