Exploring the Relationship between the Mode of Transmission and the Mileage

Jeevithiesh Duggani

08/05/2020

Executive Summary

From the dataset `mtcars`, this project explores the relationship between the `mode of transmission` and `mileage in miles per gallon (MPG)`. The project aims to answer the following two questions.

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

Import libraries

Import the necessary libraries

```
    library(ggplot2)
```

Initialize the data

Import the dataset `mtcars`.

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

Exploratory Data Analysis

Consider the average `mpg` for each level of `am`.

```
    aggregate(mpg ~ am, mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

It can be seen from the above data frame that the `average mileage` is **better** for `manual transmission` than for `automatic transmission`.

Regression Analysis

> Hypothesis : The value of mpg for manual transmission is greater than that for automatic transmission

Find the `coefficients` for the `linear regression` with only `mpg` as outcome and `am` as predictor.

```
    fit <- lm(mpg ~ am, data = mtcars)
    coefficients(summary(fit))
```

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

Find the confidence intervals for `am`.

```
    confint(fit)["am",]
```

```
##    2.5 %   97.5 %
##  3.64151 10.84837
```

The entire `confidence interval` lies above `0` and the `p-value` (0.000285) is less than `0.05`.

> Therfore the hypothesis is not to be rejected.

The value of R squared is 0.3598, i.e. only 35.9799% of the variance of `mpg` is caused by `am`.

Perform Analysis of Variance over the data.

```
    analyseVariance <- aov(mpg ~ ., mtcars)
    summary(analyseVariance)
```

```
##             Df Sum Sq Mean Sq F value         Pr(>F)
## cyl          1  817.7   817.7 116.425 0.000000000503 ***
## disp         1   37.6    37.6   5.353        0.03091 *
## hp           1    9.4     9.4   1.334        0.26103
## drat         1   16.5    16.5   2.345        0.14064
## wt           1   77.5    77.5  11.031        0.00324 **
## qsec         1    3.9     3.9   0.562        0.46166
## vs           1    0.1     0.1   0.018        0.89317
## am           1   14.5    14.5   2.061        0.16586
## gear         1    1.0     1.0   0.138        0.71365
## carb         1    0.4     0.4   0.058        0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The terms with `p-value` less than `0.05` are `cyl`, `disp`, `wt`.

Find the `linear regression` with `mpg` as outcome and `cyl`, `wt`, `disp` and `am` as predictors.

```
    fitMult <- lm(mpg ~ cyl + wt + disp + am, data = mtcars)
    coefficients(summary(fitMult))
```

```
##                 Estimate Std. Error    t value             Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 0.000000000008677574
## cyl         -1.784173258 0.61819218 -2.8861142 0.007581533437721448
## wt          -3.583425472 1.18650433 -3.0201537 0.005468412488207290
## disp         0.007403833 0.01208067  0.6128661 0.545092996564770838
## am           0.129065571 1.32151163  0.0976651 0.922919644373324077
```

The R squared is 0.8327 which is sufficiently large.

> The p values of cyl and wt are below 0.05, suggesting that these are the confunding values between mpg and am

Identify the coefficients after removing `disp` as a predictor

```
    fitFin <- lm(mpg ~ cyl + wt + am, data = mtcars)
    coefficients(summary(fitFin))
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
## am           0.1764932  1.3044515  0.1353007 8.933421e-01
```

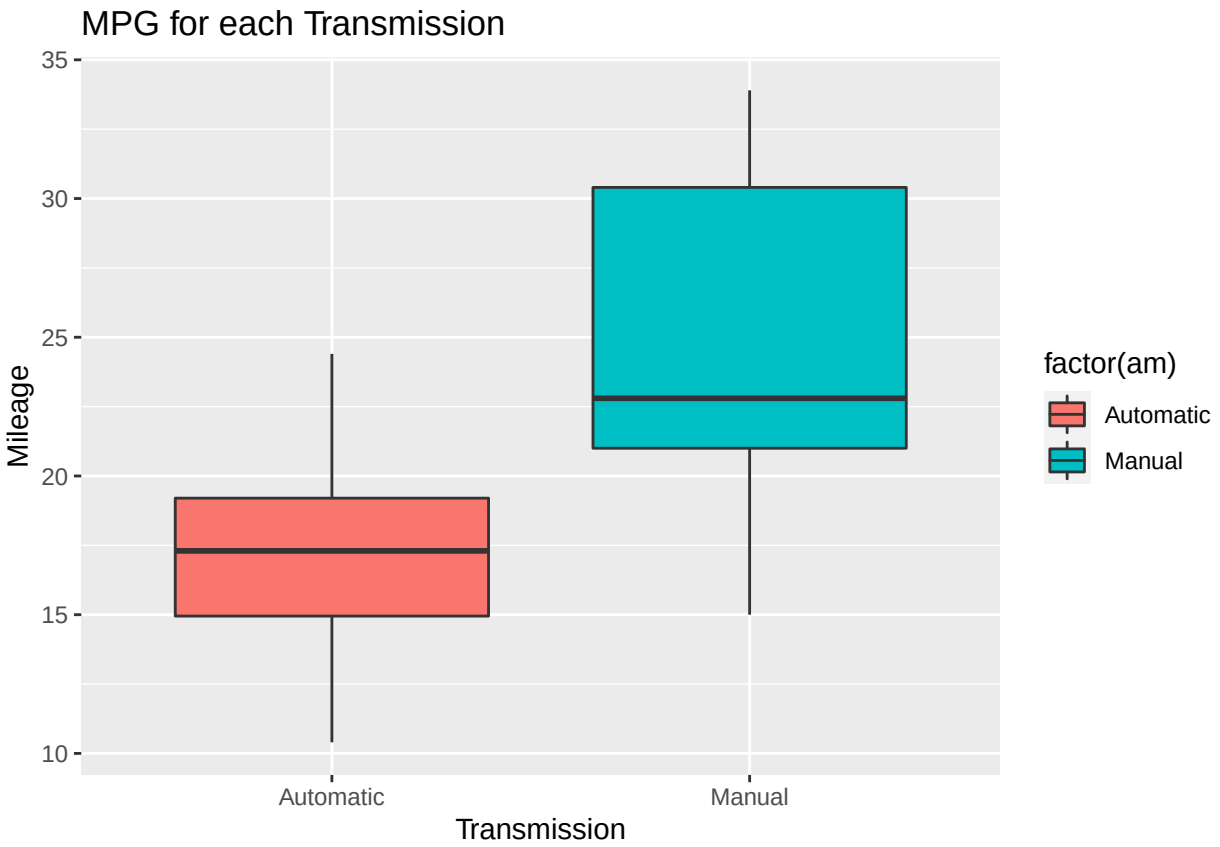
The `coefficient` for `am` is `0.1765`.

> The value of mpg increases by 0.1765 when the mode of transmission changes from automatic to manual.

Appendix

Plot the graph between `mpg` and `am`

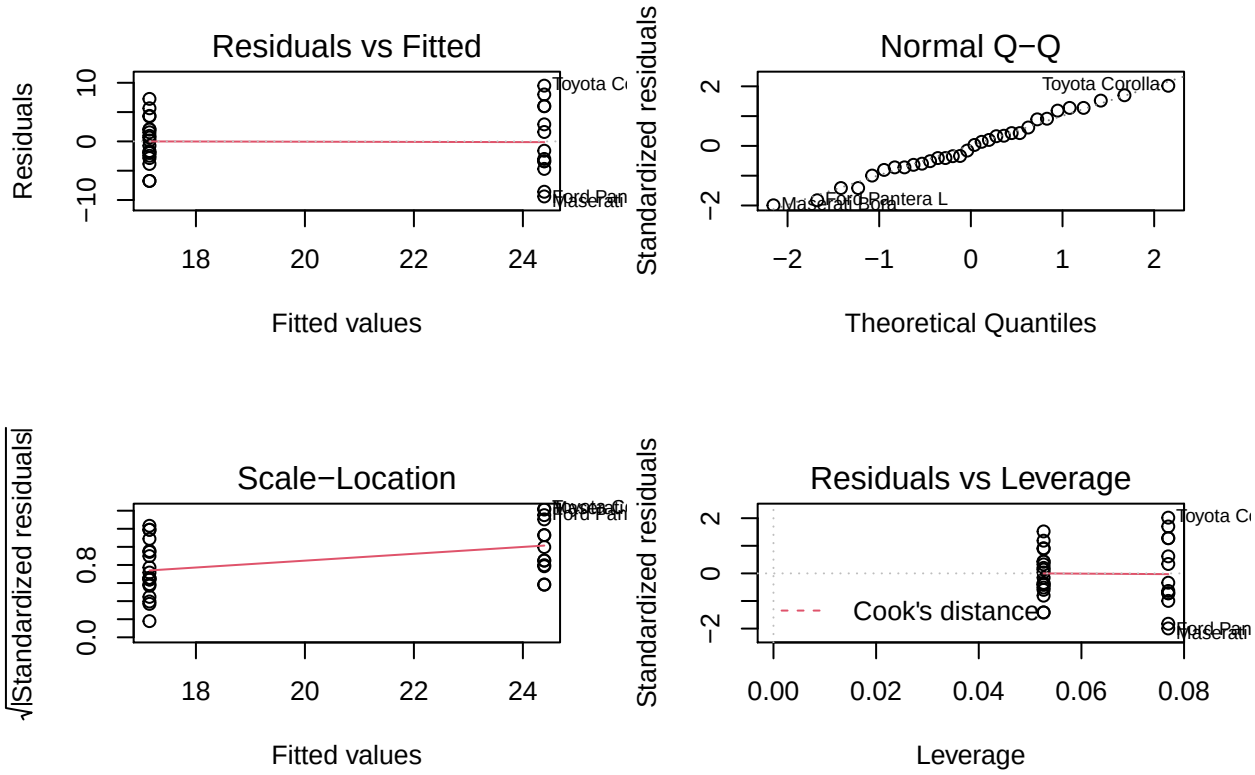
```
    mtcars$am <- as.factor(mtcars$am)
    levels(mtcars$am) <- c("Automatic", "Manual")
    graph <- ggplot(mtcars, aes(x = factor(am), y = mpg))
    graph <- graph + geom_boxplot(aes(fill = factor(am)))
    graph <- graph + labs(title = "MPG for each Transmission",
                          x = "Transmission", y = "Mileage")
    graph
```



From the above analysis, it can be initially observed that `manual transmission` is indeed **better** for `mileage`.

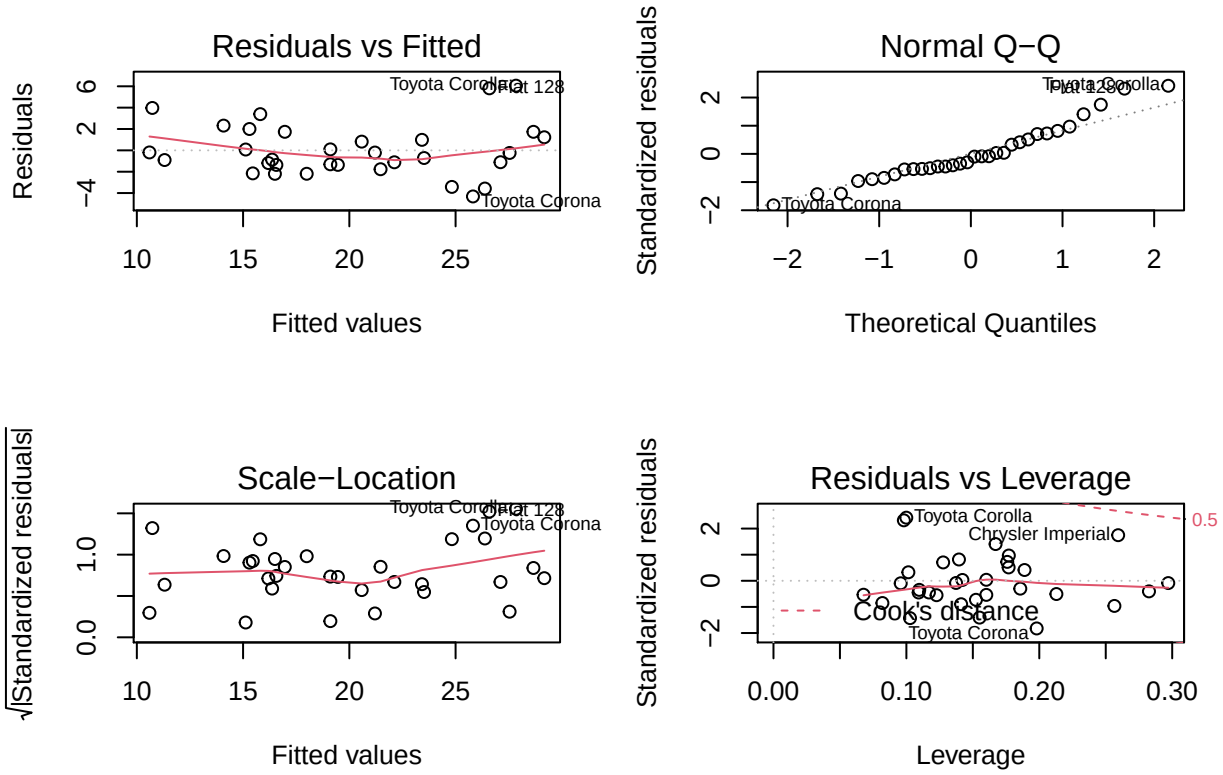
Plot for fit.

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



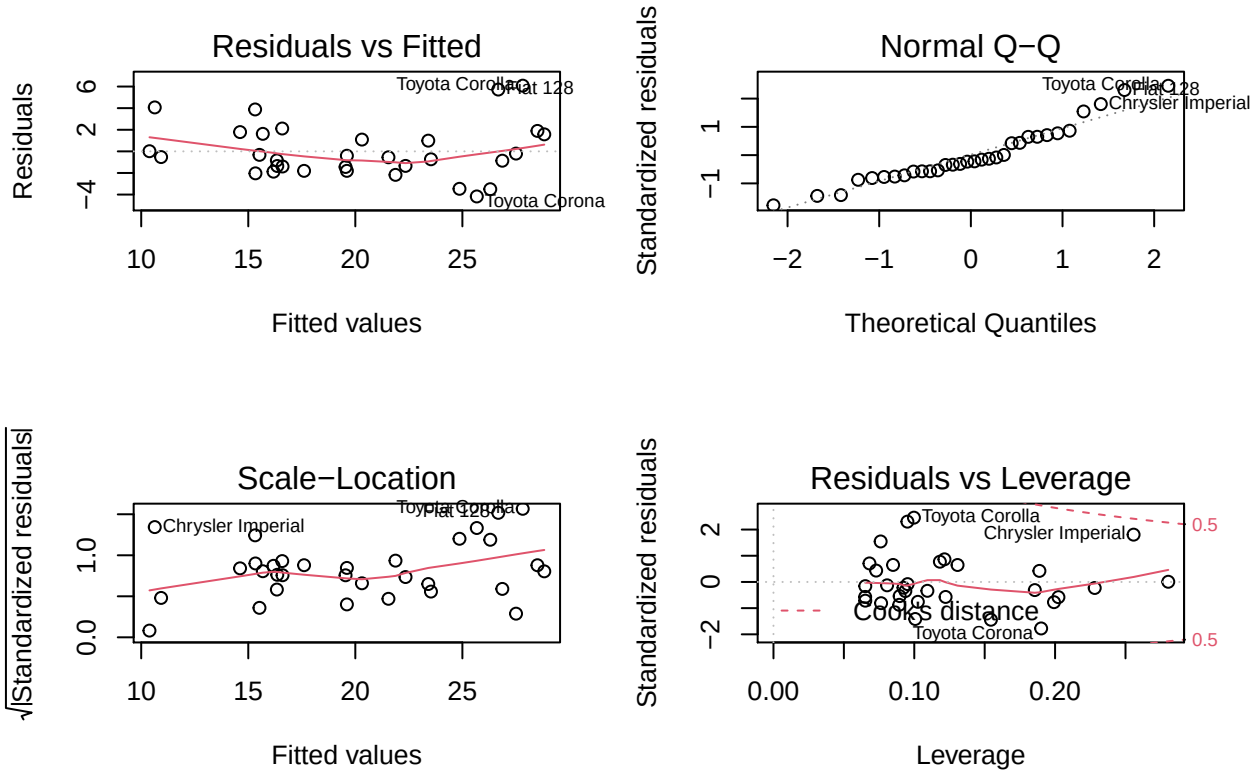
plot for fitMult.

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



plot for fitFin.

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



The `residuals vs fitted` graphs of `fitMult` and `fitFin` are very similar, indicating that the chosen confounding variables are the only confounding variables.

`**End**`