Linear Regression Models Project

Introduction

This project studys the "mtcars" data to for Motor Trend, a magazine about the automobile industry. The main questions in the studys are: 1. If automatic or manual transmission better for MPG (miles per gallon); 2. quantify the difference.

Load the data

```
data(mtcars)
```

exploratory

```
str(mtcars)
```

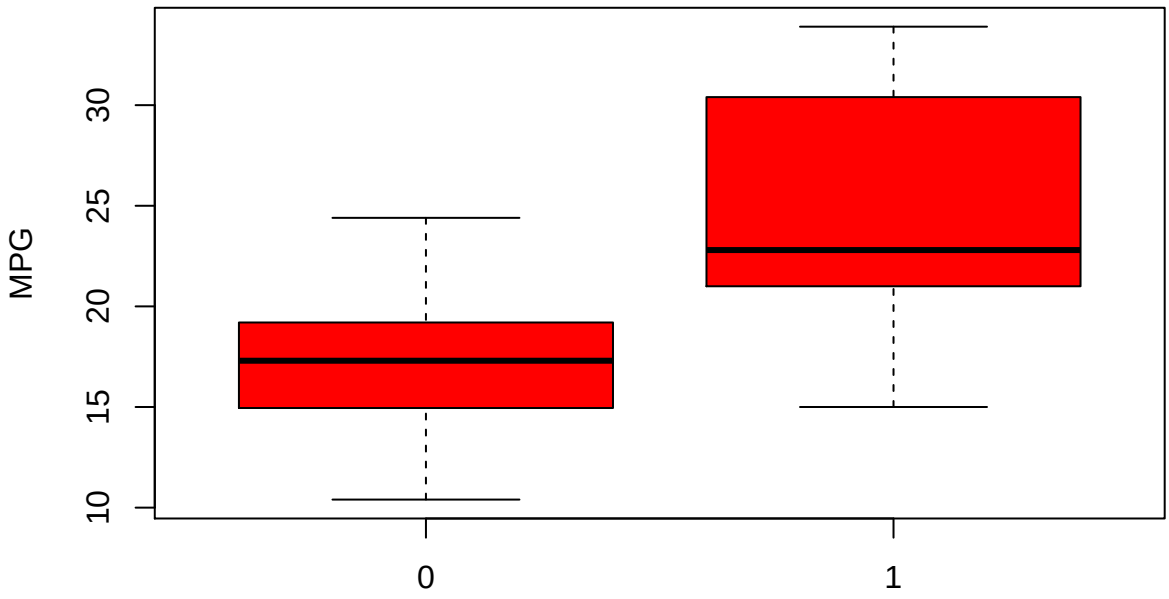
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76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

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Use plot to show the distribution difference between automatic or manual transmission:

```
boxplot(mpg ~ am, data = mtcars, col = "red", ylab = "MPG")
```



Observations:

From the prot, we can see there is different impact of automatic or manual transmission. We will use regression to quantify the difference.

Simple regression model

First, we only use very simple regression model, using only "AM" as predictor, and "MPG" as outcome.

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

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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 summary,there is 7.245 MPG increase on averge on manual transmission (AM=1) over automatic transmission (AM=0). Further check the confidence level:

```
confint(fit1)
```

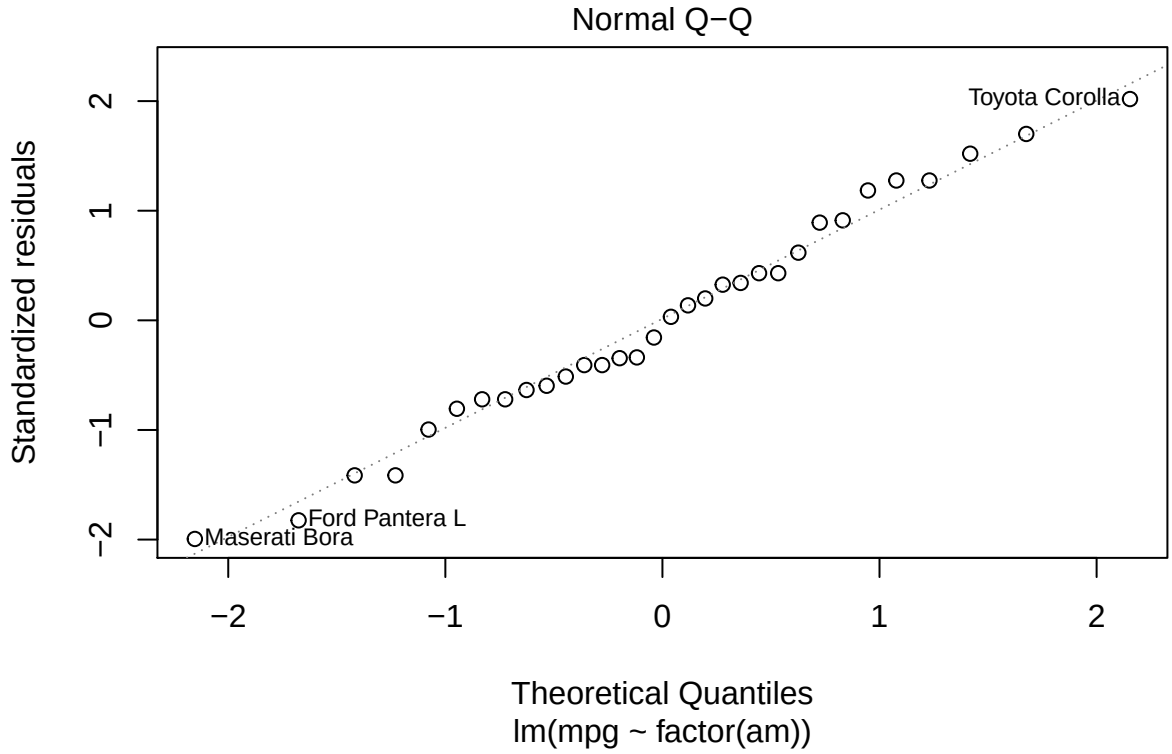
```
##                 2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## factor(am)1  3.64151 10.84837
```

Which shows the 95% confidence level is from 3.64 to 10.84, therefore, we can reject the hypothesis automatical transmission car is same as automatical transmission car on MPG, in deed, the transmission type indeed can affect the MPG.

Residual plot

Finally, we will chech the regression model to see if its residual plot is close to normal distribution.

```
plot(fit1, which=2)
```




```
mean(fit1$residuals)
```

```
## [1] -6.591949e-17
```

which shows residual is close to normal distribution, and its mean close to zero.

A more complicated model:

Let 's study second model that add some more variables to the regression model. The factors are "cyl": Number of cylinders, and "hp": Gross horsepower.

```
fit2 <- lm(mpg ~ factor(am)+factor(cyl)+hp, data=mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + factor(cyl) + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.231 -1.535 -0.141  1.408  5.322
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.29590    1.42394  19.169  < 2e-16 ***
## factor(am)1   4.15786    1.25655   3.309  0.00266 **
## factor(cyl)6 -3.92458    1.53751  -2.553  0.01666 *
## factor(cyl)8 -3.53341    2.50279  -1.412  0.16943
## hp           -0.04424    0.01458  -3.035  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 27 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7989
## F-statistic: 31.79 on 4 and 27 DF,  p-value: 7.401e-10
```

From the summary,there is 4.15786 MPG increase on averge on manual transmission (AM=1) over automatic transmission (AM=0), which is different with previous model, where the coefficant is 7.245. Since there is more factors (correcated or uncorrected) impact the MPG model.

Correction between the predictors:

```
cor(mtcars$am, mtcars$cyl)
```

```
## [1] -0.522607
```

```
cor(mtcars$am, mtcars$hp)
```

```
## [1] -0.2432043
```

We can see that there is negative correction between the "AM" and "cyl", "hp" factors, therefore, the coefficent of "AM" is negative impacted by the 2 new factors.

Confidience level

Further check the confidence level:

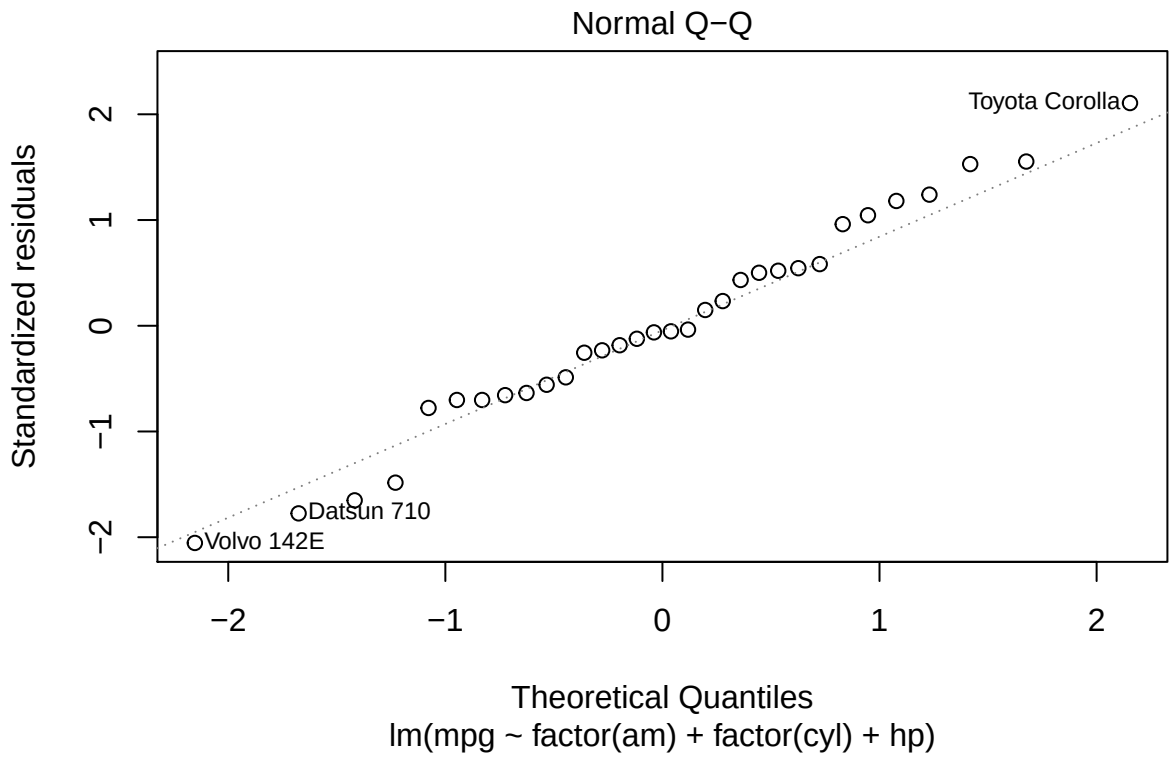
```
confint(fit2)
```

```
##                    2.5 %      97.5 %
## (Intercept)  24.37421028 30.21758829
## factor(am)1   1.57962875  6.73608418
## factor(cyl)6 -7.07929845 -0.76985854
## factor(cyl)8 -8.66871103  1.60188319
## hp           -0.07415067 -0.01433722
```

The 95% confidence interval of AM is still above zero, from 1.57 to 6.736, therefore, we can still have the conclusion that transmission type still have different impact on the MPG.

Residual plot and summary

```
plot(fit2, which=2)
```




```
mean(fit2$residuals)
```

```
## [1] 8.326673e-17
```

From the plot and brief summary, the residulal distribution is very close to normal distribution, which confirmed that our linear regression is a valid one.

Overall executive summary

From this data study of mtcars dataset, we can conclude that manual transmission better for MPG then automatic transmission. We use 2 different model toe support our conclusion, and also study the internal data correction can have some impact on multi variable linear regression's model coefficent.