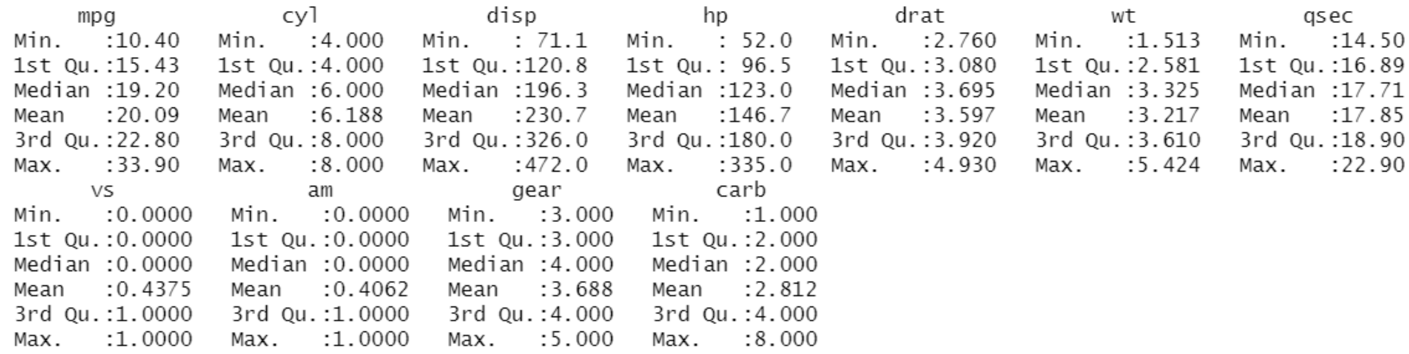
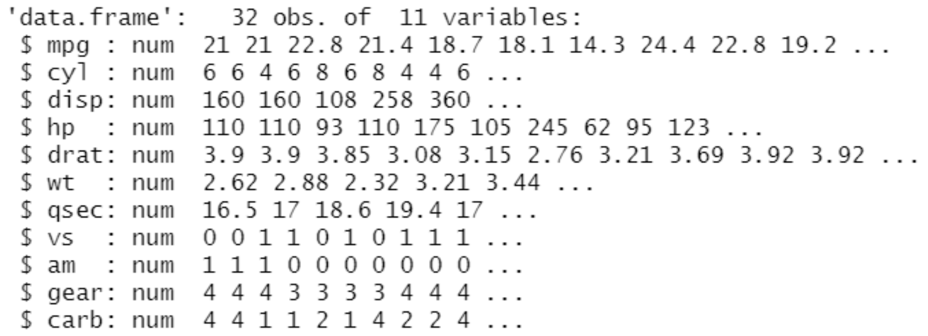
~~~
---
title: "Week 4 Project -- Regression Analaysis"
author: "Dhruv Singh"
date: "February 16, 2020"
output:
  pdf_document: default
  html_document: default
---
## PART 0: SETUP
echo settings for embedding code
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
Setting Directory
```{r dir}
getwd()
setwd("C:/Dhruv/misc/data/R_7_regression_models/wk4_logistic_reg_poisson_reg")
```
[1] "C:/Dhruv/misc/data/R_7_regression_models/wk4_logistic_reg_poisson_reg"

## Step 1: Coefficients
Loading and checking mtcars data
```{r mtcars}
data("mtcars")
summary(mtcars)
str(mtcars)
```
~~~

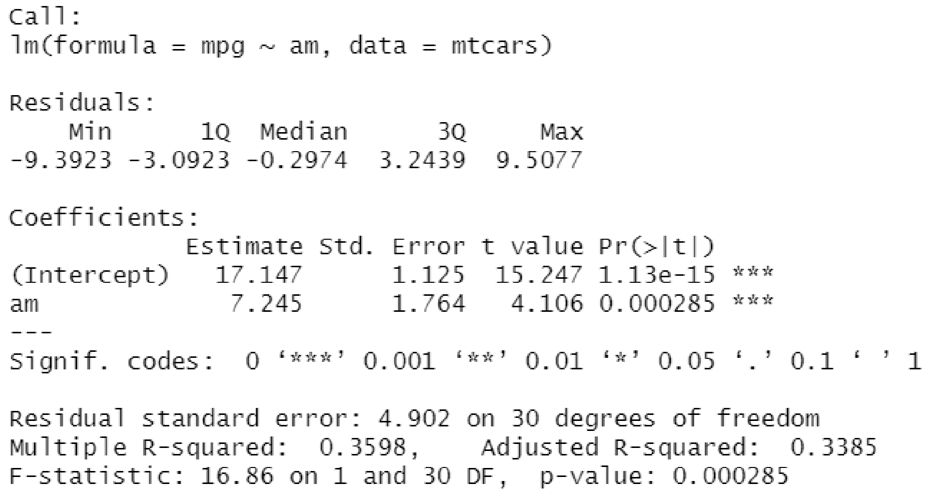


```
      mpg             cyl             disp             hp             drat             wt             qsec      
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325   Median :17.71  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90  
       vs               am              gear            carb      
 Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```



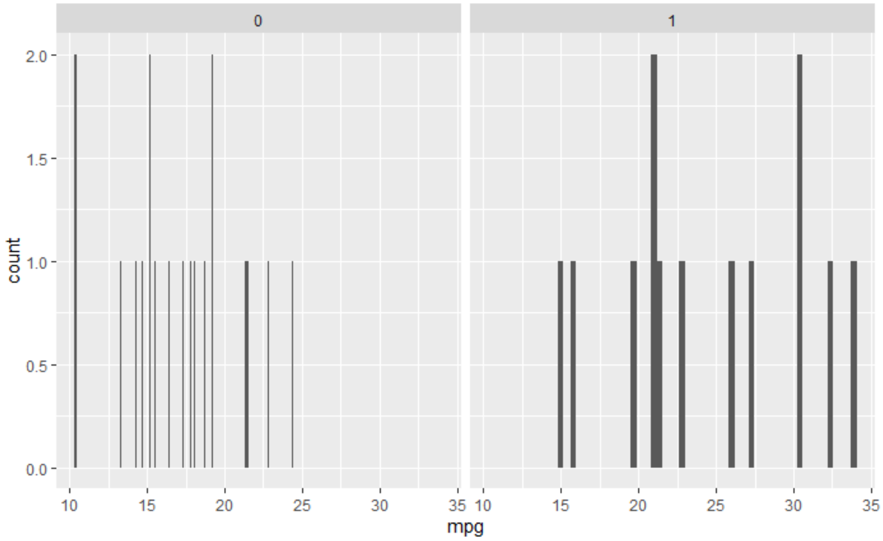
```
'data.frame':	32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ 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 ...
```

```{r am}

fit <- lm(mpg ~ am, mtcars)

summary(fit)

# a simple two variable regregression reveals that am has a significant bearing on mpg

# binary input, 0: automatic, 1: manual

# manual is related to 7 more miles per gallon on average

```

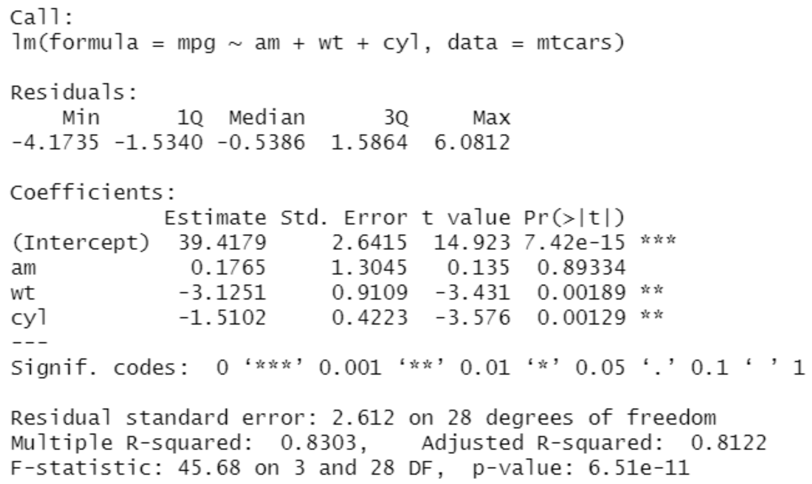


```
Call:
lm(formula = mpg ~ am, data = mtcars)

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
```

~~~
## Step 2: Exploratory data analysis

```{r eda}

library(ggplot2)

# plotting mpg against wt

p1 <- ggplot(mtcars, aes(x = mpg)) + geom_bar()

p1 + facet_wrap(~am)

# from the graph below it appears that on average, manual cars yield higher miles per gallon than automatic, counterintuitively

```
~~~



~~~
## Step 3: Model fitting

```{r regression model}

# model 1

fit1 <- lm(mpg ~ am, mtcars)

summary(fit1)

# model 2, seems to explain away the change attributable to am

# and instead attributes it to weight, and cylinders

fit2 <- lm(mpg ~ am+wt+cyl, mtcars)

summary(fit2)

```
~~~



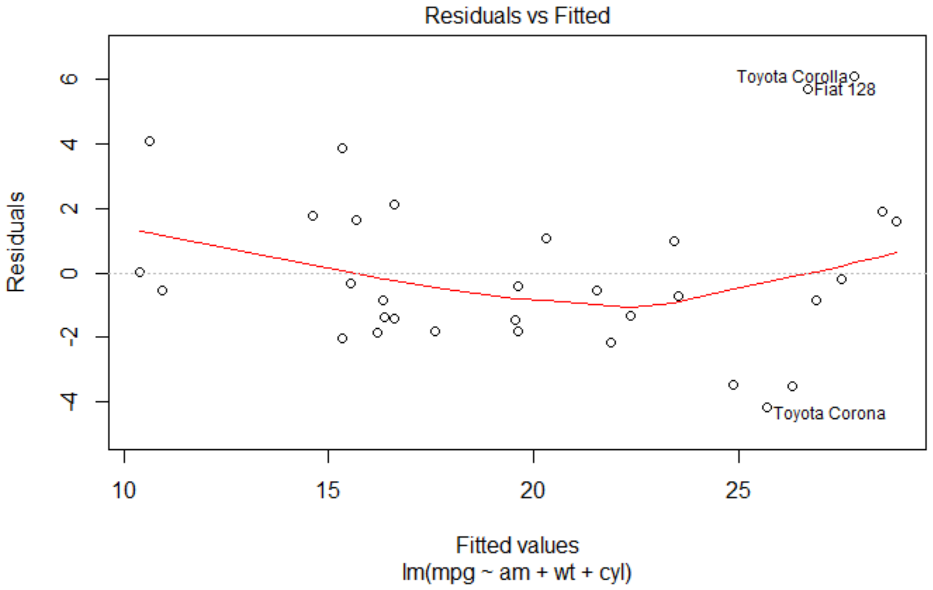
```
Call:
lm(formula = mpg ~ am + wt + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1735 -1.5340 -0.5386  1.5864  6.0812 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
am            0.1765     1.3045   0.135  0.89334    
wt           -3.1251     0.9109  -3.431  0.00189 ** 
cyl          -1.5102     0.4223  -3.576  0.00129 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.612 on 28 degrees of freedom
Multiple R-squared:  0.8303,	Adjusted R-squared:  0.8122 
F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

## Step 4: questions of interest

```{r }

# thus we can see that after controlling for other related variables such as weight and cylinders

# the size of the effect of automatic vs manual reduces, and is no longer significant

```

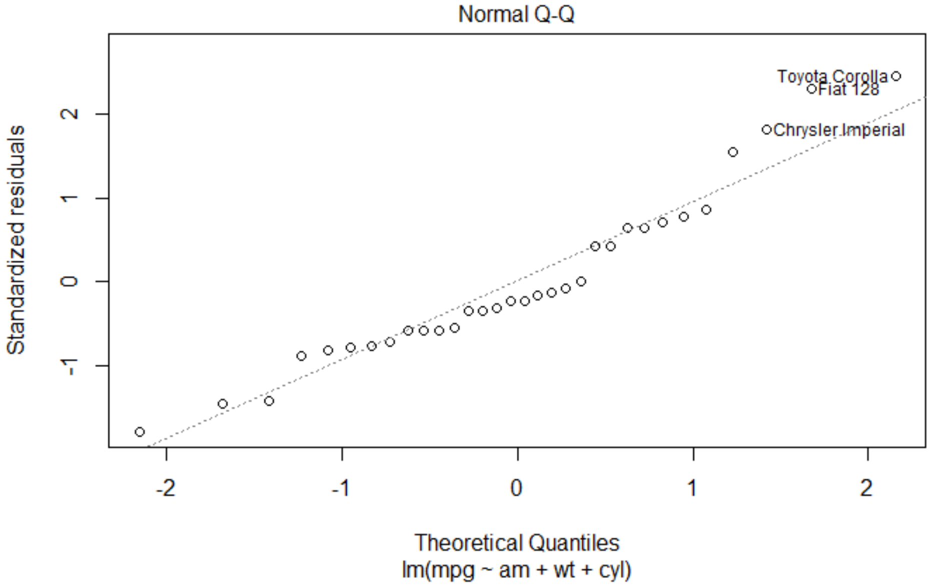
## Step 5: residual plot

```{r residual plot}

plot(fit2, which = 1)

```



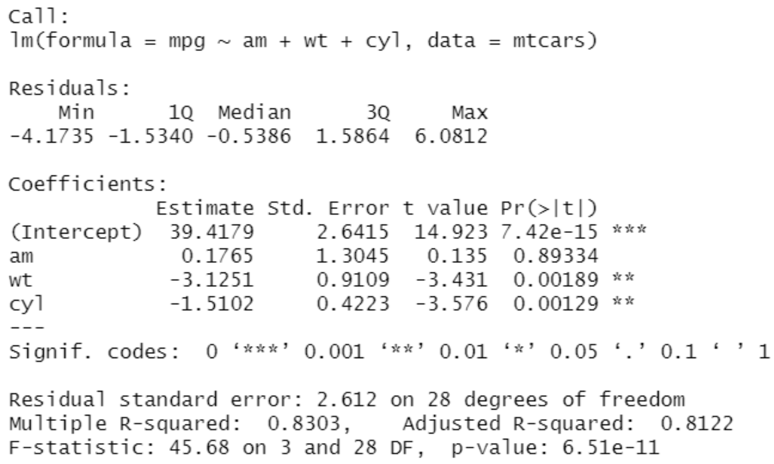
## Step 5: diagnostic plot

```{r residual plot}
plot(fit2, which = 2)
```



## Step 6: inference, uncertainty

```{r inference}
summary(fit2)
# std. error of am is 1.3 and is larger than its coefficient of 0.179
# which is clearly indication that the am predictor is not significant
```



```
Call:
lm(formula = mpg ~ am + wt + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1735 -1.5340 -0.5386  1.5864  6.0812 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
am            0.1765     1.3045   0.135  0.89334    
wt           -3.1251     0.9109  -3.431  0.00189 ** 
cyl          -1.5102     0.4223  -3.576  0.00129 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.612 on 28 degrees of freedom
Multiple R-squared:  0.8303,	Adjusted R-squared:  0.8122 
F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

## Step 7: Report Length Criteria - 2 pages or more

## Step 8: Executive Summary

```{r executive summary}

# The model fit summary and related diagnostics are a clear indication that in order to select our predictor variables

# carefully, we can turn to a variety of methods.

# Some of these include factor analysis, as a form of unsupervised learning.

# but also vif factors, to indicate which coefficients have a larger or smaller effect on the outcome

# and helps parse out autocorrelation, that is within model correlations between coefficients.

```

## Step 9: Rmd, knitr

```{r rmd knitr}

# code all written in rmd, as visible by the code chunks

# knitr package on available for installation on system

# however, i have used it before and have published to rpubs for prev assignments

# from a diff machine.

# thanks!

```