Peer graded assignment regression models

yinshu zhang

October 24, 2018

Introduction

This study is conducted to explore the relationship between a set of variables and the fuel consumption. The focus of is to address two questions.

> "Is an automatic or manual transmission better for MPG"

> "Quantify the MPG difference between automatic and manual transmissions"

Summary

By perform variable F-test and one way ANOVA test, strong evidences are observed that support relationship between transmission type and fule economy. Using multi-variable linear regression model, we conclude manula transmission results higher MPG number by 2.94 mile per gallon compares to automatic transmission, this value has 1.41 mpg standard error.

Data Exploration

We will use "mtcars" data to draw conclusions, mtcars is a 32 by 11 data frame, with columns defined as:

column name	defination
mpg	miles per gallon(US)
cyl	number of cylinders
disp	engine displacement(cu. in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	zero to 1/4 mile time in seconds
vs	Engine shape (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

A quick look at each variables

Relatinoship between MPG and AM

Correlation

From figure 1, we can see the MPG and AM variables has 0.5998 correlation value, it a indication MPG and AM has relationship, but not strong enough. Therefore, to answer first question, we need better evidences.

We will test the correlation betwen MPG and AM with alternative hypothesis "true correlation is greater than zero"

```
cor.test(mpg, am, alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  mpg and am
## t = 4.1061, df = 30, p-value = 0.0001425
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.3691544 1.0000000
## sample estimates:
##       cor
## 0.5998324
```

From the output, p-value is far away from 95% confidence interval, this is a good evidence *supporting* MPG and AM are correlated.

One way ANOVA test

To further prove the opposite is false, we can use one-way ANOVA. The null hypothesis of ANOVA is designed as the mean MPG are the same between auto and manual transmission cars, ie. either those two samples are from either from same population(or different populations with same mean), or when two sample show difference, it is due to chance.

```
anova1 <- aov(mpg ~ am, data = mtcars)
summary(anova1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## am           1  405.2   405.2   16.86 0.000285 ***
## Residuals   30  720.9    24.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
f_crit <- round(qf(.95, 1, 30),2)
```

F value of 16.86 is significiantly larger than the 95% confidence interval critical value 4.17, the result shows strong evident *rejecting* null hypophsis.

Regression

simple regression

For the sake of comparsion, let's do a single variable(factor actually) linear model.

```
mod.sim <- lm(mpg ~ am, data = mtcars)
summary(mod.sim)
```

```
##
## 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 ***
## am             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
```

To interpurte the output, the coefficient of AM is 7.245, which means changing from auto to manual, will result on average 7.245 mile/gallon fuel consumption increase. However, as expected, the model quality is low, we can't use it to predict the MPG, when only around 36% of residuals varians are explained by this model, AM is a dummy vailable after all.

seletion of independent variables

To make a multi-variable liner regression model, there are ten variables to choose from. we need to aviod chose highly correlated variables, take a quick look at first summary plot, we can see quite some variables are highly related.

This is not a big surprise, with basic car knowlege we can understand why some of independent variables are correlated. For example, more cyclder of an engine normally means bigger displacement, higher the hourse power normally results lower quater mile time.

We can categorize those ten variables into following group

1, drive train specs: cyl, disp, drat, vs, carb, drat, am, gear

2, other veichle specs: wt

3, performance measurements: qsec, hp

we can chose ourselvies and compare the results but better option is to get help from package "leaps".

Here we did a exhausded subset comparsion, with one best result from each number of variable combination. from the right plot, we can see the adjusted R squre does not show big difference after 0.7, the mallow's Cp however have one very low value compares to other combinations. Therefore, we will create regression model with independent variables of weight, quarter mile time, and auto/manual transmission.

```
mod.mul <- lm(mpg ~ wt + qsec + am, data = mtcars)
summary(mod.mul)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
```

```
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Interpretation of the output: This model tells us on every 1000 pounds increase, the veichile fuel economy will reduce by 3.9165 mpg if other two variables remain unchanged. This is expceted as the heavier the car, the wrose the fuel economy.

The next line quater mile seconds read as every second increase for a car from stand to reach quater mile, the fule economy will increase 1.2259 mpg, if other variable remain unchanged. The explaination is the lower "qsec" the fast car can travel, normally this is result of more powerful engine, which yeild worse mpg number.

Next line of am is the *answer* to second question, if other variables are unchanged, we only change transmission from auto to manual, it will improve car fuel economy by 2.9358 mpg.

Now we exame the significiant level of each of those variables. Both weight and quater mile seconds are highly significait to fuel economy, with very small p-value. However the "am" is just below 95% critical value, this tells us transmission type is related to MPG but not as sigificiant as other two variables.

Regression Residuals

Fianlly we want to check the residual plot and distribution, to verify if

1, distribution is fairly normal.

2, if multi-variable regression residual has smaller variation.

```
regfit.full <- regsubsets(mpg ~ ., data = mtcars, nbest = 1)
par(mfrow=c(1,2))
plot(regfit.full, scale = "adjr2", main="Adj R sqr")
plot(regfit.full, scale = "Cp", main="mallow's Cp")
mtext("figure 2, regression subset selection", side = 1, outer=T, line = -3)
```

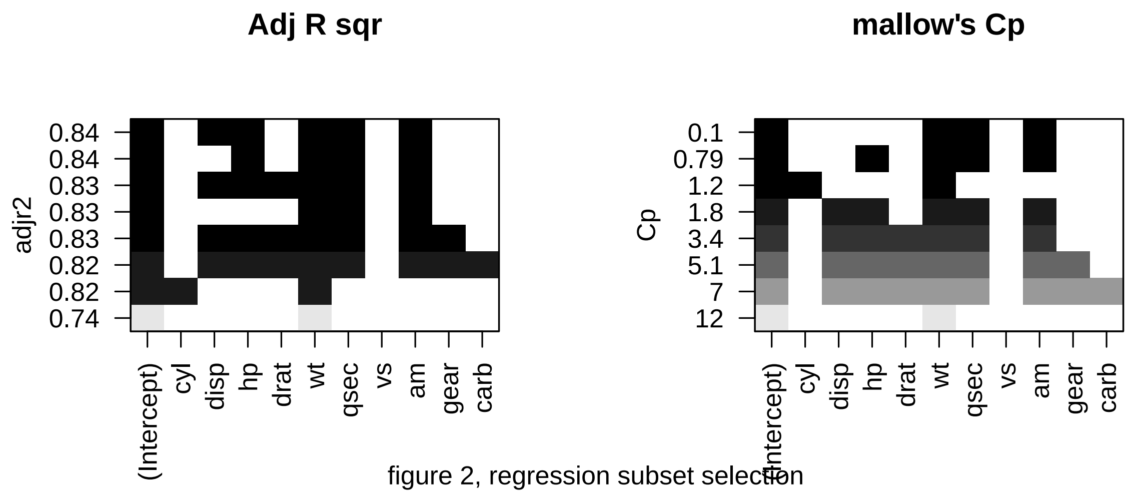
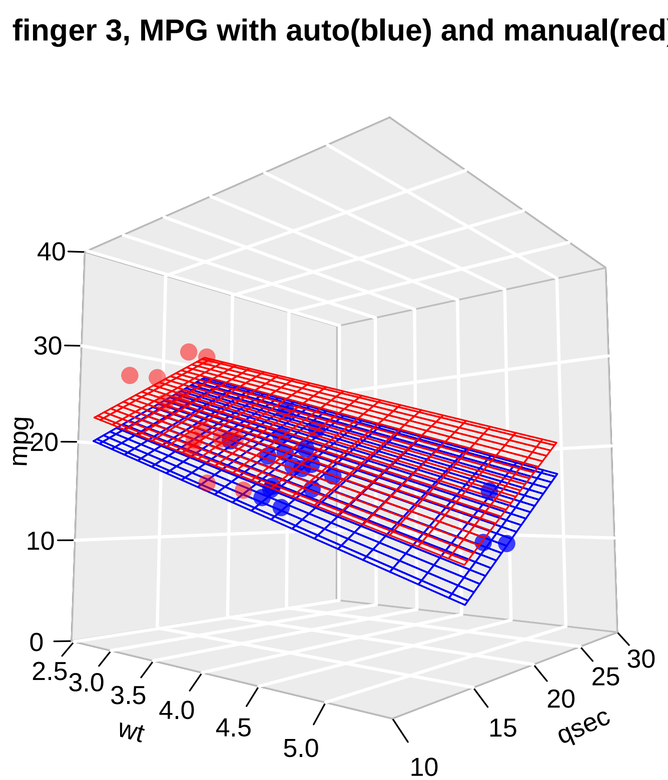


figure 2, regression subset selection

```r
par(mfrow=c(1,1))
```



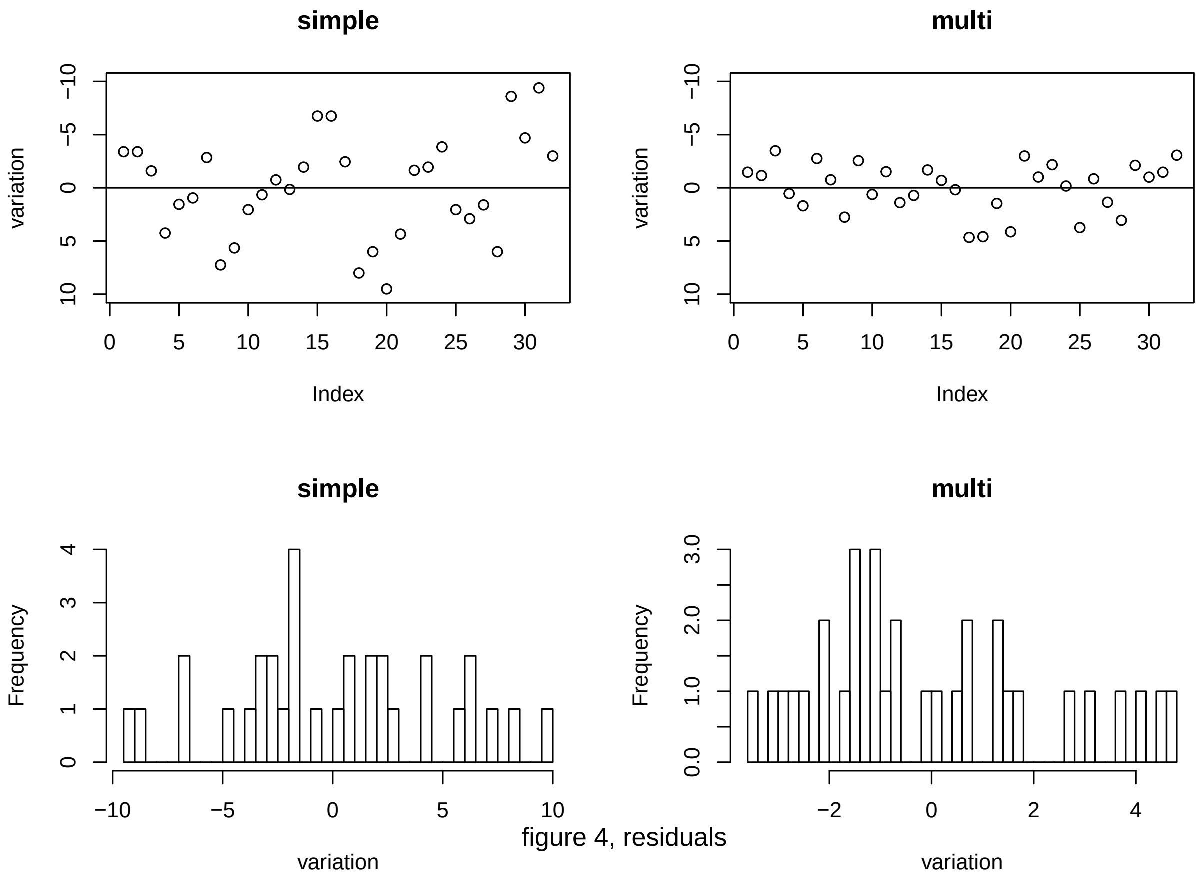


figure 4, residuals