Motor Trend

Yinning Zhang

September 13, 2018

Summary

This project uses the dataset mtcars, which is available in R package. The purpose of this project is to analyze the impact of the variables on miles per gallon and to build a regression linear model for this relationship. The results shows that the transmission type does not have significant impact on MPG. Instead, the variables of weight and number of cylinders largely explain the difference in mpg. The final regression model is built on these variables.

Preview the data

```
require(datasets)
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
```

Get an overview of correlations between the variables

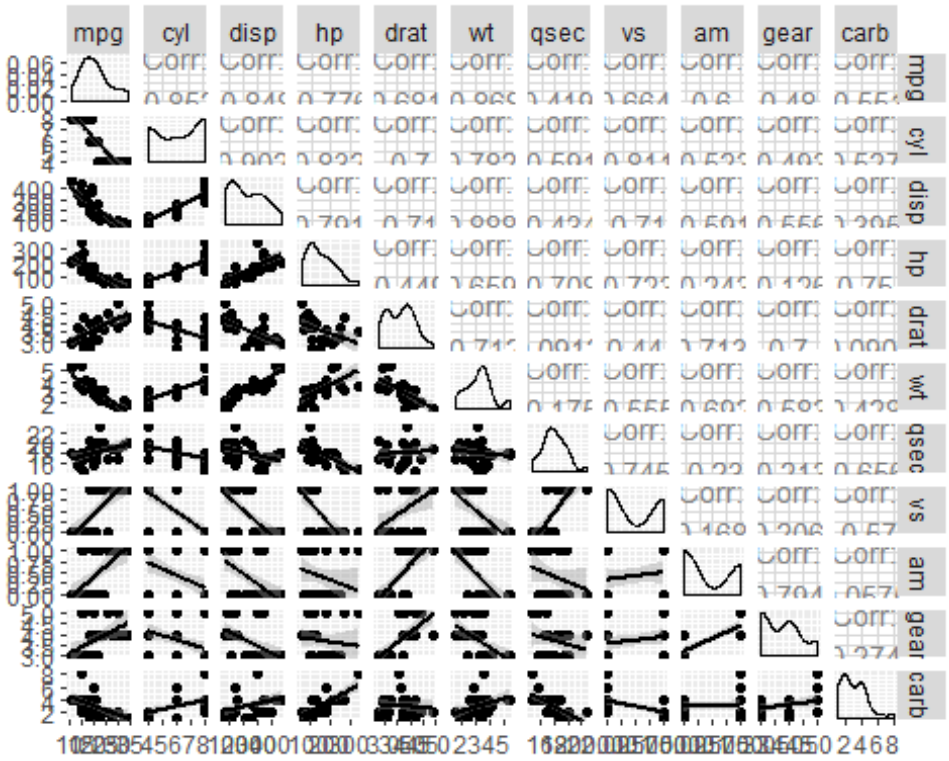
```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.4
```

```
## Loading required package: ggplot2
```

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

```
g = ggpairs(mtcars, lower = list(continuous = wrap("smooth", method = "lm")))
g
```




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

##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
```

```
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

From the correlation matrix, we notice that cars with manual transmission seem to have higher mpg than cars with auto transmission. However, this relationship can be attributed by other variables. So, we choose the variable Weight, as its P-value is 0.06, we have more confidence of this variable's correlation to mpg. Let's see if the transmission type matters when we include the variable weight.

Select variables and build the regression model

1. mpg ~ transmission+wt

```
print(summary(lm(mpg~ wt + factor(am), data = mtcars)))

##
## Call:
## lm(formula = mpg ~ wt + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## factor(am)1 -0.02362    1.54565  -0.015    0.988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

The p-value is 0.988. We fail to reject the null hypothesis. So, there is no significant difference between auto car and manual car.

2. more variables

Here, we add variables of cyl, disp, drat, and gear and use ANOVA test to see if these variables help reduce the least square residuals.

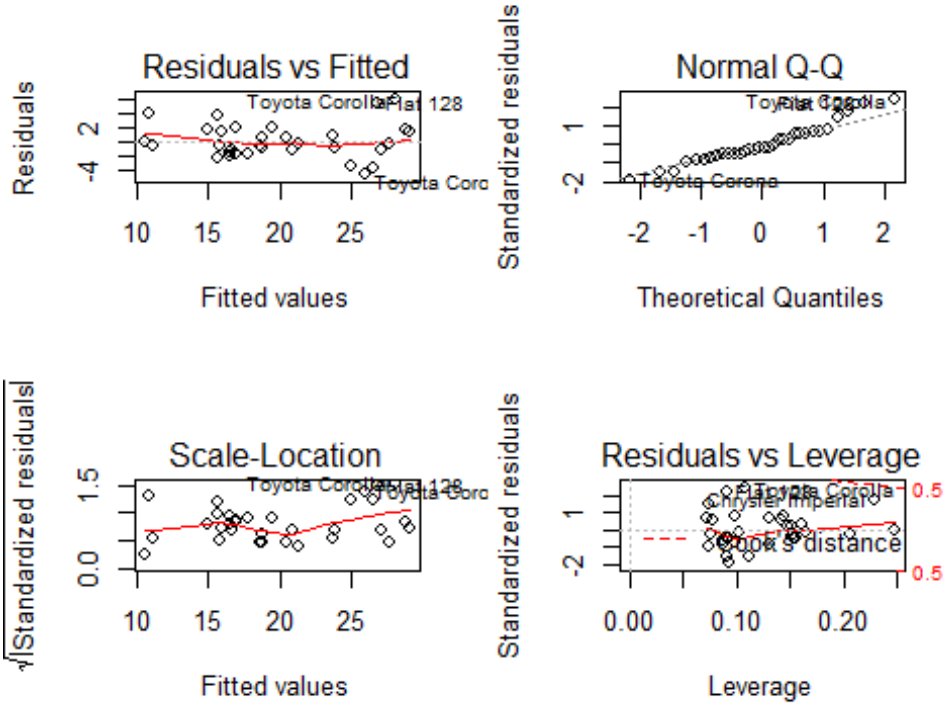
```
fit1 <- lm(mpg ~ wt, data = mtcars)
fit2 <- lm(mpg ~ wt + factor(cyl), data = mtcars)
fit3 <- lm(mpg ~ wt + factor(cyl) + disp, data = mtcars)
fit4 <- lm(mpg ~ wt + factor(cyl) + disp +drat, data = mtcars)
fit5 <- lm(mpg ~ wt + factor(cyl) + disp +drat +gear, data = mtcars)
anova(fit1, fit2, fit3, fit4, fit5)

## Analysis of Variance Table
##
```

```
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + factor(cyl)
## Model 3: mpg ~ wt + factor(cyl) + disp
## Model 4: mpg ~ wt + factor(cyl) + disp + drat
## Model 5: mpg ~ wt + factor(cyl) + disp + drat + gear
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 278.32
## 2     28 183.06  2    95.263 6.5936 0.005015 **
## 3     27 182.95  1     0.110 0.0152 0.902936
## 4     26 182.88  1     0.064 0.0088 0.925889
## 5     25 180.60  1     2.287 0.3166 0.578674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result indicates that including the variable cyl can largely reduce the sum of least sqaures of the residuals, and the P-value implies that this difference is significant. So, I will add cylinder into the regressor.

Plot the model and analyze the residuals

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



The plots shows that there is no obvious pattern within the residuals. So, this seems a good model.

Model result

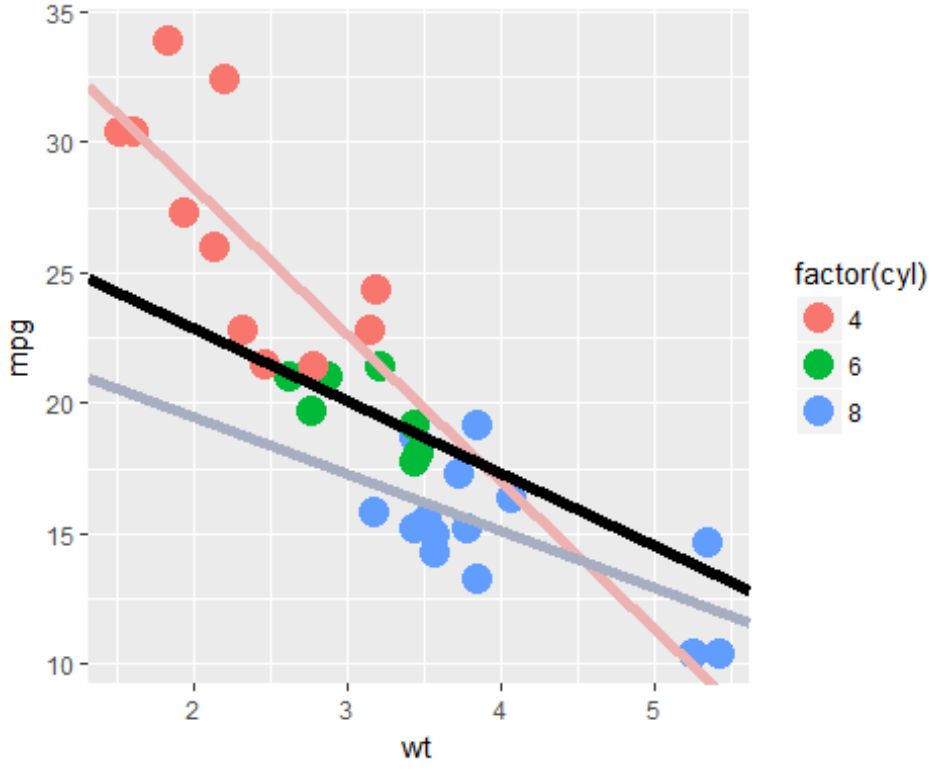
```
summary(lm(mpg ~ wt*factor(cyl), data = mtcars))$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      39.571196   3.193940 12.3894599 2.058359e-12
## wt               -5.647025   1.359498 -4.1537586 3.127578e-04
## factor(cyl)6    -11.162351   9.355346 -1.1931522 2.435843e-01
## factor(cyl)8    -15.703167   4.839464 -3.2448150 3.223216e-03
## wt:factor(cyl)6   2.866919   3.117330  0.9196716 3.661987e-01
## wt:factor(cyl)8   3.454587   1.627261  2.1229458 4.344037e-02
```

Conclusion: The estimated intercept of 4 cylinder cars is 39.57 while the intercept for 6 cylinder cars is estimated to be 39.57-11.16, and the intercept for 8 cylinder cars is 39.57-15.70. The estimated slope for 4 cylinder cars is -5.65, while the slope for 6 cylinder cars is estimated to be -5.65+2.87, and the slope for 8 cylinder cars is -5.65+3.45.

Plot the data with the fitted line and with color coded by cylinder.

```
g1=ggplot(mtcars, aes(x = wt, y = mpg, colour = factor(cyl)))
g1= g1+geom_point(size=3, colour = "black")+geom_point(size=5)
fit6 = lm(mpg ~ wt*factor(cyl), data = mtcars)
g1 = g1+geom_abline(intercept = coef(fit6)[1], slope = coef(fit6)[2], size=2,
colour = '#edb1b2')
g1 = g1+geom_abline(intercept = coef(fit6)[1]+coef(fit6)[3], slope =
coef(fit6)[2]+coef(fit6)[5], size = 2, colour ='#000000')
g1 = g1 + geom_abline(intercept = coef(fit6)[1] +coef(fit6)[4], slope =
coef(fit6)[2]+coef(fit6)[6], size = 2, colour = '#a8aec1')
print(g1)
```


35
30
25
20
15
10
mpg
2
3
4
5
wt
factor(cyl)
4
6
8