Regression Models

Nicholas Vitsentzatos

23/4/2021

Overview

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”

Loading and processing the required dataset

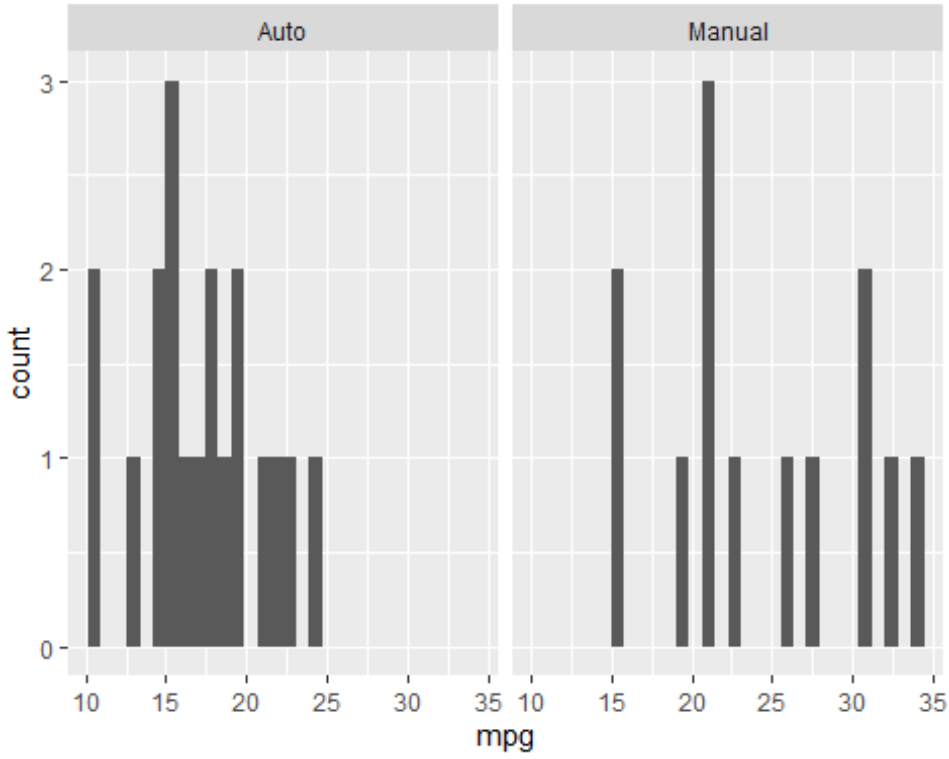
```
library(datasets)
data(mtcars)
for (i in 1:length(mtcars$am)){if(mtcars[i,"am"]==1){mtcars[i,"am"]<-
"Manual"}else{mtcars[i,"am"]<-"Auto"}}
```

In the mtcars dataset cars with automated transmission are labeled as 0 and with manual transmissions as 1

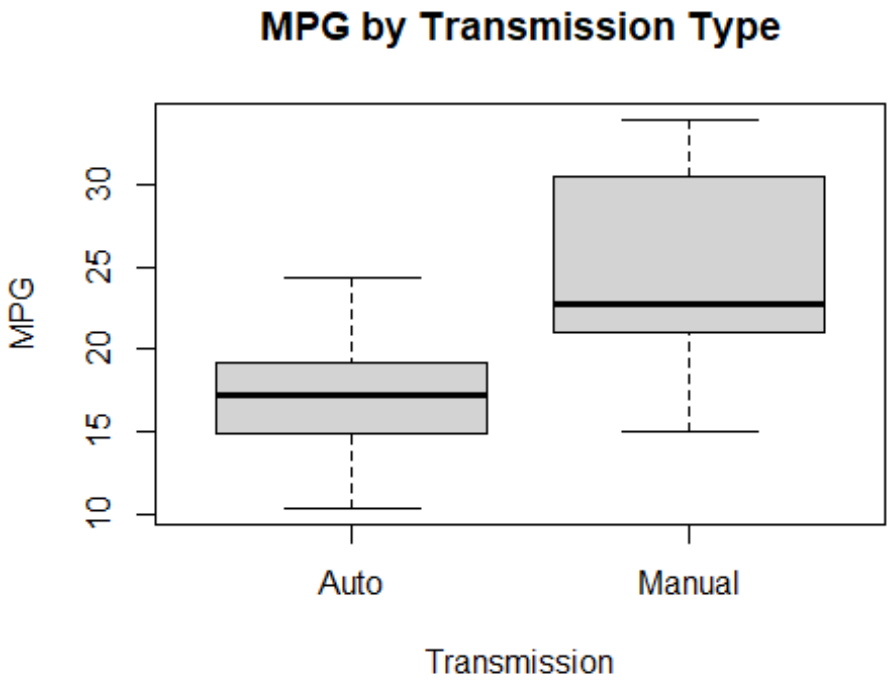
1) Is an automatic or manual transmission better for MPG

We’ll plot mpg based on transmission type to check if we can answer the question

```
library(ggplot2)
ggplot(mtcars,aes(mpg))+geom_histogram(bins =
30)+facet_grid(.~factor(am))
```




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



As we can see from the boxplots the interquartile range of automatic transmission vehicles is lower than that of the manual cars

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

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

Also the means of the different transmission types of cars are far apart from each other

```
t.test(mtcars$mpg~mtcars$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$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 Auto mean in group Manual
##             17.14737             24.39231
```

Last but no least with the p-value:0.001374 from the t-test we can conclude the that automatic transmission cars are better than manual transmission cars for MPG

2) Quantify the MPG difference between automatic and manual transmissions

We'll start with fitting a linear model solely with mpg and am

```
fit1<-lm(mpg ~ factor(am), data = mtcars)
summary(fit1)
```

```
##
## 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)Manual    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
```

R^2:0.3598. That means that am explains only the 36% of the variance of mpg

With the help of stepwise regression, R will find the best combination of variables that gives us the optimal R^2

```
summary(step(lm(mpg ~., data = mtcars)))
```

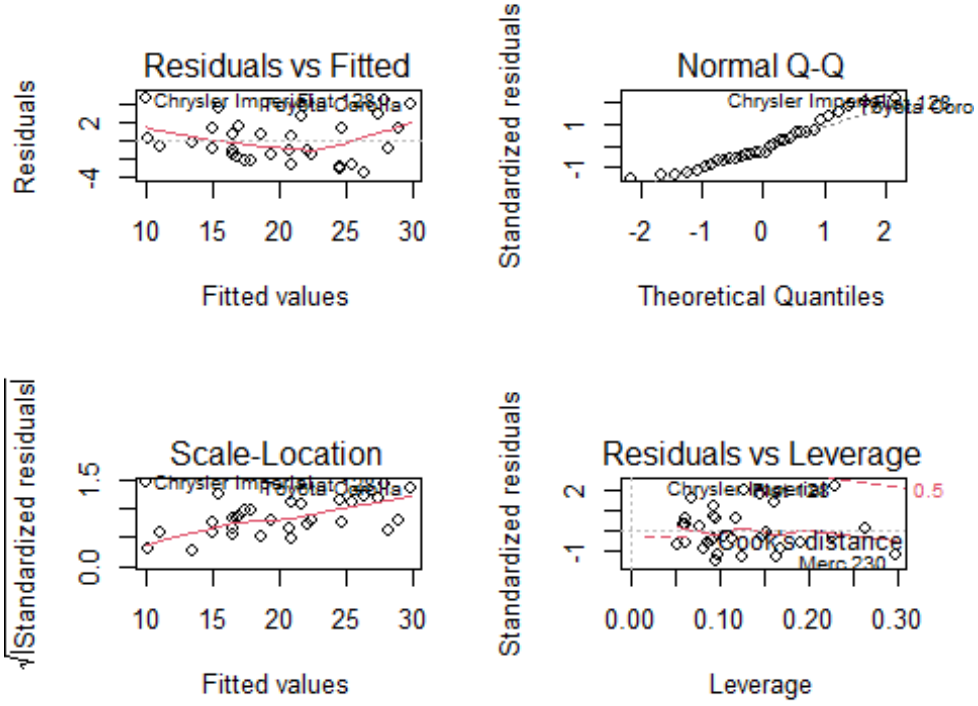
So the best model fit is wt + qsec + am with R^2:0.8336 so 86.59% of the variance is explained by this model and this can be confirmed below with the anova function producing a p-value of 1.55e-09

```
fit2<-lm(mpg ~ wt + qsec + am, data = mtcars)
anova(fit1,fit2)

## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally we can see that the residuals don't have any pattern and the Normal Q-Q plot shows that residuals are normally distributed

```
par(mfrow=c(2,2))
plot(fit2)
```



Summary

Automatic transmission cars have better mpg than manual transmission cars on average