Regression_week4_assignment

Yuan Dong

1/19/2018

Regression Models Course Project

Syposis

This is a study for Motor Trend, a magazine about the automobile industry. In order to answer following two questions: "Is an automatic or manual transmission better for MPG"; "Quantify the MPG difference between automatic and manual transmissions". We explored the relationship between a set of variables and miles per gallon (MPG) (outcome), using mtcars data set of a collection of cars. We found that automatic or manual transmission did not have significant influence for MPG. There are 0.1765 (95%CI: -2.50~2.85) in the estimated mpg change between auto and manual transmission. But because the 2.5% CI is below zero, so the influence of am is not significant, there are uncertainty in the conclusions.

Data analysis

Load R packages and dataset mtcars, check variables

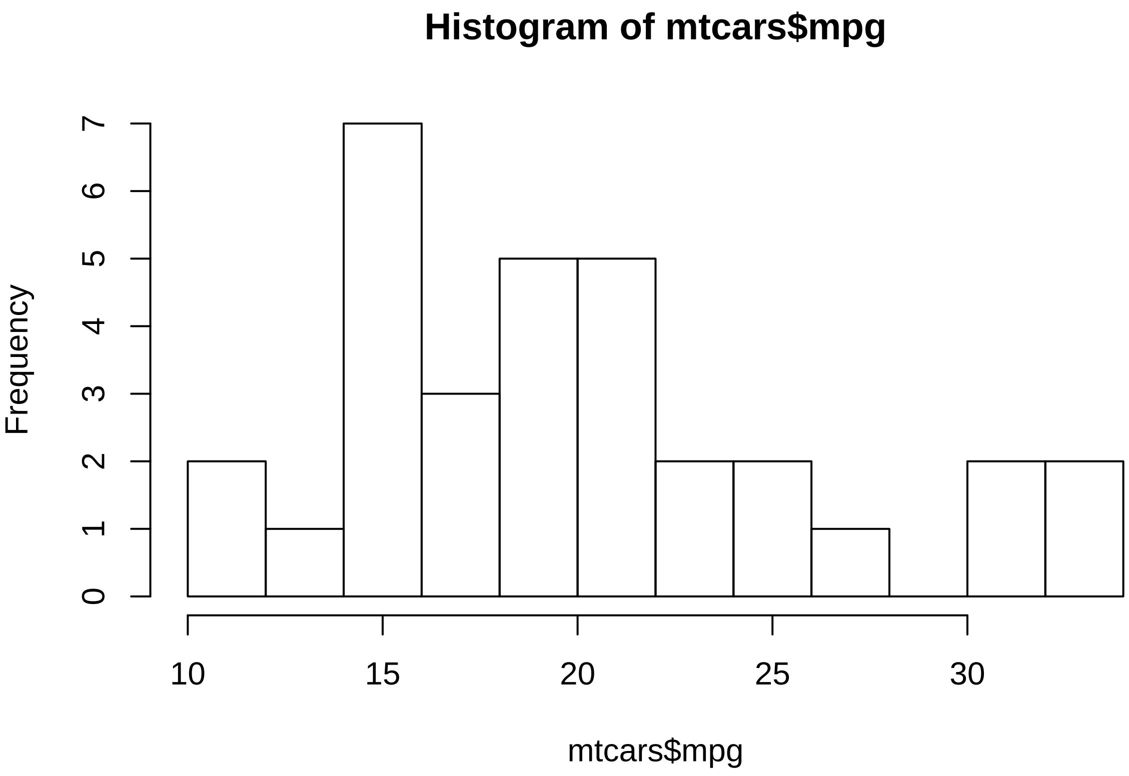
```
data(mtcars)
head(mtcars,3)
```

```
##                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
```

```
str(mtcars)
```

```
## '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 : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ 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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
#plot to explore the mpg variable (the predictor).
hist(mtcars$mpg, breaks = 10)
```



Linear regression model Selection

The hist plot of mpg variable is nearly normal, we can use linear regression model for our research. First, we discover the correlation of mpg with different variables.

```
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

We will use nested model and anova test to decide which variables to include into the final regression model. Except am (which we most interested in), we put variables into the model according to their correlation score (high to low). Then use anova to check if we should add that variable.

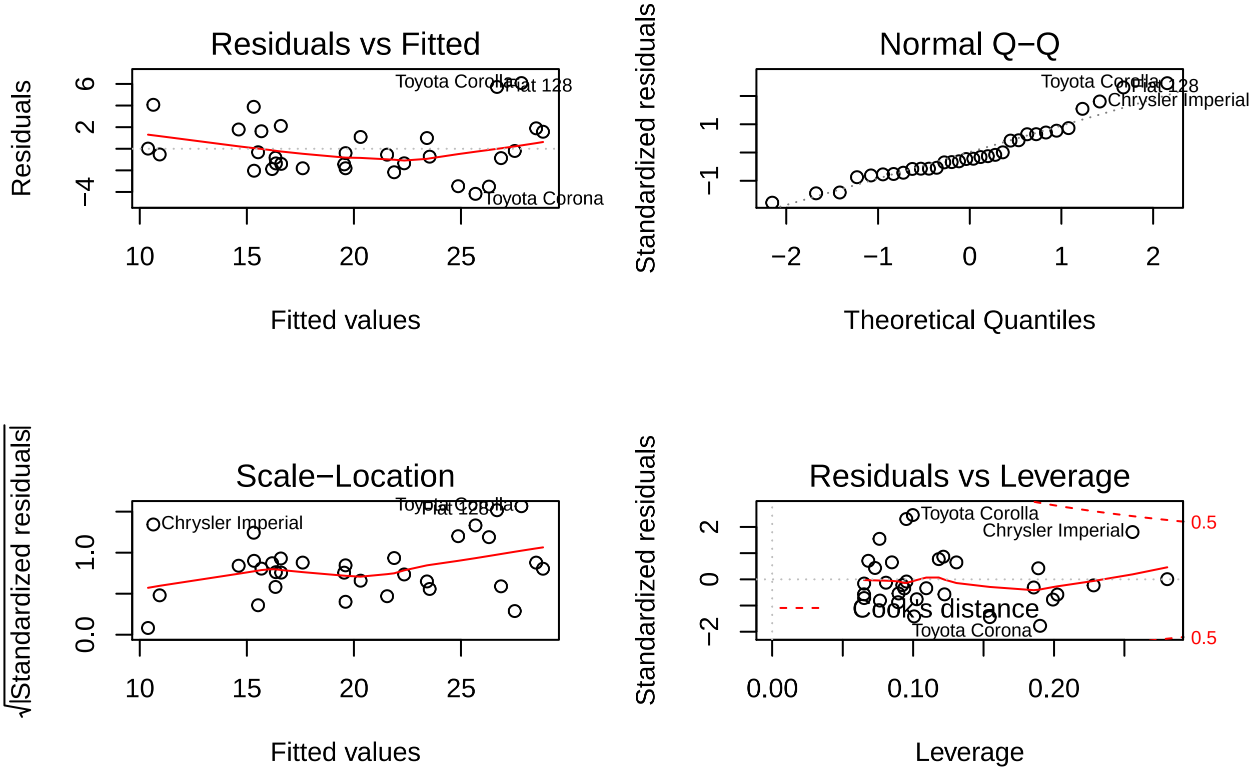
According to the anova result, we should use am, wt and cyl as variables in our regression model (fit3).

```
fit1<-lm(mpg~as.factor(am), data = mtcars)
fit2<-lm(mpg~as.factor(am)+wt, data=mtcars)
fit3<-lm(mpg~as.factor(am)+wt+cyl, data=mtcars)
fit4<-lm(mpg~as.factor(am)+wt+cyl+disp+hp, data=mtcars)
fit5<-lm(mpg~as.factor(am)+wt+cyl+disp+hp+drat+vs, data=mtcars)
fit6<-lm(mpg~as.factor(am)+wt+cyl+disp+hp+drat+vs+carb+gear+qsec, data=mtcars)
anova(fit1,fit2,fit3,fit4,fit5,fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + wt
## Model 3: mpg ~ as.factor(am) + wt + cyl
## Model 4: mpg ~ as.factor(am) + wt + cyl + disp + hp
## Model 5: mpg ~ as.factor(am) + wt + cyl + disp + hp + drat + vs
## Model 6: mpg ~ as.factor(am) + wt + cyl + disp + hp + drat + vs + carb +
##     gear + qsec
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 191.05  1     87.27 12.4257   0.00201 **
## 4     26 163.12  2     27.93  1.9881   0.16191
## 5     24 158.65  2      4.47  0.3179   0.73112
## 6     21 147.49  3     11.16  0.5296   0.66684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then we use plot to make sure our model are correct. According to the plot, residuals seemed to be automatically distributed around zero, no systemetic pattern observed. So our model is seemed to be good.

```r
par(mfrow=c(2,2))
plot(fit3)
```



Linear Regression Model Interpretation

First, summary our regression model (fit3). According to the summary, the answer to this question “Is an automatic or manual transmission better for MPG” is “NO”. Because P-value of am variable is 0.89.

The coefficient of variable am can be interpreted as: 0.1765 (95%CI: -2.50~2.85) in the estimated mpg (miles per gallon) change (intercept) between auto and manual transmission, going from auto to manual. Because the 2.5% CI is below zero, so the influence of am is not significant, there are uncertainty in the conclusions.

```
summary(fit3)
```

```
## 
## Call:
## lm(formula = mpg ~ as.factor(am) + wt + cyl, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.1735 -1.5340 -0.5386  1.5864  6.0812 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     39.4179     2.6415  14.923 7.42e-15 ***
## as.factor(am)1   0.1765     1.3045   0.135  0.89334    
## wt              -3.1251     0.9109  -3.431  0.00189 ** 
## cyl             -1.5102     0.4223  -3.576  0.00129 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122 
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

```
confint(fit3)
```

```
##                    2.5 %     97.5 %
## (Intercept)    34.007153 44.8287134
## as.factor(am)1 -2.495555  2.8485408
## wt             -4.991001 -1.2592836
## cyl            -2.375245 -0.6452459
```