Final Projecct - Regression Model Analysis (mtcars)

July 23, 2019

Synopsis (Executive Summary)

In this project we work on the `mtcars` data set and we want to explore how miles per gallon (MPG) as the outcome variable is affected by different variables, specifically evaluate the effect of automatic and manual transmissions on dependent variable MPG. The following two questions will be answered in this project:

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

Loading data and setting up environment

```
library(datasets)
data(mtcars)
library(ggplot2)
```

Describing variables in data

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. This means that the data consists of 32 observations on 11 variables.

- **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

Viewing Data Structure

Viewing structure of variables in "mtcars" data:

```
dim(mtcars)  ## 32 observations and 11 variables
```

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

```
head(mtcars) ## some observations to better understand 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) ## variable types after coersion
```

```
## '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 ...
```

Data Processing

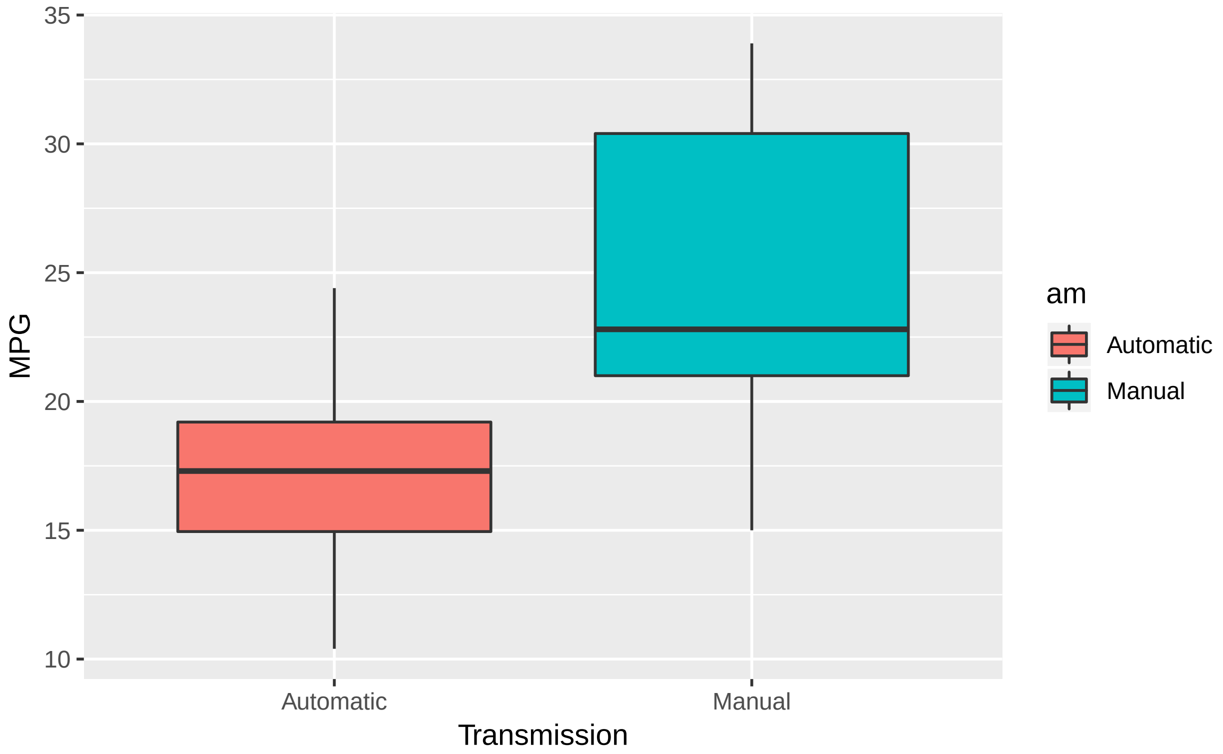
Changing categorical variables to factors, and relabeling `am` variables to `Automatic` and `Manual` for more clearancy.

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

Visualizations

Plotting the miles per gallon (MPG) for automatic and manual transmissions.

```
plot1 <- ggplot(mtcars, aes(x=am, y=mpg)) +
    geom_boxplot(aes(fill = am)) +
    xlab("Transmission") +
    ylab("MPG")
plot1
```



## Analysis

It looks like there is a definite difference in the type of transmission for MPG. Performing a t-test will help verify if the difference in means is significant.

```
auto_vs_manu_ttest <- t.test(mpg ~ am, mtcars)
auto_vs_manu_ttest
```

```
##
##  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
```

The t-test rejected the null-hypothesis that the difference in means is equal to zero, with a p-value of .0014. Therefore there is a difference in transmission type, with manual transmissions having a higher MPG.

Models of Regression Analysis

Since the project is trying to quantify the difference in MPG for automatic and manual transmissions. The best starting place is a simple linear model with transmission type as the dependent variable.

Model 1: Fitting model with just one variable (simple model):

```
basic_fit <- lm(mpg ~ am, mtcars)
summary(basic_fit)$coefficients
```

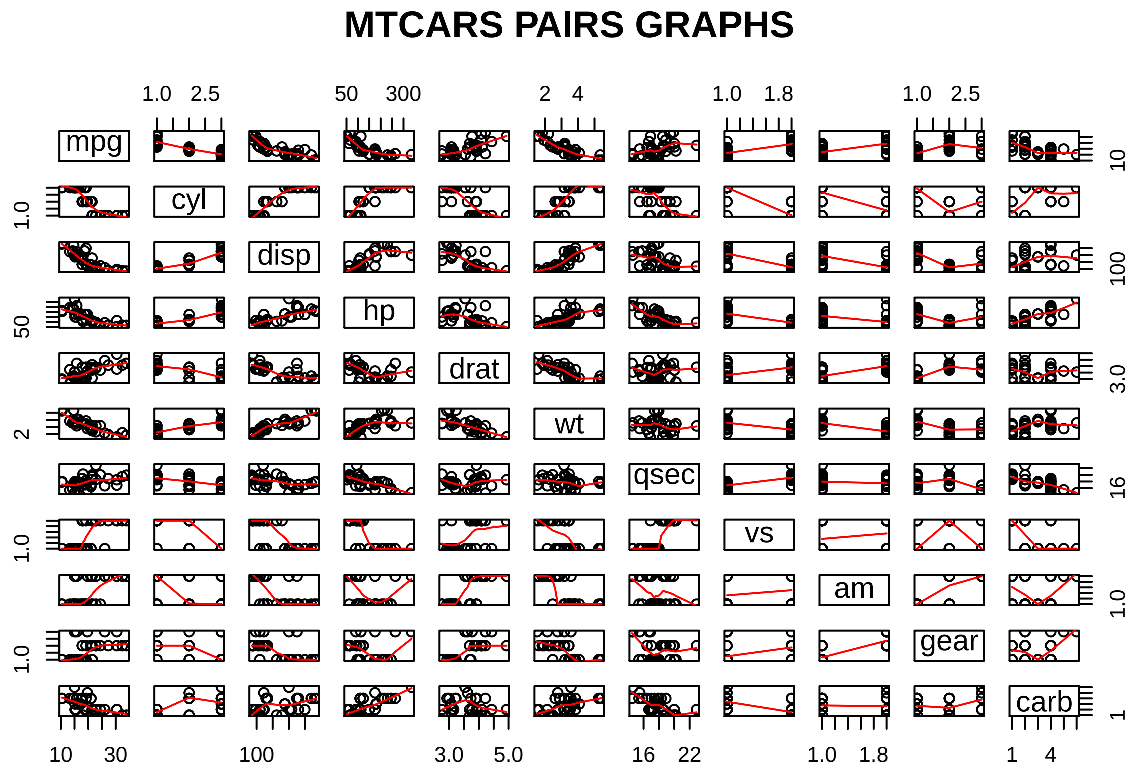
```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(basic_fit)$r.squared
```

```
## [1] 0.3597989
```

The basic linear model with `am` as the only regressor explains 36% of the variation, not a very good model. To find a model which captures more variance (better R-Squared), while significant p-values, we need to add more than just one variable to the model. However, this could be tricky too! Since the regressors can correlate with not only the predictor, but also the other regresors.

```
pairs(mtcars, panel = panel.smooth, main = "MTCARS PAIRS GRAPHS")
```



As he graph shows that MPG has correlations with other variables than just am. To obtain a more accurate model, we need predicting MPG in correlation with other variables than am. Lets use some models to evaluate the correlations.

Model 2: Fitting model using all the variables (general model-fitting all variables):

We use all variables in mtcars data to fit a general model to see R-Squared and P-Values for all variables in this case:

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

summary(fitall )$coefficients
```

```
##                 Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## amManual     1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

```r
summary(fitall)$r.squared
```

```
## [1] 0.8930749
```

As the sumamry of the fittall model shows: R-Squared value has improved about 80%, but is not able to describe the remaining variance of the MPG variable. On the other hand as the p-values show in the summary, there is no coefficient signifinact at 0.05 level. SO that we can not get a relaible conclusion about this model and its coefficients, too. Therefore, we have too search more and find a better model in between of these two models.

Model 3: Using STEP function in R

In order to do variable selection we use R-function STEP, to do variable selection.

```r
everything_fit <- lm(mpg ~ ., mtcars)
step_fit <- step(fitall,direction="both",trace=FALSE)
summary(step_fit)$coefficients
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

```r
summary(step_fit)$r.squared
```

```
## [1] 0.8658799
```

As the coeffs summary shows, the p-value column does not show significancy for coefficients in all considered variables. SO that we may want to find a better model with significant p-values, while keeping the high adjusted R-squared value in new model!

Model 4: Last Model Examination

Looking at the p-values and standard error for all variables given in general model (model 2), we decide to evaluate the following model and see the significance of p-value and R-squared in this model:

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

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

```
summary(lastModel)$r.squared
```

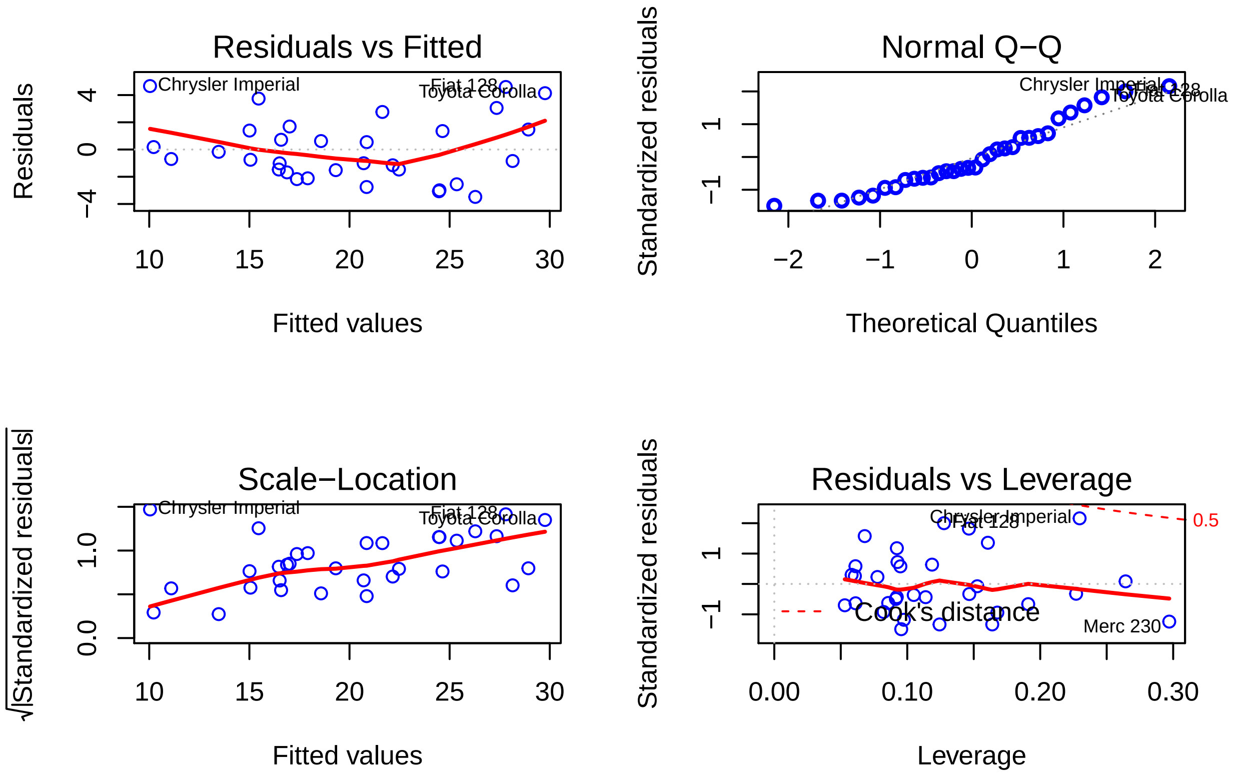
```
## [1] 0.8496636
```

As the summary of the model shows, the p-values corresponding to all the coefficients show significant at level of 0.05. Moreover, the adjusted r-squared covers good amount of variability in the model which is about 85%.

Residual Analysis:

Everything so far looks solid, but lets make sure this model fits our data well by printing the diagnostic plots.

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



The summary of the above diagnostic plots demonstrates the following: - 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.

Conclusion:

Our first exploratory data analysis plot (using box plot), along with running t-test to confirm our observation, all all show that best transmission type for MPG would have to be the manual transmission. The summary results in our last model, demonstrates that by having a manual transmission instead of an automatic the MPG will increase by 2.94 as can be seen in the best model's `amManual` coefficient. The model fit well with a $p < 0.05$ and and $R^2 = 0.85$.

The diagnostic plots on the other hand show 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`.