Regression Models - Course Project

Aiyu Li

02/14/2021

Executive Summary

This project used mtcars data to explore the relationship between a set of variables and miles per gallon (MPG). The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). In exploratary and regression anaysis, these key variables have been used: "am" - transmission (0 = automatic, 1 = manual), "wt" - Weight (1000 lbs) and "qsec" - 1/4 mile time.

From t-test, it shows that the manual transmission cars have 7.245 MPG more than automatic transmission cars.

From regression analysis: it shows by including weight of the car and acceleration, the predictability of the modeling increasing from 33.85 % to 83.36%, increasing fuel efficiency by 2.94 MPG, and the lighter manual car or heavier automatic car will increase MPG.

Exploratory Analysis

- Brief data summary from data exploratory analysis

    - 32 automobiles
    - Transmission: [13 manual, 19 automatic]
    - Miles/(US) gallon: [min 10.40, max 33.90, median 19.20]
    - Weight in 1000lbs: [max 5.424, min 1.513, median 3.325]
    - 1/4 mile time: [min 14.50, max 22.90, median 17.71]

- The MPG data distribution is approximatly normal from Histogram
- The medians are different between automatic and manual transmission
- The t.test results based mpg and transmission type

    - the null hypothesis is rejected since p-value is 0.00137 <.05 and .01
    - t statitc is -3.77
    - 95 percent confidence interval (-11.28, -3.21) does not include 0
    - mean of automatic is 17.15 and mean of manul is 24.39

Regresion analysis

Linear regression

- Intercept = 17.15 and Slope = 7.25, it means that if automatic transmission MPG starts at 17.15, the manual transmission estimated increase in MPG is 7.25.
- Miltiple R-squared is 35.98% with Adjusted R square = 33.85%
- Since P-value=0.0002, the difference is significant

Multivariable regression

- Fit a model using all the variables

    - Adjusted R-squared value is increased to 80.66%
    - model is not significant.

– necessary to find another fitted model to fit.

- Choose a model by AIC in a Stepwise Algorithm

  – variables "wt, qsec, am" selected for multivariable regaression analysis.
  – compared the linear regression, the Residual incresed from 33.85 % to 83.36%
  – manual transsmission have an average of 2.94 MPG more than autmomatic transsmission.
  – MPG of automatic transmission car estimately decreases while weight of the car decreases.

ANOVA Test

- p-value reject the null hypothesis
- weight and 1/4 mile time effect MPG
- holding weight and 1/4 mile time as constants, manual cars MPG is higher than that of automatic cars

Residuals

There are no large outliers found to affect the prediction of MPG.

Analysis conclusion

- A manual transmission provides better MPG than an automatic.
- From multivariable regression model, the fuel efficiency of manual cars is about 2.94 MPG higher than automatic cars.

APPENDIX: codes, tables and charts

for exploratory data analysis

```
## Load and exploratory data , hidden results
require(data.table); require(ggplot2); require(sqldf); require(MASS); data(mtcars); dim(mtcars)
sqldf('select sum(case when am=1 then 1 else 0 end) as manual ,
                    sum(case when am=0 then 1 else 0 end) as automatic
                    from mtcars')
summary(mtcars); mtcars$am <- as.factor(mtcars$am); levels(mtcars$am) <-c("Automatic","Manual")

## t.test
mpg.Automatic <- mtcars[mtcars$am == "Automatic",]$mpg
mpg.Manual <- mtcars[mtcars$am == "Manual",]$mpg
t.test(mpg.Automatic, mpg.Manual)
## Histgram of by Transmission type
ggplot( data = mtcars, aes(mpg, fill=am)) + geom_histogram() + facet_grid(.~am) +
        labs ( x = "MPG", y = "Frequency", title = "Histogram: MPG Distribution by Transmission Type") +
        theme(plot.title = element_text(color="grey25", face="bold", hjust=0.5))
```
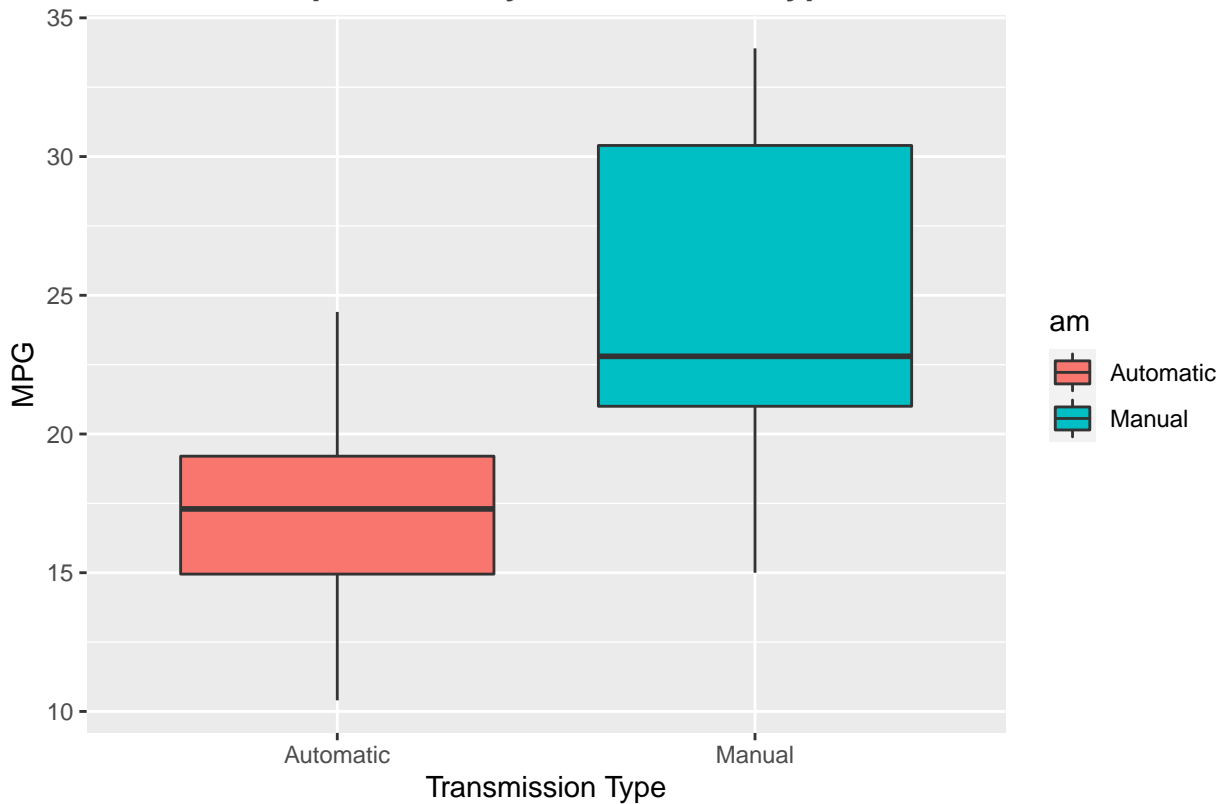
Histogram: MPG Distribution by Transmission Type
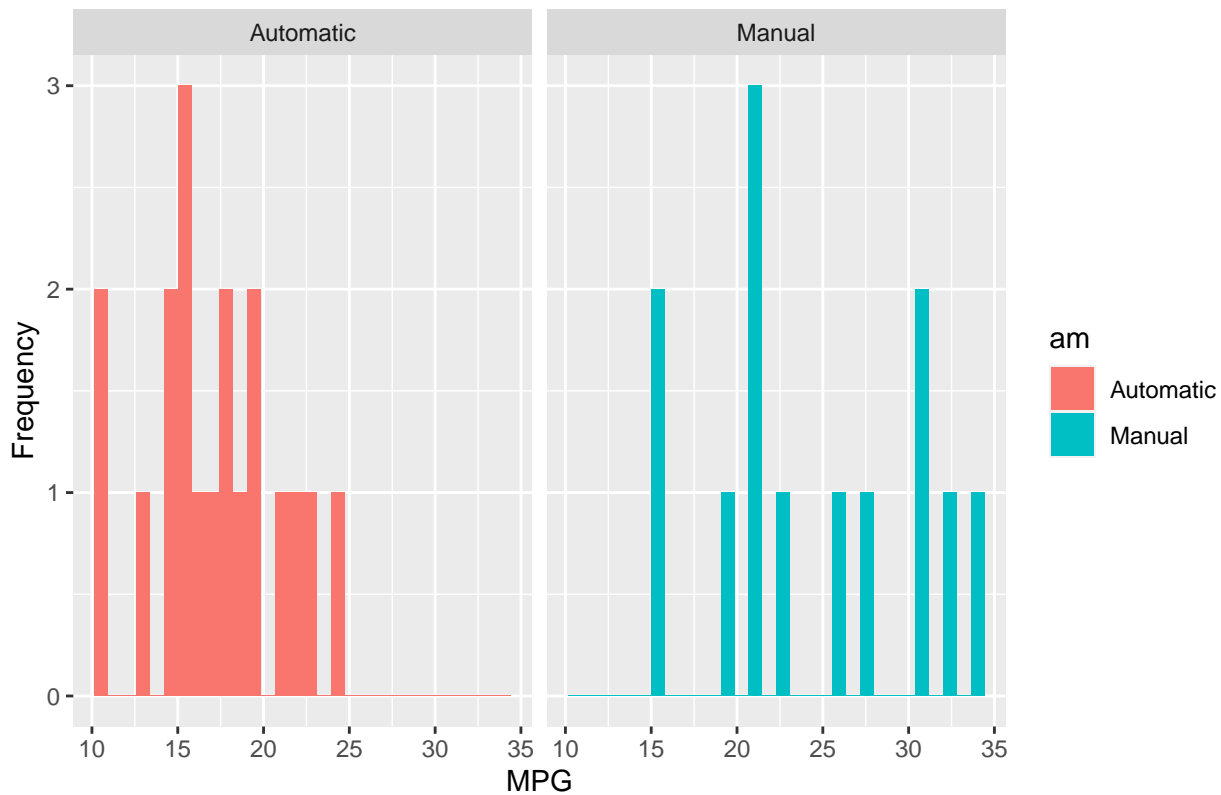

```
## Box Plot by Transmission Type
ggplot(aes(x = am, y = mpg), data = mtcars)  +  geom_boxplot(aes(fill = am)) +
        labs(x = "Transmission Type", y = "MPG", title = "Boxplot: MPG by Transmission Type") +
        theme(plot.title = element_text(color="grey25", face="bold",hjust=0.5))
```

for linear regression

```r
## Linear regression
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)
```

```
##
## 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
```

for multivariable regression

```r
# fit with all variables
fit1 <- lm(mpg ~ . , data = mtcars)
summary(fit1) # hidden results


# Choose a model by AIC in a Stepwise Algorithm
fit2 <- stepAIC(fit1, direction="both", trace=FALSE)
summary(fit2)
```

for ANOVA

```r
# Anova test
anova(fit, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

for residuals

```
#Residual Plots
par(mfrow = c(2,2))
plot(fit2)
```