Analysis of Motor Trend Car Road Test Data with Regressional Models (A Coursera course project)

Ye Li

September 27, 2015

Executive Summary

This project aims at practicing the regressional model selection and results interpretation through analyzing the dataset "mtcars" in the R data package. Through the analysis, two questions need to be answered:

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

To answer the two questions, we first used some exploratory analysis to determine which variables are most relevant to changes in the fule efficiency (mpg). Then, several multivariate regressions were performed to select the model which provides sufficient fitting and optimized variance. The necessity of adding transmission type (am) as a regressor was evaluated and discussed. Last, the selected model is used to explain the changes in MPG quantitatively.

Data Exploratory

The "mtcars" data contains 32 observations and 11 variables about fuel consumption and 10 aspects of automobile design from the 1974 Motor Trend US magazine. The variables observed are mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb. Accroding to the "str" results (not shown), all variables are numbers. Among them, the vs (V or Straight Engine), am(automatic or manual transmission) should be considered factors; and the cyl(number of cyclinders), gear, and carb (number of carburetors) should be considered integers not continuous numbers. More details about the variables can be found in the article by R.R. Hocking published in Biometrics, 32(1), 1976, 1-49, Link.

The summary of correlations between pairs of variables is listed in the Appendix A. The variables have correlations with mpg stronger than 0.8 ("+") are cyl, disp, and wt. It's also worth noting that the correlation between cyl & disp, cyl & hp, cyl & vs, and wt and disp are also stronger than 0.8. Therefore, when performing regrssional analysis to mpg v.s. other variables, the wt, cyl, and disp should be considered as priototies and the dependency among the three variables should be considered too. The correlation between mpg & am is 0.5998, which is not sufficient to determin if am should be a good regressor for mpg. Further tests need to be performed.

To explore the relationship among the most interested variables, a box plot and a scatter plot are shown in figures in the Appendix A. As shown in Fig 1-1, the differences in fuel efficiency related to manual or automatic transmission appear to be rather significant for cars with 4 or 6 cylinders but not clear for 8 cylinder cars. In Fig 1-2 (figure g2 not shown), the fuel efficiency appears to be increasing when the car weight and displacement increase. The two figures further demonstrated potential dependencies of the fuel efficiency (mpg) to the three variables, wt, disp, and cyl. The effect of the transmission type remains unclear.

Regressional Model Selection

Possible Regressional Models

Based on the exploratary analysis above, a step-by-step multivariate regression is performed to the fuel efficiency (mpg) v.s. wt, disp, and cyl. And the variable transmission type(am) is also considered since the

exploratory analysis didn't give a clear message about its relevancy.

After fitting multiple nested models with the above variables, the variance analysis of the models is shown in anova5 table in Appendix B. Based on this ANOVA table, the model 3, mpg ~ wt + disp + wt*disp is optimized in minimizing the variation. Adding factor cyl and adding factor am as regressors are both rejected due to high Pr(>F) value. (We also ran the ANOVA table for adding am without cyl (anova4 in Appendix B, not printed). The result also reject the addition of am as a regressor.)

Resigudal Plot and Diagnosis

The residual plot of the selected model is plotted as a figure in Appendix C. The residuals vs Fitted and the scale-location plot do not show any clear patterns. The Normal Q-Q plot shows that the data points of Fiat 128, Pontiac Firebird and Toyota Corolla deviates from the normal distribution a bit. And the Chrysler Imperial, Pontiac Firbird and Toyota Corolla may cause comparatively high coefficient change when deleting them from the modeling. The conclusion can also be confirmed by the dfbetas and dfhatvalues calcualted in Appendix C.

Interpretation of the Selected Model

As shown in the summary of the selected model printed in Appendix D, our model estimates an expected -6.496 decrease in fuel efficiency (mpg) for every 1000 lb increase in car weight when the displacement is held at 0. It also estimates that an expected -0.056 decrease in fuel efficiency (mpg) for every 1 cu. in. increase in displacement when the car has no weight. The increased mpg per 1000 lb car weight by each cu. in. displacement is 0.012. These interpretation may not carry any physical meaning considering the car machnics.

The intercept of the model 44.082 also does not carry any phsyical meaning since it's not possible to have a car with 0 lb weight.

The compariosn betweent the regressional model with mpg ~ wt + disp + wt * disp and mpg ~ wt + disp + wt * disp + am demonstrates that adding the transmission type (manual or automatic) as a factor regressor to the model does not improve the model significantly (r.square does not change much). And the t value and the Pr(>|t|) show that the mpgs of manual transmission cars are NOT significantly higher than the mpgs of automatic transmission ones.

Conclusion

Based on this limited dataset (32 observations in total), the multivariate regression model explaining the fuel efficiency (mpg) the best is the model mpg ~ wt + disp + wt * disp. The transmission type (am) is not a significant regressor and should not be included in the model. There is NO enough evidence in this data set to demonstrate if the manual and automatic transmission gives better fuel efficiency (mpg). Even if we add the mpg factor in the selected regression model, it shows that the difference attribute to the transmission type (am) is NOT significant. With the selected model, the fuel efficiency is best explained quantitatively by the model shown in Appendix D.

Appendix: R codes and Figures

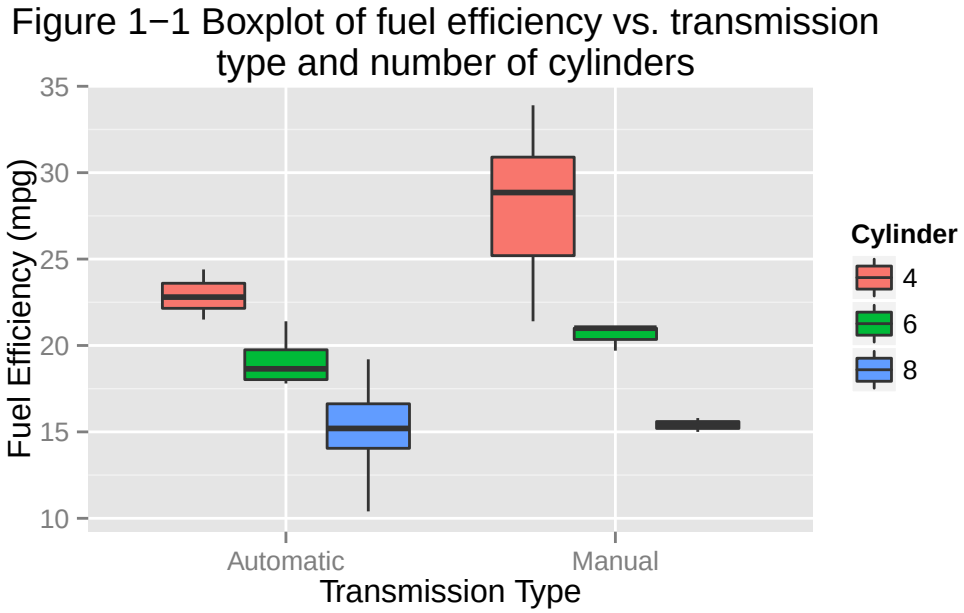
A. R code and figure for data exploratory

```
# load data and preview
data(mtcars)
# str(mtcars) not to print due to file length
```

```r
cor_all <- cor(mtcars)
symnum(cor_all)
```

```
##      m cy ds h dr w q v a g cr
## mpg  1
## cyl  + 1
## disp + *  1
## hp   , +  ,  1
## drat , ,  ,  . 1
## wt   + ,  +  , ,  1
## qsec . .  .  ,      1
## vs   , +  ,  , .  . , 1
## am   . .  .    ,  ,     1
## gear . .  .    ,  .     , 1
## carb . .  .  ,    . , .     1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

```r
library(ggplot2)
g1 <- ggplot(data=mtcars, aes(x=factor(am),y=mpg))
g1 <- g1 + geom_boxplot(aes(fill=factor(cyl)))
g1 <- g1 + labs(title ="Figure 1-1 Boxplot of fuel efficiency vs. transmission \n type and number of cy
g1 <- g1 + scale_x_discrete(label= c("Automatic", "Manual"))
g1 <- g1 + scale_fill_discrete(name = "Cylinder", labels = c ("4", "6", "8") )

g2 <- ggplot(data=mtcars, aes(x=wt,y=mpg))
g2 <- g2 + geom_point(aes(color=disp))
g2 <- g2 + labs(title ="Figure 1-2 Effect of car weight and displacement \n on fuel efficiency", x = "Ca
g2 <- g2 + scale_color_continuous(name = "Displacement (cu. in.)" )

g1
```



B. R code and figure for regression models

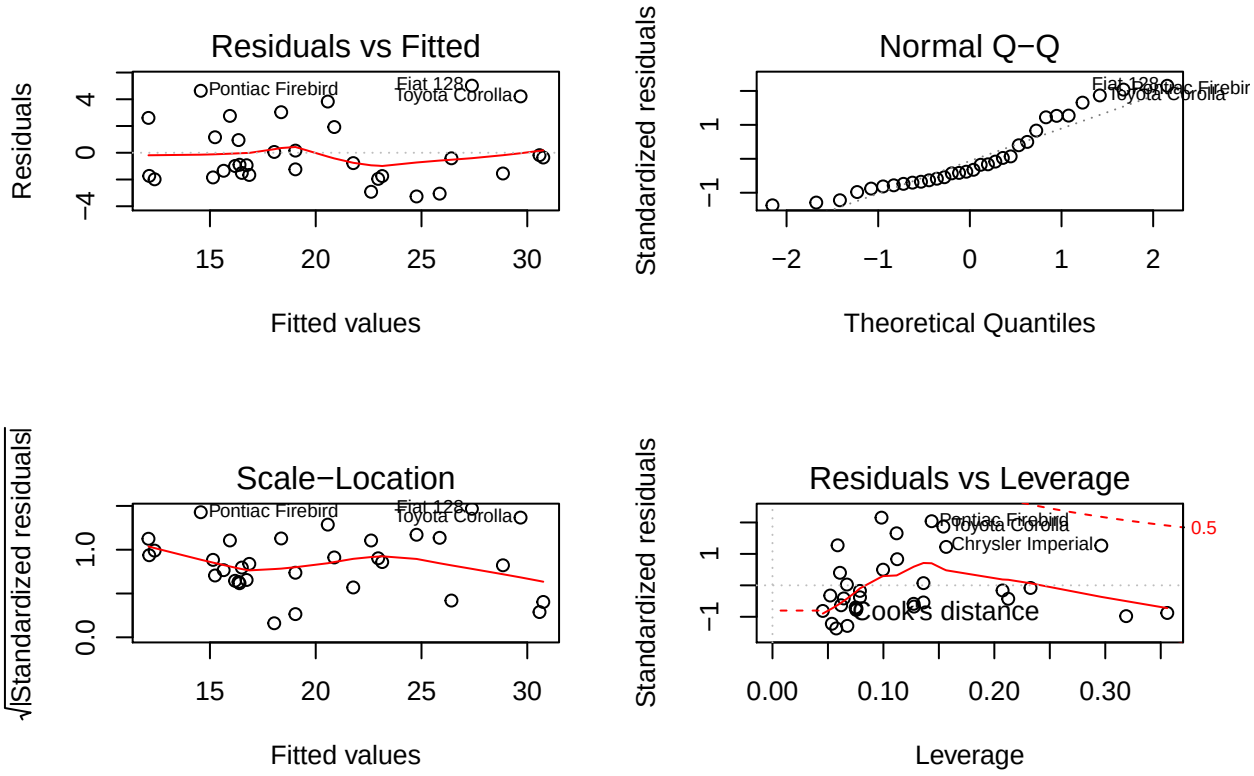
```r
#classify some variables to factors and integers as analyzed in the exploratory section
library(reshape2)
suppressMessages(library(dplyr))
mtcars1 <- mutate(mtcars, am = as.factor(am), vs = as.factor(vs), cyl = as.integer(cyl),
                  gear = as.integer(gear), carb = as.integer(carb))

# fit multiple models to compare
fit_wt <- lm(mpg ~ wt, data = mtcars)
fit_disp <- lm(mpg ~ disp, data = mtcars)
fit_cyl <- lm(mpg ~ factor(cyl), data = mtcars)
fit_am <- lm(mpg ~ factor(am), data = mtcars)
fit_wt_disp <- lm(mpg ~ wt + disp, data = mtcars)
fit_wt_disp_int <- lm(mpg ~ wt + disp + wt * disp, data = mtcars)
fit_wt_disp_int_am <- lm(mpg ~wt + disp + wt*disp + factor(am), mtcars)
fit_wt_disp_int_cyl <- lm(mpg ~ wt + disp + wt* disp + factor(cyl), data = mtcars)
fit_wt_disp_int_cyl_am <- lm(mpg ~ wt + disp + wt* disp + factor(cyl) + factor(am), data = mtcars)
fit_all <- lm(mpg ~., data = mtcars1)

#analyze the variance of several models to select one
anova4 <- anova(fit_wt, fit_wt_disp, fit_wt_disp_int, fit_wt_disp_int_am)
anova5 <- anova(fit_wt, fit_wt_disp, fit_wt_disp_int, fit_wt_disp_int_cyl, fit_wt_disp_int_cyl_am)

print(anova5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + disp
## Model 3: mpg ~ wt + disp + wt * disp
## Model 4: mpg ~ wt + disp + wt * disp + factor(cyl)
## Model 5: mpg ~ wt + disp + wt * disp + factor(cyl) + factor(am)
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 278.32
## 2     29 246.68  1    31.639  5.1991 0.031392 *
## 3     28 168.75  1    77.934 12.8063 0.001449 **
## 4     26 158.08  2    10.667  0.8764 0.428678
## 5     25 152.14  1     5.942  0.9764 0.332549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# confirm decision with variance inflation factor analysis
library(car)
vifall <- vif(fit_all)
```

C. R code and figure for residual analysis

```r
par(mfrow = c(2,2))
plot(fit_wt_disp_int)
```




```
#residual diagnostics
dfbetas_selected <- round(dfbetas(fit_wt_disp_int)[, 2], 3)
dfhatvalues <- round(hatvalues(fit_wt_disp_int)[1:32], 3)
```

D. R code and summary of the selected model

```
##
## Call:
## lm(formula = mpg ~ wt + disp + wt * disp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.267 -1.677 -0.836  1.351  5.017
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.081998   3.123063  14.115 2.96e-14 ***
## wt          -6.495680   1.313383  -4.946 3.22e-05 ***
## disp        -0.056358   0.013239  -4.257  0.00021 ***
## wt:disp      0.011705   0.003255   3.596  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.455 on 28 degrees of freedom
## Multiple R-squared:  0.8501, Adjusted R-squared:  0.8341
## F-statistic: 52.95 on 3 and 28 DF,  p-value: 1.158e-11
```