Regression Models Course Project -

Effects of automatic and manual transmission on MPG

Executive Summary

In this report, we examined the mtcars data set and explored the relationship between a set of variables and miles per gallon (MPG) (outcome). In particaularly, we are interested in these two questions:
- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

From our analysis, we conclude that Manual transmission is better for MPG than Automatic transmission. After adjusting the confounders (Number of cylinders, Gross horsepower and vehicle weight), Manual transmissoin has an MPG 1.809 higher than Automatic transmission.

Data Processing

We first load mtcars data set and convert several variables into factors

```
data(mtcars)
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")) # 0=Automatic, 1=Manual
```

Exploratory Analysis

To help us understand the effect of transmission type on MPG, we should look at the mean of MPG for each transmission type and box plot (Appendix: Plot 1)

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

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

Inference

From the boxplot we can see that the mean MPG of manual cars is higher than automatic cars by about 7 and we want to use t test to check if there is significant MPG difference between auto and manual.

```
ttest <- t.test(mpg ~ am, data=mtcars)
ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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
```

The p value of 0.0013736 indicates there is significant difference of MPG between automatica and manual cars. Now to quantify this, we will use regression analysis.

Regression Analysis

Now let's look at the linear relationship between MPG and transmission type.

```
basemodel <- lm(mpg ~ am, data=mtcars)
summary(basemodel)
```

```
##
## 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
```

The results show that the average MPG for automatic cars is 17.147 and MPG of manual cars is 7.245 higher. The R square is 0.3598 so this model only explains about 36% of the variance. We need to build a multivariate linear model.

We perform stepwise model selection in order to select significant predictors for the final, best model.The step function will perform this selection by calling lm repeatedly to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods using AIC algorithm. This ensures that we have included useful variables while omitting ones that do not contribute significantly to predicting mpg.

```
initialmodel <- lm(mpg ~ ., data = mtcars)
bestmodel <- step(initialmodel, direction = "both", trace=0)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, 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 ***
## 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 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## 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
```

The best model obtained from the above computations shows that variables, cyl, wt and hp as confounders and am as the independent variable

The adjusted R-squared value of 0.84 indicates more than 84% of the variability is explained by the above model.

Now let's compare the base model with only am as only predictor variable and the best model with other confounder variables.

```
anova(basemodel, bestmodel)
```

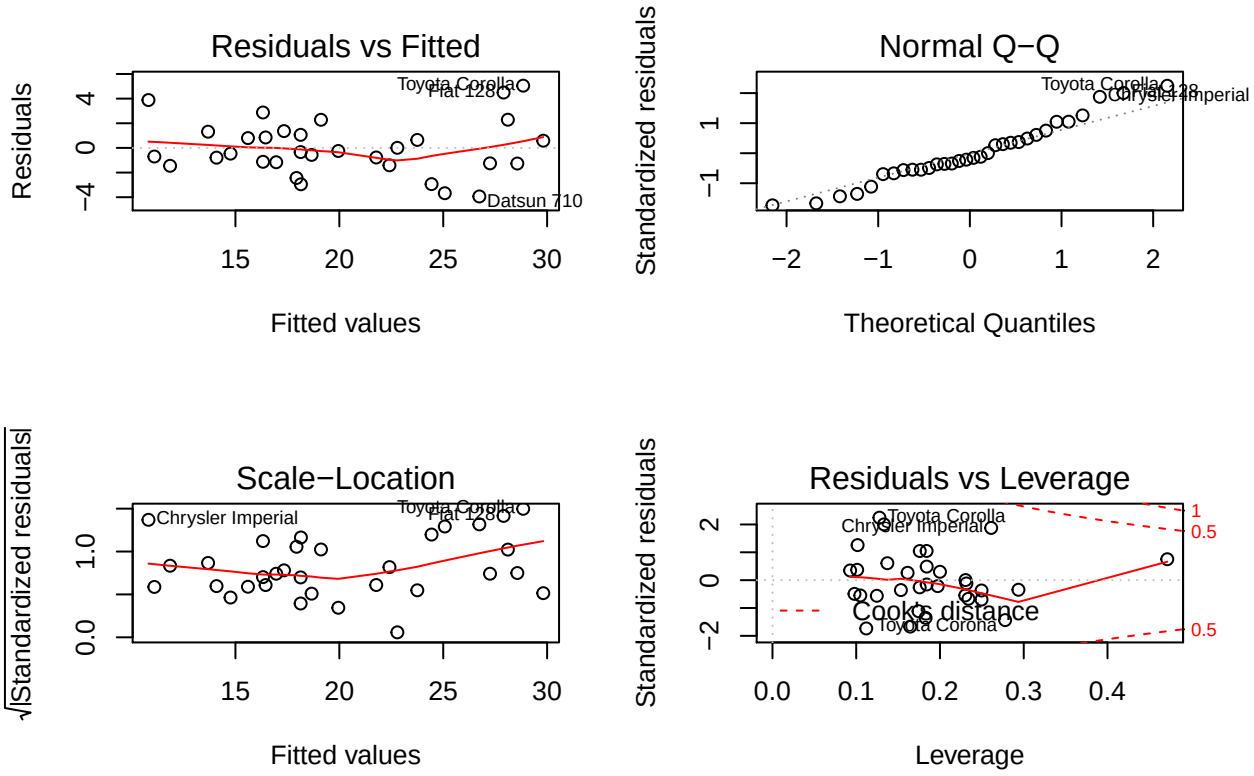
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The pvalue is highly significant, so we can reject the null hypothesis that cyl, wt and hp don't contribute to the model.

Residuals and Diagnostics

In this section, we examine the residual plots of our regression model along with computation of regression diagnostics for our liner model. This will help us in examining the residuals and finding leverage points to find any potential problems with the model.

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



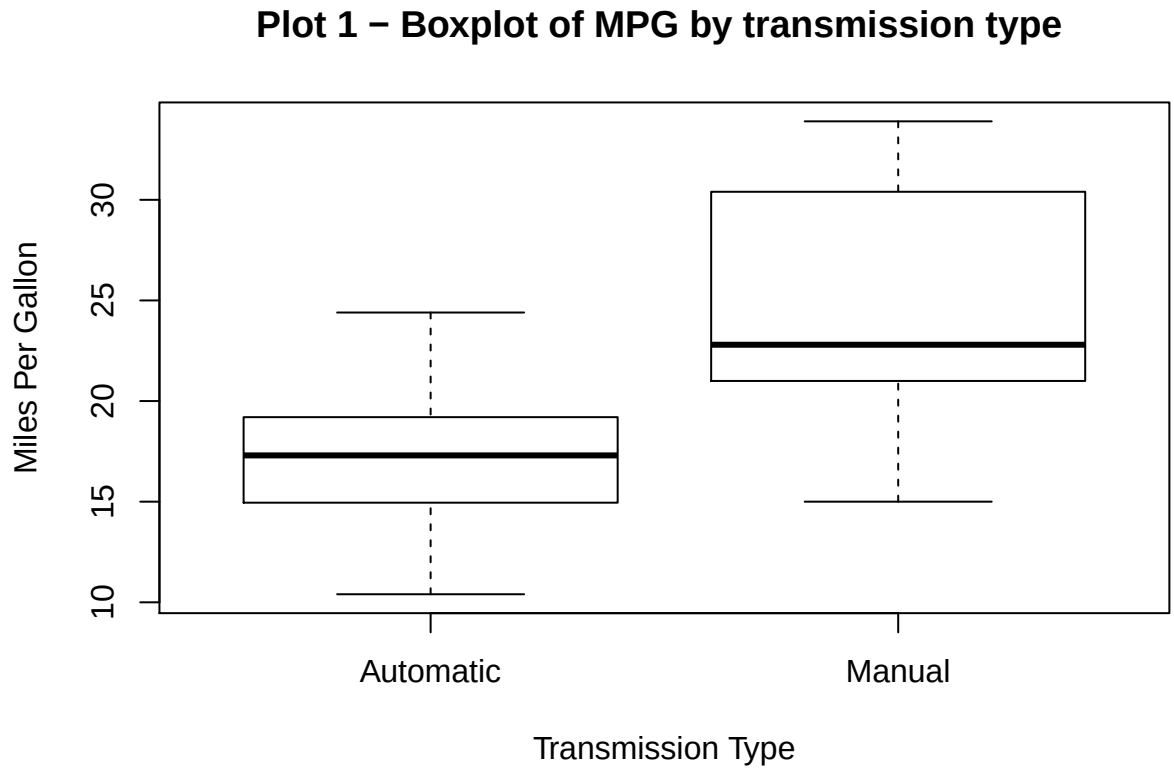
Following observations are made from the above plots

- The points in the Residuals vs. Fitted plot are randomly scattered on the plot that verifies the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- All points are within the Cook's distance line in the Residuals vs Leverage plot. Therefore there is no influential case that we need to exclude.

Appendix

Plot 1 - Boxplot of MPG by transmission type

```
boxplot(mpg ~ am, data = mtcars, ylab = "Miles Per Gallon", xlab = "Transmission Type",
        main="Plot 1 - Boxplot of MPG by transmission type ")
```



Plot 2

- Pairs Plot

```
pairs(mpg ~ ., data = mtcars)
```