Johns Hopkins University Coursera: Regression Models (Project 1: 2015 February)

Executive Summary

MPG difference between automatic and manual transmission is statistically analyzed for the 3 predictors of vehicular *weight*, *displacement*, and *horsepower* independently as well as the 3 multivariables on linear regression models. For each of the 3 predictors, both summary and regression plot are shown, where data points for automatic and manual transmission are colored 'blue' and 'red' respectively. The conclusions are:

- MPG difference between automatic and manual transmission is insignificant when the predictor is *weight* because the respective correlations with MPG are nearly the same with -0.98 and -0.97.
- Among the 3 predictors, vehicular *weight* ranks the best, based on its consistently low sigma and high absolute correlation of 95.8%.
- Linear regression of including all the 3 predictors results in, as expected, degraded correlations whose highest absolute correlation of 84.6% is lower than the worst of the 3 single predictors at 88.4% by *displacement*.

Overview

This course project for Regression Models by Johns Hopkins University Coursera is to analyze the [mtcars][mtcars_doc] data (included in the the base version of R) using linear regression models.

Assignment Context

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Assignment Question

Take the [mtcars][mtcars_doc] data set and write up an analysis to answer their question using regression models and exploratory data analyses.

Your report must be:

- Written as a PDF printout of a compiled (using knitr) R markdown document.
- Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures.
- Include a first paragraph executive summary.

Summary of Data Preparation and Analysis

Detailed preliminary analyses (not shown for brevity) indicate the use of 3 predictors of *disp*, *hp*, and *wt* for predicting *mpg* by *am* of automatic (0) or manual (1) transmission. The analyses below start with the following data reduction:

```
# initialize
library( datasets )
str( mtcars )
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
# reduce mtcars to a subframe consisting of mpg, disp, hp, wt, and am
subfm <- with( mtcars , data.frame( MPG = mpg , DSP = disp , HP = hp , WT = wt , TRNSM = am ) )
```

On the reduced subframe, linear regressions are first applied per predictor based only on *TRNSM* or $am = 0$ or 1. Correlation of Coefficients are examined for their closeness. The closer the values are, the less *MPG* depends on the vehicular *TRNSM*.

The plots are then generated for each of the predictor showing points of both *TRNSM* in 'blue' and 'red' for $am = 0$ and 1, respectively. The resulting linear regression line is also superimposed.

Summary of Linear Regressions

Linear regressions are generated for *MPG* vs *displacement*, *horsepower*, *weight*, and the 3 multi-variates as follows:

1. MPG vs Displacement Regression

Three sets of linear regression are generated for $am = 0$, 1, and both:

```
fitDSPa <- lm( MPG ~ DSP , subset( subfm , TRNSM == 0 ) )
fitDSPm <- lm( MPG ~ DSP , subset( subfm , TRNSM == 1 ) )
fitDSP <- lm( MPG ~ DSP , subfm )
smryDSP <- summary( fitDSP , correlation = TRUE )
```

Correlation of Coefficients for $am = 0$ and 1:

```
corDSPa <- summary( fitDSPa , correlation = TRUE )$correlation[ 2 , 1 ]
corDSPm <- summary( fitDSPm , correlation = TRUE )$correlation[ 2 , 1 ]
```

2. MPG vs Horsepower Regression

Three sets of linear regression are generated for $am = 0$, 1, and both:

```
fitHPa <- lm( MPG ~ HP , subset( subfm , TRNSM == 0 ) )
fitHPm <- lm( MPG ~ HP , subset( subfm , TRNSM == 1 ) )
fitHP <- lm( MPG ~ HP , subfm )
smryHP <- summary( fitHP , correlation = TRUE )
```

Correlation of Coefficients for $am = 0$ and 1:

```
corHPa <- summary( fitHPa , correlation = TRUE )$correlation[ 2 , 1 ]
corHPm <- summary( fitHPm , correlation = TRUE )$correlation[ 2 , 1 ]
```

3. MPG vs Weight Regression

Three sets of linear regression are generated for $am = 0$, 1, and both:

```
fitWTa <- lm( MPG ~ WT , subset( subfm , TRNSM == 0 ) )
fitWTm <- lm( MPG ~ WT , subset( subfm , TRNSM == 1 ) )
fitWT <- lm( MPG ~ WT , subfm )
smryWT <- summary( fitWT , correlation = TRUE )
```

Correlation of Coefficients for $am = 0$ and 1:

```
corWTa <- summary( fitWTa , correlation = TRUE )$correlation[ 2 , 1 ]
corWTm <- summary( fitWTm , correlation = TRUE )$correlation[ 2 , 1 ]
```

4. MPG vs Weight, Displacement, and Horsepower Regression

A multi-variate linear regression is generated for both *am*:

```
fitWDH <- lm( MPG ~ WT + DSP + HP , subfm )
smryWDH <- summary( fitWDH , correlation = TRUE )
```

The matrix of Correlation of Coefficients for both *am*:

```
corWDH <- summary( fitWDH , correlation = TRUE )$correlation
```

Summary of Sigmas and Correlations

The following summary table shows the predictor *WT* having the lowest absolute *correlation difference* between $am = 0$ and 2, the lowest *sigma* and the highest absolute *correlation* between *MPG* and predictor:

```
prj1Smry <- matrix( , nrow = 3 , ncol = 3 )
colnames( prj1Smry ) <- c( '|cor(am=0)-cor(am=1)|' , '       sigma' , '|correlation|' )
rownames( prj1Smry ) <- c( 'MPG vs CYL' , 'MPG vs HP' , 'MPG vs WT' )

prj1Smry[ 1 , ] <- c( abs( corDSPa - corDSPm ) , smryDSP$sigma , abs( smryDSP$correlation[ 2 , 1 ] ) )
prj1Smry[ 2 , ] <- c( abs( corHPa - corHPm ) , smryHP$sigma , abs( smryHP$correlation[ 2 , 1 ] ) )
prj1Smry[ 3 , ] <- c( abs( corWTa - corWTm ) , smryWT$sigma , abs( smryWT$correlation[ 2 , 1 ] ) )
round( prj1Smry , 3 )  # round to 3 digits
```

```
##            |cor(am=0)-cor(am=1)|        sigma |correlation|
## MPG vs CYL                 0.074        3.251         0.884
## MPG vs HP                  0.107        3.863         0.908
## MPG vs WT                  0.009        3.046         0.958
```

Thus, the vehicular *weight* is a good predictor of its *MPG* irrespective of *transmission* type.

Including all 3 predictors in a linear regression results in worse correlations, as in the following correlation matrix:

```
corWDH
```

```
##             (Intercept)         WT        DSP         HP
## (Intercept)   1.0000000 -0.8464699  0.6371733 -0.3729788
## WT           -0.8464699  1.0000000 -0.7970839  0.1549379
## DSP           0.6371733 -0.7970839  1.0000000 -0.5953597
## HP           -0.3729788  0.1549379 -0.5953597  1.0000000
```

Even the highest absolute correlation of 84.6% here is lower than the lowest of the above 3 regressions for a single predictor at 88.4% since including additional variables will increase the actual standard errors of coefficient estimates of other correlated predictors.

Appendix: Linear Regression Models and Plots

A1. *MPG* vs *Displacement* Regression

Correlation of Coefficients for $am = 0, 1$, and their absolute difference:

```
corDSPa  # correlation of mpg with dsp for am = 0
```

```
## [1] -0.9380793
```

```
corDSPm  # correlation of mpg with dsp for am = 1
```

```
## [1] -0.8636306
```

```
abs( corDSPa - corDSPm )  # absolute difference
```

```
## [1] 0.07444866
```

```
abs( smryDSP$correlation[ 2 , 1 ] )  # absolute correlation for all am
```

```
## [1] 0.8840414
```

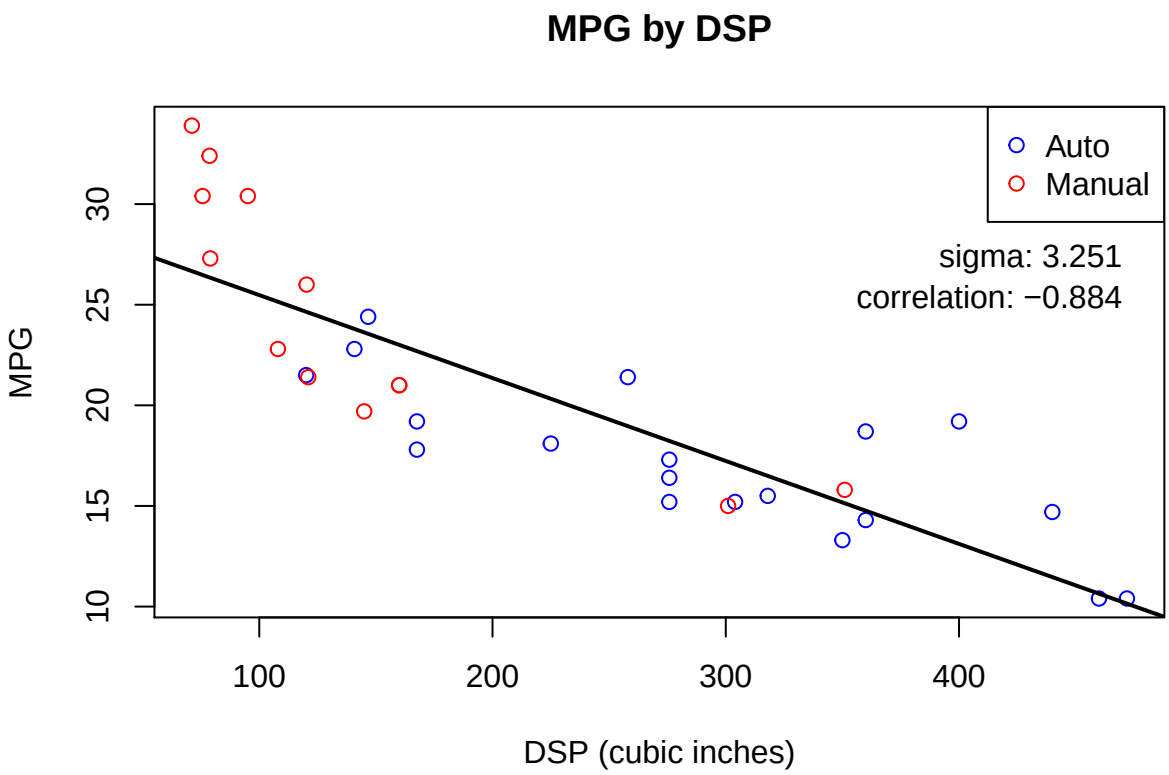

Figure 1 Regression plot of *mpg* with *dsp* for both *am*.

A2. *MPG* vs *Horsepower* Regression

Correlation of Coefficients for $am = 0, 1$, and their absolute difference:

```
corHPa  # correlation of mpg with hp for am = 0
```

```
## [1] -0.950361
```

```
corHPm  # correlation of mpg with hp for am = 1
```

```
## [1] -0.8435284
```

```
abs( corHPa - corHPm )  # absolute difference
```

```
## [1] 0.1068326
```

```
abs( smryHP$correlation[ 2 , 1 ] )  # absolute correlation for all am
```

```
## [1] 0.9084744
```

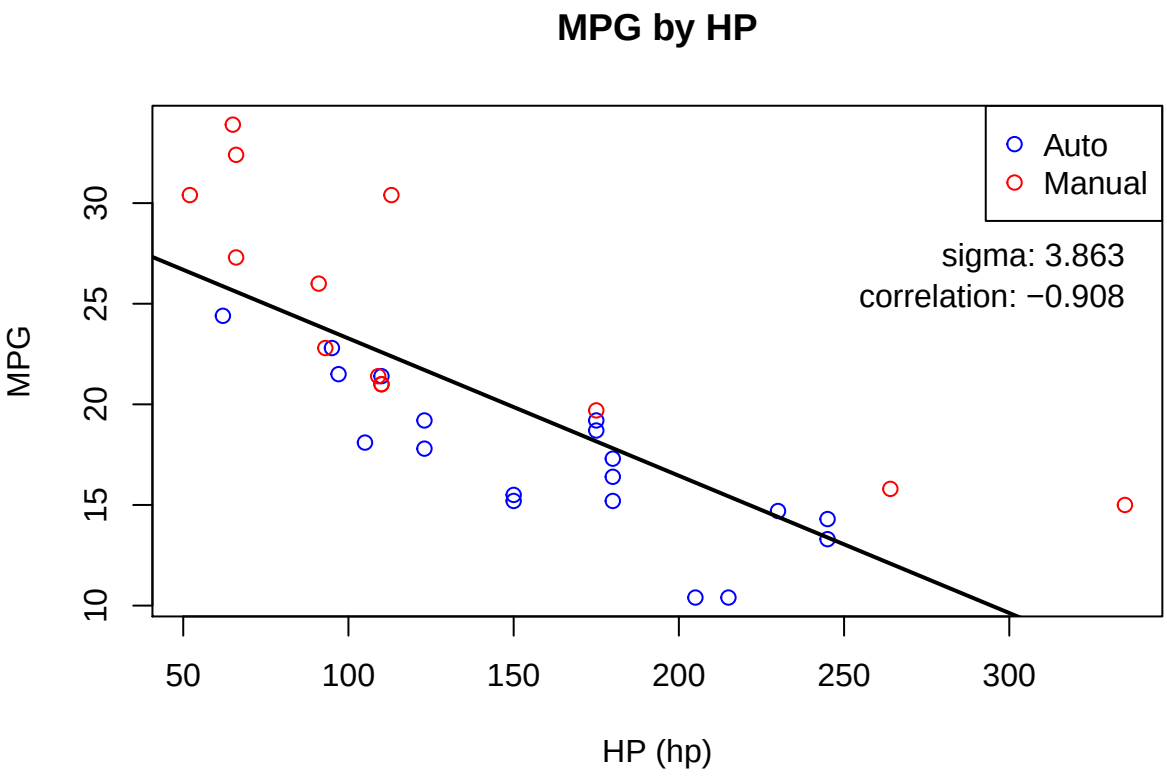

Figure 2 Regression plot of *mpg* with *hp* for both *am*.

A3. *MPG* vs *Weight* Regression

Correlation of Coefficients for $am = 0, 1$, and their absolute difference:

```
corWTa  # correlation of mpg with wt for am = 0
```

```
## [1] -0.980436
```

```
corWTm  # correlation of mpg with wt for am = 1
```

```
## [1] -0.9710803
```

```
abs( corWTa - corWTm )  # absolute difference
```

```
## [1] 0.009355701
```

```
abs( smryWT$correlation[ 2 , 1 ] )  # absolute correlation for all am
```

```
## [1] 0.9580005
```

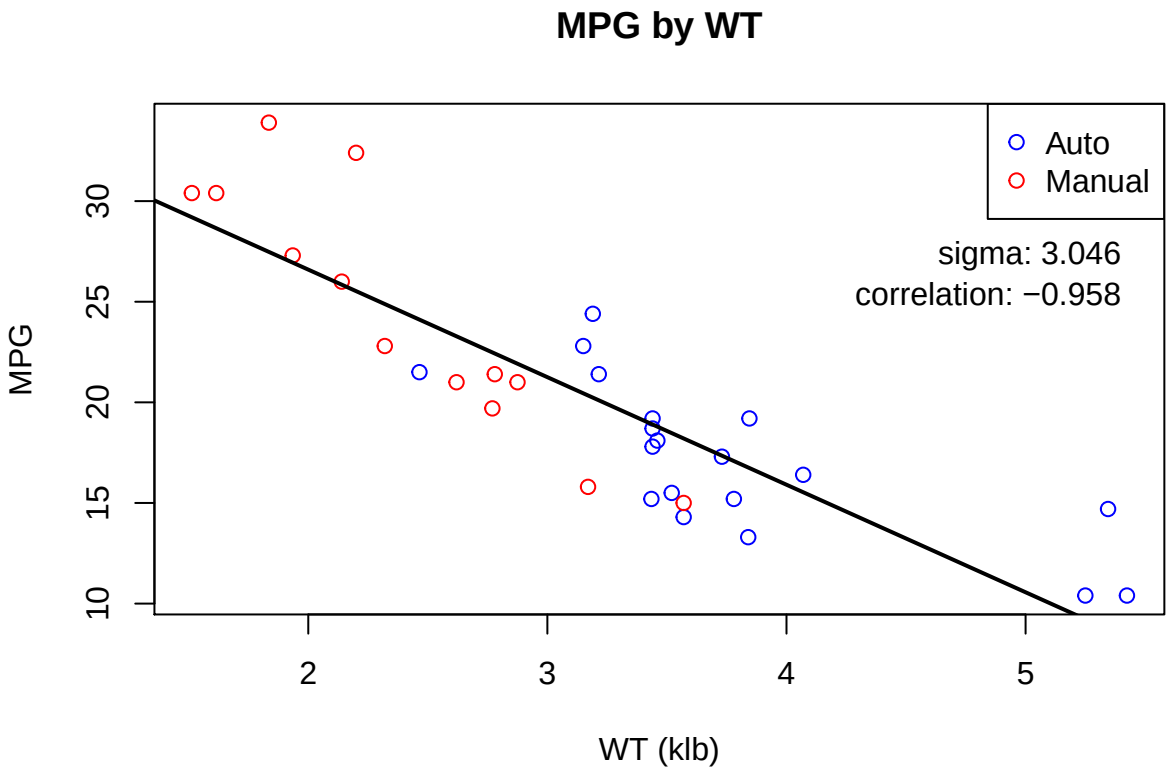

Figure 3 Regression plot of *mpg* with *wt* for both *am*.