Regression Models Project

Mark S.

Thursday, May 21, 2015

I. Executive Summary

The goal of this project is to analyze vehicle data (mtcars data set) with respect to MPG. We wish to answer the following questions:

1. Is automatic or manual transmission better for MPG?
2. How much different is MPG between automatic and manual transmission vehicles?

Using a linear model to fit the data, we are 95% confident that there is no significant difference in mpg based on transmission type. According to the model, we see a difference of 2.9 ± 1.4 mpg between transmission types, but our analysis shows that this is not significant enough to claim a difference between transmission types with respect to MPG.

II. Exploratory Data Analysis

We would like to know which variables influence MPG, and we would like to quantify the difference in MPG by transmission type. We'll first inspect the correlations of all variables, and take a look at the MPG for each transmission type. A level plot of the correlation matrix is a usefual way to visualize these correlations, and a boxplot can help us visualize the difference between each transmission type. **Please see appendix for these plots.**

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

Note that many variables (predictors) are correlated with mpg, and many correlations exist between the predictors themselves. This hints that a model with few predictors may be possible. Based on the boxplot, it appears that there is a clear difference in MPG for each transmission type. However, the difference we see may be due to other confounding variables. In order to make an accurate comparison between transmission types, we must account for the other variables. To do this, we need to model the data.

III. Model Selection

Refer to the appendix for a detailed explanation of the model selection. The model selected takes the form: $\text{mpg} = 9.618 + 2.936 \times \text{am} + 1.226 \times \text{qsec} - 3.917 \times \text{wt}$. The following summarizes the model:

```
fit2<- lm(mpg~am+qsec+wt,df)
(summary(fit2)$coef)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## am           2.935837  1.4109045  2.080819 4.671551e-02
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
```

```
round(summary(fit2)$adj.r.squared,3)
```

```
## [1] 0.834
```

```
round(cor(fit2$residuals,select(df,-mpg,-am,-qsec,-wt)),3)
```

```
##         cyl  disp     hp  drat     vs   gear   carb
## [1,] -0.033 0.047 -0.105 0.056 -0.001 -0.016 -0.141
```

The model explains approximately 83% of the variance of the mpg data using the predictors am, qsec, and wt. This model tells us that, all else being equal, a vehicle with a manual transmission will have 2.9 ± 1.4 mpg greater than that of a vehicle with an automatic transmission.

IV. Results and Conclusions

The difference in MPG for transmission types is less when we take into account other confounding variables. To quantify the difference in MPG for transmission types under the new model, we will use the residuals after accounting for wt and qsec. A t-test will allow us to determine if there is a significant difference between transmission types:

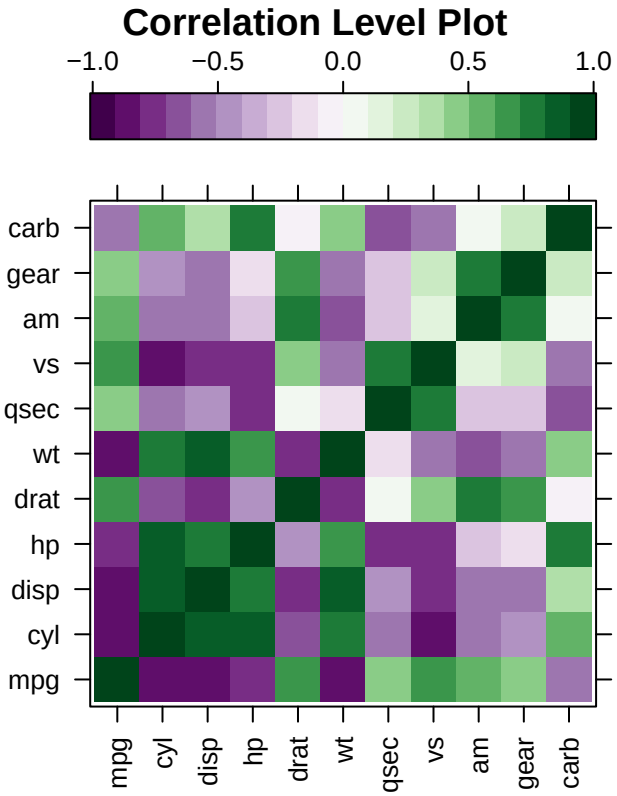
```
mpg.adj<- resid(lm(mpg~wt+qsec,df1))

#t-test for
#adjusted mpg for each transmission type
t.test(mpg.adj[df1$am=="Auto"],mpg.adj[df1$am=="Man"])
```

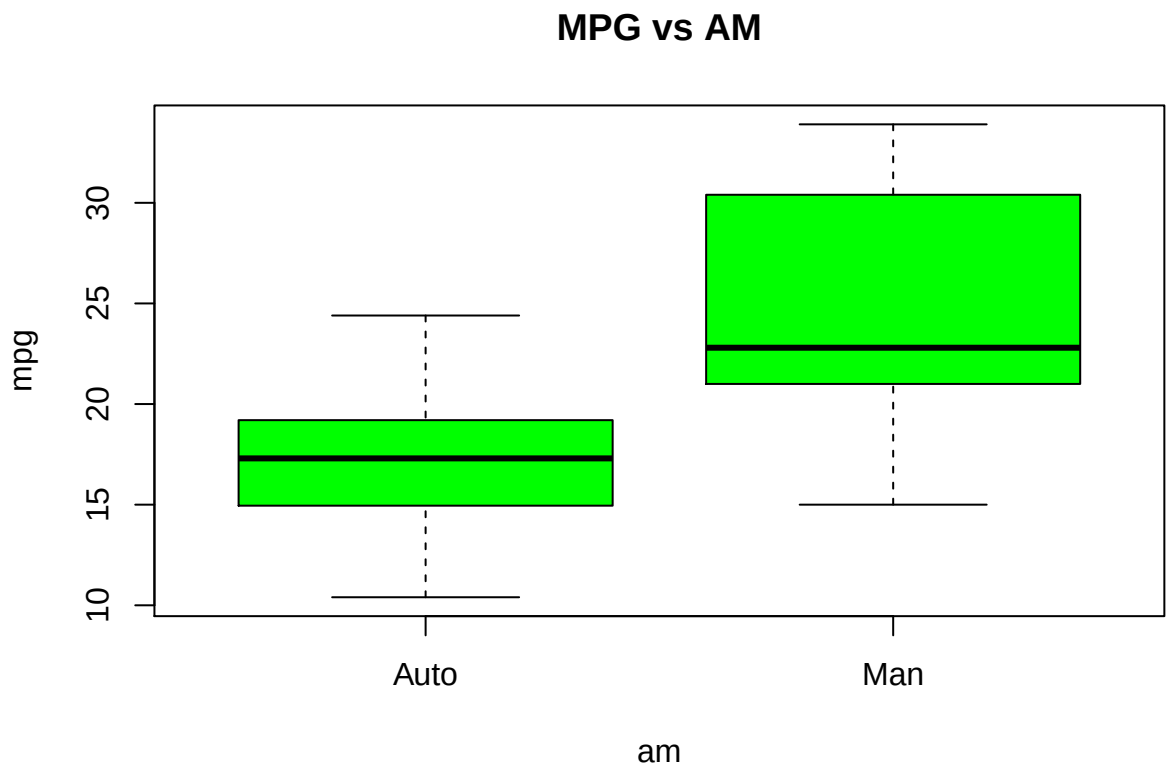
```
##
##  Welch Two Sample t-test
##
## data:  mpg.adj[df1$am == "Auto"] and mpg.adj[df1$am == "Man"]
## t = -1.2924, df = 25.983, p-value = 0.2076
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.9925121  0.6821366
## sample estimates:
##  mean of x  mean of y
## -0.4692950  0.6858927
```

This t-test compares the difference in means of mpg of each transmission type, when accounting for the wt and qsec predictors. The test results in a p-value of 0.20 which is greater than 0.05, and thus we fail to reject the null hypothesis. **Based on these results, all else being equal, we are 95% confident that there is not a significant difference in MPG between vehicles with automatic or manual transmissions.**

V. Appendix

A. Exploratory Data Analysis Plots The first plot is a levelplot, used to help visualize the correlations of variables in the data. The second plot is a boxplot of the mpg data for each transmission type.





B. Model Selection The exploratory analysis suggests a model with few predictors may be possible. The model selection process will begin with a base model with one predictor (transmission type). Additional predictors will be added one at a time, according to their correlation with the residuals of the current model. The adjusted r^2 value and variable inflation factors will be monitored to avoid having too many unnecessary predictors in the model.

We start by fitting the base model with mpg as the outcome and transmission type (am) as the predictor:

```
fit0<- lm(mpg~am,df1)
summary(fit0)$coef
```

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

```
round(summary(fit0)$adj.r.squared,3)
```

```
## [1] 0.338
```

This model quantifies the relationship we see in the boxplot. The intercept coefficient represents the mpg we should expect for a vehicle with an automatic transmission. The "amMan" coefficient is the additional mpg we add to the intercept if the vehicle has a manual transmission. The model equation looks like this: $y = 17.147 + 7.245x$, where x represents transmission type ($x = 0$ for automatic, $x = 1$ for manual), and y is the expected mpg. If the transmission is automatic ($x = 0$), then we expect 17.147 mpg for the vehicle. If the

transmission is manual ($x = 1$), then we expect 24.392 mpg for the vehicle. The model, as is, only explains approximately 34% of the variance of mpg, so we need to add more predictors.

To select the next variable, we first look at which variables are most correlated with the model residuals:

```
round(cor(fit0$residuals,select(df,-mpg,-am)),3)
```

```
##         cyl   disp     hp  drat     wt  qsec    vs  gear   carb
## [1,] -0.673 -0.616 -0.788 0.317 -0.565 0.696 0.704 0.005 -0.732
```

Horsepower (hp) is most correlated, so it is added to the model:

```
fit1<- lm(mpg~am+hp,df)
(summary(fit1)$coef)
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 26.5849137 1.425094292 18.654845 1.073954e-17
## am           5.2770853 1.079540576  4.888270 3.460318e-05
## hp          -0.0588878 0.007856745 -7.495191 2.920375e-08
```

```
round(summary(fit1)$adj.r.squared,3)
```

```
## [1] 0.767
```

```
round(cor(fit1$residuals,select(df,-mpg,-am,-hp)),3)
```

```
##         cyl   disp  drat     wt  qsec    vs  gear   carb
## [1,] -0.142 -0.127 0.148 -0.265 0.096 0.227 0.105 -0.158
```

The model is improving, but there are still correlations between the model residuals and predictors. This process is repeated, and eventually the predictors am, qsec, and wt are selected for the model (see Appendix for model diagnostics):

```
fit2<- lm(mpg~am+qsec+wt,df)
(summary(fit2)$coef)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## am           2.935837  1.4109045  2.080819 4.671551e-02
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
```

```
round(summary(fit2)$adj.r.squared,3)
```

```
## [1] 0.834
```

```
round(cor(fit2$residuals,select(df,-mpg,-am,-qsec,-wt)),3)
```

```
##         cyl  disp     hp  drat     vs   gear   carb
## [1,] -0.033 0.047 -0.105 0.056 -0.001 -0.016 -0.141
```

The final model takes the form: $\text{mpg} = 9.618 + 2.936 \times \text{am} + 1.226 \times \text{qsec} - 3.917 \times \text{wt}$, where $\text{am} = 0$ for automatic transmission and $\text{am} = 1$ for manual transmission.

C. Model Diagnostics The diagnostic plots show that the residuals appear to be normally distributed, with no observable pattern. The normal Q-Q plot shows a few outliers that the model does not fit particularly well, but the Residual vs Leverage plot shows that they will not significantly change the model by being removed: (**Note: The plots were distorted via knitr, and it may be difficult to read the details**)