Regression_Mtcars

Emmanuel Okyere Darko

3/28/2021

Excutive Summary

This project will focus on a data collection of cars, exploring the relationship between a set of variables and miles per gallon (MPG) as outcome variable. Interested questions for this project are

- Is automatic or manual transmission better for MPG?
- Quantify the MPG difference between manual and automatic transmission

---

> Data Description A data frame with 32 observations on 11 (numeric) variables.

- [, 1] mpg Miles/(US) gallon
- [, 2] cyl Number of cylinders
- [, 3] disp Displacement (cu.in.)
- [, 4] hp Gross horsepower
- [, 5] drat Rear axle ratio
- [, 6] wt Weight (1000 lbs)
- [, 7] qsec 1/4 mile time
- [, 8] vs Engine (0 = V-shaped, 1 = straight)
- [, 9] am Transmission (0 = automatic, 1 = manual)
- [,10] gear Number of forward gears
- [,11] carb Number of carburetors

Exploratory Analysis

```
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 ...
```

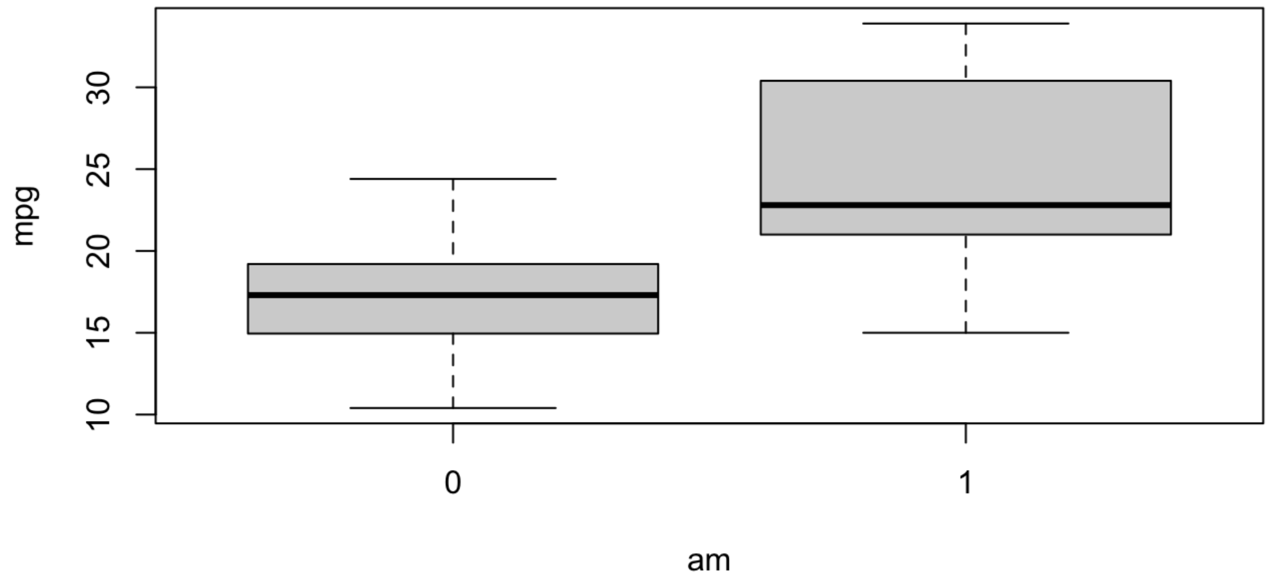
> Change data types

- Cylinder is a count variable, therefore convert to factor `4, 6 or 8`
- gear also as factor varible (3,4 or 5)
- Transmission (am) as factor - `(0 = automatic, 1 = manual)`

```
mtcars[,c(2,9,10)] <- sapply(mtcars[,c(2,9,10)] , factor)
```

Since there are no complete overlaps in the boxplot below, Transmission type appears to have better relationship with mpg

```
with(mtcars, boxplot(mpg~am))
```



Model

We will fit a linear linear model with only cylinder as the predictor varible

Transmission (am) as factor - `(0 = automatic, 1 = manual)`

```
fit.cyl <- lm(mpg~am, data= mtcars)
summary(fit.cyl)
```

```
##
## 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 ***
## am1            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
```

Transimission type is a significant predictor of MPG with based on the p-values. Automatic transmission appears to consume fewer MPG with Manual transmission expected consume and 7.2 more gallons per 1 mile. However, the low R-Squared shows that approximately `36%` of the variation in the model is explained by the Transmission variable and our co-efficients may not be reliable.

Let’s compare this to a full model

```
fit.full <- lm(mpg~., data= mtcars)
summary(fit.full)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.2015 -1.2319  0.1033  1.1953  4.3085 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 15.09262   17.13627   0.881   0.3895  
## cyl6        -1.19940    2.38736  -0.502   0.6212  
## cyl8         3.05492    4.82987   0.633   0.5346  
## disp         0.01257    0.01774   0.708   0.4873  
## hp          -0.05712    0.03175  -1.799   0.0879 .
## drat         0.73577    1.98461   0.371   0.7149  
## wt          -3.54512    1.90895  -1.857   0.0789 .
## qsec         0.76801    0.75222   1.021   0.3201  
## vs           2.48849    2.54015   0.980   0.3396  
## am1          3.34736    2.28948   1.462   0.1601  
## gear4       -0.99922    2.94658  -0.339   0.7382  
## gear5        1.06455    3.02730   0.352   0.7290  
## carb         0.78703    1.03599   0.760   0.4568  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.616 on 19 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8116 
## F-statistic: 12.13 on 12 and 19 DF,  p-value: 1.764e-06
```

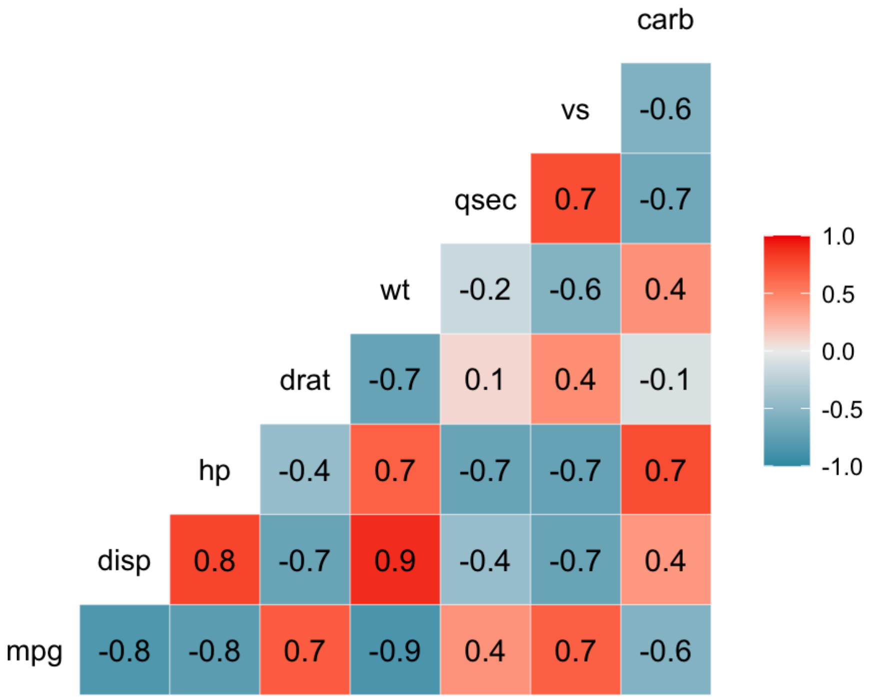
The full model gives a relatively high Adjusted R-squared `88.5%` but has less significant variables, implying that there could be overfitting.

> Selecting the best model All numerical variables have at least moderate correlation with mpg

```r
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```r
df_cor <- na.omit(mtcars[, c(1,3:8,11)])
ggcorr(df_cor, label = T)
```



We can use the step fubction to select the best model, we will use backward stepwise regression, removing one variable a time untill we reach the best model (when AIC stops decreasing)

```
# To see output, specify trace = True
fit.best <- step(fit.full, direction = "backward", trace = F)
summary(fit.best)
```

```
##
## 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 ***
## am1           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
```

Based on the AIC, the best model has predictors: `wt, am, qsec` Now we can fit a parsimonious model. This model now gives a Adjusted Rsq of 83.3 with a low p-value indicatiing the model significance.

Hypothesis test for model selection This test also suggest that the best.fit model is significant than the full model

```
anova(fit.cyl, fit.best, fit.full)
```

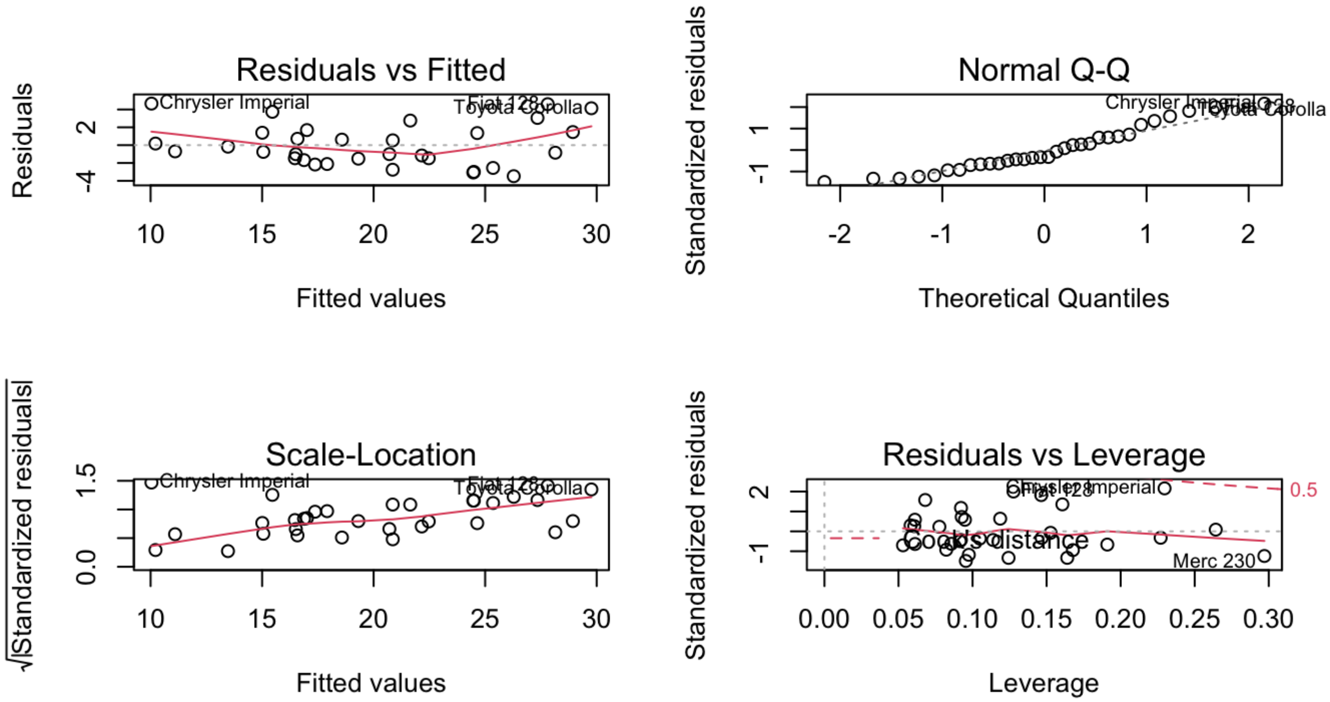
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 40.2941 1.463e-07 ***
## 3     19 130.05  9     39.23  0.6369    0.7524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Diagnoistics

- The model diagnostics shows that the best fit model has linear pattern, indicating the linearity was captured by the mode
- residuals appear normaly distributed as shows on the Normal-Q-Q plot as well
- The Scale-Location shows that the best fit model has approximately constant variance
- Residual vs Leverage captures influetial points, in particular, there is no outliers in our model

You can read more here (https://data.library.virginia.edu/diagnostic-plots/)

```
par(mfrow = c(2,2))
plot(fit.best)
```



Sum of residual is also zero, all assumptions for a linear regression are adequately satisfied

```
sum(fit.best$residuals)
```

```
## [1] 1.665335e-15
```

Conclusion

- Is automatic or manual transmission better for MPG? This question can be answered based on the model, that Automatic transmission appears to consume fewer MPG with Manual transmission expected consume additional 2.9 MPG
- Quantify the MPG difference between manual and automatic transmission could be hard to answer as different models could predict different results