Coursera Regression Project

Zhongyi Lin

Saturday, April 25, 2015

Abstract

In this mini project, the relation of two variables, automatic/manual variable (am) and miles per gallon variable (mpg), is investigated to answer the questions “Is an automatic or manual transmission better for MPG?”. Also the the mpg difference between automatic and manual transmissions is quantatively defined. Techniques of exploratory data analysis are used in order to achieve the goal.

Analysis

Single variable regression

At the very beginning, we firstly import the mtcars data and take a quick look at the head of it.

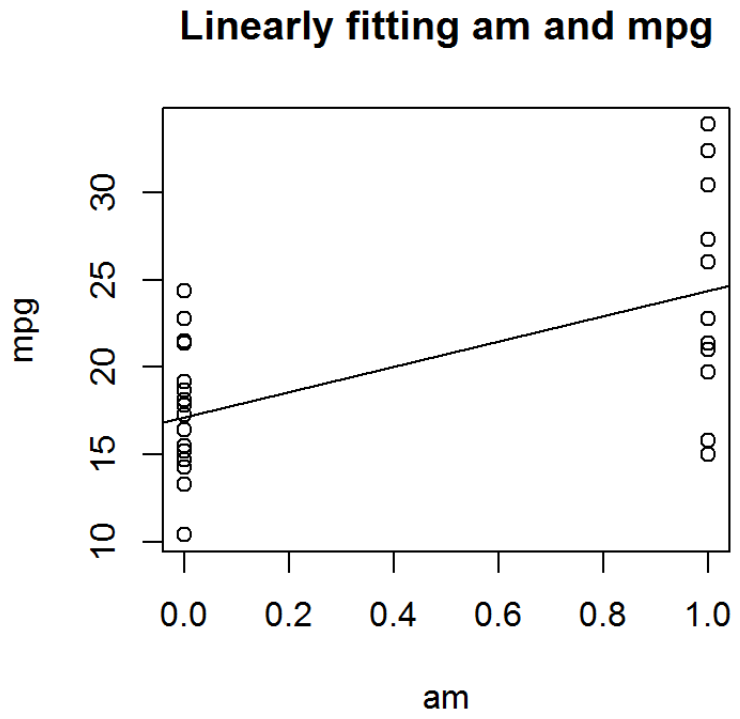
```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

It can be observed that the automatic/manual variable (am) is binary. Setting am as the predictor and mpg as the outcome, we first try linear regression with binary independent variables on these two data sets then show the coefficients and plot the fitting line of the data.

```
mpg <- mtcars$mpg
am <- mtcars$am
fit <- lm(mpg~am)
fit$coefficients
```

```
## (Intercept)          am
##   17.147368    7.244939
```

```
plot(am,mpg,main="Linearly fitting am and mpg")
abline(fit)
```



From the distribution of the scattered points in the figure we can see that there is a significant difference on Miles per Gallon bewteen automatic and manual car. As indicated by the formula of the model,

$$mpg_i = b_0 + b_1 am_i + e_i,$$

the difference between mpg(0) and mpg(1) is equal to b_1, the slope, the value of which is equal to 7.245 as shown above. This is to say, a manual car is expected to drive 7.245 miles more than a automatic car by burning one gallon of gasoline.

To have a more detailed look at the result, we plot the results of the fitting to examine the normality of the residuals.

```
par(mfrow=c(2,2))
plot(fit)
```

(See Fig.1 in appendix) It can be seen that the scattered points of the residual are randomly distributed in both sides of the 0 line. These points roughly form a horisontal band around the 0 line, and there is no single point standing out of the others. These characteristics means that the assumption of linearity is reasonable, the variances of the error terms are equal, and homoskedasticity is guaranteed.

Multi-variable analysis

However, data provided in the mtcars package is of different aspects which describe a car as an integration. This means attributes other than am probably also contributes to the mpg variable. Thus a multi-variable analysis is necessary. Linear regression with multiple variable is conducted below.

```
fit_mul <- lm(mpg~.,data=mtcars)
summary(fit_mul)$coefficients
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Obviously am and wt are the two variables contribute to mpg the most. It is intuitively understandable that heavier cars cost more fuel for driving the same mileage. Besides no sign inversion is shown in the result.

Similar to previous analysis, therefore, the difference between mpg(0) and mpg(1) is equal to 2.520.

We also check the assumption of linearity and homoskedasticity, which are both passed. (See Fig.2 in Appendix)

Comparison

An ANOVA analysis can be applied here to compare the single varible and multi-variable regression.

```
anova(fit,fit_mul)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     21 147.49  9     573.4 9.0711 1.779e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-value larger than 1 and p-value less than 0.05 reject the null hypothesis of the anova test and suggest that the multi-variable model is better than the single variable one in this case.

Conclusion

In this mini project, single variable and multi-variable linear regression are applied to solve the two given problems. Regression diagonostics of both models are checked to guarantee the correct application. After comparison a conclusion is drawn that multi-variable linear regression is better in this analysis.

Appendix

Fig.1 Results of the single variable linear regression

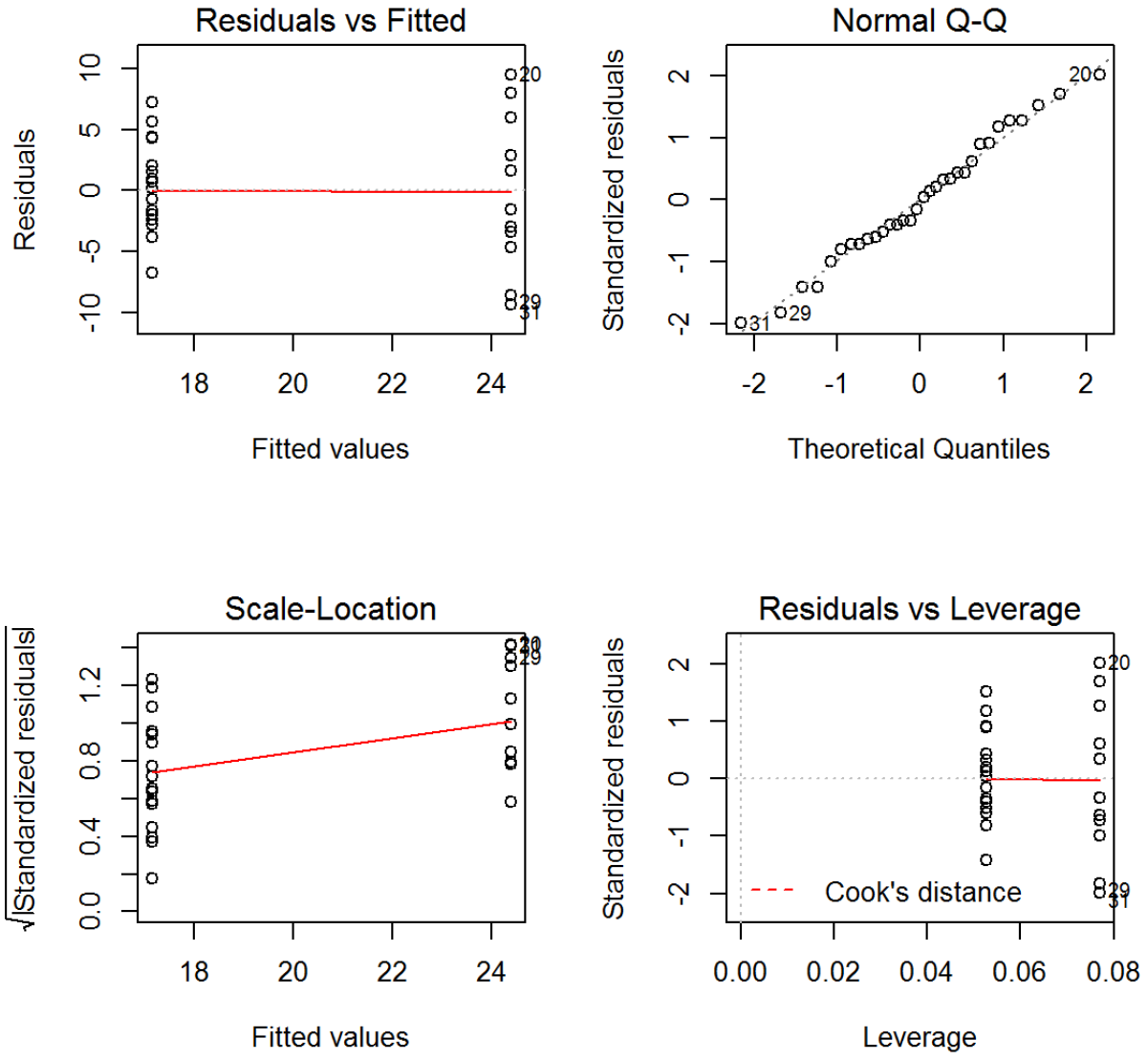
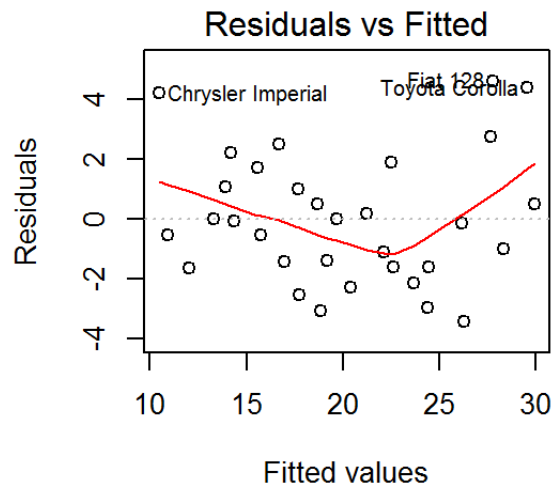
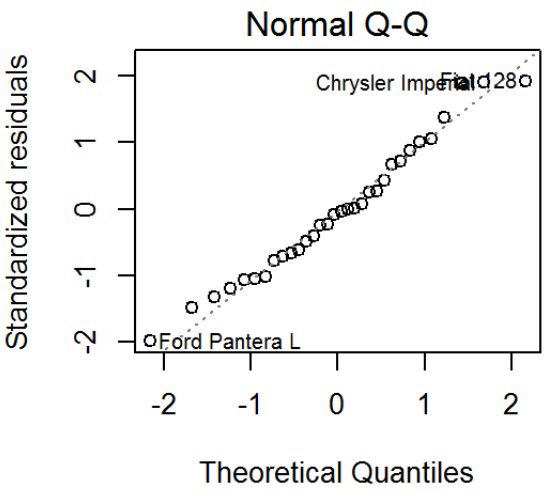
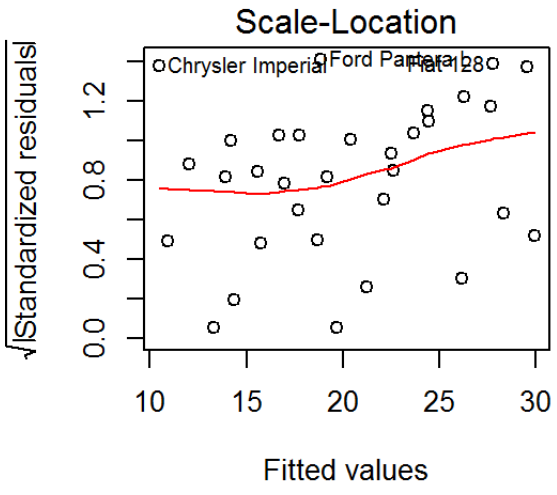
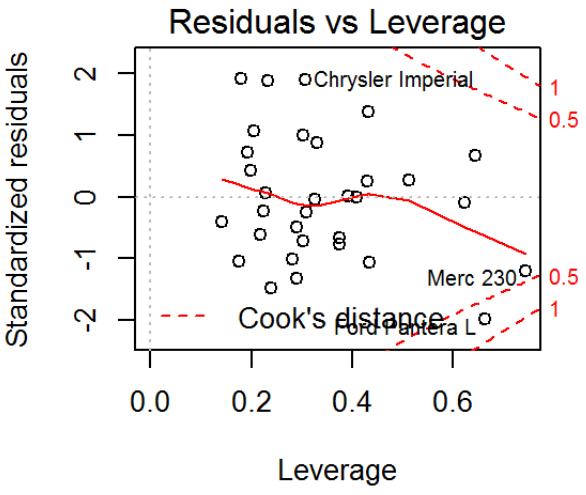


Fig.2 Results of the multiple variable linear regression


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Ford Pantera L


Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Ford Pantera L
Fiat 128


Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Merc 230
Ford Pantera L
Cook's distance
1
0.5
0.5
1