Analysis on the relationship between a set of variables and fuel consumption (miles per gallon) in the automobile industry

Thawatchai Phakwithoonchai - 1/15/2020

Executive Summary

The objective of this assignment is to study and perform the analysis based on `mtcar` database in order to determine the responses for the following inquiries:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

The analysis result indicated that **" Automatic transmission caused more fuel consumption (MPG) than manual transmission about 7.2 miles per gallon based on the simple linear model (mpg ~ am). "** However, the simple linear model showed the poor model fit. Multivariable model fit would improve the model quality by adding number of cylinders (cyl), gross horsepower (hp), and weight (wt) into the model.

Load the data and libraries

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). Dataset consists of 32 observations with 11 variables.

Perform the exploratory data analysis

The data correlation can be determined by plotting in order to visualized the effects of transmission type related to the fuel consumption. The plotting result is showed in the *figure01*. The plotting results apparently indicated that manual transmission caused more the average fuel consumption than automatic transmission.

Establish the key assumption

1. All sample observations is independent and identically distributed (i.i.d.).
2. Normal distribution can be verified by:
   i) Histogram plot based on transmission type. The plotting results as shown in *figure02* indicated the mtcars dataset tended to follow a normal distribution.

ii) Shapiro-Wilk's method. It is based on the correlation between the data and the corresponding normal scores. The result indicated that p-value > 0.05; therefore, the distribution of the data are not significantly different from normal distribution. (Normality was valided)

3. Variances of fuel consumption are different in terms of transmission type.

Statistalcal inference

Student t-test was performed to verified the hypothesis whether if any significant difference in the fuel consumption between automatic and manual transmission in the mtcars dataset or not.

The result indicates the p-value is 0.001374, which < 0.05. Therefore, it could be rejected the null hypothesis or there was an evidence to suggest that **the fuel consumption from automatic transimission was significantly less than from the manual transmission.**

Simple linear regression

Because the fuel consumption was pertained to transmission type; therefore, the linear model could be fit and determined the inference results.

However, after the residual analysis was plotted as shown in *figure03*, it was found that the simple linear regression indicated the poor model fit; therefore, model adjustment by multivariable regression was required to improve the model accuracy.

Multivariable regression

Data correlation was explored by plotting in order to visualized the correlation of variables. The result indicated that cyl, disp, hp, drat, wt, and vs, may have strong correlation to the fuel consumption as shown in the *figure04*.

The nested model testing was performed for model selection. The result indicated the model with necessary variable were fit1, fit3 and fit5. Let's eliminate the insignificant variables and test the new model again in order to get a parsimonious explanatory model.

The result indicated that re-fit model (fit5a) should be selected. Furthermore, residual analysis was required by plotting to examine any heteroskedacity between the fitted and residual values; as well as to check for any non-normality as shown in *figure05*. The "Residuals vs Fitted" plot showed that the residuals were homoscedastic, while "Normal Q-Q" plot showed the normally distribution with the exception of a few outliers.

Supporting Appendix

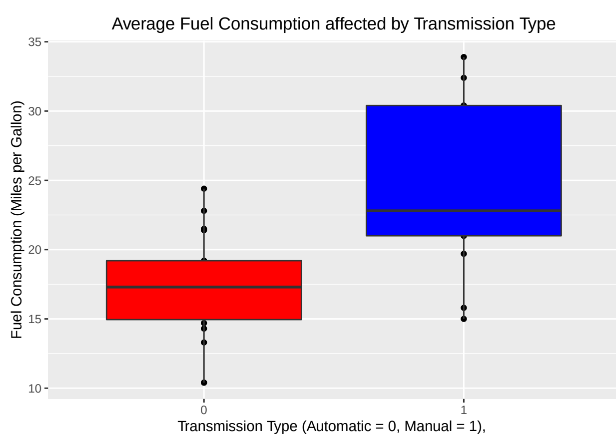
```r
# Load the data and libraries
data("mtcars")
library(ggplot2)
```

```r
# Prepare and transform the data
str(mtcars)
```

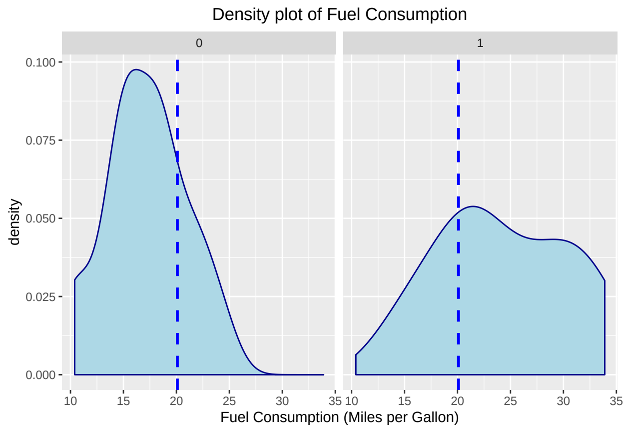
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76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

```r
mtcars$am.label <- factor(mtcars$am, labels = c("Automatic",
    "Manual"))
```

```r
# Figure01 - Explore the data
qplot(factor(am), mpg, data = mtcars, main = "Average Fuel Consumption affected by Transmission Type",
    xlab = "Transmission Type (Automatic = 0, Manual = 1), ",
    ylab = "Fuel Consumption (Miles per Gallon)") + geom_boxplot(aes(group = am),
    fill = c("red", "blue")) + theme(plot.title = element_text(hjust = 0.5))
```




```r
# Figure02 - Explore the data
norm.plot <- ggplot(data = mtcars, aes(x = mpg))
norm.plot + geom_density(color = "darkblue", fill = "lightblue") +
    facet_grid(. ~ am) + geom_vline(aes(xintercept = mean(mpg)),
    color = "blue", linetype = "dashed", size = 1) + labs(title = "Density plot of Fuel Consumption",
    x = "Fuel Consumption (Miles per Gallon)") + theme(plot.title = element_text(hjust = 0.5))
```




```r
# Normality test for data
shapiro.test(mtcars$mpg[mtcars$am == 0])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg[mtcars$am == 0]
## W = 0.97677, p-value = 0.8987
```

```r
shapiro.test(mtcars$mpg[mtcars$am == 1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg[mtcars$am == 1]
## W = 0.9458, p-value = 0.5363
```

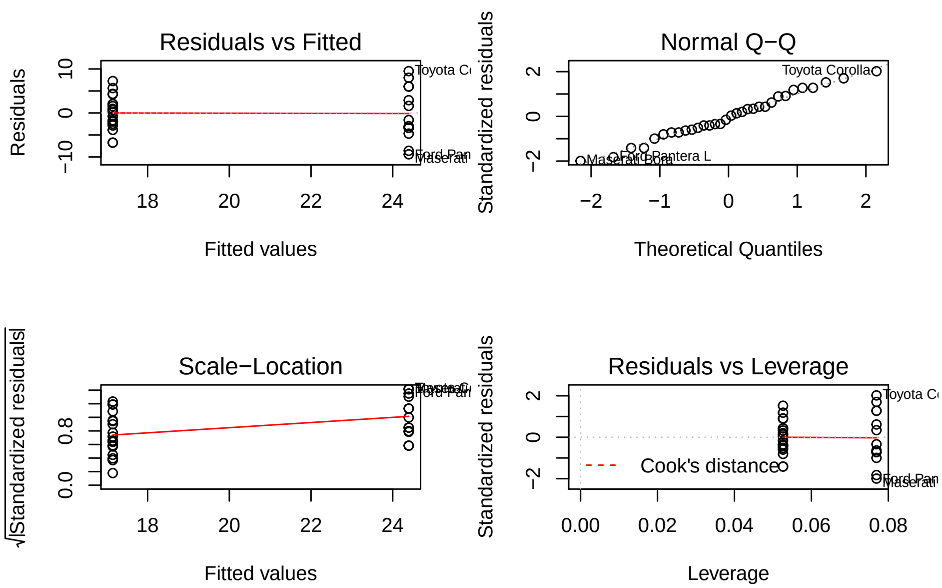
```r
# Statistical inference for transmission type
t.test(mpg ~ am.label, data = mtcars, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am.label
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

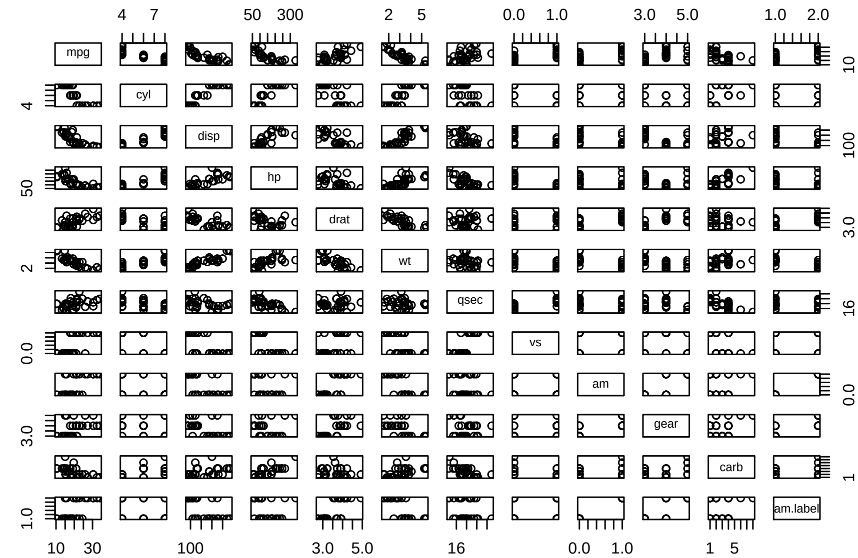
```r
# Fit the simple linear model
fit <- lm(mpg ~ factor(am), data = mtcars)
summary(fit)
```

```
##
## 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
```

```
# Figure03 - Residual analysis of simple linear regression
par(mfrow = c(2, 2))
plot(fit)
```




```
# Figure04 - Pair plot of dataset
pairs(mpg ~ ., data = mtcars)
```




```r
# Determine the correlation of variable
mtcars$am.label <- NULL
cor(mtcars)[1, ]
```

```
##        mpg        cyl       disp         hp       drat         wt       qsec
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594  0.4186840
##         vs         am       gear       carb
##  0.6640389  0.5998324  0.4802848 -0.5509251
```

```r
# Nested model testing
fit0 <- lm(mpg ~ am, data = mtcars)
fit1 <- update(fit0, mpg ~ am + cyl, data = mtcars)
fit2 <- update(fit0, mpg ~ am + cyl + disp, data = mtcars)
fit3 <- update(fit0, mpg ~ am + cyl + disp + hp, data = mtcars)
fit4 <- update(fit0, mpg ~ am + cyl + disp + hp + drat, data = mtcars)
fit5 <- update(fit0, mpg ~ am + cyl + disp + hp + drat + wt,
    data = mtcars)
fit6 <- update(fit0, mpg ~ am + cyl + disp + hp + drat + wt +
    vs, data = mtcars)
anova(fit0, fit1, fit2, fit3, fit4, fit5, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + drat
## Model 6: mpg ~ am + cyl + disp + hp + drat + wt
## Model 7: mpg ~ am + cyl + disp + hp + drat + wt + vs
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
```

```
## 1     30 720.90
## 2     29 271.36  1    449.53 68.0021 1.837e-08 ***
## 3     28 252.08  1     19.28  2.9167  0.100574
## 4     27 216.37  1     35.71  5.4025  0.028894 *
## 5     26 214.50  1      1.87  0.2829  0.599667
## 6     25 162.43  1     52.06  7.8757  0.009783 **
## 7     24 158.65  1      3.78  0.5717  0.456945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Nested re-model testing
fit3a <- update(fit0, mpg ~ am + cyl + hp, data = mtcars)
fit5a <- update(fit0, mpg ~ am + cyl + hp + wt, data = mtcars)
anova(fit0, fit1, fit3a, fit5a)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + hp
## Model 4: mpg ~ am + cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 71.3976 4.619e-09 ***
## 3     28 220.55  1     50.81  8.0698  0.008458 **
## 4     27 170.00  1     50.56  8.0295  0.008603 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Figure05 - Residual analysis of multivariable linear
# regression
par(mfrow = c(2, 2))
plot(fit5a)
```