Regression Models Course Project

Pingping

8/5/2020

Introduction

In this report, we explored the relationship between a set of variables and miles per gallon (MPG) (outcome). We are interested in solving these following two questions:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

Data Analysis

```
mtcars <- datasets::mtcars
dim(mtcars)
```

```
## [1] 32 11
```

This dataset has 32 observations and 11 variables. Among the variables, mpg is the outcome, and the others are 10 variables that may or may not influence the mpg.

Is an automatic or manual transmission better for MPG?

```
summary(mtcars[mtcars$am == 0, ]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
summary(mtcars[mtcars$am == 1, ]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

As we can see from the boxplot (in the appendix), automatic cars (am = 0) on average have a smaller mpg than manual cars (am = 1).

Quantify the MPG difference between automatic and manual transmissions

1. Simple linear regression model

First, We will do a simple linear regression model fit of the mtcars data using use mpg as the outcome and am as the variable.

```
lm_simple <- lm(mpg ~ factor(am), data=mtcars)
summary(lm_simple)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(lm_simple)$adj.r.squared
```

```
## [1] 0.3384589
```

From the result, We can see that the adjusted R squared value is only 0.338, which means that only 33.8% of the regression variance can be explained by this simple linear model. We may need to take other variables in the mtcars dataset into consideration in the modelling.

2. Multivariable regression model

```
lm_multi <- lm(mpg ~ ., data=mtcars)
summary(lm_multi)$coef[ ,1]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs          am        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
```

```
summary(lm_multi)$adj.r.squared
```

```
## [1] 0.8066423
```

With all the variables taken into account, we got a multivariable linear model, with an adjusted R-squareof 0.8066. This multivariable linear regression model fits better than the single linear model. However, not all the variables contribute to the outcome mpg evenly. Therefore, we'll use anova to tests whether the model terms are significant and try to find a better model to fit the data.

```
anova(lm_multi)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## cyl        1 817.71  817.71 116.4245 5.034e-10 ***
## disp       1  37.59   37.59   5.3526  0.030911 *
```

```
## hp          1   9.37    9.37   1.3342  0.261031
## drat        1  16.47   16.47   2.3446  0.140644
## wt          1  77.48   77.48  11.0309  0.003244 **
## qsec        1   3.95    3.95   0.5623  0.461656
## vs          1   0.13    0.13   0.0185  0.893173
## am          1  14.47   14.47   2.0608  0.165858
## gear        1   0.97    0.97   0.1384  0.713653
## carb        1   0.41    0.41   0.0579  0.812179
## Residuals 21 147.49    7.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results suggested that cyl, disp and wt showed significant influence (P value < 0.05) to the modelling. Hence, we'll choose these variables plus the am to fit the model.

```r
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
lm_final <- lm(mpg~cyl+wt+disp+am, mtcars)
summary(lm_final)$coef[ ,1]
```

```
##  (Intercept)          cyl           wt         disp     amManual
## 40.898313414 -1.784173258 -3.583425472  0.007403833  0.129065571
```

To examine any heteroskedacity between the fitted and residual values, and to check for any non-normality, we'll plot the residuals (in the appendix). As we can see from the residual fitted graph, the residuals are homosekdastic, and they have similar variance.

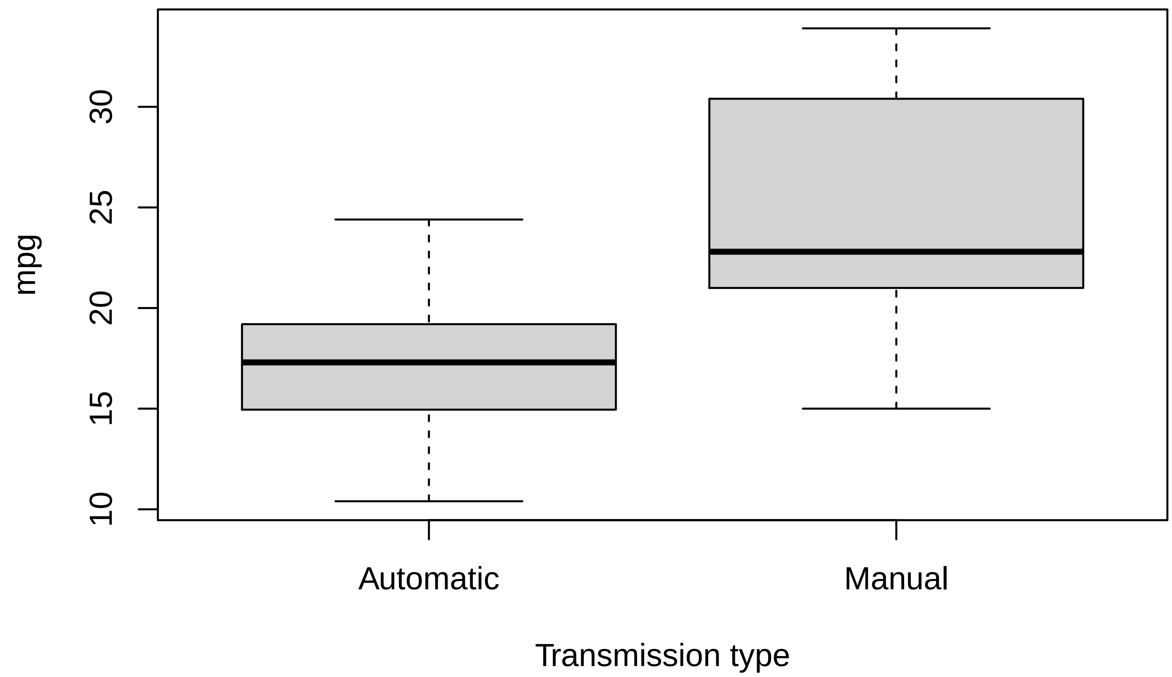
Conclusion

With all the other same setting, manual transmission cars on average have 0.129 miles per gallon more than automatic cars.

Appendix

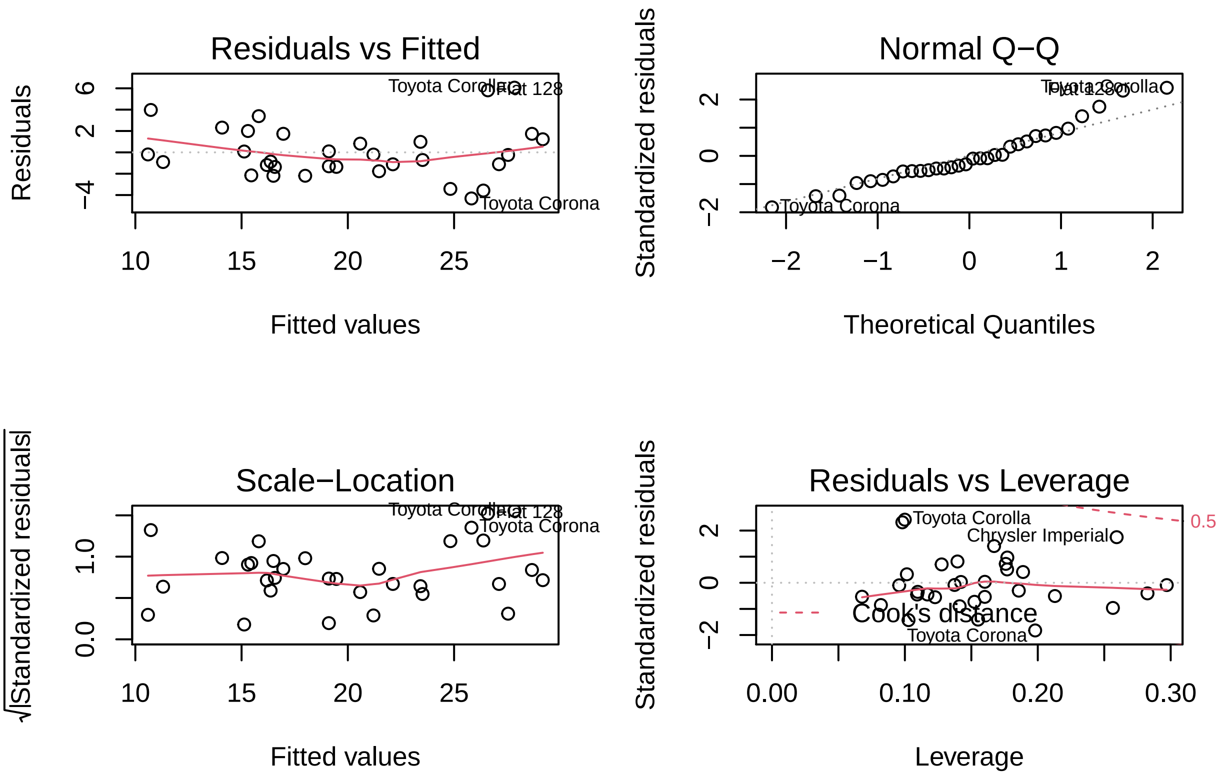
1. Boxplot of automaic and muanual car mpg

```r
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type")
```



2. Residual plots

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



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