Regression models

Davyd

May, 2018

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:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

Load Data

Load required packages, dataset, and convert categorical variables to factors.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

data(mtcars)
head(mtcars, n=3)
dim(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

Exploratory Data Analysis

See Appendix I Box plot comparing MPG between Automatic and Manual transmission. The results show a significant increase in MPG for manual transmission when compared to automatic transmission.

Statistical Inference

T-Test between transmission type and MPG

```
testResults <- t.test(mpg ~ am)
testResults$p.value

## [1] 0.001373638
```

The T-Test rejects the null hypothesis that there is no difference in MPG for both transmission types.

```
testResults$estimate

## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The estimated difference between the two transmission types is 7.24494 MPG in favour of manual transmission.

Regression Analysis

Fitting the model for the data

```
fullModelFit <- lm(mpg ~ ., data = mtcars)
summary(fullModelFit)  # results hidden
summary(fullModelFit)$coeff  # results hidden
```

Since none of the p-values are below 0.05, we cannot conlude that there is any statistical significance.

Selecting variables which are most statistically significant

```
stepFit <- step(fullModelFit)
summary(stepFit) # results hidden
summary(stepFit)$coeff # results hidden
```

The new model has 4 variables (cylinders, horsepower, weight, transmission). The R-squared value of 0.8659 confirms that this model explains about 87% of the variance in MPG. The p-values also are statistically significantly because they have a p-value less than 0.05. The coefficients conclude that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.03. Further increasing the cylinders to 8 with decrease the MPG by 2.16. Increasing the horsepower is decreases MPG 3.21 for every 100 horsepower. Weight decreases the MPG by 2.5 for each 1000 lbs increase. A Manual transmission improves the MPG by 1.81.

Residuals & Diagnostics

Residual Plot

See Appendix II

The plots show that:

1. The randomness of the Residuals vs. Fitted plot supports the assumption of independence
2. The points of the Normal Q-Q plot following closely to the line conclude that the distribution of residuals is normal

3. The Scale-Location plot random distribution confirms the constant variance assumption
4. Since all points are within the 0.05 lines, the Residuals vs. Leverage concludes that there are no outliers

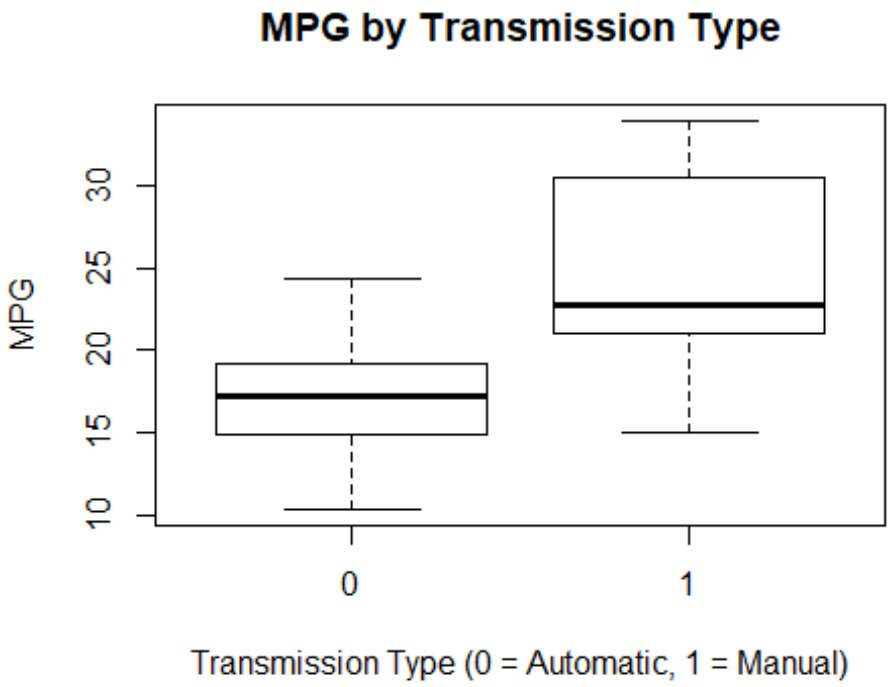
```
sum((abs(dfbetas(stepFit)))>1)
```

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

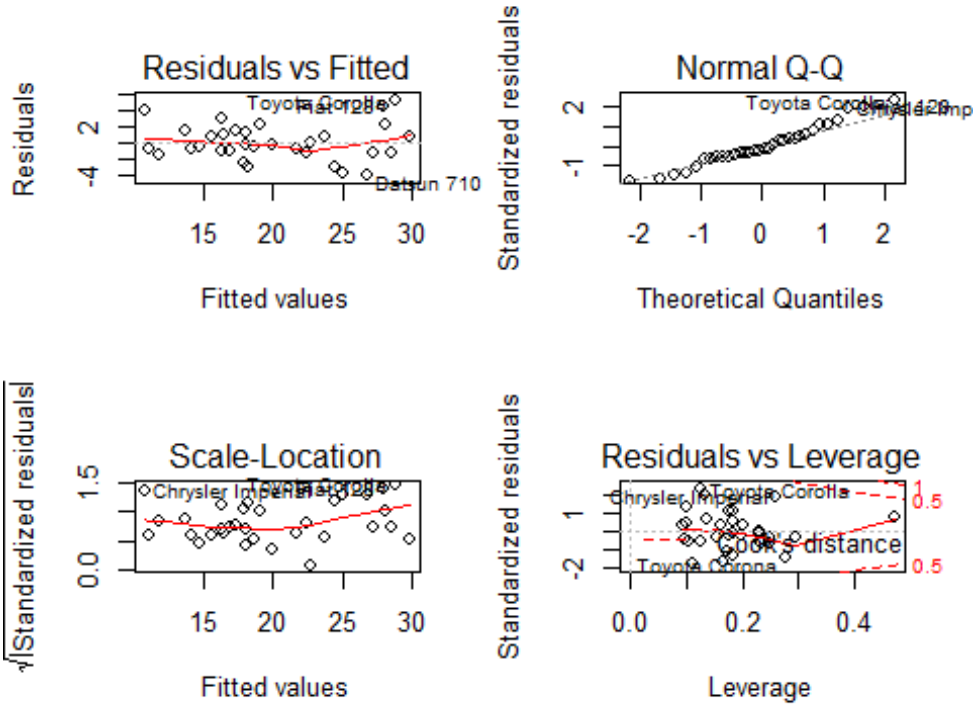
Conclusion

There is a difference in MPG between transmission type. A manual transmission will have a slight advantage in MPG. However, weight, horsepower, & number of cylinders are more statistically significant when determining MPG.

Appendix I



Appendix II


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Datsun 710
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Chrysler Imperial
Toyota Corolla
Toyota Corona