Coursera Regression Models Project

Rajat Hada

10/14/2019

Setup

Load packages

```
library(data.table)
library(ggplot2)
library(leaps)
library(printr)
```

Load data

```
data("mtcars")
```

Executive Summary

Here we will see how mileage (MPG: Miles per Gallon) is affected by type of transmission (Automatic or Manual).We will analyze following:

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

Preliminary Exploratory Data Analysis

The data of this project are extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973/74 models).

```
dim(mtcars)
```

```
## [1] 32 11
```

The data consists of 32 observations on 11 variables.

```
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

- **mpg**: Miles/(US) gallon
- **cyl**: Number of cylinders
- **disp**: Displacement (cu.in.)
- **hp**: Gross horsepower
- **drat**: Rear axle ratio
- **wt**: Weight (lb/1000)
- **qsec**: 1/4 mile time
- **vs**: V/S
- **am**: Transmission (0 = automatic, 1 = manual)
- **gear**: Number of forward gears
- **carb**: Number of carburetors

Data Processing

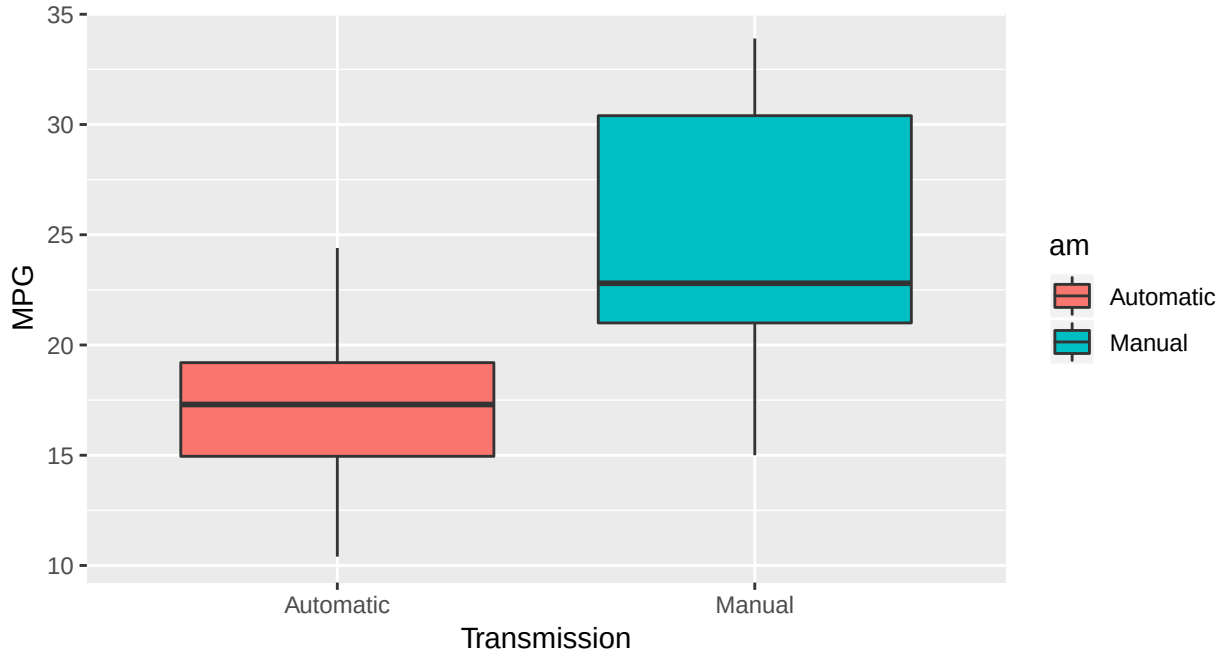
Change the data type of all categorical variable as Factor data type.

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic","Manual"))
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

Data Visualizations

Lets plot MPG("Mileage") vs am("Transmisson Type"), to understand the distribution.

```
ggplot(mtcars, aes(x=am, y=mpg)) +
    geom_boxplot(aes(fill = am)) +
    xlab("Transmission") +
    ylab("MPG")
```



The plot show that manual transmissions have higher MPG.

Performing a t-test will help verify if the difference in means is significant.

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

```
##
##  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
```

p-value is less than 0.05. So, it rejects Null Hypothesis. Therefore there is a difference in transmission type, with manual transmissions having a higher MPG.

Lets use some model to evaluate the correlations.

Linear Regression Fitting

Simple Linear Regression Model

Lets start the linear model with only transmission type as the independent variable.

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

```
##
## 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 p-value are low (0.000285) and R-Squared is 0.3385.

Before making any conclusions on the effect of transmission type on fuel efficiency, we will look at the variances between several variables in the dataset.

Lets fitting all parameters of mtcars.

```
fitall <- lm(mpg ~ .-1, mtcars)
summary(fitall)
```

```
##
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## cyl4     23.87913   20.06582   1.190   0.2525
## cyl6     21.23044   18.33416   1.158   0.2650
## cyl8     23.54297   18.22250   1.292   0.2159
## disp      0.03555    0.03190   1.114   0.2827
## hp       -0.07051    0.03943  -1.788   0.0939 .
## drat      1.18283    2.48348   0.476   0.6407
## wt       -4.52978    2.53875  -1.784   0.0946 .
## qsec      0.36784    0.93540   0.393   0.6997
## vs1       1.93085    2.87126   0.672   0.5115
## amManual  1.21212    3.21355   0.377   0.7113
## gear4     1.11435    3.79952   0.293   0.7733
## gear5     2.52840    3.73636   0.677   0.5089
## carb2    -0.97935    2.31797  -0.423   0.6787
## carb3     2.99964    4.29355   0.699   0.4955
## carb4     1.09142    4.44962   0.245   0.8096
## carb6     4.47757    6.38406   0.701   0.4938
## carb8     7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.9914, Adjusted R-squared:  0.9817
## F-statistic:   102 on 17 and 15 DF,  p-value: 1.979e-12
```

All the p-values are greater than 0.05. By including all the varaibles it increased the R-Squared value but it hurts the prediction. So, we have to meet somewhere in the middle.

Lets use the R function STEP to do the variable selection.

STEP function

```
bestFit <- step(fitall,direction="both",trace=FALSE)
summary(bestFit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am - 1, 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|)
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amAutomatic   9.6178     6.9596   1.382 0.177915
## amManual     12.5536     6.0573   2.072 0.047543 *
## ---
## 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:  0.9879, Adjusted R-squared:  0.9862
## F-statistic: 573.7 on 4 and 28 DF,  p-value: < 2.2e-16
```

The Residual standard error of this model is 2.459 on 28 degrees of freedom. The Adjusted R-Squared value has increased to 0.9862.

Final Model Examination

Now we fit the model "mpg ~ wt + qsec + am" as final examination model.

```
lastModel <- lm(mpg ~ wt + qsec + am, data = mtcars)
summary(lastModel)
```

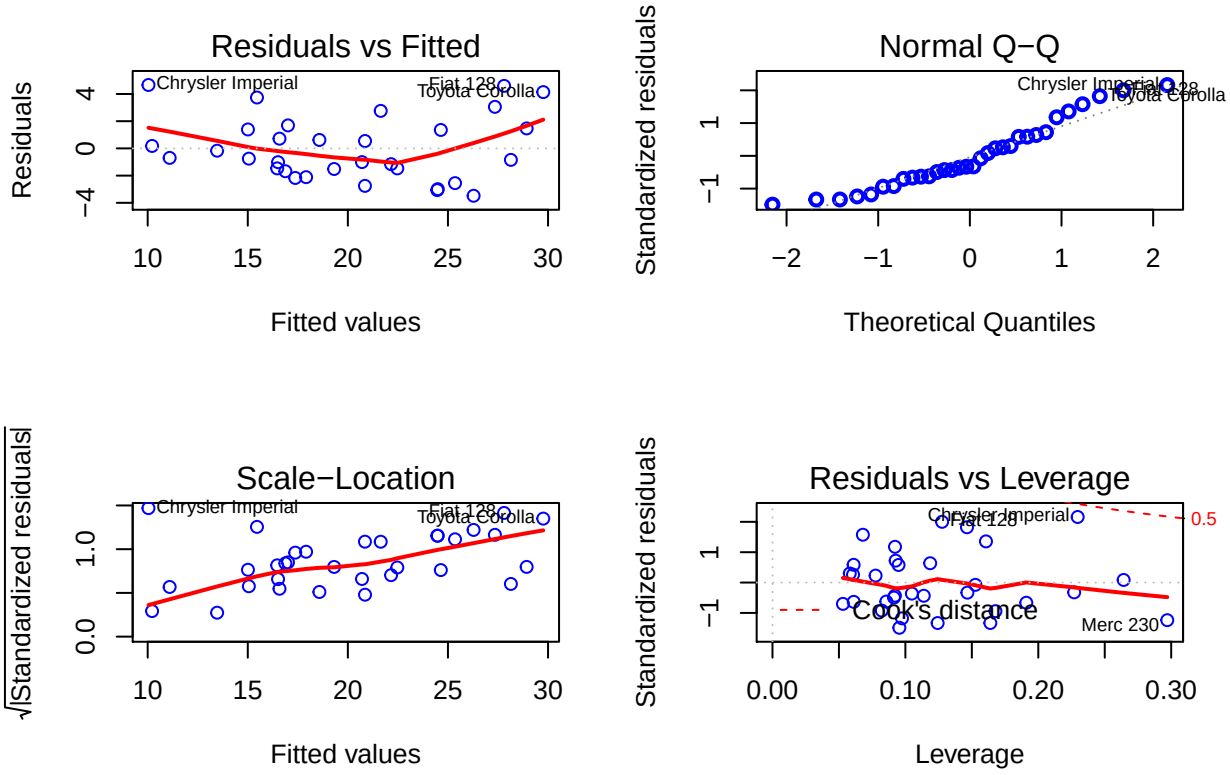
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      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
```

Residual Analisys

The resulting final model examination is dependant not only on the transmission (am), but also weight (wt) and 1/4 mile time (qsec).

Now, Lets look into diagonstic plots.

```
par(mfrow = c(2,2))
plot(lastModel, col = "blue", lwd = 2)
```



- Residuals vs Fitted: The points are randomly scattered, but may have a slight non-linear relationship.
- Normal Q-Q: The points pass normality, they deviate slightly from the diagonal, but they follow the diagonal fairly close.
- Scale-Location: The upward slope line is worrisome, the residues spread slightly wider.
- Residuals vs Leverage: No high leverage points.

Conclusions

The best transmission type for MPG would have to be the manual transmission. Its confirmed by the t-test, as well as our final linear model. By having a manual transmission instead of an automatic the MPG will increase by 2.94.

The model fit well with a $p < 0.05$ and adjusted $R^2 = 0.83$, but the diagnostic plots did warn us that something may be missing in our model. I believe the true cause for these trends are do to the small sample size with little overlap on the parameters `wt` and `qsec`.