Regression Models Course Project

Executive Summary

Here we used the `mtcars` dataset which contains the set of variables with related to miles per gallon (MPG). They are particularly interested in the following two questions:

- "Is automatic or manual transmission better for MPG"
- "Quantify this difference"

Data Analysis

I analysis the dataset in dffrent statistical method.

Load and Test Data

```
library(datasets)
data(mtcars)
mpgData <- with(mtcars, data.frame(mpg, am))
mpgData$am <- factor(mpgData$am, labels = c("Automatic", "Manual"))
manual    <- as.list(subset(mtcars, am == 1, select = mpg))[[1]]
automatic <- as.list(subset(mtcars, am == 0, select = mpg))[[1]]
```

Basic Analysis

```
summary(mpgData[mpgData$am == "Automatic",])
```

```
##       mpg               am
##  Min.   :10.4   Automatic:19
##  1st Qu.:14.9   Manual   : 0
##  Median :17.3
##  Mean   :17.1
##  3rd Qu.:19.2
##  Max.   :24.4
```

```
summary(mpgData[mpgData$am == "Manual",])
```

```
##       mpg               am
##  Min.   :15.0   Automatic: 0
##  1st Qu.:21.0   Manual   :13
##  Median :22.8
##  Mean   :24.4
##  3rd Qu.:30.4
##  Max.   :33.9
```

The initial comparison is simply the summary statistics between automatic and manual.The "Manual" transmission cars on average get 7.3(24.4-17.1) MORE miles to the gallon than automatic cars do.

Basic Linear Model

```
fit <- lm(mtcars$mpg ~ as.integer(am), data=mpgData)
summary(fit)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ as.integer(am), data = mpgData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.90       2.63    3.77  0.00072 ***
## as.integer(am)     7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

From summary `Automatic` transimission cars as the baseline, so the Intercept of 17.14.The coefficient for the `Manual` transmission is interepreted as 'how many more miles per gallon on average do you get by switching from an Automatic to a Manual' which is 7.24.The R-Squared is 0.338, which is quite low. That means this model does not fit the data terribly well.

Improved Linear Model

Now i try to improve the linear model by using `ANOVA` The Analysis Of Variance, popularly known as the ANOVA, can be used in cases where there are more than two groups.

```
fit2 <- update(fit, mtcars$mpg ~mtcars$am + mtcars$wt)
fit3 <- update(fit, mtcars$mpg ~mtcars$am + mtcars$wt + mtcars$qsec)
fit4 <- update(fit, mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec + mtcars$cyl)
anova(fit, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mtcars$mpg ~ as.integer(am)
## Model 2: mtcars$mpg ~ mtcars$am + mtcars$wt
## Model 3: mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec
## Model 4: mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec + mtcars$cyl
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
## 2     29 278  1       443 71.22 4.7e-09 ***
## 3     28 169  1       109 17.55 0.00027 ***
## 4     27 168  1         2  0.24 0.62706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is clear that the fit3 model, mpg being predicted by am, wt, and qsec gives us the strongest evidence to reject the null hypothesis.

```
summary(fit3)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec,
##     data = mpgData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## mtcars$am      2.936      1.411    2.08  0.04672 *
## mtcars$wt     -3.917      0.711   -5.51    7e-06 ***
## mtcars$qsec    1.226      0.289    4.25  0.00022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

This is much better. Now we have an adjusted R squared of 0.834,meaning this model explains 88.36% of the variance making for a much better fit.

```
bestfit <- update(fit, mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec + mtcars$am*mtcars$wt)
anova(fit3, bestfit)
```

```
## Analysis of Variance Table
##
## Model 1: mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec
## Model 2: mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec + mtcars$am:mtcars$wt
##   Res.Df RSS Df Sum of Sq  F Pr(>F)
## 1     28 169
## 2     27 117  1        52 12 0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After trial and error (not detailed here for sake of space) I have discovered that the interaction am*wt should be included to create my bestfit model.

```
summary(bestfit)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$qsec +
##     mtcars$am:mtcars$wt, data = mpgData)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.508 -1.380 -0.559  1.063  4.368
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.723      5.899    1.65  0.11089
## mtcars$am             14.079      3.435    4.10  0.00034 ***
## mtcars$wt             -2.937      0.666   -4.41  0.00015 ***
## mtcars$qsec            1.017      0.252    4.04  0.00040 ***
## mtcars$am:mtcars$wt   -4.141      1.197   -3.46  0.00181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.08 on 27 degrees of freedom
## Multiple R-squared:  0.896,  Adjusted R-squared:  0.88
## F-statistic: 58.1 on 4 and 27 DF,  p-value: 7.17e-13
```

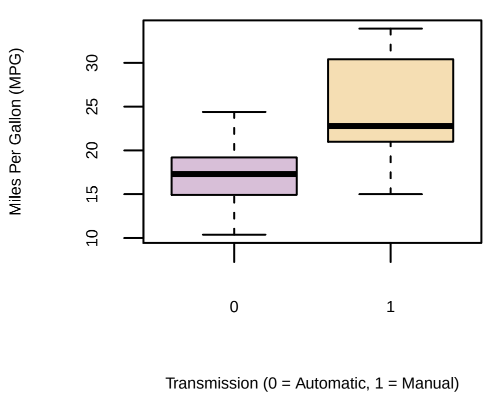
Now we can see we have an adjusted R squared of 0.8804 and small p-values for the coefficients, so this model explains 88.04% of the variance in our data providing a good fit with statistically significant coefficients.

Results

So from the `Linear model` and `Basic analysis` gave us manual transmission is better than automatic for MPG, which increased by 7.2449,and from `Improved Linear Model` difference between an automatic and manual car, a manual transmission car gets 14.079 more miles to the gallon than an Automatic - 4.141 * the weight. - Now answer to Q1 can say manual transmission is better than automatic - Now answer to Q2 manual transmission is better than automatic for MPG, which increased by 7.2449.

APPENDIX

Boxplot `am` vs `mpg`



Residuals

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