Multiple Linear Regression Case Study

proschool
An IMS Initiative

Contents

Description of the data

The Dataset *mtcars* is available as part of the R datasets and was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The objective of this analysis is to see how we could determine the mpg (Miles/gallon) based on other aspects of the car. The variables available in the dataset are given below.

Variable description:

Variable	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburettors

The *head* function displays the top 10 records of the data passed

```
> 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
```

To know the variables type we can use the function *str.* It provides the type of variables we are dealing with. It can be observed that all variables are numeric

```
> 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 ...
```

Data exploration

Summary of the data

The *summary* function provide the descriptive statistics for all the variables as show below.

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

It helps to understand the distribution of the variables & missing values if any.

Correlation analysis

Correlation measures the relative strength of the linear relationship between two variables. It ranges between –1 and 1. Closer to –1 implies negative linear relationship, closer to 1, implies stronger linear relationship & closer to 0 implies weaker linear relationship.

The *cor* function provides the correlation matrix of the data in R.

```
> cor(mtcars)
            mpg        cyl       disp         hp        drat         wt        qsec         vs          am       gear        carb
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958 -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157  0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953 -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870 -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059 -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

It can be observed almost all the variables are highly correlated among themselves

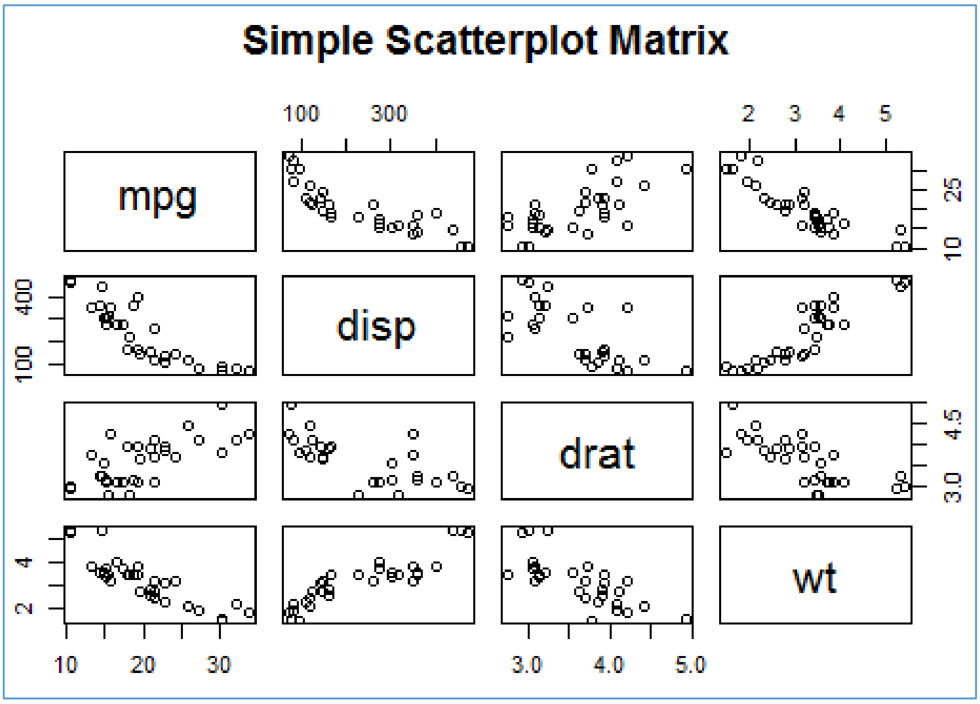
Data plots of the data

Simple scatter plots can be plot & understood using the following script. Data plots are useful to observe a pattern/distribution of the variables. Multi plots helps to understand relationship between variables. The relation technically explains change in one variable with respect to change in another variable.
In R there are plenty of options for interactive plots. It is totally up to the user which one he wants.

hist produces the histogram
plot produces the scatter plots
barplot produces the barplots
boxplot produces the boxplots

The below script plots pairwise scatterplots for the variables mentioned

```
pairs(~ mpg + disp + drat + wt, data = mtcars, main = "Simple Scatterplot Matrix")
```



Linear regression model building of the data

For building a good model, a thorough analysis of the variables should be performed. The distribution of the variables, correlation among them etc. This will help to define the formula to be used. Below we define a regression model where the mileage is regressed by the number of cylinders and the weight of the vehicle.

```
> fit <- lm(mpg~ cyl + wt, data = mtcars)
> summary(fit)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.2893 -1.5512 -0.4684  1.5743  6.1004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
cyl          -1.5078     0.4147  -3.636 0.001064 **
wt           -3.1910     0.7569  -4.216 0.000222 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.568 on 29 degrees of freedom
Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

The model explains 83% of the variation in *mpg* based on *cyl* & *wt*. Both are significant. Explanation of the output terms is given below.

Summary of model with explanation for all the statistics

- Estimated Coefficient

The estimated coefficient is the value of slope calculated by the regression. It might seem a little confusing that the Intercept also has a value, but just think of it as a slope that is always multiplied by 1

- Residuals

The residuals are the difference between the actual values of the variable you're predicting and predicted values from your regression.

- Significance Stars

The stars are shorthand for significance levels, with the number of asterisks displayed according to the p-value computed. The more punctuation there is next to your variables, the better.
Blank = bad, Dots = pretty good, Stars = good, More Stars = very good

- Standard Error of the Coefficient Estimate

Measure of the variability in the estimate for the coefficient. Lower means better but this number is relative to the value of the coefficient

- Residual Std. Error / Degrees of Freedom

The Residual Std. Error is just the standard deviation of your residuals. The Degrees of Freedom is the difference between the number of observations included in your training sample and the number of variables used in your model (intercept counts as a variable).

- R-squared

Metric for evaluating the goodness of fit of your model. Higher is better with 1 being the best. Corresponds with the amount of variability in what you're predicting that is explained by the model

- adjusted R-squared

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

- p-value

In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Model Diagnostics

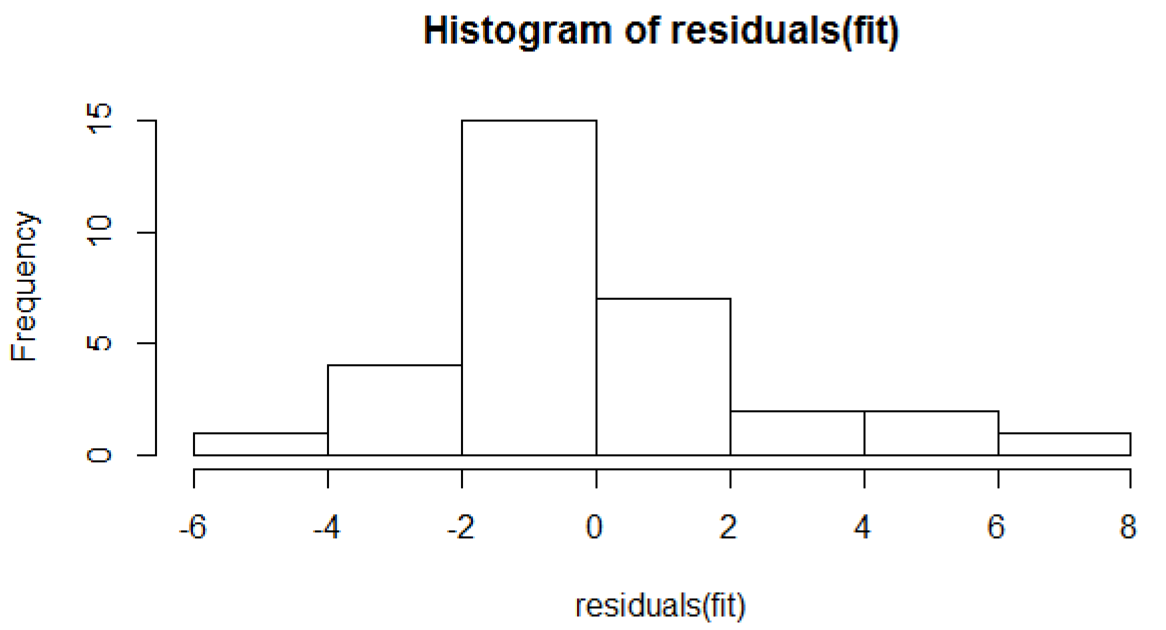
1. There should not be any autocorrelation of errors

```
> durbinWatsonTest(fit)
 lag Autocorrelation D-W Statistic p-value
   1       0.1302185      1.671096   0.284
 Alternative hypothesis: rho != 0
```

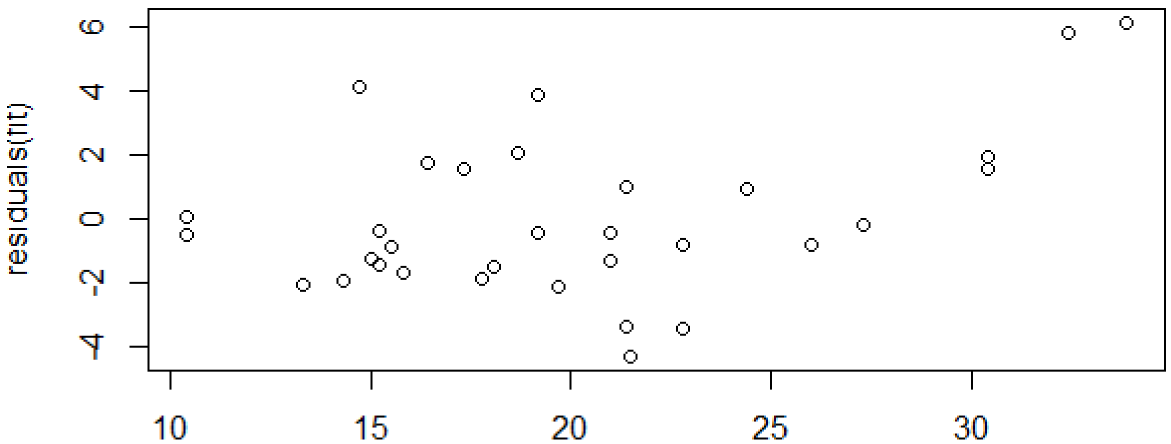
The DW values range between 0 to 4. The DW Statistic between 1.5 to 2 indicates no autocorrelation between errors. Since the DW Statistic for our model is 1.671096 we can assume there exists no autocorrelation in our errors.

2. The errors should be distributed N~(0,s^2). Homoscedasticity of variance i.e. constant variance

```
> hist(residuals(fit))
```

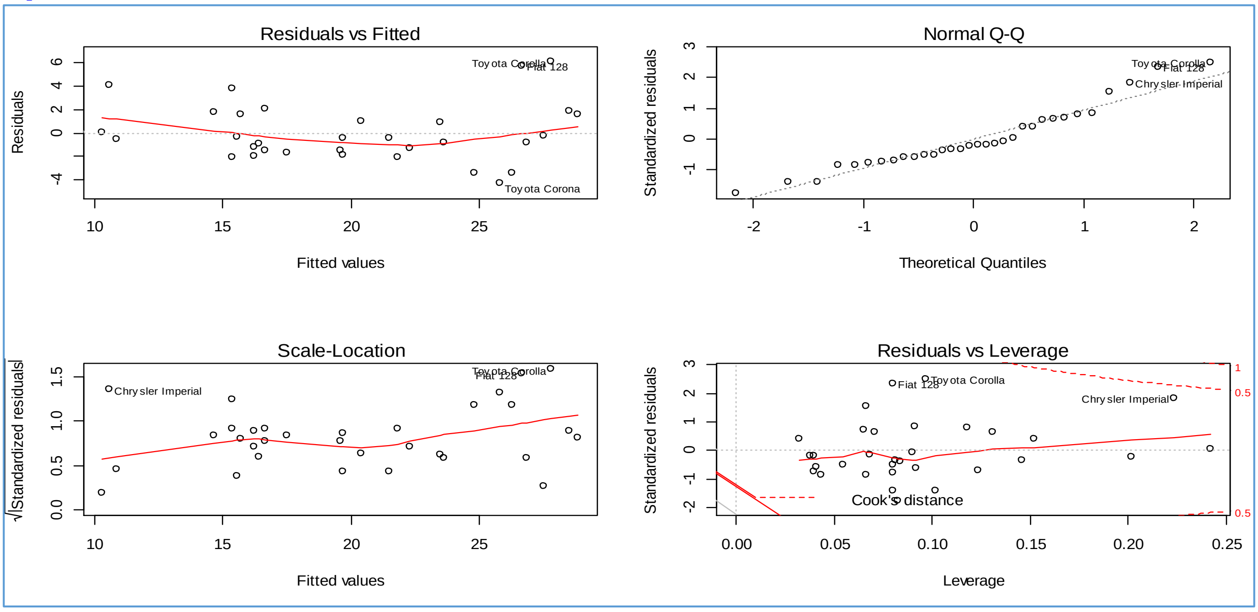



```
> plot(mtcars$mpg, residuals(fit))
```



Model plots of the data

```
> par(mfrow = c(2, 2))
> plot(fit)
```



- The first plot gives an idea of whether there is any curvature in the data. If the red line is strongly curved, a quadratic or other model may be better.
- The second plot is to check whether the residuals are normally distributed.
- The third plot is used to check if the variance is constant (i.e., if the standard deviation among the residuals appears to be about constant).
- The last plot is used to check to see if there were any overly influential points.