regression modeling - motor trend

jeffzfw

June 20, 2015

Executive Summary

Motor Trend is 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" We will use regression models and exploratory data analyses to mainly explore how automatic and manual transmissions features affect the MPG feature. T-test will show the performance difference between automatic and manual transmission. Then, we will fit several linear regression models and select the one with highest Adjusted R-squared value. ## Exploratory Data Analysis

```
library(ggplot2)
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

After loading the data set `mtcars`,We like to do some visual explorations.check the Appendix,the boxplot(fig1) shows the difference between automatic transmisssion and manual transmission seems to be true. ## Inference Let's check the averge MPG for automatic transmissions:

```
auto_trans <- subset(mtcars, am == 0)
mean(auto_trans$mpg)
```

```
## [1] 17.14737
```

Average MPG for manual transmissions:

```
man_trans <- subset(mtcars, am == 1)
mean(man_trans$mpg)
```

```
## [1] 24.39231
```

perform a t-test between two groups:

```
r1 <-t.test(man_trans$mpg,auto_trans$mpg)
```

By the t-test,p-value 0.0013736 is less than 5%,and confidence interval 3.2096842 - 11.2801944 doesn't contain 0. so we can say that the manual cars are better than automatic cars. But using the transmission alone is not enough to quantify the difference for specific cases. We would like to see linear regression of miles per gallon using the single transmission indepent variable.

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

The p-value 0.000285 shows relation between mpg and transmission is significant.But R-squared 0.3597989 shows the variance explained is low.

Now, we like to check full model that contains all the variables:

```
fullM <- lm(mpg ~ .,data = mtcars)
sf <- summary(fullM)
```

This model has R-squared value of 0.8690158,but the coeffients are not at 0.05 significant level.

Then, we use backward selection to select some statistically significant variables.

```
bestM <- step(fullM, trace= 0)
summary(bestM)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

So,the model we choose is "mpg ~ wt + qsec + am".All of the coefficients are significant at 0.05 significant level and the R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable.

Residual analysis and diagnostics

According to Appendix fig2,the diagnostic plots show the residuals are normally distributed and homoskedastic.

Appendix

fig1

```
boxplot(mpg ~ am, data=mtcars, xlab="Transmission Type(0 = Auto, 1 = Manu)", ylab="Miles per Gallon",ma
```

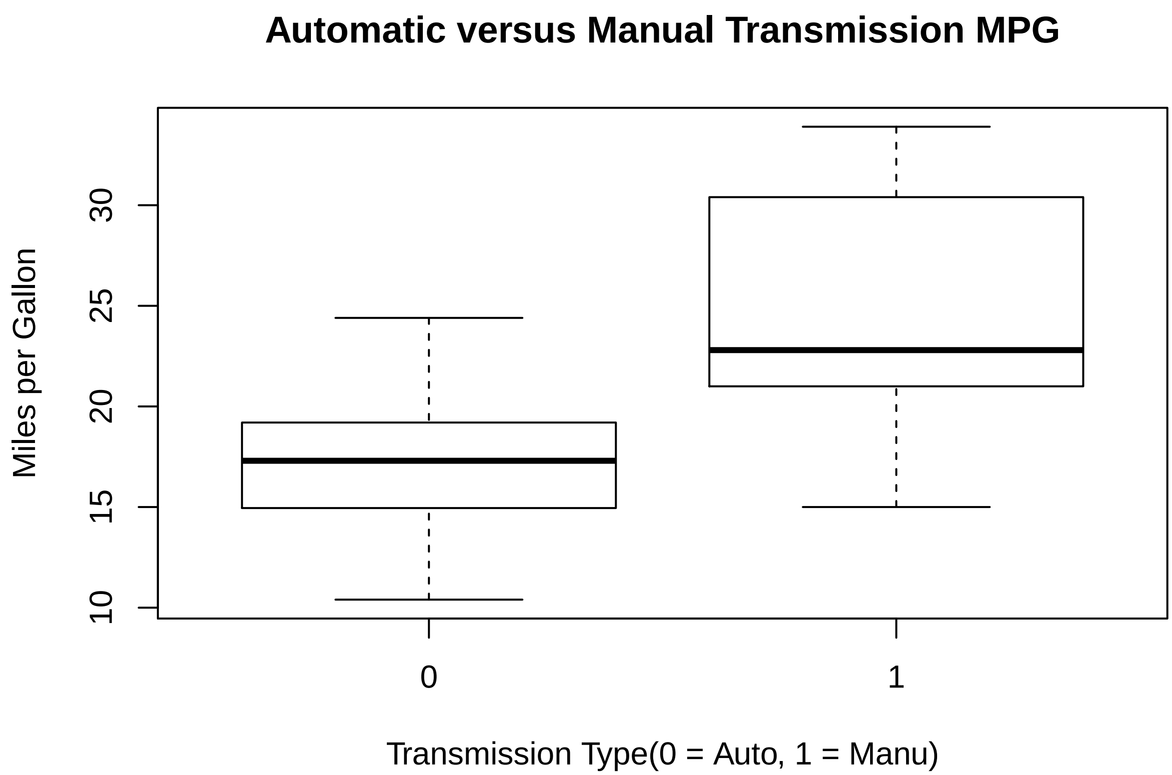
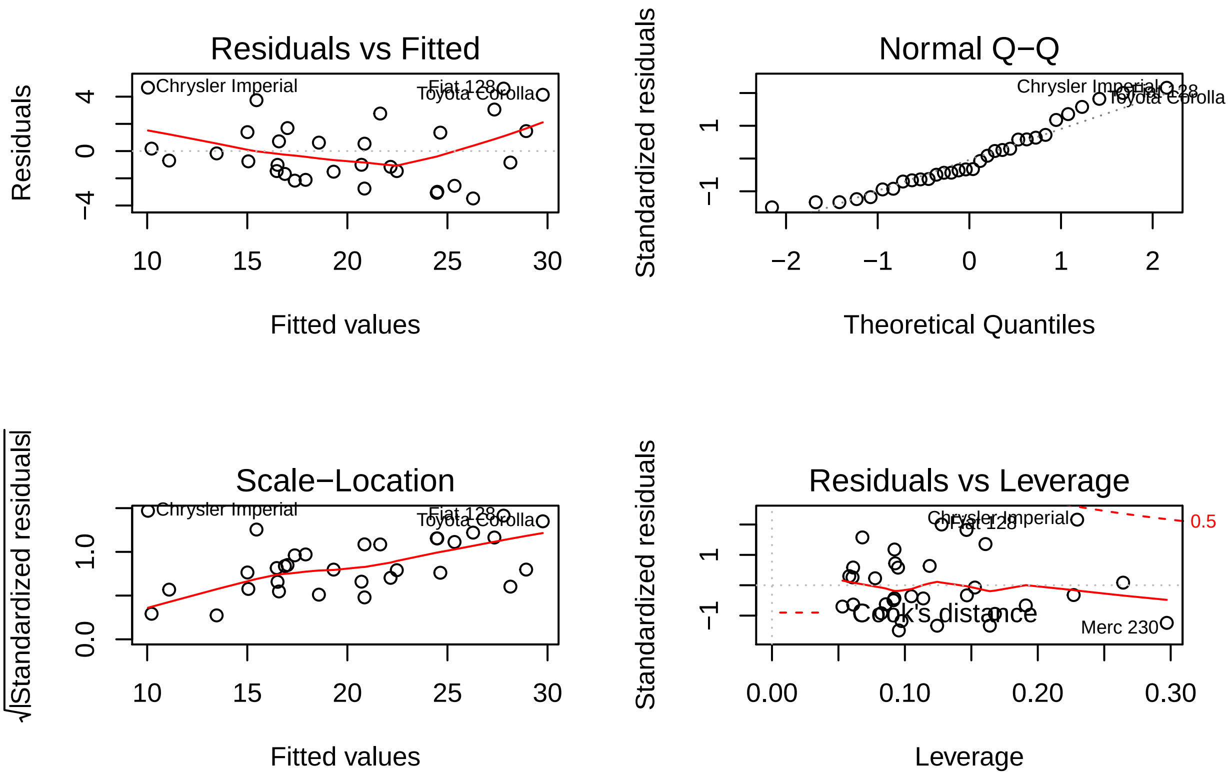


fig2

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


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5