Relationship between MPG and Transmission

October 9, 2014

Executive Summary

This analysis will explore the relationship between between a set of variables (predictors) and miles per gallon (MPG) (outcome) by looking at a data set of a collection of cars. Of particular interested is whether an automatic or manual transmission is better for MPG. The analysis will also quantify the MPG difference between automatic and manual transmission cars.

The analysis detailed in this paper performed a backward stepwise model selection, eliminating one variable from the linear model (based on the highest p-value) until all remaining models are statistically significant. A 95% confidence interval is run on the final model and the residuals are evaluated.

Exploratory Data

The variables in the mtcars dataset are:
- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (lb/1000)
- qsec: 1/4 mile time
- vs: V/S (engine shape) - am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Mean MPG by transmission type

```
##   am   mpg
## 1  0 17.15
## 2  1 24.39
```

The mean mpg for manual cars is higher than the mean mpg for automatic cars. This preliminary check indicates that manual cars get better miles per gallon.

Sample of the data

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1

Based on a review of the boxplot, mpg appears to be higher for manual transmission cars than for automatic transmission cars. The graphs for the explatory analysis can be found in the Appendix.

We reviewed the pairwise plot of all variables and see a strong linear relationship between mpg and each of disp, hp, wt. However, disp and hp also appear to have a strong linear with wt so we would not expect to see all of these variables in the final model.

Multiple Linear Regression Model

In order to find the best predictors for MPG, we will start with a model containing all predictor variables and cycle through the variables eliminating those which add no value to the linear model. We want a model where all remaining predictors are significant. We will do this by eliminating one predictor at a time (based on the highest p-value for each variable). The p-value will be compared to 5% to determine whether it is statistically significant.

Fit regression line for mpg using all variables as predictors

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788  0.6573  0.51812
## cyl         -0.11144    1.04502 -0.1066  0.91609
## disp         0.01334    0.01786  0.7468  0.46349
## hp          -0.02148    0.02177 -0.9868  0.33496
## drat         0.78711    1.63537  0.4813  0.63528
## wt          -3.71530    1.89441 -1.9612  0.06325
## qsec         0.82104    0.73084  1.1234  0.27394
## vs           0.31776    2.10451  0.1510  0.88142
## am           2.52023    2.05665  1.2254  0.23399
## gear         0.65541    1.49326  0.4389  0.66521
## carb        -0.19942    0.82875 -0.2406  0.81218
```

We see that cyl has the since highest p-value of 0.9161 so it will be removed.

Second, running a model without cyl, we find that vs the highest p-value (0.8433) so it is removed from the model.

Third, from the model which excludes cyl and vs, we find carb has the highest p-value of 0.747.

Fourth, from the model which excludes cyl, vs and carb, we remove gear since it has the highest p-value (0.6196).

Fifth, from the model which excludes cyl, vs, carb and gear, we remove drat since it has the highest p-value (0.4624).

Sixth, from the model which excludes cyl, vs, carb, gear and drat, we remove disp since it has the highest p-value (0.299).

Seventh, from the model which excludes cyl, vs, carb, gear, drat and disp, we remove hp since it has the highest p-value (0.2231).

```
fit_nohp <- lm(mpg ~ wt + qsec + am, data=mtcars)
summary(fit_nohp)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## am             2.936     1.4109   2.081 4.672e-02
```

```
max_pvalue <- max(summary(fit_nohp)$coef[2:4,4])
max_pvalue
```

```
## [1] 0.04672
```

Finally we see that the remaining p-values are all less than 0.05 and therefore statistically significant. At this stage, we cease removing variables from the linear model. Our final model includes transmission, weight and 1/4 mile time as predictors for miles per gallon.

The linear model is:

mpg(hat) = 9.6178 + -3.9165 * weight + 1.2259 * 1/4 mile time + 2.9358 * tranmission.

```
low <- summary(fit_nohp)$coef[4,1] - abs(qt(0.025, df = 28)) *
summary(fit_nohp)$coef[4,4]
high <- summary(fit_nohp)$coef[4,1] + abs(qt(0.025, df = 28)) *
summary(fit_nohp)$coef[4,4]
```

Interpretation of Results

All else held constant, the model predicts that manual transmission cars get 2.936 mpg more than an automatic car, on average. We are 95% confident that the model predicts that the mpg for manual cars is 2.8401 points to 3.0315 higher than for automatic transmission cars.

About 0.3598 of the variability in mpg is explained by transmission alone and 0.8497 is explained by a combination of transmission, weight and 1/4 mile time.

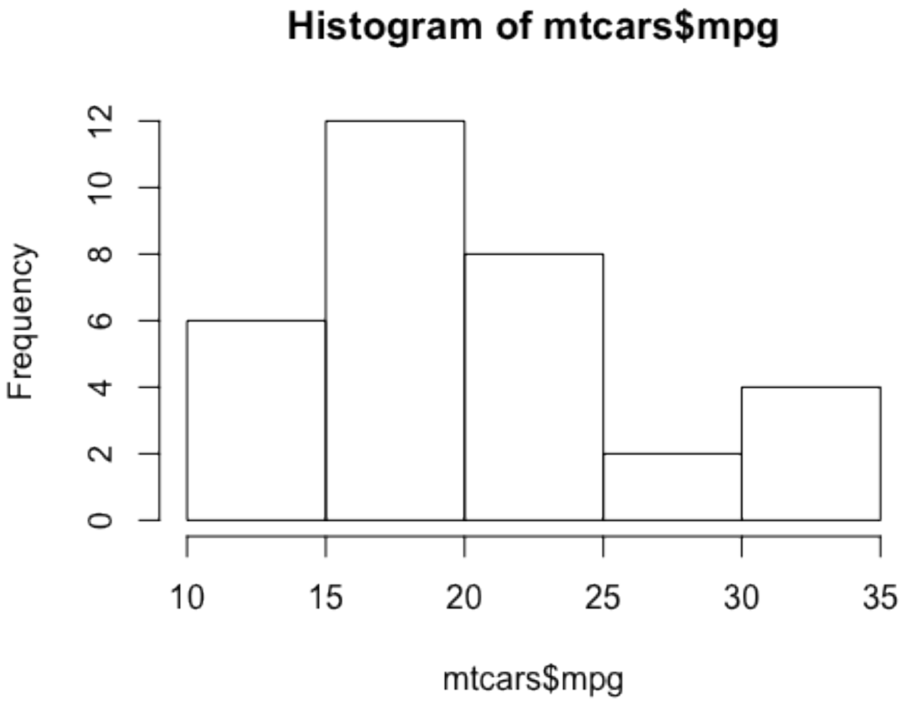
Diagnostics of the Residuals

We expect the residuals from our final linear model to be nearly normal with a mean of 0 and to have constant variance. The mean of the residuals is 5.2042 × 10-17. The residual plot (in the Appendix) shows that the residuals appear to be randomly disributed around 0. We check normality using the normal probability plot (in the Appendix). This plot shows a nearly normal distribution.

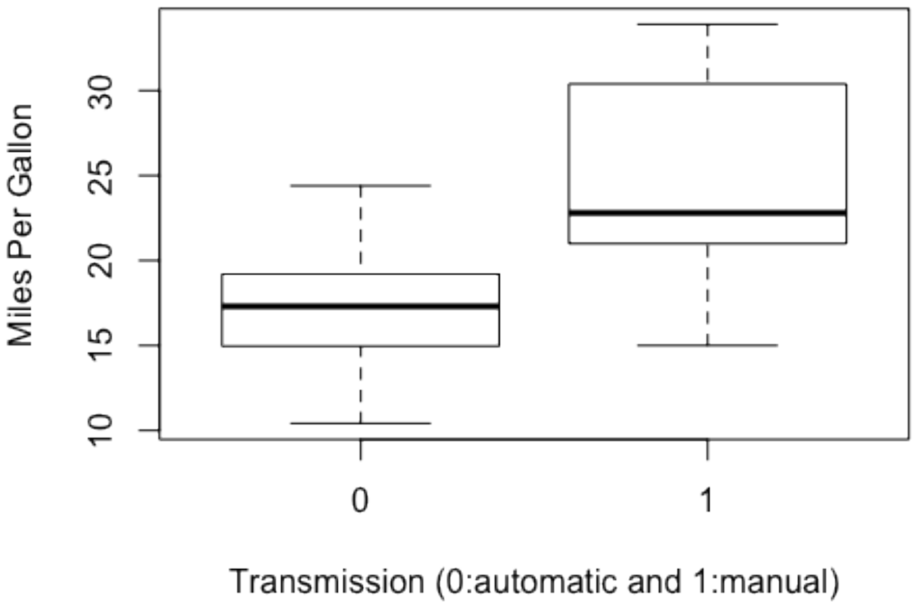
~

Appendix

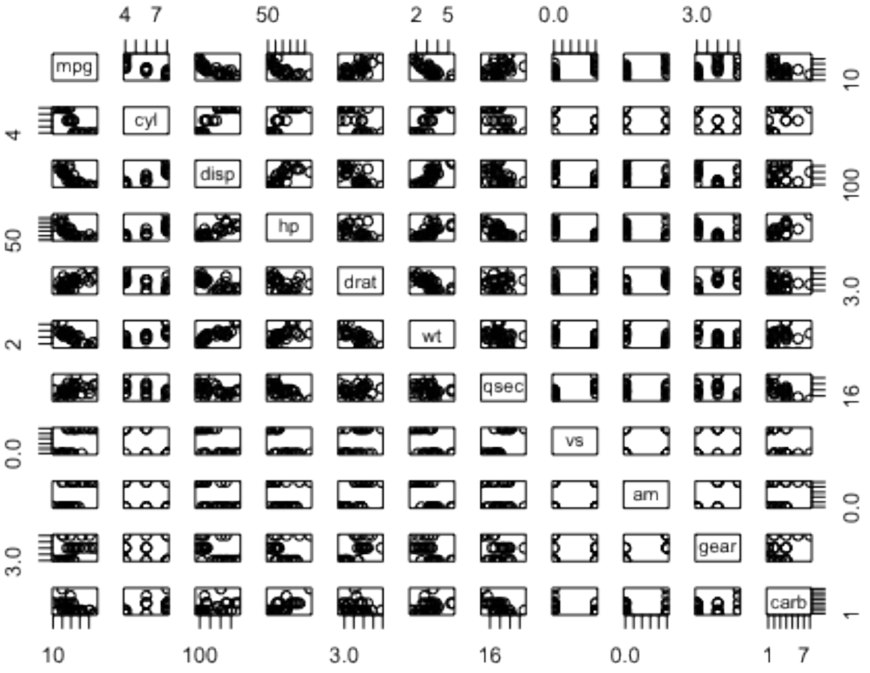
Histogram



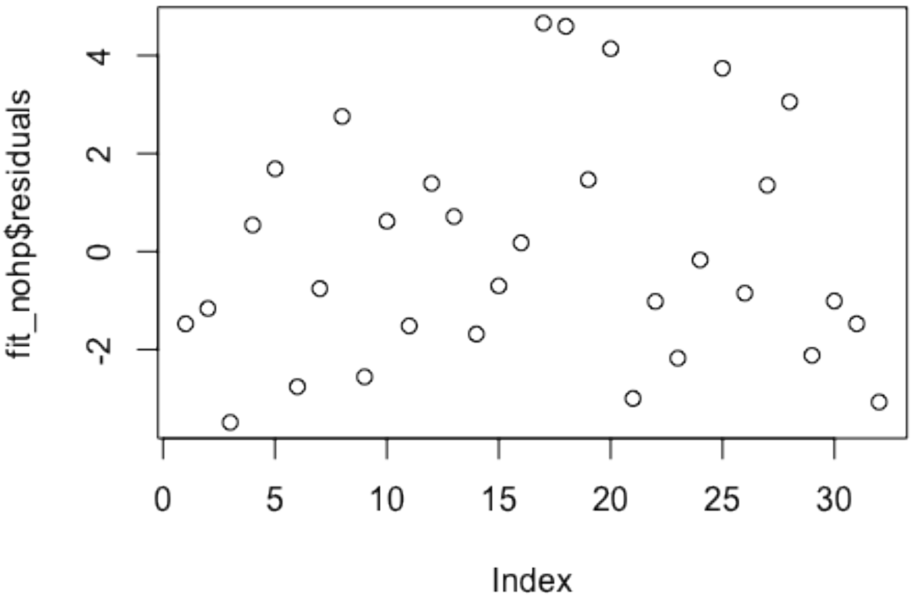
Boxplot



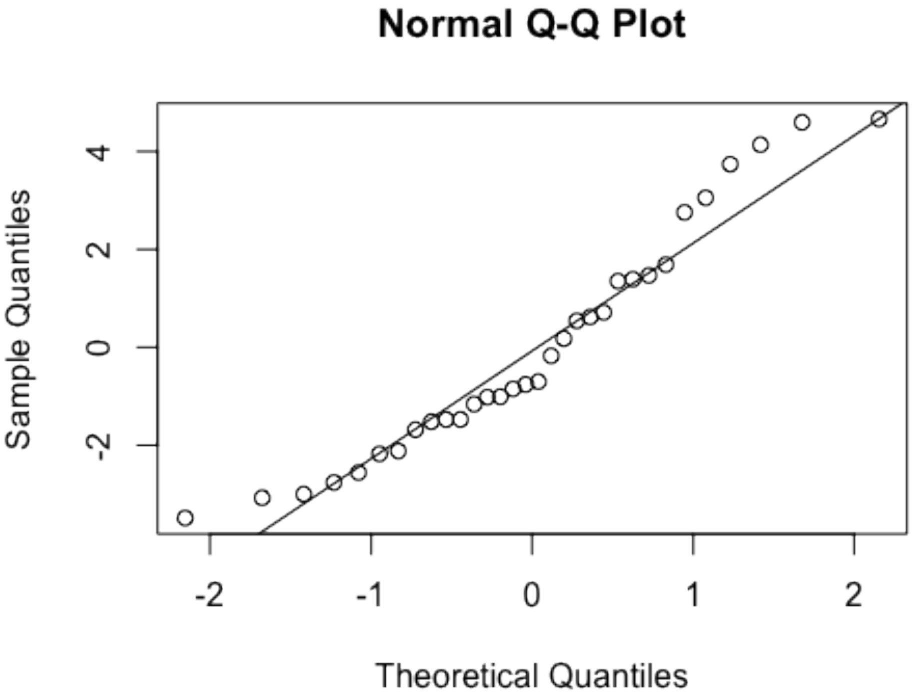
Pairs Plot



Residuals Plots



Normal Probability Plot of the Residuals


Normal Q-Q Plot
Sample Quantiles
4
2
0
-2
-2
-1
0
1
2
Theoretical Quantiles