Coursera Regression Models: MPG Analysis

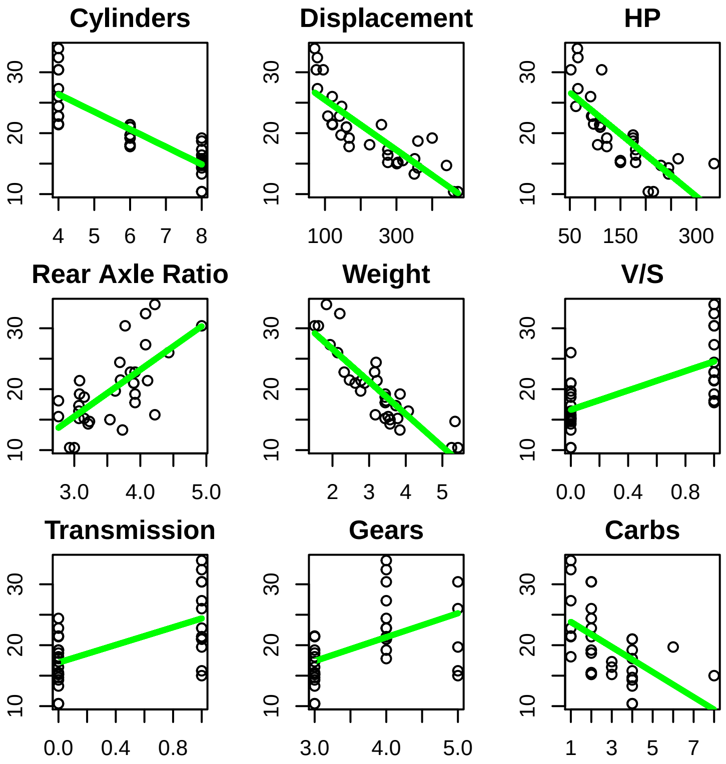
Saturday, May 23, 2015

Executive Summary

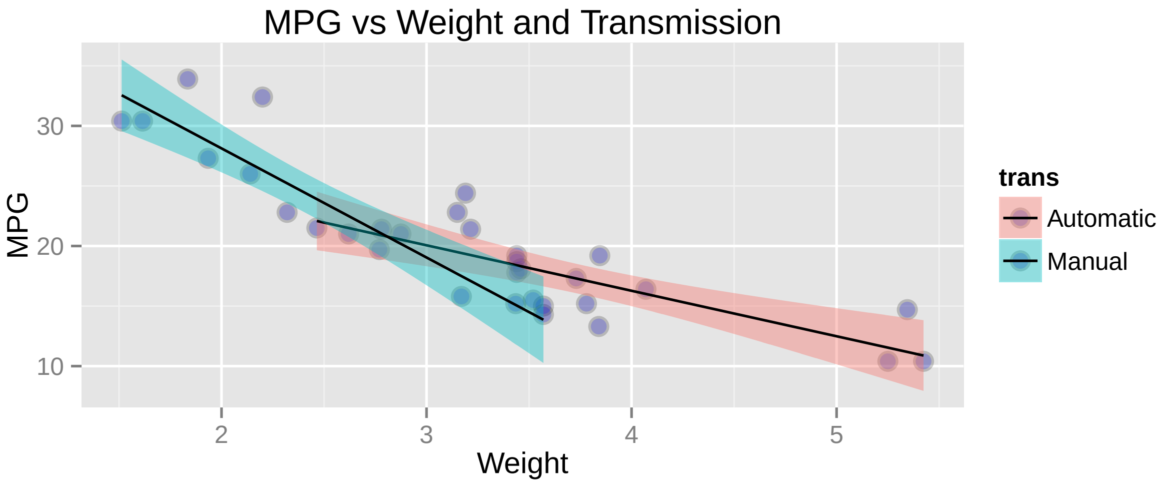
Motor Trend magazine is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). More specifically, they are interested in whether automatic or manual transmission cars get better gas mileage and would like to see the differences quantified. Using regression analysis on the mtcars dataset, we can show that on average cars with manaual transmissions get better gas mileage than automatic (24.4 vs 17.1), but that there are more important variables that impact mpg than just the type of transmission, such as vehicle weight, size of engine, etc…

Exploratory Data Analysis

First we take a look at all of the different measurements and factors vs. MPG to get an idea of which variables could be influential on the mpg. We could have used a pairs plot, but this showed all combinations of variables. Since we are only interested in MPG, I limited the plots to only MPG as the Y variable.



We can look at the regression lines to see that cylinders, displacement, hp, and number of carbuerators are all negatively correlated with mpg witht the other variables all positively correlated. Intuitively, weight seems like it would be a major contributor to mpg, so we next take a look at mpg vs weight for manual and automatic transmissions:



In this plot we can see that overall manual transmission cars may get better gas mileage, but this could be more a reflection of their weight than their transmission. We even see for the heavier, manual cars, that the automatic transmision cars with the same weight actually get better gas mileage.

In order to better understand the variables that most influence mpg, the next step was to look at a linear regression including all variables.

```
fit1=lm(mpg ~ ., data=mtcars)
summary(fit1)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```
fit1R2=summary(fit1)$r.squared
```

This model has an R-squared of 0.869, but the large P-values for each variable suggest that the model is overfit. The next step is to evaluate different combinations of the variables to find a better fitting model. The next iteration looks at only the transmission, since this was the initial question. Then additional variables were added with the resulting R squared and assocaited P-values reviewed to evaluate the model fit. Looking at the exploratory graphs, both weight and displacement have the steepest slope and therefore are good candidates for the next iterations of the model. The iterations of the model are listed in the appendix.

Linear Models

Model	Variables	R-Squared
1	All	0.869

Model	Variables	R-Squared
2	am	0.36
3	am,wt	0.753
4	am,wt,cyl	0.83
5	wt, cyl	0.83
6	wt,cyl,disp	0.833

Looking at the model summary of the last three models:

Model 4: Transmission, Weight and cylinders

```
summary(fit4)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## am           0.1764932  1.3044515  0.1353007 8.933421e-01
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
```

Looking at the model with transmission, weight and cylinders, the P-value for transmissions is large, relative to the weight and cylinders - so I've removed it from the model in the next iteration.

Model 5: Weight and Cylinders.

```
summary(fit5)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 39.686261  1.7149840 23.140893 3.043182e-20
## wt          -3.190972  0.7569065 -4.215808 2.220200e-04
## cyl         -1.507795  0.4146883 -3.635972 1.064282e-03
```

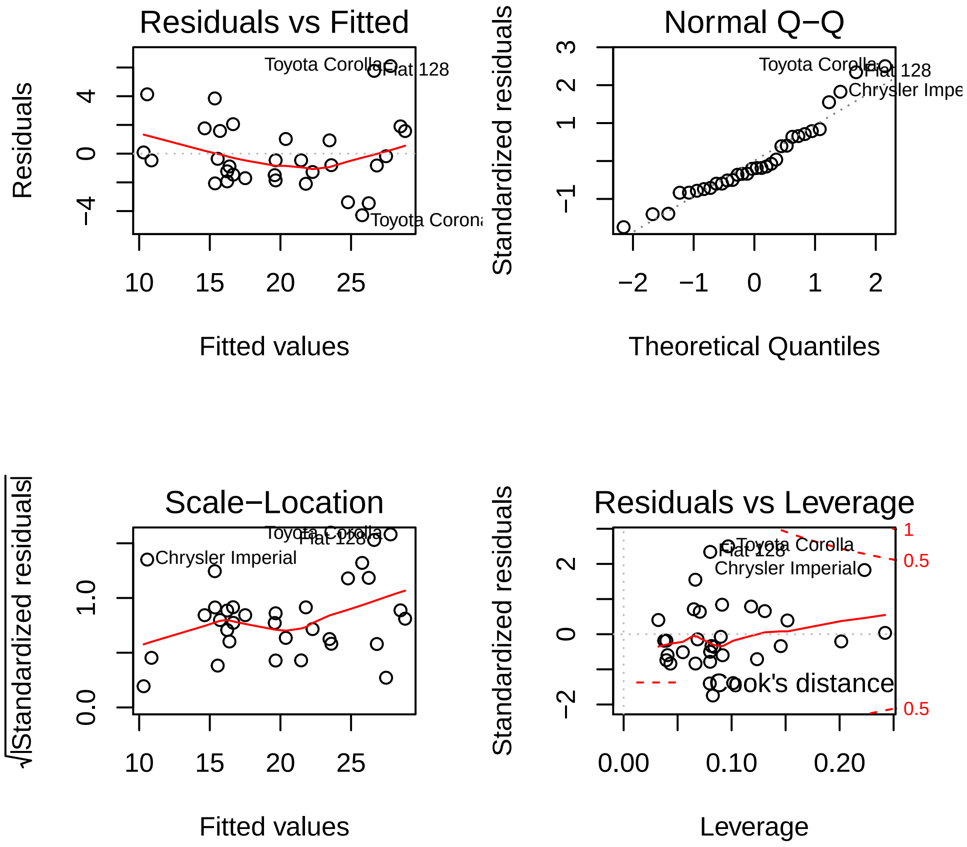
This model looks good, but maybe the addition of another variable (displacement) will improve it yet further.

```
summary(fit6)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 41.107677641 2.84242604 14.4621802 1.620381e-14
## wt          -3.635677016 1.04013753 -3.4953811 1.595519e-03
## cyl         -1.784943519 0.60711048 -2.9400638 6.511676e-03
## disp         0.007472925 0.01184472  0.6309079 5.332173e-01
```

The P-value for the additional variable is also large relative to the weight and cylinders. The impact to R-squared was minimal, so we can remove displacement and choose fit5 as the final model.

Looking at the residual plots for the final model:



Nothing stands out in the residual plots as concerning. The residuals vs. fitted values are relatively evenly spread. The Q-Q plot is close to the identity line, but is showing a slight, cyclical pattern that could reflect one of the factor variables. Nothing stands out in the other two plots as an outlier or cause for further model refinement.