- **Regression Models - Coursera Course Project**
- **Executive Summary**

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions: 1. Is an automatic or manual transmission better for MPG;2. Quantify the MPG difference between automatic and manual transmissions.

- This project will answer these questions using simple linear model & multivariate linear model.
- **Exploratory Data Analysis**
- Boxplot to show the difference between types of transmission and mpg (0 = automatic, 1 = manual). Manual transmission seems to have better mpg. See Appendix, figure 1.
- The manual transmission cars higher mpg mean by 7.245 mpg's.

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

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

- T-test to confirm the previous evidence. Since the p-value is 0.001374, the null hypothesis is rejected and it can be said that manual transmission cars have better mpg than automatic transmission cars.

```
automatic_t <- mtcars[mtcars$am == 0,]
manual_t <- mtcars[mtcars$am == 1,]
t.test(automatic_t$mpg, manual_t$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic_t$mpg and manual_t$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

- **Simple Linear Regression. Variables: mpg, am.**

```
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 ***
## am             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
```

- Manual transmission cars have 7.245 mpg's more than automatic transmission but this simple linear regression model only explains 35.9% of the variance (R-squared).
- **Multivariate Regression**
- Correlation between the other variables and mpg.

```
data(mtcars)
sort(cor(mtcars)[1,])
```

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

- There is a strong correlation between mpg, wt, cyl and hp, so we could try to add these variables to the previous simple regression and compare the models using ANOVA.

```
fit0 <- lm(mpg~am, data = mtcars)
fit_m1 <- lm(mpg~am+wt, data = mtcars)
fit_m2 <- lm(mpg~am+wt+cyl, data = mtcars)
fit_m3 <- lm(mpg~am+wt+cyl+hp, data = mtcars)
anova(fit0, fit_m1, fit_m2, fit_m3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.2925  5.39e-09 ***
## 3     28 191.05  1     87.27 13.8611 0.0009165 ***
## 4     27 170.00  1     21.05  3.3432 0.0785534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to these p-values, the wt variable has more statistical relevance to be added to the multivariate regression along with am variable. Let's summarize this model to confirm it.

- **Summary of the multivariate regression. Variables: mpg, am, wt.**

```
fit_m1 <- lm(mpg ~ am + wt, data = mtcars)
summary(fit_m1)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## am          -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

- This multivariate regression models explains 75.2% of the variance (R-squared) against only the 35.9% of the simple linear model.
- The residuals of the multivariate regression seem to be random and normally distributed (see Appendix, figure 2)

- **Conclusions**

Using a simple linear model we were able to answer the two questions for the Motor Trend Magazine: 1. Manual transmission has better MPG; 2. Manual transmission cars have 7.245 mpg's more than automatic transmission.

However this simple linear model only explains the 35% of the variance and it was necessary to add the weight variable to reach a 75%.

- **Appendix**

Figure 1. Boxplot "MPG by Transmission Type"

```
boxplot(mpg~am, data = mtcars,xlab = "Transmission Type",ylab = "Miles per Gallon",main = "MPG by Transmission Type")
```

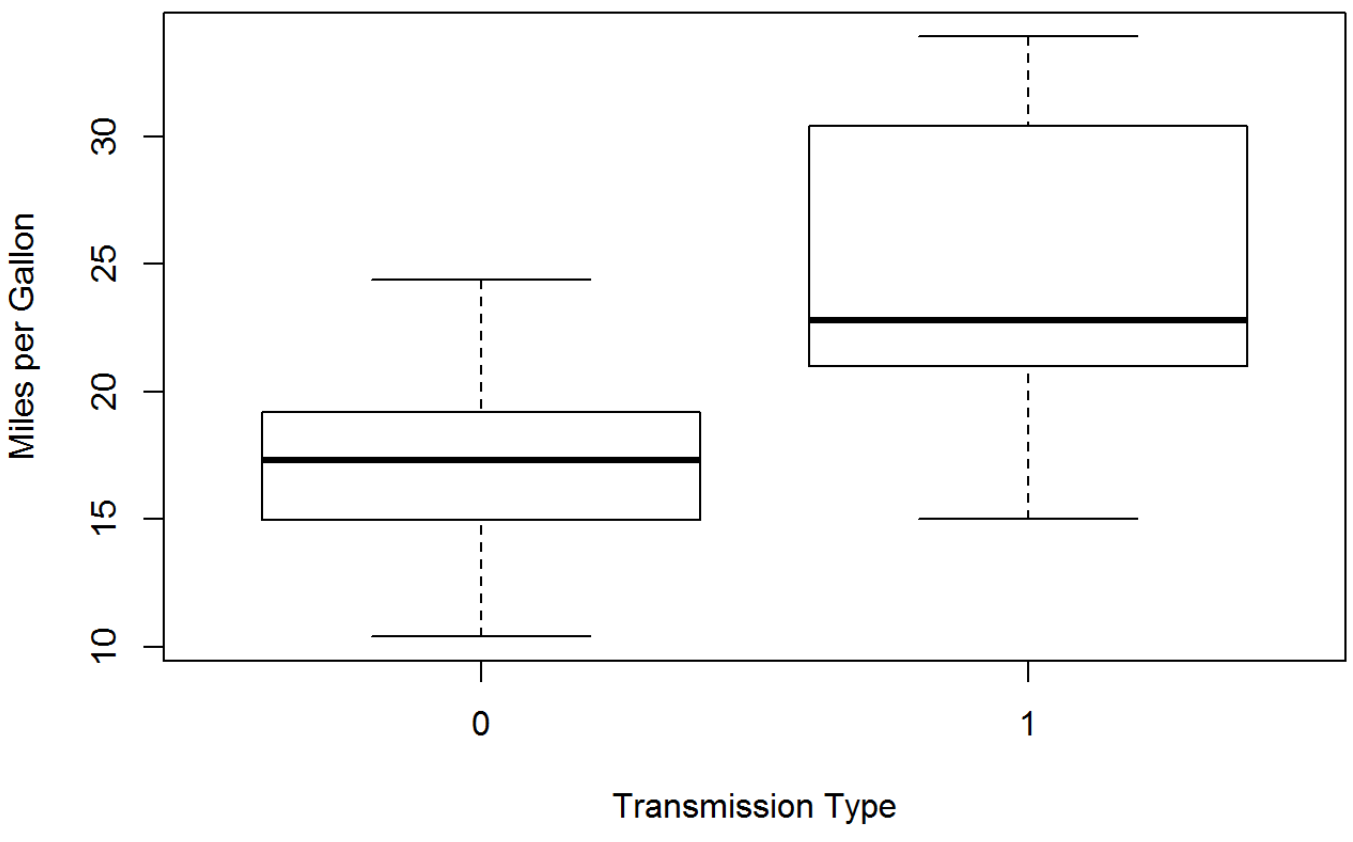

Figure 2. Residual Plot for the multivariate model.

```
par(mfrow = c(2,2))
plot(fit_m1)
```