Exp:2- Correlation and Regression

Correlation Coefficient

The **correlation coefficient** of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related.

Karl Pearson's Correlation

The most commonly used type of correlation is Pearson correlation, named after Karl Pearson, introduced this statistic around the turn of the 20th century.Pearson's r(x ,y) measures the linear relationship between two variables, say X and Y. A correlation of 1 indicates the data points perfectly lie on a line for which Y increases as X increases. A value of -1 also implies the data points lie on a line; however,Y decreases asX increases. The formula for r(x,y) is

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

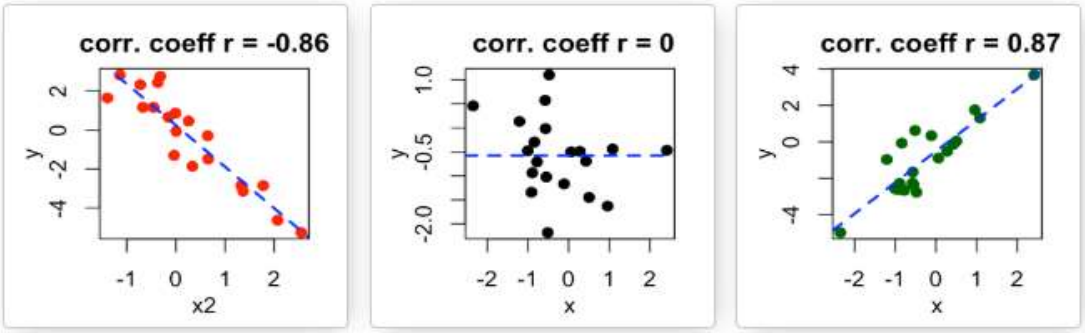
Also,if the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

The Pearson correlation has two assumptions:

1.The two variables are normally distributed. We can test this assumption using
   a.A statistical test (Shapiro-Wilk)
   b.A histogram
   c.A QQ plot

2.The relationship between the two variables is linear. If this relationship is found to be curved, etc. we need to use another correlation test. We can test this assumption by examining the scatterplot between the two variables.

Correlation coefficient is comprised between -1 and 1:



R-code:

To calculate Pearson correlation, we can use the cor() function. The default method for cor( ) is the Pearson correlation. Getting a correlation is generally only half the story, and you may want to know if the relationship is statistically significantly different from 0.

- H_0: There is no correlation between the two variables: $\rho = 0$
- H_a: There is a nonzero correlation between the two variables: $\rho \neq 0$

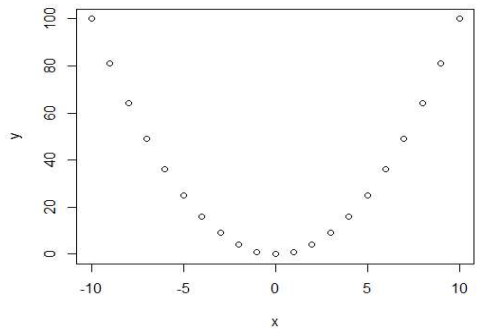
To assess statistical significance, you can use cor.test() function.

Example-1

How the Pearson measure is dependent on the data distribution assumptions (in particular linearity), observe the following deterministic relationship:

$$y = x^2.$$

```
> x=seq(-10,10,1)
> y=x^2
> plot(x,y)
> cor(x,y)
[1] 0
```



The R code below computes the correlation between mpg and wt variables in mtcars data set:

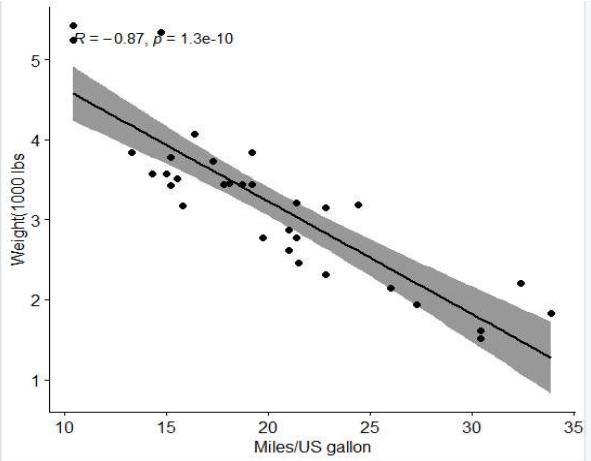
```
> my_data=mtcars
```

```
> head(my_data,6)
                   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
```

Visualize your data using scatter plots

To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll use the ggpubr R package.

```
> library("ggpubr")
>ggscatter(my_data,x="mpg",y="wt",add="reg.line",conf.int=TRUE,cor.coef=TRUE
,cor.method="pearson",xlab="Miles/US gallon",ylab="Weight(1000 lbs")
```



Preleminary test to check the test assumptions

1. Is the covariation linear? Yes, from the plot above, the relationship is linear. In the situation where the scatter plots show curved patterns, we are dealing with nonlinear association between the two variables.
2. Are the data from each of the 2 variables (x, y) follow a normal distribution?
   - Use Shapiro-Wilk normality test –> R function: shapiro.test()
   - and look at the normality plot —> R function: ggpubr::ggqqplot()

Shapiro-Wilk test can be performed as follow:

- Null hypothesis: the data are normally distributed
- Alternative hypothesis: the data are not normally distributed

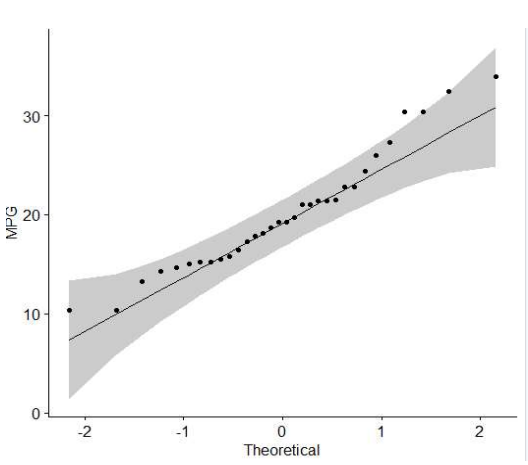
```
# Shapiro-Wilk normality test for mpg
>shapiro.test(my_data$mpg)    # => p = 0.1229
# Shapiro-Wilk normality test for wt
```

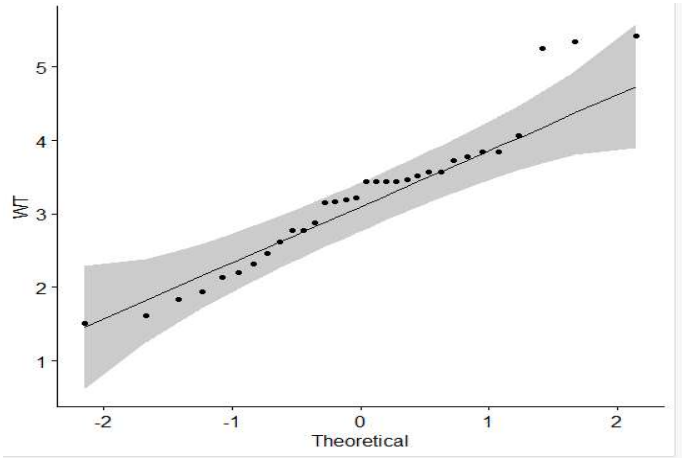
shapiro.test(my_data$wt) # => p = 0.09

From the output, the two p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

Visual inspection of the data normality using **Q-Q plots** (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution.

```
>ggqqplot(my_data$mpg,ylab ="MPG")
>ggqqplot(my_data$wt, ylab = "WT")
```





From the normality plots, we conclude that both populations may come from normal distributions.

Note that, if the data are not normally distributed, it's recommended to use the non-parametric correlation, including Spearman and Kendall rank-based correlation tests.

Pearson correlation test

Correlation test between mpg and wt variables:

```
> cor.test(my_data$wt,my_data$mpg,method="pearson")

        Pearson's product-moment correlation

data:  my_data$wt and my_data$mpg
t = -9.559, df = 30, p-value = 1.294e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9338264 -0.7440872
```

```
sample estimates:
       cor
-0.8676594
```

In the result above :

- t is the t-test statistic value (t = -9.559),
- df is the degrees of freedom (df= 30),
- p-value is the significance level of the t-test (p-value = 1.29410^{-10}).
- conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [-0.9338, -0.7441]);
- sample estimates is the correlation coefficient (Cor.coeff = -0.87).

Interpretation of the result

The **p-value** of the test is 1.29410^{-10}, which is less than the significance level alpha = 0.05. We can conclude that wt and mpg are significantly correlated with a correlation coefficient of -0.87 and p-value of 1.29410^{-10} .

Spearman's Correlation Coefficient in R

When the data contains no ties,ρ can be found by taking the difference of the ranked values using the following equation:

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right] \qquad \text{[Read the symbol (as 'Rho'.]}$$

Where, $\sum d^2$ = Sum of squares of differences of ranks between paired items in two series n = Number of paired items'

When the data contain ties, the following equation can be used:

$$\rho = 1 - \frac{6(\sum d^2 + C.F_i)}{n(n^2-1)}$$

Spearman's rank correlation coefficient can easily be calculated by again using the cor.test() function in R. The only difference from the Pearson example is the method argument.

```
> cor.test(x=cars$speed,y=cars$dist,method='spearman')

        Spearman's rank correlation rho

data:  cars$speed and cars$dist
S = 3532.8, p-value = 8.825e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8303568
```

Warning in cor.test.default(x = cars$speed, y = cars$dist, method = "spearman"): Cannot compute exact p-value with ties.

Spearman's ρ is reported as .83, slightly higher than Pearson's r of .80, indicating a stronger correlation than previously estimated with Pearson's correlation. Since Pearson's r measures the linear association between two variables, it is likely the outlier in the top right corner of the first graph was causing under-reported values as linear relationships can be extremely affected by outliers.

Obtain the rank correlation coefficient for the following data :

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

```
> x=c(68,64,75,50,64,80,75,40,55,64)

> y=c(62,58,68,45,81,60,68,48,50,70)
> cor.test(x,y,method="spearman")

        Spearman's rank correlation rho

data:  x and y
S = 73.326, p-value = 0.09542
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5555979
Warning message:

In cor.test.default(x, y, method = "spearman") :
  Cannot compute exact p-value with ties

> cor.test(x,y,method="spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  x and y
S = 73.326, p-value = 0.09542
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5555979
```

Linear Regressions

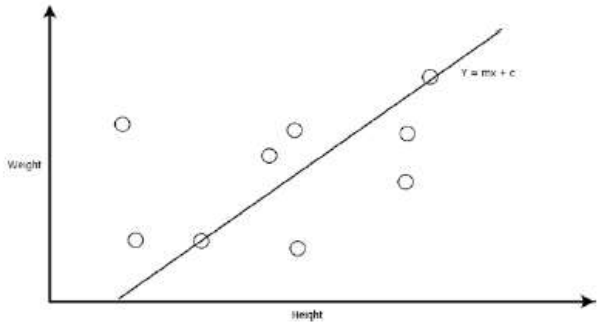
Regression analysis is a form of predictive modelling techniques that identify the relationships between dependent and independent variables(s). The technique is used to find causal effect relationships between variables.

The benefit of using regression analysis is that it identifies the significant relationships between dependent and independent variables and the strength of the impact of multiple independent variables on independent variables.

Linear regression finds the relationship between one dependent variable and one independent variable using a regression line.

The linear regression equation is $y = b_0 + b_1 x$

Where y is the dependent variable and x is the independent variable.



To calculate the slope, you can use

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

To calculate the intercept, you can use

$$b_0 = \bar{y} - b_1 \bar{x}$$

If $b_1 > 0$, x and y have a positive relationship.
If $b_1 < 0$, x and y have a negative relationship.

Steps to Establish a Regression

A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.

The steps to create the relationship is −

- Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation using these
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the weight of new persons, use the **predict()** function in R.

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for **lm()** function in linear regression is −

lm(formula,data)

Following is the description of the parameters used −

- **formula** is a symbol presenting the relation between x and y.
- **Data** is the vector on which the formula will be applied.

Create Relationship Model & get the Coefficients

```
> x <- c(151,174,138,186,128,136,179,163,152,131)

> y <- c(63,81,56,91,47,57,76,72,62,48)
> relation=lm(x~y)
> print(relation)

Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
     61.380        1.415

> print(summary(relation))
```

```
Call:
lm(formula = x ~ y)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0529 -2.4833 -0.0912  1.3774 10.0562

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.3803     7.2653   8.448 2.94e-05 ***
y             1.4153     0.1089  12.997 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.712 on 8 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

```
> cor(x,y)
[1] 0.9771296
```

Therefore the regression line is y=61.3803+1.4153x where slope is 1.4153 and intercept is 61.3803.Also there is a positive correlation between x and y.

Interpretation of R-Squared

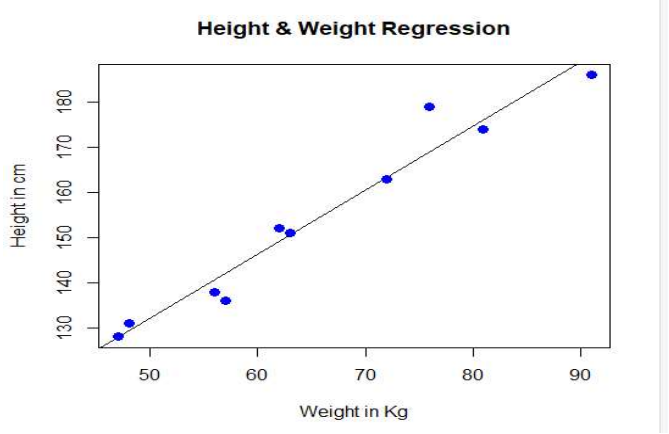
R-squared can take any values between 0 to 1. The R-squared value from the summary is 0.9548, suggesting (correctly here) that X is a good predictor of Y. That is an r-squared of 95% reveals that 95% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

However, it is not always the case that a high r-squared is good for the regression model. The quality of the statistical measure depends on many factors, such as the nature of the variables employed in the model, the units of measure of the variables, and the applieddata transformation. Thus, sometimes, a high r-squared can indicate the problems with the regression model.

A low r-squared figure is generally a bad sign for predictive models. However, in some cases, a good model may show a small value.

Visualize the Regression Graphically

```
> png(file="linear regression.png")
> plot(y,x,col="blue",main ="Height & Weight Regression",abline(lm(x~y)),cex =
1.3,pch = 16,xlab ="Weight in Kg",ylab ="Height in cm")
```



R - Multiple Regression

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is −

y = a + b1x1 + b2x2 +...bnxn

Following is the description of the parameters used −

- **y** is the response variable.
- **a, b1, b2...bn** are the coefficients.
- **x1, x2, ...xn** are the predictor variables.

We create the regression model using the **lm()** function in R. The model determines the value of the coefficients using the input data. Next we can predict the value of the response variable for a given set of predictor variables using these coefficients.

lm() Function: This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for **lm()** function in multiple regression is −

```
lm(y ~ x1+x2+x3...,data)
```

Following is the description of the parameters used −

- **formula** is a symbol presenting the relation between the response variable and predictor variables.
- **Data** is the vector on which the formula will be applied.

Example:

```
input <- mtcars[,c("mpg","disp","hp","wt")]
```

```
> model=lm(mpg~disp+hp+wt,data=input)
>
> model

Call:
lm(formula = mpg ~ disp + hp + wt, data = input)

Coefficients:
(Intercept)         disp           hp           wt
  37.105505    -0.000937    -0.031157    -3.800891
```

Based on the above intercept and coefficient values, we create the mathematical equation.

```
Y = 37.15+(-0.000937)*x1+(-0.0311)*x2+(-3.8008)*x3
```

We can use the regression equation created above to predict the mileage when a new set of values for displacement, horse power and weight is provided.

For a car with disp = 221, hp = 102 and wt = 2.91 the predicted mileage is −

```
Y = 37.15+(-0.000937)*221+(-0.0311)*102+(-3.8008)*2.91 = 22.7104
```

References:

1. Correlation. (n.d.). Retrieved from www.mathsisfun.com/data/correlation.html.
2. Covariance. (n.d.). Retrieved from http://mathworld.wolfram.com/Covariance.html.
3. Kabacoff, R. (n.d.). Multiple (Linear) Regression. Retrieved from www.statmethods.net/stats/regression.html.