Regression Models Course Project

Vinod Manghani

November 23, 2017

Automatic Or Manual: Which Is Better In Terms Of MPG

1. Summary

In this study, We are trying to understand whether an automatic or manual transmission is better for mpg. The dataset mtcars[1] used in the study was extracted from the 1974 Motor Trend US magazine.Based on a two sided t-test, there is a statistical significant difference between automatic and manual transmitted cars in terms of mpg.The regression model selected using backward selection with AIC as the criteria suggests that holding other variables in the model constant, manual transmitted cars' mpg is 1.80921 more than automatic transmitted cars.

2. Loading And Processing Data

```
data(mtcars)
```

Converting categorical variables into factors

```
mtcars$am <- as.factor(mtcars$am)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
mtcars$vs <- as.factor(mtcars$vs)
```

3. Exploratory Data Analysis

According to figure 1(see appendix), manual cars are better than automatic cars in term of MPG. Figure 2(see appendix) suggests that variables "cyl","disp","hp" and "wt" are also highly correlated with mpg.

4. Statistical Inference

The two sample t-test confirmed this difference. The null hypothesis H_0 is:There is no difference between automatic car and manual car in terms of MPG.

```
TTest<-t.test(x=mtcars[mtcars$am==0,]$mpg,y=mtcars[mtcars$am==1,]$mpg,paired=FALSE)
TTest$p.value
```

```
## [1] 0.001373638
```

Based on this p value and using 0.05 as the criteria, the null hypothesis should be rejected, suggests a statistical significant difference between automatic car and manual car in terms of MPG.

5. Model Selection

Backward selection method with AIC as criteria is usd to choose a regression model.

```
model<-lm(mpg~.,data=mtcars)
Selected.model<-step(model,direction="backward",k = 2)
```

Table 1(see appendix) is the summary table of the selected model. According to it, the adjusted r-squared is 0.8659, suggest that 86% of the variation in mpg is explained by the selected model.
The coefficient of "am" in the selected model 1.80921 interpreted as: holding "wt", "cyl" and "hp" constant, manual transmitted cars' mpg is 1.80921 more than automatic transmitted cars.

6. Model Comparision

Comparing the selected model with models contain only a subset of variables of it using Analysis of Variance(ANOVA).

```
Null.model<-lm(mpg~am,data=mtcars)
model1<-lm(mpg~am+cyl,data=mtcars)
model2<-lm(mpg~am+cyl+hp,data=mtcars)
anova(Null.model,model1,model2,Selected.model)[2:4,6]
```

```
## [1] 1.387517e-08 2.164196e-03 9.081408e-03
```

The analysis suggests that selected model is statistical significantly different from all the other models and variables "cyl","hp" and "wt" do contribute to the variation of mpg in the dataset.

7. Residual Analysis

Figure 3(see appendix), a plot of fitted values of the selected model vs the residuals, suggests that the residuals are independently distributed.Figure 4(see appendix), a residual QQ plots, suggests that the residuals are normally distributed.

8. Diagnositics

The leverages for the dataset is calculated and based on a criteria 0.5 there are no outlier.

```
leverage<-hatvalues(Selected.model)
sum(leverage>0.5)
```

```
## [1] 0
```

Appendix

Figure 1.

```
boxplot(mtcars$mpg~mtcars$am,col=c("blue","red"),xlab="Transmission",ylab="mpg",main="Transmission and
```
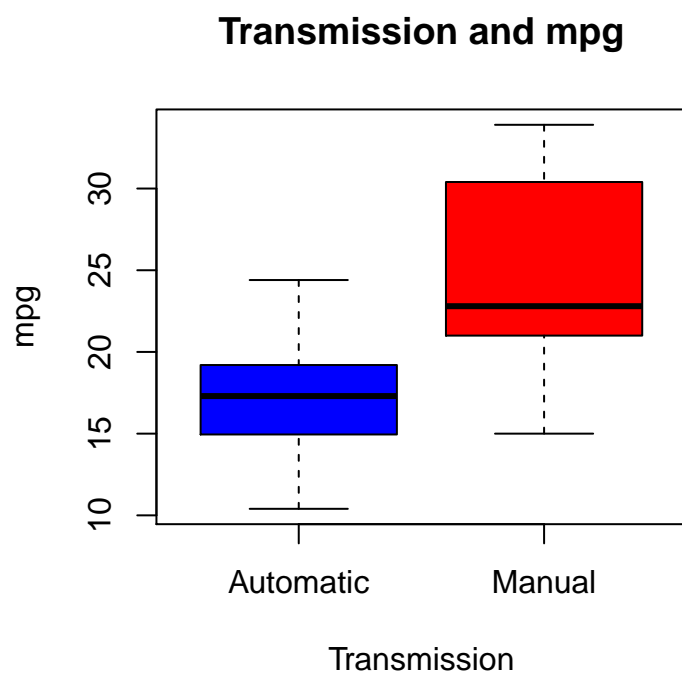
Transmission and mpg

Figure 3. and Figure 4.

```r
par(mfrow=c(1,2))
plot(Selected.model$fitted.values,Selected.model$residuals,pch=19,main="Fitted value VS Residuals")
abline(0,0,col="Red",lwd=3)
qqplot(Selected.model$fitted.values,Selected.model$residuals,pch=19,main="Residual QQ plot")
```
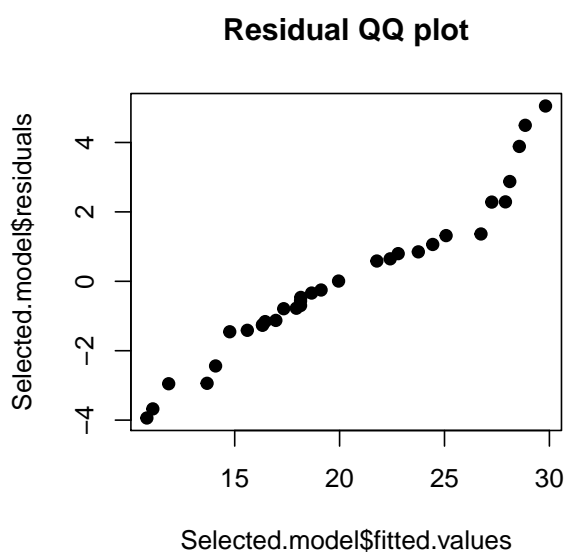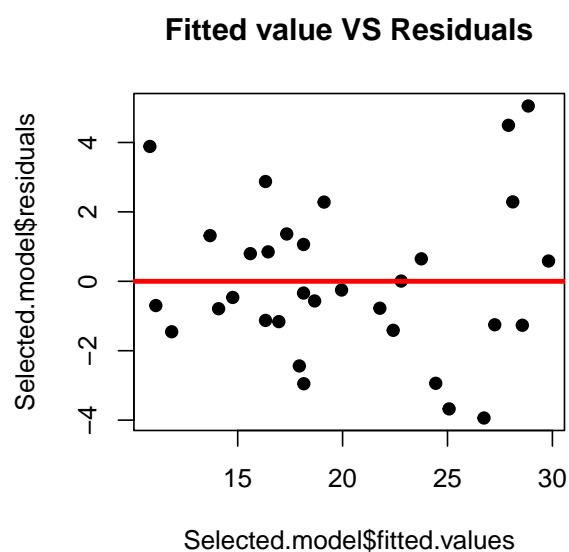
Fitted value VS Residuals

Residual QQ plot

Figure 2.

```
plot(mtcars,panel=panel.smooth,pch=19,lwd=3,main="Relationships among variables")
```


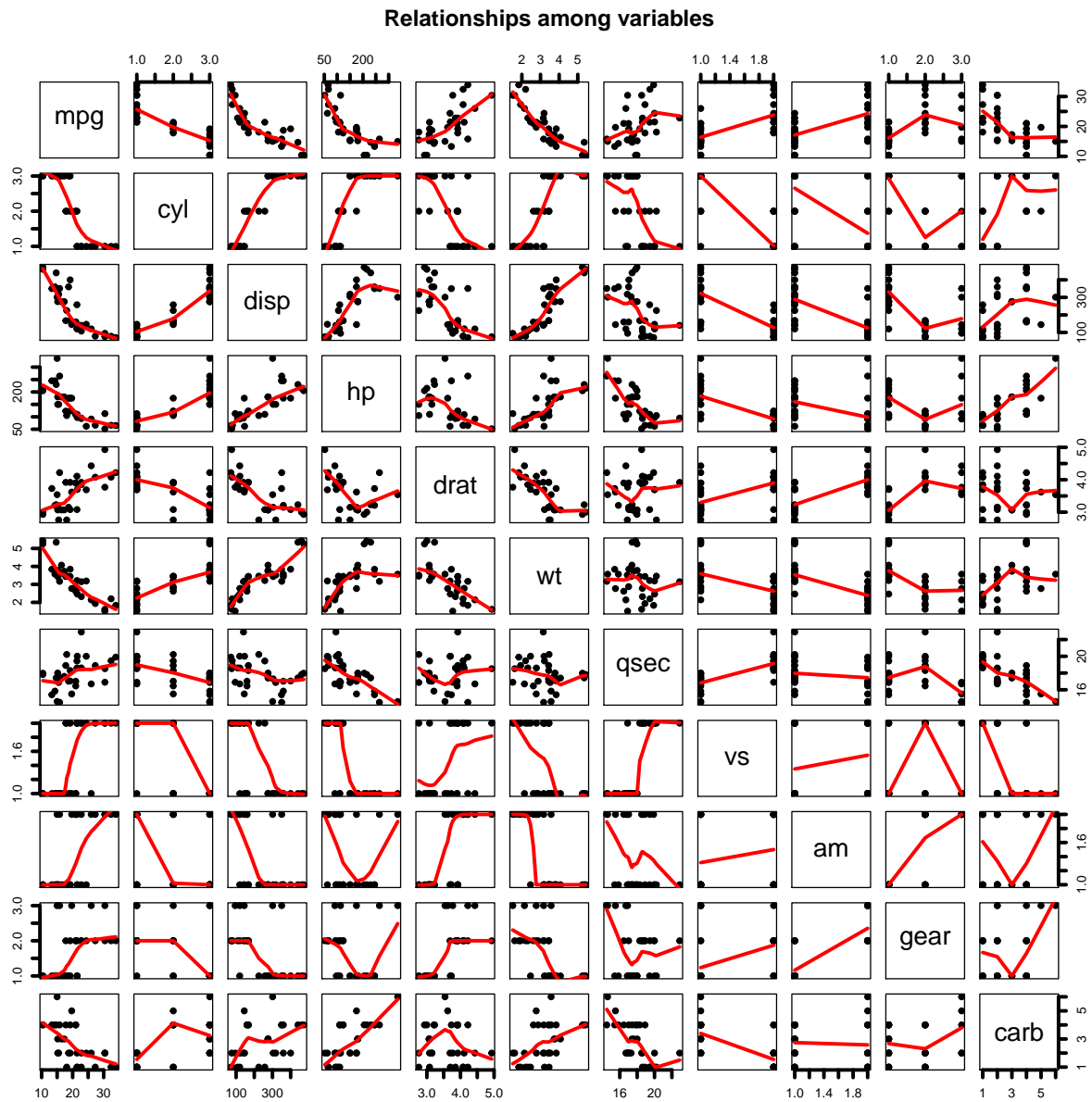**Relationships among variables**

Table 1.

```
summary(Selected.model)
```

```
##
## 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 **
## am1          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
```