Regression Models Project

Dev Das

October 20, 2020

Executive Summary

The purpose of this report is to analyze the mtcars dataset provided in R, and to conduct regression analysis. Looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). The two main questions we are looking at, are as follows:

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

Exploratory Data Analysis

```
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
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
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
```

From the analysis above we can see that there are 32 rows of data with 11 columns (variables). A quick summary of each variable can give us basic statistics that will aid our exploration. Forexample we can see that the "am" column ranges from 0 to 1 which tells us that cars are either automatic or manual.

Fitting the model

Since this is a multivariable regression problem we would first like to see what values are highly correlated with mpg to determine which variables are most likely to affect our models. We can do this by using a correlation matrix:

```
result <- cor(mtcars)
round(result, 2)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

We can define "high" correlation as variables with R-values greater than $|0.6|$ in our case, so we can limit our model to the cyl, disp, hp, drat, wt, vs, and am variables. We'll fit multiple models to start and look at them all below:

```r
fit0 <- lm(mpg ~ am, data = mtcars) #Just looking at transmission

fit1 <- lm(mpg ~ wt + cyl + disp, data = mtcars) # Only very highly correlated variables

fit2 <- lm(mpg ~ wt + hp + cyl + disp + am, data = mtcars) # Variables with correlation |0.6| or more

summary(fit0)
```

```
##
## 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 ***
## am             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
```

Looking at the first fit, the coefficients tell us that with manual transmission MPG increases by 7.245 mpg howeveer the adjusted R^2 value in this case is only 0.3385. So the coefficients obtained here seem to be biased

If we fit the model to only the very highly correlated variables we get the following

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4035 -1.4028 -0.4955  1.3387  6.0722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.107678   2.842426  14.462 1.62e-14 ***
## wt          -3.635677   1.040138  -3.495  0.00160 **
## cyl         -1.784944   0.607110  -2.940  0.00651 **
## disp         0.007473   0.011845   0.631  0.53322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.595 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.8326, Adjusted R-squared:  0.8147
## F-statistic: 46.42 on 3 and 28 DF,  p-value: 5.399e-11
```

The adusted R^2 value has increased to 0.8147 with negative coefficients for weight and cylinders meaning that they will decrease our mileage values

If we combine the two models and do an anova on the three:

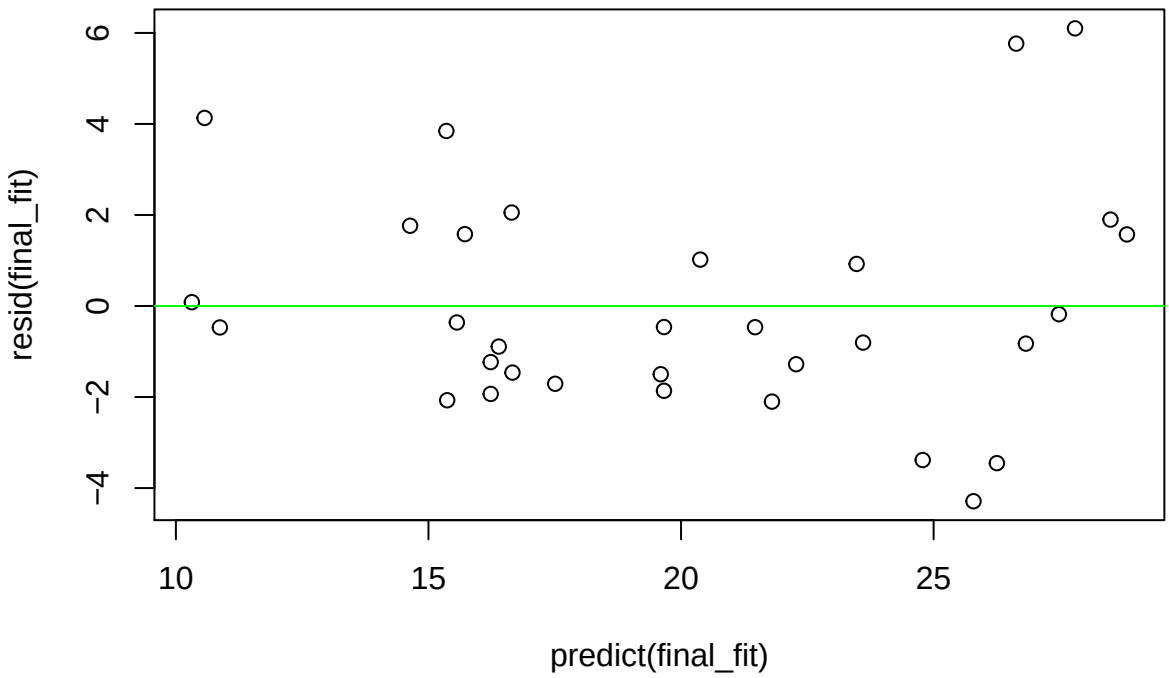
```
anova(fit0, fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + cyl + disp
## Model 3: mpg ~ wt + hp + cyl + disp + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 188.49  2    532.40 42.4305 6.494e-09 ***
## 3     26 163.12  2     25.37  2.0221    0.1527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the F value here is only significant for fit1 so that would be our best indicator for mpg, if we remove the disp variable.

So our new model will be the following, and we can look at a residual plot as well:

```
final_fit <- lm(mpg ~ wt + cyl, data = mtcars)
plot(predict(final_fit), resid(final_fit))
abline(h=0, col = "green")
```



The residual plot does not have a distinctive pattern so we can assume the model is linear

Conclusion

From our final model above we have learned that MPG is mostly influenced by vehicle weight and number of cylinders. Manual transmission is better than auto transmission for MPG, however with the given data we cannot determine the difference between the two types of transmissions on MPG.