Peer-graded Assignment: Regression Models Course Project

JF Ee

7/2/2021

Executive Summary

EDA and multiple linear regression model conducted on the mtcars dataset, with mpg being the dependent variable. Boxplot shows that automatic cars has a better mpg. However, stepwise algorithm was adopted for the regression model and a three variables model (wt, qsec and am) was produced. From the coefficients, we see that generally manual cars have a higher mpg, but it is dependent on the weight and acceleration speed of the cars as well.

Exploratory Data Analysis (EDA)

There are 11 variables in the dataset. By using the cor() function, we can see cyl, disp, hp, wt and carb are negatively correlated to mpg. Furthermore, we also see cyl, disp, hp and wt are strongly correlation to mpg as well.

The variable am should be a factor type instead, with 0 being automatic and 1 being manual. Hence data conversion is required here.

```
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
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
```

```
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
paste("Correlation:");cor(mtcars$mpg,mtcars[,-1])
```

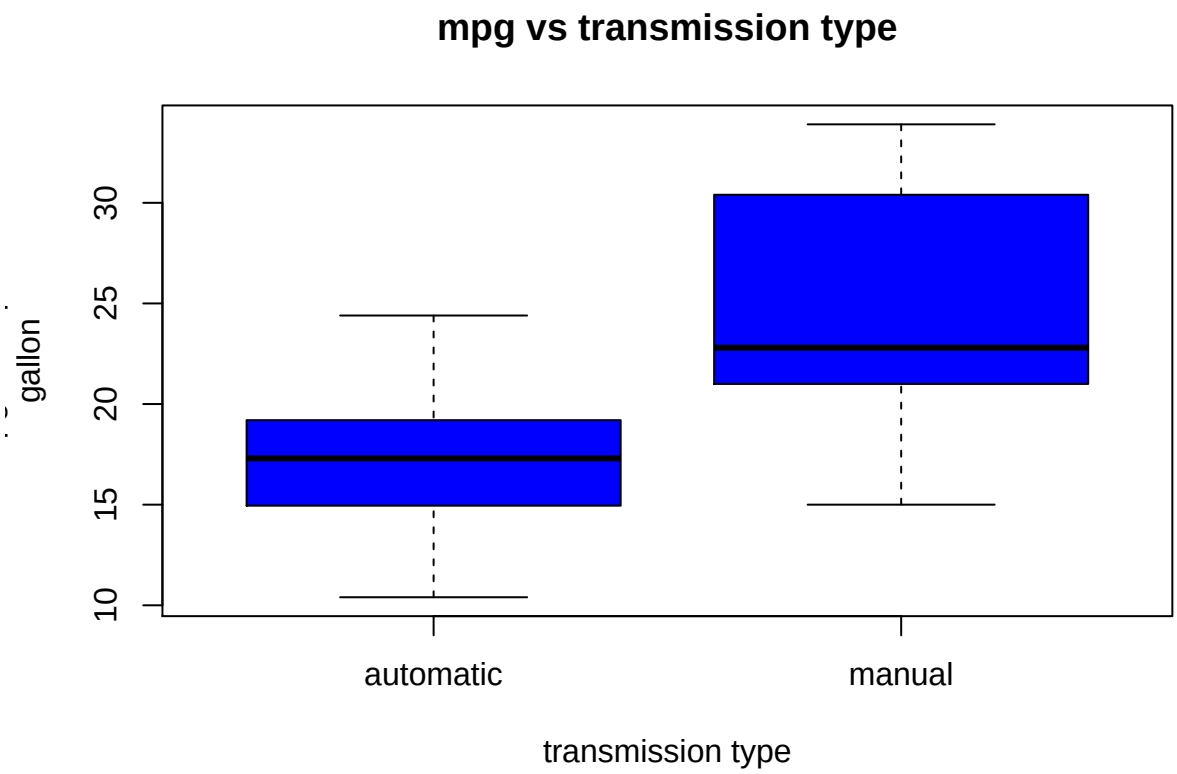
```
## [1] "Correlation:"
```

```
##            cyl       disp         hp      drat         wt     qsec        vs
## [1,] -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684 0.6640389
##             am      gear       carb
## [1,] 0.5998324 0.4802848 -0.5509251
```

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c('automatic','manual')
```

Box plot & T-test

From the box plot, we can see a relationship that automatic car has a better mpg. We can test the hypothesis with a t-test. P value from the t-test is 0.001374 and hence we can reject the null hypothesis.




```
t.test(mtcars$mpg~mtcars$am,conf.level=0.95)
```

```
##
##  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 automatic    mean in group manual
##                17.14737                24.39231
```

Multiple linear regression

Stepwise algorithm is adopted to choose the best model. We see that a three variable model is produced here, with wt, qsec and am variables. The model has a variance of 0.85. From the coefficients, we see that: - Every increase in weight (wt), mpg decreases by -3.9165 - Every increase in qsec, mpg increases by 1.2259 - Manual transmission increases mpg by 2.9358 compared to automatic tranmission

```
stepmodel = step(lm(data = mtcars, mpg ~ .),trace=0,steps=10000)
summary(stepmodel)
```

```
##
## 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
```

To further optimise the model, we can examine mpg~wt+qsec correlation with am. We see the variance increase to 0.8946 in this optimised model.

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

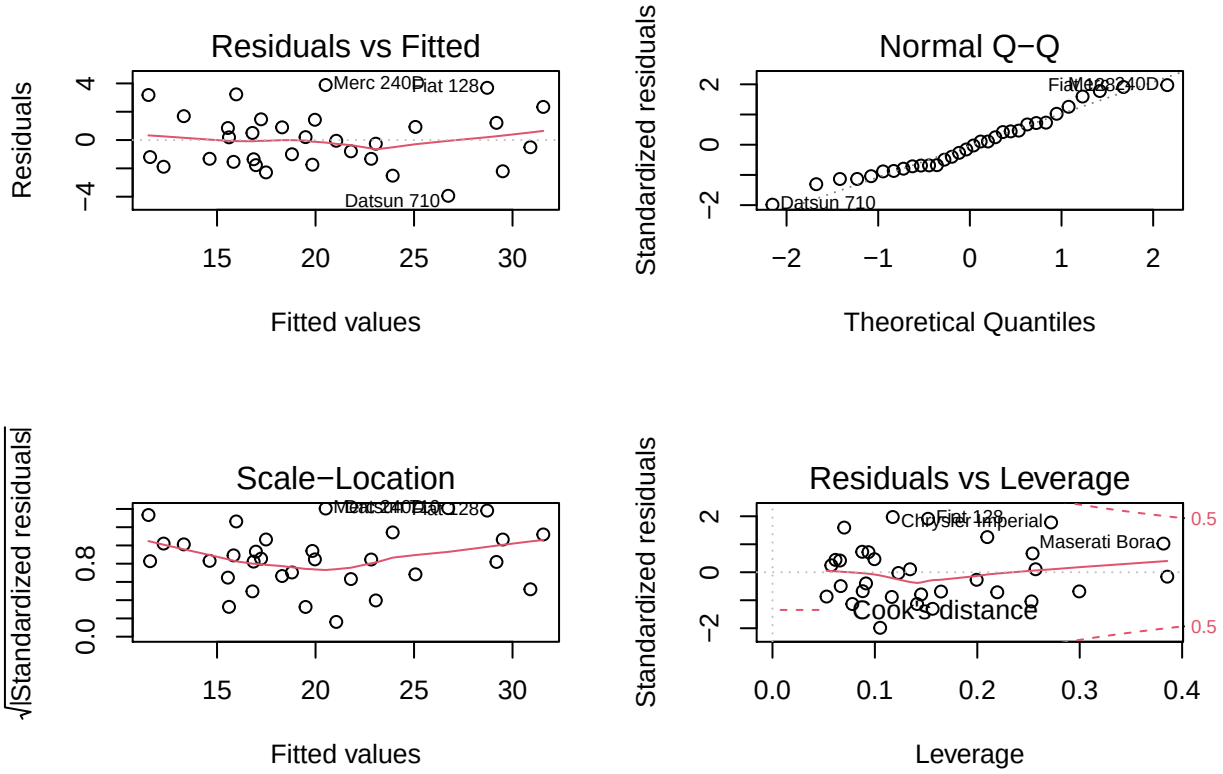
```
##
## Call:
## lm(formula = mpg ~ factor(am):wt + factor(am):qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9361 -1.4017 -0.1551  1.2695  3.8862
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                13.9692     5.7756   2.419  0.02259 *
## factor(am)automatic:wt     -3.1759     0.6362  -4.992 3.11e-05 ***
## factor(am)manual:wt        -6.0992     0.9685  -6.297 9.70e-07 ***
## factor(am)automatic:qsec    0.8338     0.2602   3.205  0.00346 **
## factor(am)manual:qsec       1.4464     0.2692   5.373 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.097 on 27 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.879
## F-statistic: 57.28 on 4 and 27 DF,  p-value: 8.424e-13
```

Conclusion

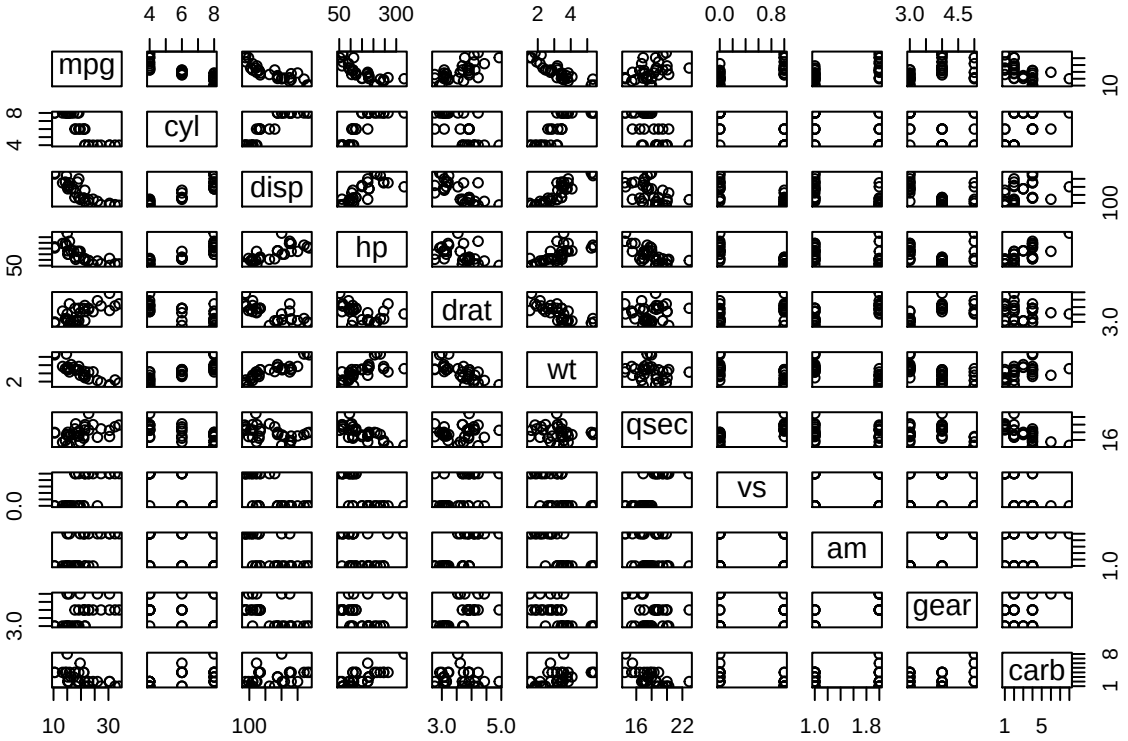
From the coefficients, we see that: - When weight increased by 1000 pounds, the mpg factor decreases by 3.1759 for automatic cars and decreases by 6.0992 for manual cars - Hence, as weight increases, consumers

should choose manual cars. - As acceleration speed increases, the mpg factor increases by 0.8338 for automatic cars and 1.446 for manual cars. - Hence, the lower acceleration speed, holding weight constant, manual cars are more efficient. - Mpg is largely determined by weight, acceleration and transmission. As such, looking at the weight and acceleration of the car, then consumer can decide whether manual or automatic cars better for them.

Appendix 1: Residual check and diagnostics plot



Appendix 2: Scatterplots



Appendix 3: Density and histogram