Regression Project

Junyang Liu

11/5/2016

Regression Methods Project: Effects of Automatic and Manual Transmission on MPG of Cars

Executive Summary:

This is a course project from Coursera, offered by Johns Hopkins University. In this report, we will investigate a data set(`mtcars`) from `MASS` package and address the following questions: 1. Is an automatic or manual transmission better for MPG? 2. Quantify the MPG difference between automatic and manual transmissions. This analysis discovers that there is a significant difference in MPG between automatic transmission and manual transmission. from the model we pick, with everything holding fixed,by switching from automatic transmission to manual transmission, the MPG on average is going to increase 1.80921 units.

Getting and Cleaning&Transforming the data

First, we load the data from `MASS` package and change the class of some variables to factors.

```r
library(MASS)
cars<-mtcars
cars$am<-factor(cars$am, labels = c("Auto", "Manual"))
cars$vs<-factor(cars$vs)
cars$cyl<-factor(cars$cyl)
cars$carb<-factor(cars$carb)
cars$gear<-factor(cars$gear)
```

Exploratory Data analysis

From the plot, we can see that `cyl`, `hp`,`wt`, and `am` are target variables. to varify that, I uesed a `step` function to confirm.

```r
fullmodel<-lm(mpg~.,data = cars)
```

```r
summary(step(fullmodel,direction = "both"))
```

```r
bestmodel<- lm(mpg ~ cyl + hp + wt + am,data = cars)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

From the summary, we can see that the Residual standard error: 2.41 on 26 degrees of freedom, the Adjusted R^2 is 0.84, meaning that about 84% of the error is explained through the regression. `mpg` will decrease by 2.49683 units for every 1000 lb increase in `wt`, while hoding other predictors fixed. `mpg` decreases 0.03211 units for every `hp` increase, while hoding other predictors fixed.in this cass, this is negligible. `mpg` decreases 3.03134 units if `cyl` increase from 4 to 6,while hoding other predictors fixed. `mpg` decreases 2.16368 units if `cyl` increase from 6 to 8,while hoding other predictors fixed. `mpg` increases 1.80921 units if `am` increase from auto to manual, while hoding other predictors fixed.

Model Analysis

now we have our theoretical `bestmodel`. We need to verify that if the predictors in this model is indeed significant enough to be included. I used `mpg` against `am` as a base model.and used `anova` test to test its significance.

```r
basemodel<-lm(mpg ~ am, data = cars)
anova(fullmodel,basemodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
##   Res.Df   RSS  Df Sum of Sq      F   Pr(>F)
## 1     15 120.4
## 2     30 720.9 -15   -600.49 4.9874 0.001759 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the the anova summary, the every low P-value suggests taht we reject the null hypothesis that `cyl`, `hp`, and `wt` are not significant contributors to the `bestmodel`. Therefore, our `bestmodel` have all the significant contributors.

Residual Control

Now, we need to check if the residuals is constant, normally distributed,and wether it has large outliers. From the Q-Q plot(in Figure3) we see that the residuals are approximately normal becasue they are close to a line. Residuals are approximately constant, because from the Fitted.value vs Residual plot,I did not see an obvicious pattern. and the Scale-Location plot confirms it is contant variance, and randomly distributed. Lastly, the Leverage-Residuals plot shows that there is no extreme outliers since all the points lies in the 0.5 bands line.

Appendix:Plots and Figures

Figure 1, Car MPG by Transmission type

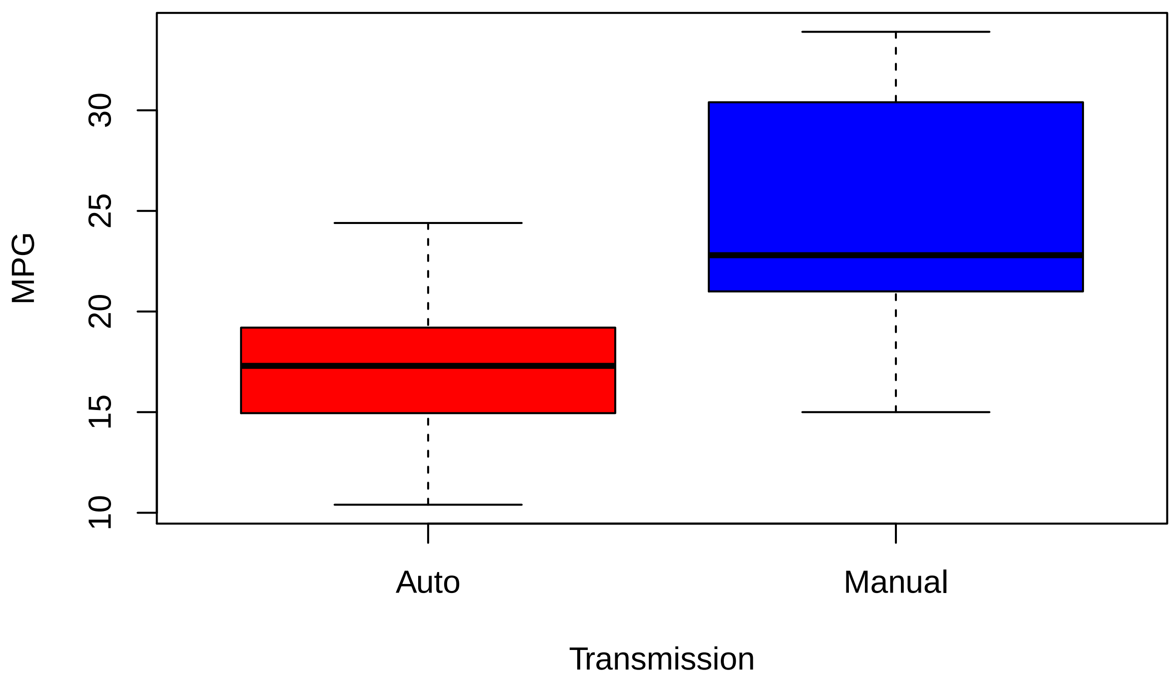
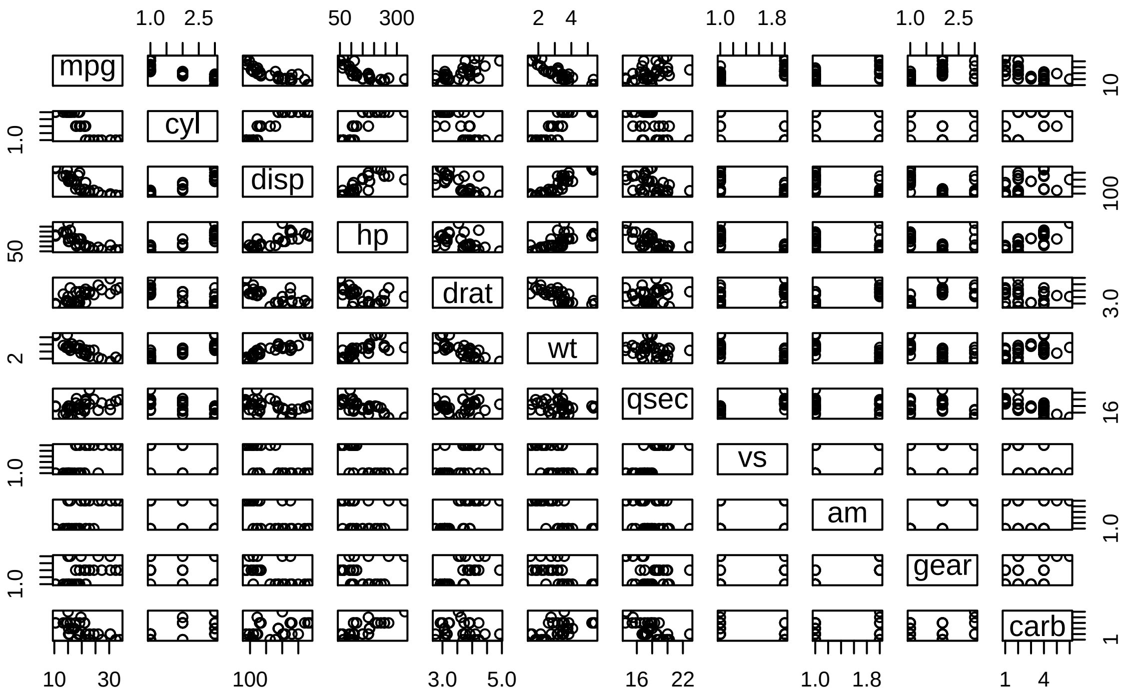


Figure 2


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb

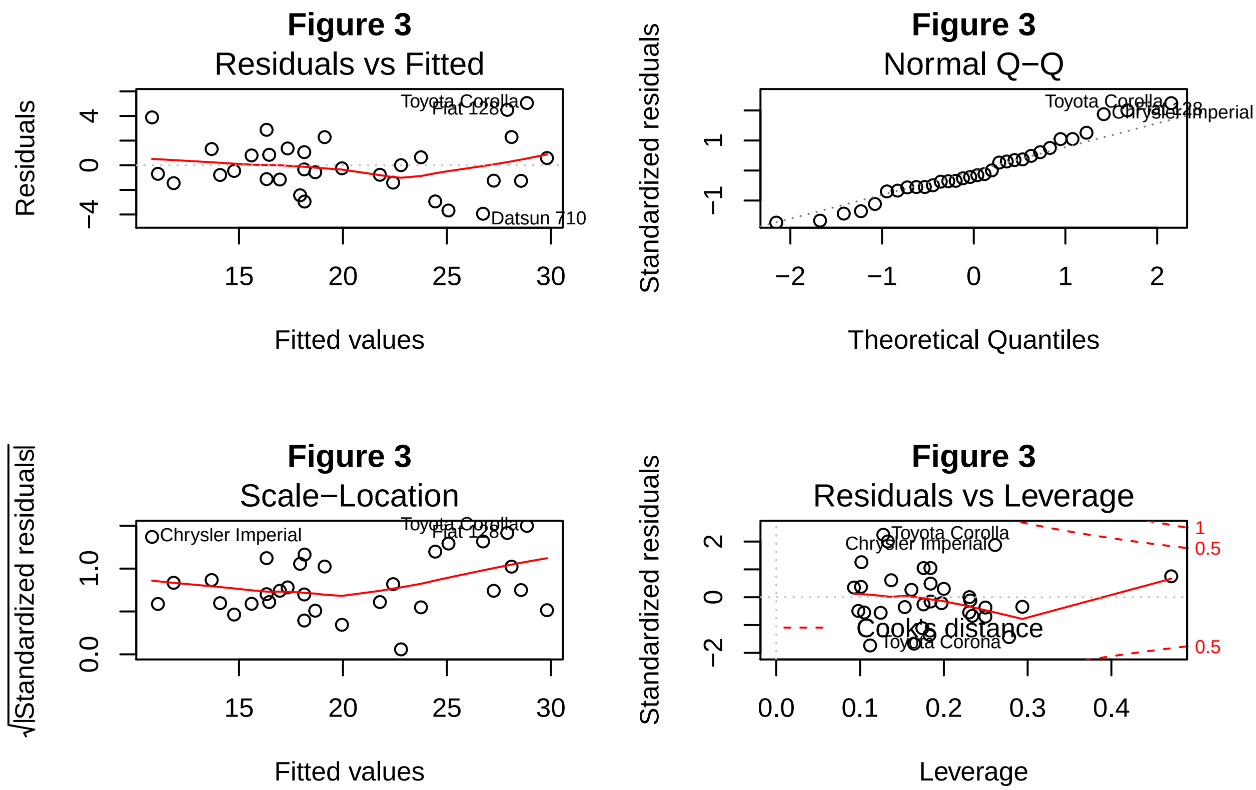


Figure 3
Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Figure 3
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Figure 3
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Figure 3
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5