Regression — Course Project

Motor Trend – How MPG Get Impacted by Transmission Type

December 26, 2015

Executive Summary

I have analyzed mtcars dataset and explored a set of car features and their impact on MPG. From common sense, transmission type matters MPG. So, I started with a single variable(am) regression, and then find ways to improve the model by including other relevant features. The final model is a three variables linear regression model on MPG, and am is part of the selected features. With residual diagnosis, I believe my model would be one of the best models to describe MPG by car features. Also, with the model, we can say automatic transition is better for MPG, and MPG will go up by 2.94 gallon when changing to manual transmission and holding other features constant.

1 Exploratory Analysis – The first look at Mtcars

1.1 mtcars dataset

- The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).
- A data frame with 32 observations on 11 variables.

str(mtcars) *Figure 1 mtcars data strucutre* * mpg: miles per gallon * cyl: number of cylinders (factor, 4,6,8) * disp: displacement (cu.in.) (numerical) * hp: gross horsepower (numerical) * drat: rear axle ratio (numerical) * wt: weight (1000 pounds) (numerical) * qsec: 1/4 mile time (numerical) * vs: V/S, V-engine or Straight engine (factor, V,S) * am: transmission type (factor, 0=automatic, 1=manual) * gear: number of forwards gears (factor, 3,4,5) * carb: number of carburetors (factor, 1,2,3,4,5,6,7,8)

1.2 Data Transformation

To get a more accurate model, first, I transform continuous variable to factor when possible.

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

1.3 MPG on Transmission

A simple plot(*Figure 2*) of mpg on am(transmission) shows manual transmission has higher MPG performance.

1.4 Thoughts and Method

Based on common sense, we know other car attributes, such as weight, displacement, horsepower, etc, may affect MPG performance at the same time(refer to *Figure 3 pair chart*). With that, I think only examine a single variable am may lead to biased outcome. Thus, my method would be 1) start with a simple regression of MPG on am, but add other variables bit by bit to achieve the best fitting model

2 Model Selection - fit multiple model

To start with, I did a single variable regression "mpg on am" (lm(mpg~am) *Figure 4*). Although the coefficients are significantly different from 0, the R square is only 0.3385.

To improve, I used the **step** method to find driving variables. This method selected four variables for me, wt + qsec + hp + am. (*Figure 5*)

null.model <- lm(mpg~am, data=mtcars); summary(null.model) full.model <- lm(mpg~.,data=mtcars) selected.model<-step(null.model, scope=list(lower=null.model, upper=full.model), direction="forward")

To verify, accordingly, I did a regression with the four variables above. lm(mpg~wt + qsec + hp + am). The outcome(*Figure 6*) showed a good adjusted r square value(0.8368), but the coefficient of hp is not significant.

Therefore, I did the regression once again with three variable, lm(mpg~wt + qsec + am). This model(*Figure 7*) also gives a good adjusted r square value of 0.8336, and all the variables coefficients are significant.

3 Residual Analysis and diagnosis

Refer to (*Figure 8*). All the diagnosis plots look good to me. * The residuals vs. fitted plot shows no pattern * The Normal QQ plot shows points are along the line, meaning residuals are roughly following normal distribution * The Scale-Location plot shows points are randomly distributed * The residuals vs. leverage plot implies that there are no outliers.

4 Conclusion and answer the questions

lm(mpg~wt + qsec + am) is the best model to describe car features impact on mpg, and the above residual analysis confirms this model is reliable.

To answer the questions, Q1 "Is an automatic or manual transmission better for MPG" Q2 "Quantify the MPG difference between automatic and manual transmissions" Answer: Yes, automatic transmission is better for MPG. Since the coefficient of "am"" in the final model is 2.94, meaning MPG will go up by 2.94 gallon when changing to manual transmission and holding other features constant.

Appendix

Figure 1 - structure of mtcars dataset

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Figure 2 - MPG on Transmission(0=automatic, 1=manual)

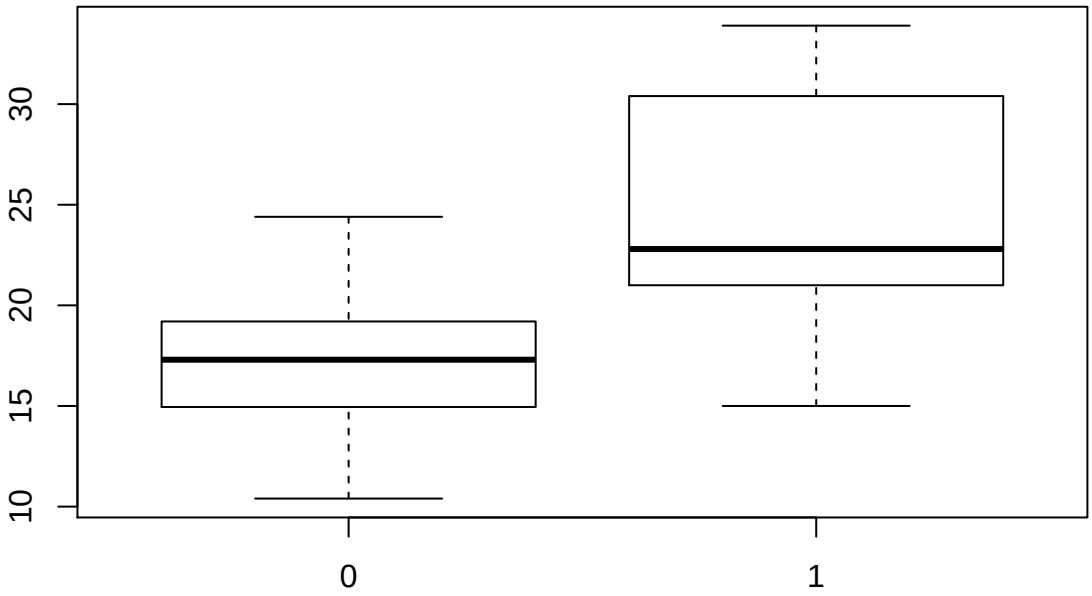

Figure 3 - Pair chart

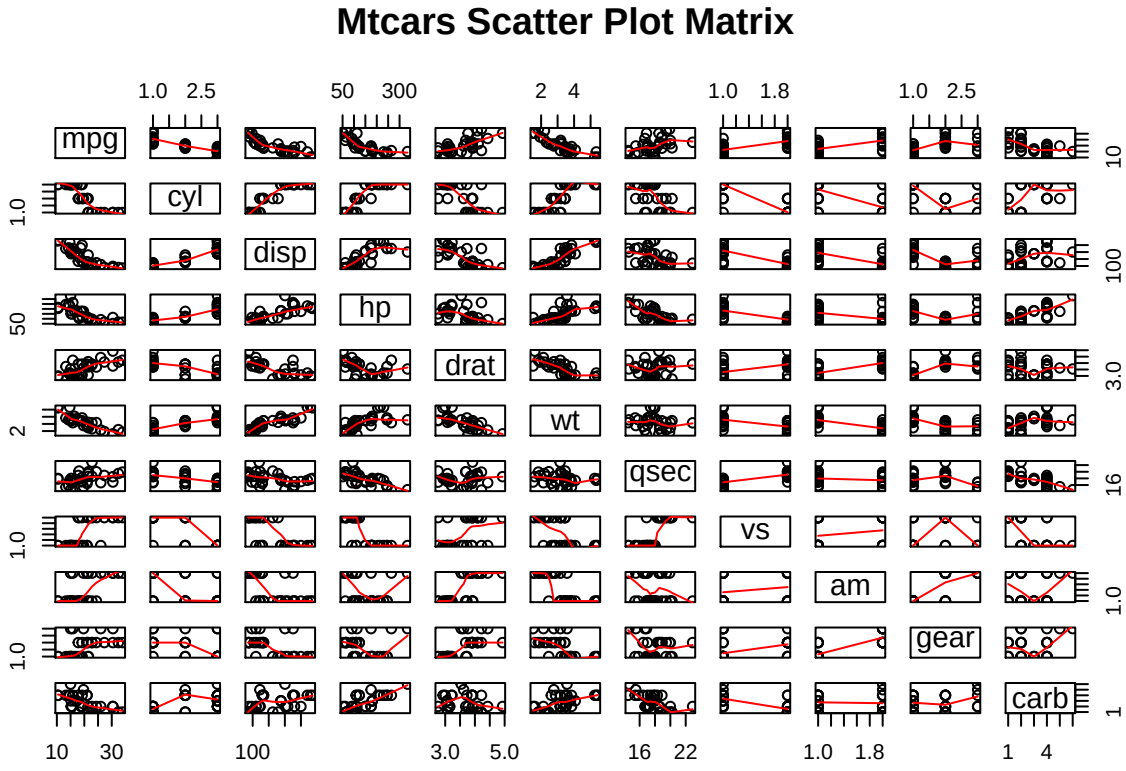

Figure 4 - Single variable regression

```r
summary(lm(mpg~am, data=mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Figure 5 – model seletion with step method

null.model <- lm(mpg~am, data=mtcars); summary(null.model) full.model <- lm(mpg~.,data=mtcars); selected.model<-step(null.model, scope=list(lower=null.model, upper=full.model), direction="forward")

Step: AIC=61.52 mpg ~ am + hp + wt + qsec

Figure 6 – regression on four variables that selected by "step"

lm(formula = mpg ~ wt + qsec + hp + am, data = mtcars) Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 17.44019 9.31887 1.871 0.07215 . wt -3.23810 0.88990 -3.639 0.00114 ** qsec 0.81060 0.43887 1.847 0.07573 . hp -0.01765 0.01415 -1.247 0.22309
am1 2.92550 1.39715 2.094 0.04579 * — Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.435 on 27 degrees of freedom Multiple R-squared: 0.8579, Adjusted R-squared: 0.8368 F-statistic: 40.74 on 4 and 27 DF, p-value: 4.589e-11

Figure 7 – improved regression model (final model)

```
##
## Call:
## lm(formula = mpg ~ wt + hp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## wt          -2.878575   0.904971  -3.181 0.003574 **
```

```
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## am1          2.083710   1.376420   1.514 0.141268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Figure 8 – residual analysis

```
par(mfrow = c(2,2));
plot(lm(mpg ~ wt + hp + am, data=mtcars))
```