Regression Model Course Project

Siyang Ni

1/12/2021

Executive Summary

This is the course project for the Regression Model Class at John Hopkins University. This project examines the `mtcars` data set in R, and explores how miles per gallon (mpg) is affected by different variables. This report specifically addresses one question: Is automatic or manual transmission vehicles better for fuel economy. **The analysis indicates that compared to automatic transmission vehicles, manual transmission vehicles have a higher mpg value.** However, further model fitting indicates that **the difference of mpg between automatic and manual transmission vehicles can be explained by the number of cylinders, horse power and vehicle weight rather than the transmission type per se.**

Data Preperation and Wrangling

```
# Load the dataset and setting environment
library(tidyverse)
data(mtcars)
```

```
# Transform certain variables into factors
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am, labels=c("Automatic","Manual"))
```

Exploratory Data Analysis

From both the mean and median we can see that manual transmission vehicles have higher values of mpg. Fig 1 confirms what we see from the median and mean comparison. Fig 1 also shows that mpg values of automatic transmission cars roughly follow a normal distribution, with values concentrate at around 16-17, while mpg values of manual transmission cars display larger variance, with values concentrate at around 22-23.

```
# Split mpg by transmission type and show its mean and median
aggregate(mpg~am, data=mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

```
aggregate(mpg~am, data=mtcars, median)
```

```
##          am  mpg
## 1 Automatic 17.3
## 2    Manual 22.8
```

[Fig 1 about here]

To further investigate if there is a significant difference between the mpg value of automatic and manual transmission cars, a t-test is done. The result show that there is a significant difference (P<.01) between the mpg value of automatic and manual transmission cars. This sets the stage for further regression analysis.

```
# t-test
m_mpg <- mtcars$mpg[mtcars$am=='Manual']
a_mpg <- mtcars$mpg[mtcars$am=='Automatic']

t <- t.test(m_mpg, a_mpg)
t$p.value
```

```
## [1] 0.001373638
```

Model Fitting

Given the nature of the data, Classical linear model seems to be a good fit theoretically. However, given `mpg` in practice cannot take negative values, Poisson model might also be a competitive choice if we transform the `mpg` variable into all integers. Nevertheless, this type of transformation loses some information in the outcome variable. Poisson model is also not as easily interprtable as classical linear model. Following the parsimonious principle, we use the classical linear model.

The ANOVA test compared 11 models, starting from the crude model where we only have information of the mean difference. Then, each newer model adds a new variable to the old model sequentially. Eventually, the last model is the full model that contains all variables in the mtcars dataset. According to the model comparison result, We determined that the best-fit classical linear model should include transmission type (`am`), cylinder(`cyl`), horse power (`hp`), and vehicle weight (`wt`) as regressors.

```
# Linear Model
fitl <- lm(mpg ~ am + cyl + hp + wt + disp + drat + qsec + vs + gear + carb, data=mtcars)

# Test model Fit
anova(fitl, test="F")
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15 50.4745 3.597e-06 ***
## cyl        2 456.40  228.20 28.4297 7.890e-06 ***
```

```
## hp         1  67.30   67.30  8.3840   0.01110 *
## wt         1  46.17   46.17  5.7524   0.02992 *
## disp       1   0.62    0.62  0.0768   0.78541
## drat       1   0.31    0.31  0.0384   0.84726
## qsec       1   8.89    8.89  1.1081   0.30916
## vs         1   2.18    2.18  0.2719   0.60964
## gear       2   5.02    2.51  0.3128   0.73606
## carb       5  13.60    2.72  0.3388   0.88144
## Residuals 15 120.40    8.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we fit the best fit model to our data. The model fitting result indicates that if **controlling for cylinder(`cyl`), horse power (`hp`), and vehicle weight (`wt`), transmission type is not a significant predictor of vehicles' fuel economy (`mpg`) at P<.05 significant level.** In other words, the difference of fuel economy between automatic and manual transmission vehicles can be explained away by the number of cylinders, horse power and vehicle weight.

```
# Choose the best fit model
fitl_best <- lm(mpg ~ am + cyl + hp + wt, data=mtcars)
summary(fitl_best)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## amManual     1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Now we do a diagnostic check of our model. The diagnostic plots are in Fig 2. The residual plot shows no observable pattern. This means that our model fit does not presents detrimental errors. We do see few data points (as it is indicated by the Cook's Distance Plot in Fig 2), but they have acceptable leverage, thus shouldn't be able to exert large influence on our model fit.

[Fig 2 about here]

Conclusion and Discussion

This analysis aims at identifying which type of transmission gives better mpg. First, our exploratory analysis shows that there is a significant difference in mpg between automatic and manual transmissions vehicles. However, our regression analysis indicates that while controlling the number of cylinders, horse power and vehicle weight, transmission type is no longer a valid predictor for vehicle fuel economy.

Therefore, for ordinary consumers, buying a manual transmission car is still a good option for fuel saving. However, we must point out that this fuel saving should not be attributed to transmission type, but other factors that are potentially correlated to transmission type, such as number of cylinders, horse power and vehicle weight.

Appendix

Fig 1. Violin Plot of MPG by Transmission Type

```
g = ggplot(data=mtcars, aes(am, mpg, color=am))
g = g + geom_violin()
g = g + stat_summary(fun.data=mean_sdl, geom="pointrange", color="blue")
g = g + labs(title="Violin plot of MPG by transmission type",
             x = "Transmission Type", y = "Miles Per Gallon")
g
```

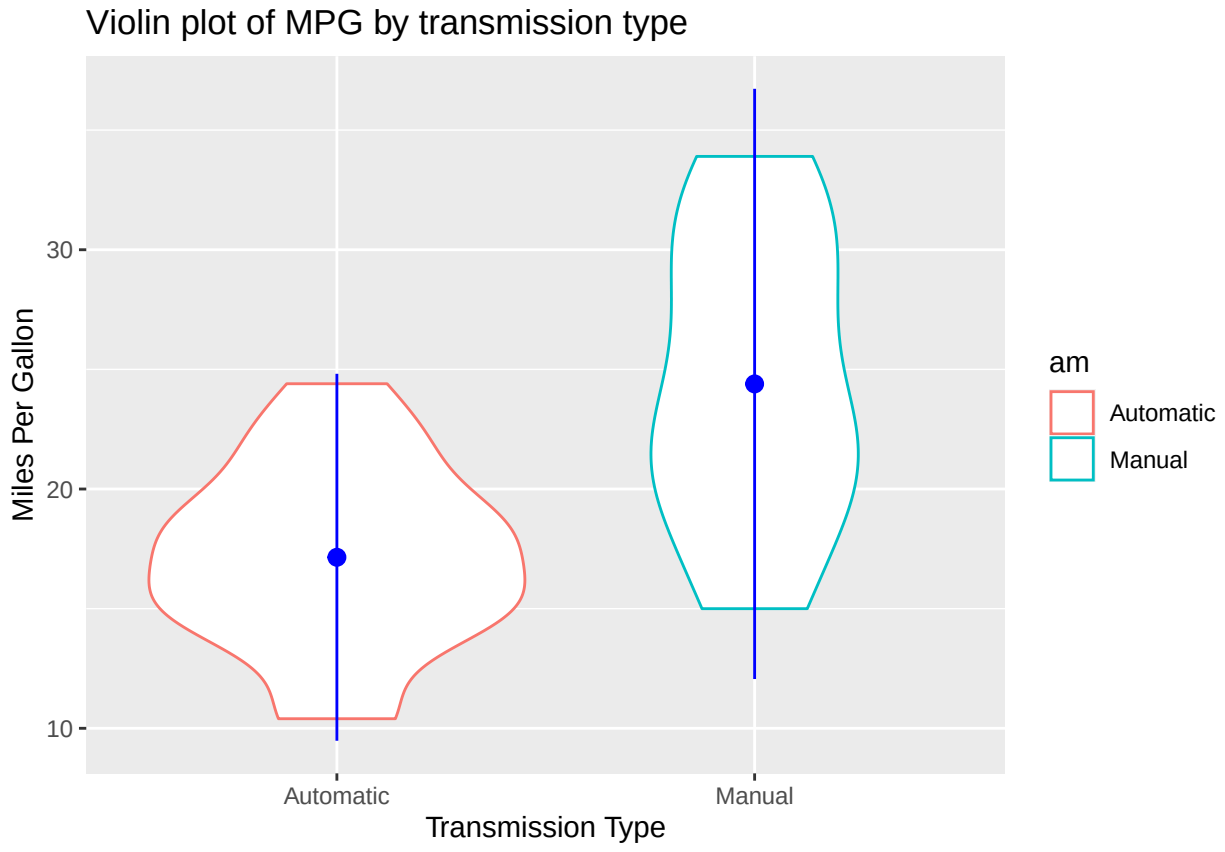

Fig 2. Model Diagnostic Plots

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