Regression Models Course Project

yhuai

2018/4/22

Coursera Data Science Regression Models Course Project

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1.Is an automatic or manual transmission better for MPG?

2.Quantify the MPG difference between automatic and manual transmissions

Overview

```
library(datasets)
library(MASS)
library(ggfortify)
```

```
## Loading required package: ggplot2
```

```
library(car)
```

```
## Loading required package: carData
```

```
data("mtcars")
str(mtcars)
```

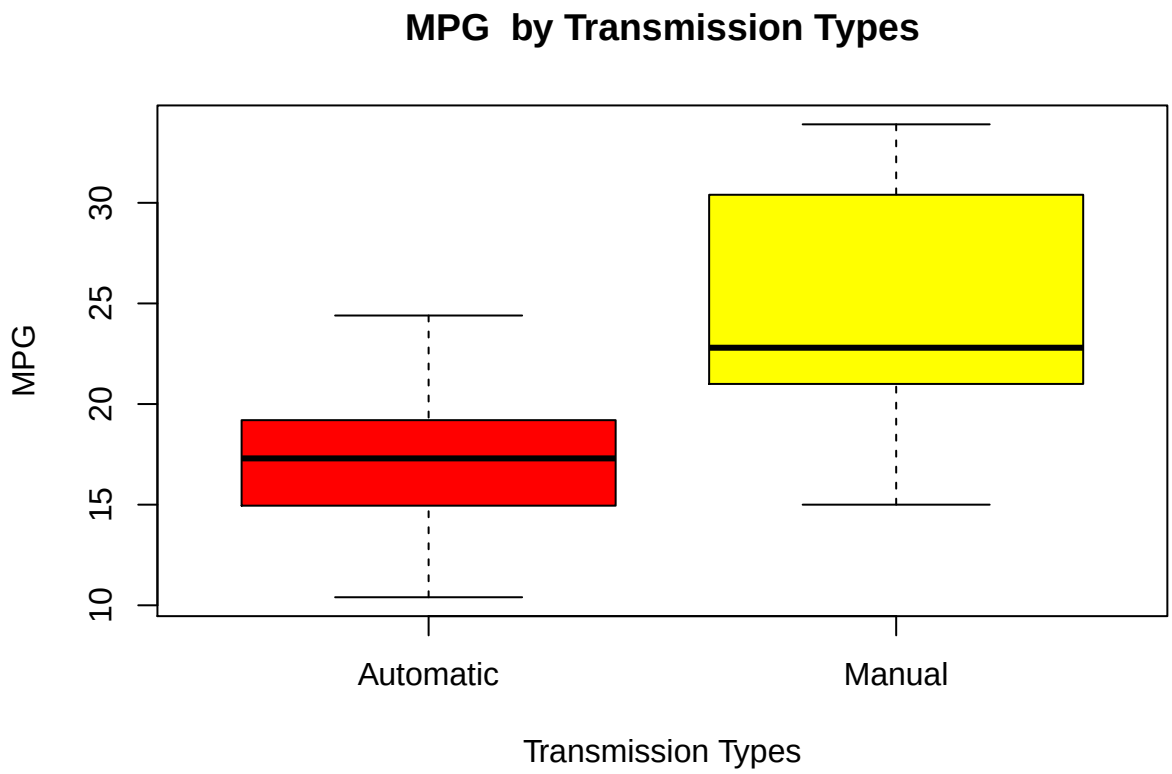
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 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 ...
```

From the result of str function, we need to understand the meaning of variables: mpg–miles per gallon | cyl–Number of cylinders | disp–Displacement | hp–horsepower |drat–Rear axle ratio | wt–weight(lb/1000) | qsec–1/4 mile time | vs–V/S | am–transmission(0 is auto, 1 is manual) | gear–gears | carb–carburetors

Exploratory Data Analysis

First of all, we want to compare the automatic and manual simply in a boxplot:

```
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual"))
boxplot(mtcars$mpg ~ mtcars$am, xlab="Transmission Types", ylab="MPG", main="MPG  by Transmission Types
```

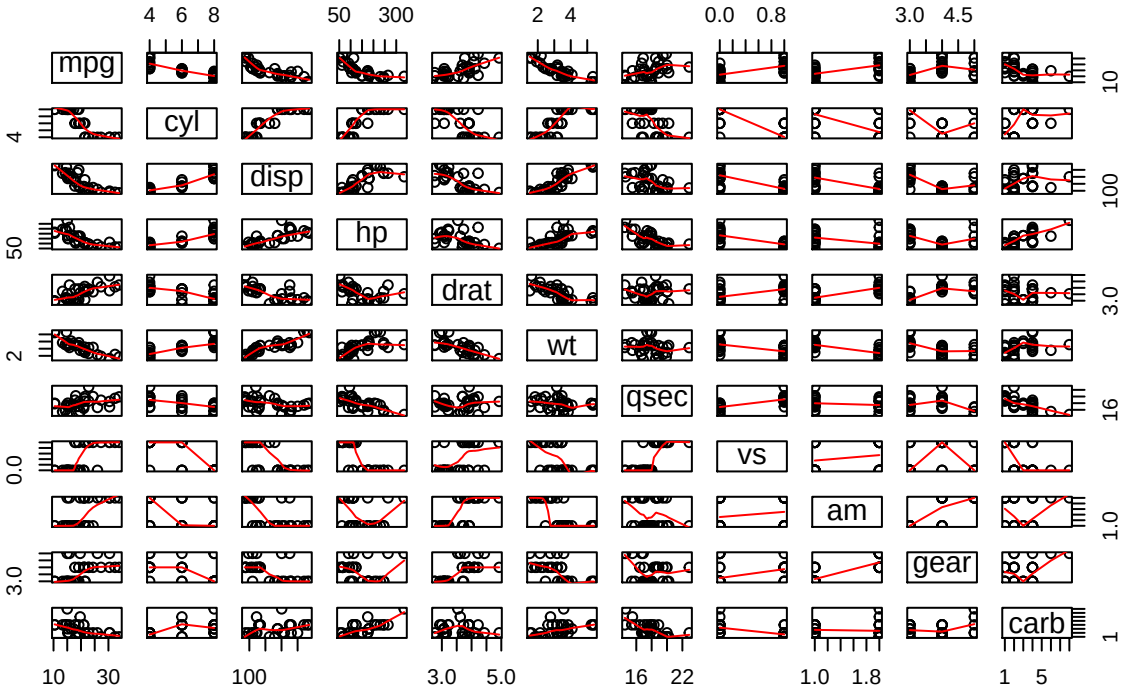


From the boxplot, we got preliminary opinion that manual clearly have more MPG than auto. Now let's go deeply.

Regression Analysis

```
pairs(mpg ~ ., data = mtcars, panel=panel.smooth, main="Overview")
```

Overview




```
fit1 <- lm(mpg ~ am,data = mtcars); summary(fit1)
```

```
##
## 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 ***
## amManual       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
```

From the pairs, we can find that "mpg" may has linear relationsip with multiple variables, so our model must consider multiple parameters. Here we take advantages from AIC test and ANOVA test. Taking into account the statistical fit of the model and the parameters used to fit the model, the smaller the AIC value, the more preferential the model is selected. This shows that with fewer parameters a sufficient degree of fitness can be obtained.

```r
my_lms <- lm(mpg ~. ,data = mtcars)
stepAIC(my_lms)
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
```

```
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
```

```
## (Intercept)           wt         qsec     amManual
##       9.618       -3.917        1.226        2.936
```

From the result, we choose "lm(mpg ~ wt + qsec + am)" with the lowest ACI, but we also need to check "lm(mpg ~ hp + wt + qsec + am)" which has pretty lower ACI value. Then, check them with our base model "fit1" by ANOVA.

```
fit2 <- lm(mpg ~ wt + qsec + am, data = mtcars)
fit3 <- lm(mpg ~ hp + wt + qsec + am,data = mtcars)
anova(fit1,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ hp + wt + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 46.5228 1.787e-09 ***
## 3     27 160.07  1      9.22  1.5551    0.2231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results from ANOVA shows that "hp" is not necessary and our first choice "Model2" is great. From summary function, the R-squared value shows it is fitting and P-value shows it is significant, give us reliance of our model.

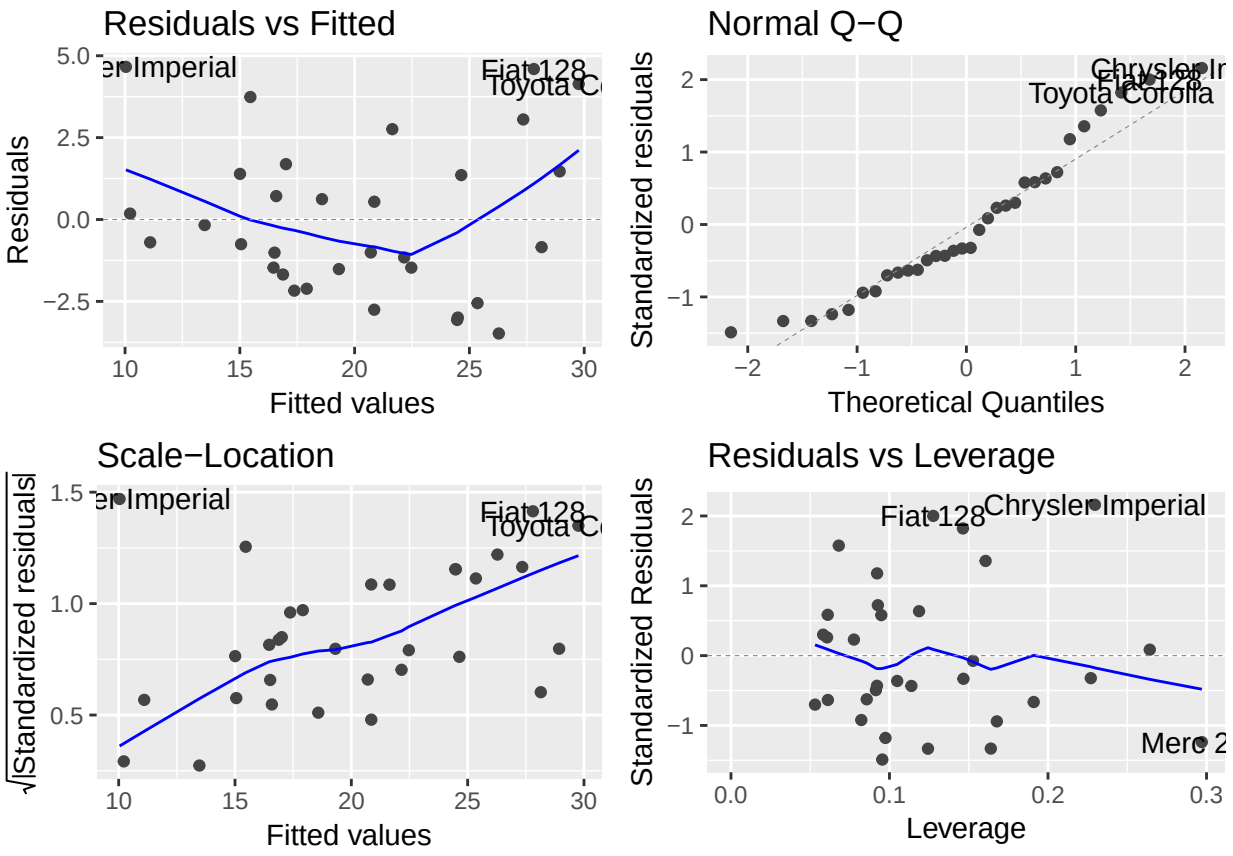
```
summary(fit2)
```

```
##
## 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 ***
## amManual      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
```

Residual and Diagnostics

Residual and diagnostics are important. We start from the plot:

```
autoplot(fit2)
```



The Residuals vs Fitted – The abscissa is the Y value in the fitted equation, and the ordinate is the residual value. The closer the fitting curve is to 0, the smaller the error between the fitting function and the sample point, and the better the model.

The Normal Q-Q – The abscissa is the quantile of the standard normal distribution, and the ordinate is the sample value. Whether the points on the Q-Q diagram are approximately in the vicinity of a straight line is used to identify whether the sample data approximates a normal distribution.

The Scale-Location –If the invariant variance hypothesis is met, the points around the horizon should be randomly distributed in this plot.

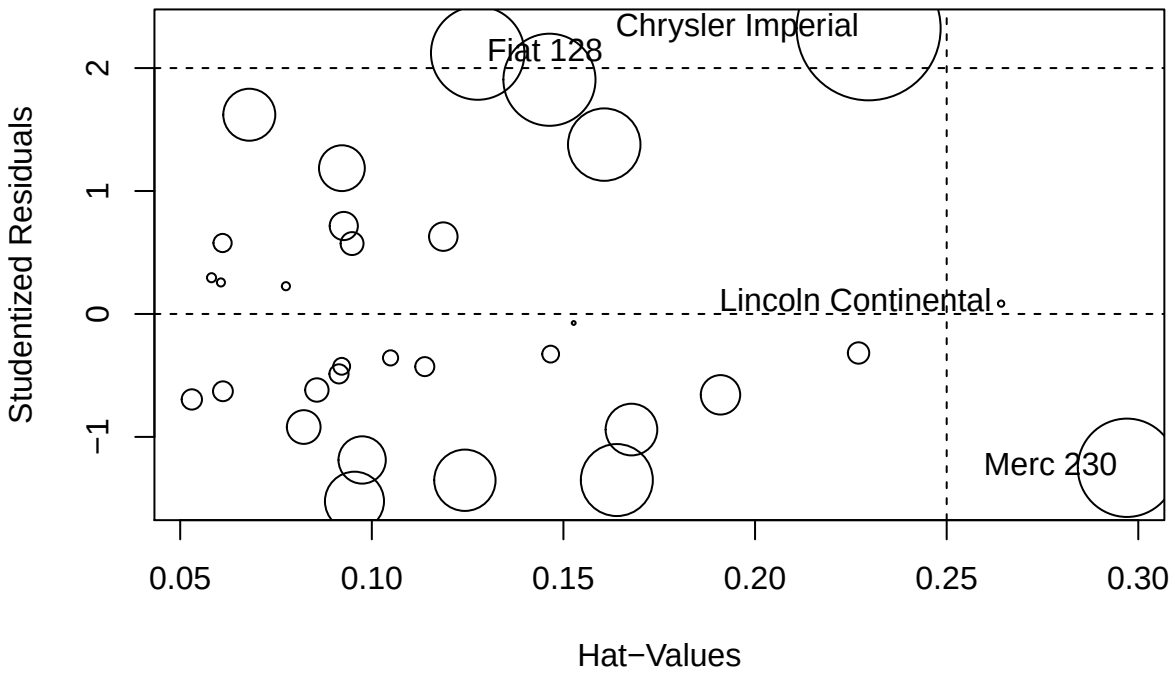
The Residual vs Leverage – there are some outliers, high leverage and strong influence points fall outside the confidence interval

We have to find out the special points for improving our model in the future:

```
outlierTest(fit2)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##                   rstudent unadjusted p-value Bonferonni p
## Chrysler Imperial 2.323119           0.027949      0.89437
```

```
influencePlot(fit2)
```




```
##                          StudRes       Hat        CookD
## Merc 230            -1.25110559 0.2970422 0.1620826668
## Lincoln Continental  0.08383783 0.2642151 0.0006541961
## Chrysler Imperial    2.32311937 0.2296338 0.3475974030
## Fiat 128             2.12214584 0.1276313 0.1464019096
```

Inference

t-test also give us some useful information:

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

From the result of t-test we can say manual and automatic transmissions have significantly different influence on mpg, manual gets 7.245 more mpg(average).

Conclusion

To sum up, manual transmission is better for MPG. The manual transmission cars have 2.9358 more MPG than automatic transmission cars(When “wt” and “qsec” stay in constant)