Analysing relationship between Miles per gallon and Transmission for mtcars dataset

Reza Nirumand

Sunday, September 28, 2015

Executive Summary

In this document we have analysed mtcars dataset which consist of 32 observations and 11 variables. 13 cars in this dataset have manual transmissions and 19 have automatic.
At the first glance (Fig.2) considering only transmission variable we could see that manual transmission is more fuel efficient having average 7.2 better mileages per gallon. But manual transmission is not only variable describing the MPG variable (Fig.3,Fig4,Fig5).
To *Quantify the MPG difference between automatic and manual transmissions*, a more through analysis using multivariable regression shows that, the variable MPG could be better described as linear model consisting predictors am ,hp ,wt, qsec.
We estimate that with having other variables(hp,wt,qsec) fixed, using manual transmissions will result in between .03 to 2.9 more miles per gallon.
Our model's r.squared shows that the 83% of outcome can be explained by linear relationship of regressors.

Exploratory data analysis

Considering the documentations i have decided to convert the variables "am"={0,1},"vs"={0,1} to factor variables.

To get an insight into the data we have ploted the relationship between MPG and some of the variables(which we used in our Model).Fig.1 - Fig.5 shows plot of mtcars pairwise variables along their correlations. The variable MPG has the average 20.09 and standard deviation 6.03.
Also We would like to see if cars with automatic and manual transmission are from the same population. Since the number of observation is not alot we have used t-test. The null hypothesis is the MPG of the automatic and manual transmissions are from the same population.

Since the p-value is 0.00137, we reject our null hypothesis. So, the automatic and manual transmissions are from different populations. And the mean for MPG of manual transmitted cars is about 7 Miles per gallon more than that of automatic transmitted cars.

Model Selection

As we build our model only based on trasnmission type (mpg~am), We can see that the adjusted R squared value is only 0.34 which indicates that only 34% of the regression variance can be explained by our model. In order to find a parsimonious model, we will use *nested model* technique. That means we will begin with one regressor and will add regressors one-by-one, comparing the result for each model using anova test. But Considering the correlation matrix (Fig.1) finding the best subset of regressors requires exhaustive search for the best subsets of the variables.

Before Selecting the right model using anova test, we need to first verify the assumptions required for the anova test. The assumption for anova test is that the model's Residual are approximately Normal. To validate the assumption we have used the ***Shapiro-Wilk test***. The null hypothesis on this test is that the distribution is approximately normal.

Considering Shapiro-Wilk test, all models's p-value are bigger than alpha=0.05 , so we failed to reject the normality hypothesis, hence the models are valid for anova test. Considering the anova test we will **choose**

the model 4 since it shows significant change in RSS.(altought p-value is bigger than alpha, it has better adjusted.r.squared).

```
## [1] "Models adjusted r squared:"
```

```
## [1] "Mod1= 0.3385" "Mod2=0.767"   "Mod3=0.8227"  "Mod4=0.8368"
## [5] "Mod5=0.8267"  "Mod6=0.8196"
```

Regression Diagnostics

To determin whether our selected model fit to data adequately represents our data, we will do regression diagnostics(see Fig.6):
- The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot and verify the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a almost constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top and top right of the plots.
- The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

Results

Under 95% confidence interval we could mention followings:

- **Adjusted.R Squared**=0.8368: It means 84% of mpg(miles per gallon) is explained by linear relationship with regressors("am"" and "hp and"wt").

- Compared to automatic transmission, Manual transmission shows siginificant change in milles per gallon (between 0.06 and 5.79 miles/gallon) having fixed other variables.

- we estimated that, increasing 100 pounds in **Weight**, will result in reduction of miles per gallon between 0.14 and 0.51 having other variables fixed. (note: slop/10). Hint: [, 6] wt:Weight (lb/1000) variable was provided in the mtcars dataset as kilopounds(kip).

- Having other variables fixed Horspower does not show significant linear effect on miles per gallon.
- Having other variables fixed qsec(1/4 mile time) does not show significant linear effect on miles per gallon.
  Conisdering that we have achived the highest r.squared using mentioned model and also horsepower and qsec are not statistically significant, it might be a hint that linear model is not a proper model for this dataset.

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 17.44019110  9.3188688  1.871492 0.072149342
## am1          2.92550394  1.3971471  2.093913 0.045790788
## hp          -0.01764654  0.0141506 -1.247052 0.223087932
## wt          -3.23809682  0.8898986 -3.638726 0.001141407
## qsec         0.81060254  0.4388703  1.847021 0.075731202
```

Appendix

please note all codes for generating this report can be found in **here**

Fig.1: scatterplot of pairwise variable & correlations

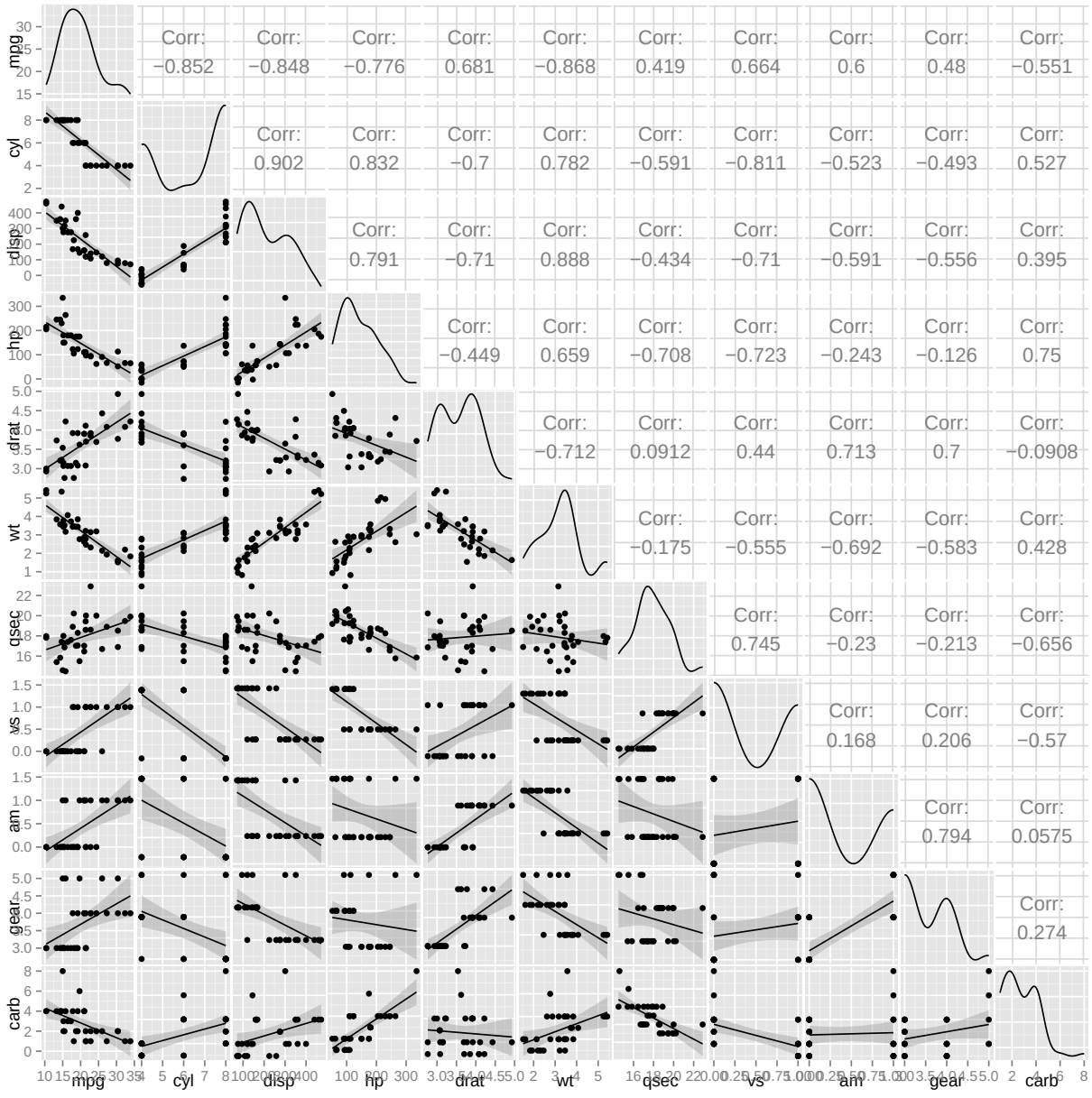

Fig.2:Comparison of mileages per gallon
by transmission types

Fig.3:Comparison of mileages per gallon
and horsepower
by transmission types

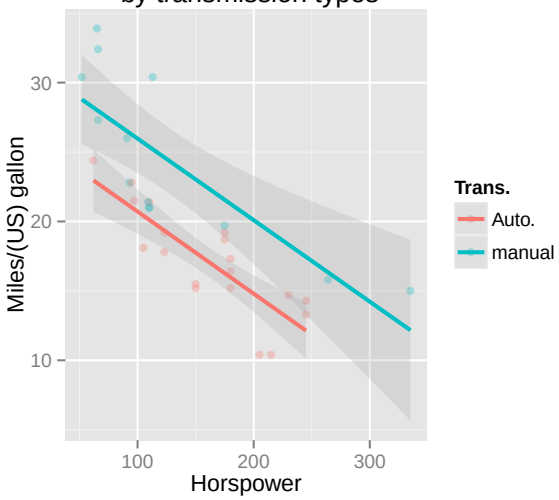

Fig.4:Comparison of mileages per gallon
and weigth
by each transmission types

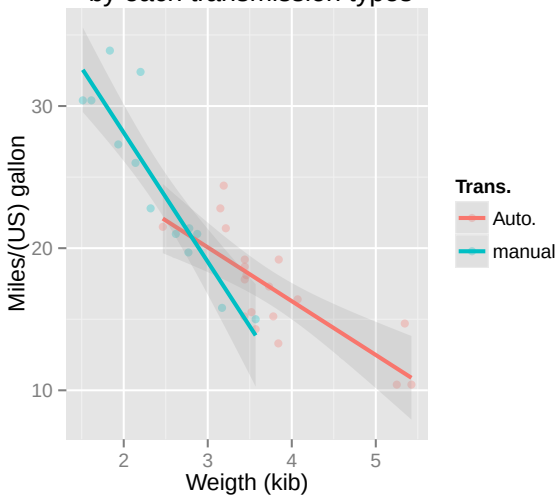

Fig.5:Comparison of mileages per gallon
and 1/4 mile time
by transmission types

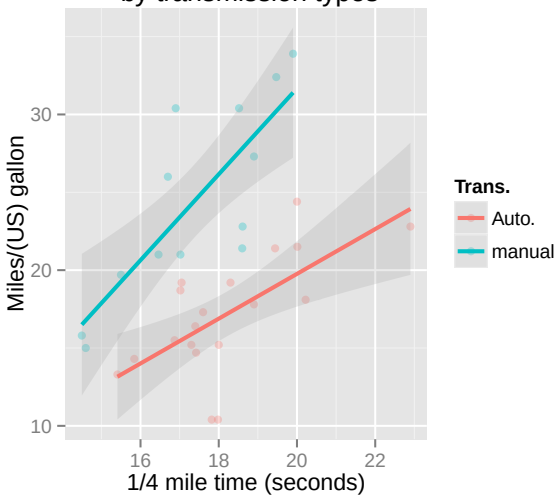

Fig.6:regression model dignostics