INTRODUCTION TO DATA ANALYTICS

REPORT

Pearson's Product-Moment Correlation analysis for paired data

INSTRUCTOR: **DR.SREEJA SR**

DONE BY

GROUP-13

NAME	ROLL NO
MANAM SAMPATH KUMAR REDDY	S20190010114
LIKHITH BHARADWAJ ANNEPU	S20190010008
A NAGA SAI LOKESH REDDY	S20190010005
A PRANAY	S20190010001
ARYAN NIKHIL PHULKAR	S20190010011

INDEX

NOTE:- ALL WORKS ARE EQUALLY DISTRIBUTED AMONG US.

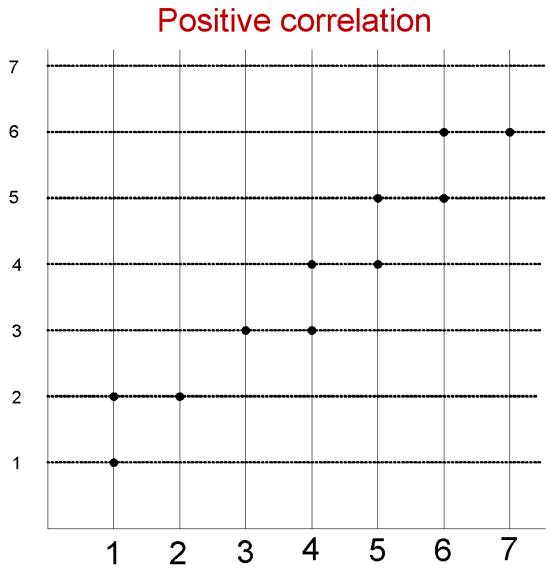
CORRELATION

- In statistics, the word correlation is used to denote some form of association between two variables.
  Ex-Weight is correlated with height.

- The correlation may be positive, negative, or zero.

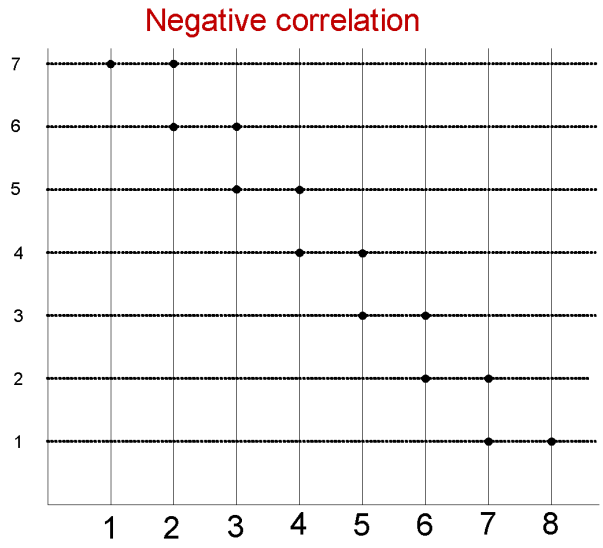
- **Positive correlation:** If the value of attribute A increases with the increase in the value of attribute B and vice-versa.

  Ex- Relation between rate of change of velocity and acceleration.



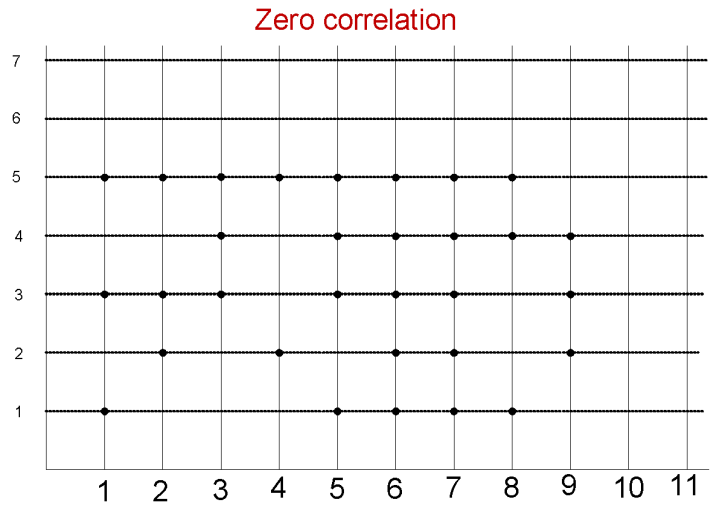
- **Negative correlation:** If the value of attribute A decreases with the increase in the value of attribute B and vice-versa.

Ex-Relation between the speed of the train and time taken to reach the destination.



- **Zero correlation:** When the values of attribute A varies at random with B and vice-versa.

  Ex- Relationship between the amount of tea drunk and level of intelligence.



Correlation Coefficient

- Correlation coefficient is used to measure the degree of association.
- It is usually denoted by r.
- The value of r lies between +1 and -1.
- Positive values of r indicate positive correlation between two variables, whereas, negative values of r indicate negative correlation.
- r = +1 implies perfect positive correlation, and otherwise.
- The value of r nearer to +1 or -1 indicates a high degree of correlation between the two variables.
- r = 0 implies, there is no correlation.
- There are three methods known to measure the correlation coefficients

  • Karl Pearson's coefficient of correlation-
  This method is applicable to find correlation coefficient between two numerical attributes

  • Charles Spearman's coefficient of correlation-
  This method is applicable to find correlation coefficient between two ordinal attributes.

  • Chi-square coefficient of correlation-
  This method is applicable to find correlation coefficients between two categorical attributes.

- Karl Pearson's coefficient of correlation is also called Pearson's Product Moment Correlation.

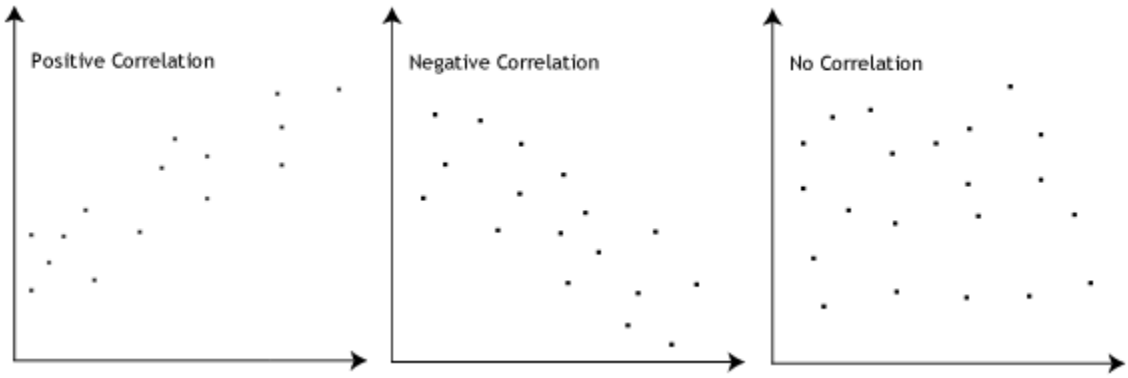
Pearson Product-Moment Correlation

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates

how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

- Values which Pearson correlation coefficient take

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



THE t DISTRIBUTION

(Used for testing significance of correlation)

1. To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = (\bar{x}-\mu)/(\sigma/\sqrt{n})$
2. The Central Limit Theorem requires the known value of σ a priori.
3. However, in many situations, σ is certainly no more reasonable than the knowledge of the population mean μ.
4. In such a situation, only the measure of the standard deviation available may be the sample standard deviation.
5. It is natural then to substitute for σ. The problem is that the resulting statistics are not normally distributed!

6. The - distribution is to alleviate this problem. This distribution is called a student's t-distribution or simply t-distribution.
If "$\bar{x}$ is the mean of a random sample of size # taken from a normal population having the mean μ and , S^2 is the variance of the sample then

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is a random variable having the t distribution with the parameter v =n − 1

About Dataset

- Data is about life expectancy and various factors.
- Columns
    - Year - YearStatus - Status of the given country (Developing / Developed)
    - Life expectancy - in years
    - Adult Mortality - Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
    - Infant deaths - Number of Infant Deaths per 1000 population
    - Alcohol - Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
    - Percentage expenditure - Expenditure on health as a percentage of Gross Domestic Product per capita(%)
    - Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
    - Measles - Measles - number of reported cases per 1000 population
    - BMI - Average Body Mass Index of entire population
    - Under five death - Number of under-five deaths per 1000 population
    - Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)
    - Total expenditure - General government expenditure on health as a percentage of total government expenditure (%)
    - Diphtheria - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
    - HIV/AID - Deaths per 1 000 live births HIV/AIDS (0-4 years)
    - GDP - Gross Domestic Product per capita (in USD)

- Population - Population of the country
- thinness 1-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%)
- Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling - Number of years of Schooling(years)

Dataset preprocessing

- Describing data
  - Here in this data set we have different attributes of the countries.
  - There are not many values with null values.
  - All the missing values are continuous numerical.
  - Observing data I found many meaningless outliers so I am using median for filling NA.
  - I am solving this country wise to avoid mistakes.
- Observation
  - I still found some missing data may be because of
    - countries data for all null values are null for every year.
    - Many countries have the first value as null and this method doesn't fill the first null entry.
- Solution
  - Doing the same thing but Year wise this time.
  - Found a good result.
  - Came up with good data without any missing values .
- Dealing with Outliers
  - Observing the boxplot of all columns observations are
    - There are some values which should be calculated out of 1000 there are some wrong data having more than 1000
      - Infant_Deaths
      - Measles
      - Under_five_deaths

- BMI values are very unrealistic, general values are around 40 but in this case even the median is greater than 40 so preferred to remove that column.

  - Result
    - Came up with good data without any missing values .
- CODE (Clear comments also available in code)

```
sapply(da, function(x) sum(is.na(x)))
for (i in unique(da$Country)){
  for (j in 1:ncol(da[da$Country == i,])){
    da[da$Year == i,][,j] = ifelse(is.na(da[da$Year ==
i,][,j]),ave(da[da$Year == i,][,j],FUN =
function(x)median(x,na.rm = TRUE)),da[da$Year ==
i,][,j])}}
sapply(da, function(x) sum(is.na(x)))
for (i in unique(da$Year)){
  for (j in 1:ncol(da[da$Year == i,])){
    da[da$Year == i,][,j] = ifelse(is.na(da[da$Year ==
i,][,j]),ave(da[da$Year == i,][,j],FUN =
function(x)median(x,na.rm = TRUE)),da[da$Year ==
i,][,j])}}
sapply(da, function(x) sum(is.na(x)))
for(i in 4:ncol(da)){
  boxplot(da[,i],main=colnames(da)[i])
}
da = da[da$infant.deaths < 1001,]
da = da[da$Measles < 1001,]
da = da[da$under.five.deaths < 1001,]
da$BMI = NULL
```

SUMMARY OF CODE

In this part, I went through data **country wise** and tried to fill data using **median** but I observed in some countries complete data is null. So, in better choice, I went through **year wise** and filled data using **median**.

Later, I removed Outliers like **BMI** values which are normal at 40's and **Infant death** rate which cannot be greater than **1000**. Then I found more than 50% of **BMI** values are null so I have **removed the BMI column**.

Correlation analysis

Degree of correlation:

- ➢ It is used to measure the degree of association .It is denoted by r
- ➢ r value lies between -1 and 1 i.e $-1 \leq r \leq 1$
- ➢ The value of r nearer to +1 or -1 indicates a high degree of correlation between the two variables.
- ➢ r = 0 implies, there is no correlation

From the problem statement we have to check whether there is a correlation between `Life expectancy` and other variables like `Alcohol ,Adult mortality, infant deaths ,percentage expenditure,measles,BMI,polio,GDP,Schooling,Population ,Diphtheria.`

Formula for Pearson Product-Moment Correlation

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

Now let's discuss how to write code for calculating correlation

From the formula we can see that we should calculate mean of variable x and y M_x and M_y

Correlation between Life Expectancy and Infant deaths

CODE

```
n=nrow(da)

# calculating degree of correlation for Life
Expectancy and infant Deaths
x=da["Life.expectancy"]
y=da["infant.deaths"]
sigma_x=sum(x)/n
sigma_y=sum(y)/n

x1=(x-sigma_x)

y1=(y-sigma_y)

x2=sum(x1 *x1)
y2= sum(y1 *y1)
num=sum(x1*y1)
denom=sqrt(x2* y2)
r=num/denom

r
```

Here we are taking Life Expectancy into variable x and infant deaths into variable y
Now we are calculating mean on variable x and y
After that we will calculate numerator and denominator part of the formula and divide both of them,and display the value of r

PLOT;

```
 plot(da$Life.expectancy,da$infant.deaths,xlab='Life expectancy',ylab='Infants deaths')
```

After calculating we will get r=-0.3929808
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

Formula for t-test

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

CODE:

```
t=(r/(sqrt(1-(r*r))))*(sqrt(n-2))

df=n-2;
```

After calculating we get t=-20.984
Degrees of freedom=n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero. And conclude that correlation is significant

CALCULATING R AND T VALUE USING cor.test() BUILTIN METHOD

```
test <- cor.test(da$Life.expectancy, da$infant.deaths)
test
```

As we can see that the calculated value for above function and using default function are same

Output :

```
        Pearson's product-moment correlation

data:  da$Life.expectancy and da$infant.deaths
t = -20.984, df = 2411, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4262007 -0.3587023
sample estimates:
       cor
-0.3929808
```

Same results are expected for all further calculations

COEFFICIENT OF DETERMINATION

The coefficient of determination or r^2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get

$r^2 = 0.1544339$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and infant deaths only explains about 15% of the variability in infant deaths.

Correlation between Life Expectancy and Alcohol

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Alcohol"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Alcohol,xlab='Life
expectancy',ylab='Alcohol')
```

After calculating we will get r =0.3987168
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=21.348
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero. And conclude that correlation is significant

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.1589751$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and alcohol only explains about 15% of the variability in Alcohol.

Correlation between Life Expectancy and Adult Mortality

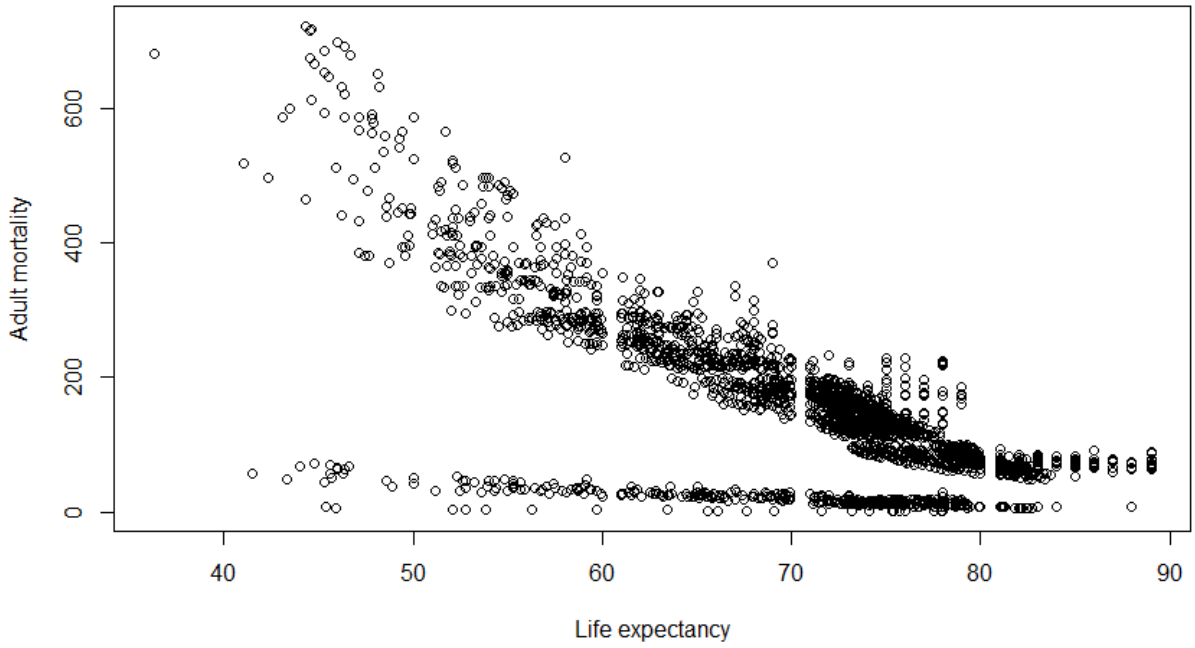
The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Adult.Mortality"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Adult.Mortality,xlab='Life expectancy',ylab='Adult mortality')
```



After calculating we will get r = -0.7155585
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=-50.297
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero. And conclude that correlation is significant

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.512024$

CONCLUSION

Here the value of the coefficient of Determination is moderate and the correlation between life expectancy and adult mortality explains about 51% of the variability of adult mortality.

Correlation between Life Expectancy and Percentage Expenditure

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["percentage.expenditure"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$percentage.expenditure,xlab
='Life expectancy',ylab='Percentage Expenditure')
```

After calculating we will get r = 0.3798795

To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=20.164
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value.. We get
$r^2 = 0.1443084$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and Percent Expenditure only explains about 14% of the variability in Percent Expenditure.

Correlation between Life Expectancy and Measles

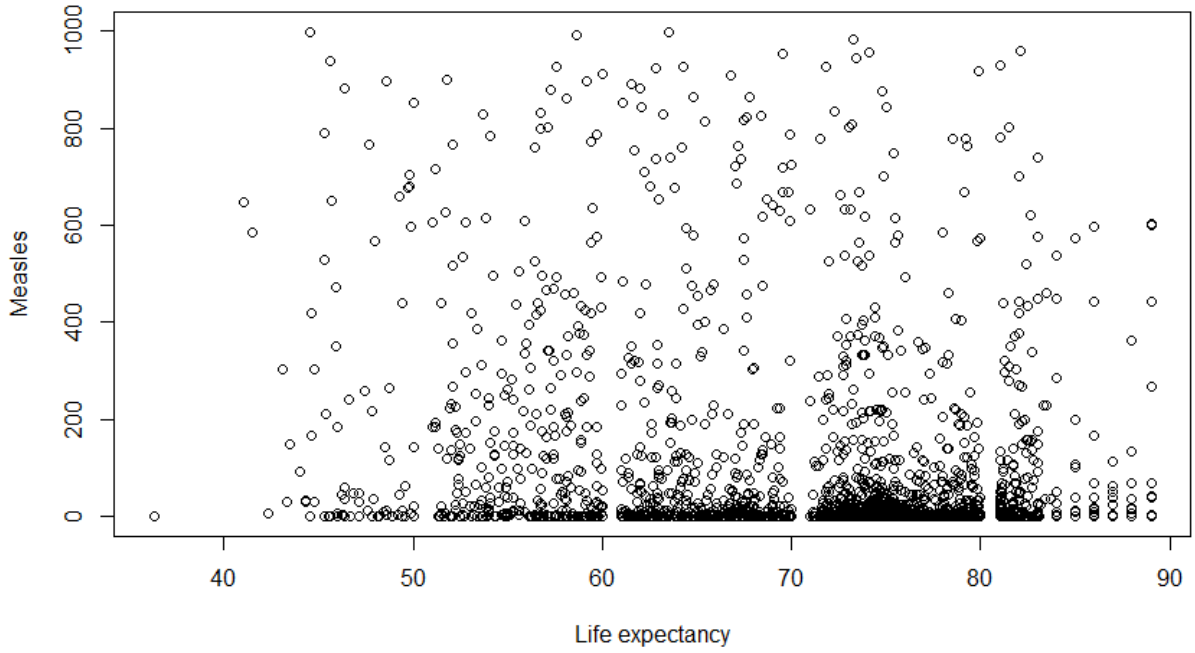
The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Measles"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Measles,xlab='Life
expectancy',ylab='Measles')
```



After calculating we will get r =-0.2064817
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=-10.362
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero. And conclude that correlation is significant

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
r^2 = 0.04263469

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation only explains about 4% of the variability in Measles.

Correlation between Life Expectancy and Total Expenditure

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Total.expenditure"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Total.expenditure,xlab='Lif
e expectancy',ylab='Total expenditure')
```

After calculating we will get r =0.2009415
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=10.072
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.04037749$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and total expenditure only explains about 4% of the variability in total expenditure.

Correlation between Expectancy and GDP

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["GDP"]
```

Now we have to check for correlation

PLOT

```
 plot(da$Life.expectancy,da$GDP,xlab='Life
expectancy',ylab='GDP')
```

After calculating we will get r =0.4263814
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=23.146
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

Result:- As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.1818011$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and GDP only explains about 18% of the variability in GDP.

Correlation between Life Expectancy and Population

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Population"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Population,xlab='Life expectancy',ylab='Population')
```

After calculating we will get r =0.02586578
To check whether this correlation is significant or not we use t-test
It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=1.2705
Degrees of freedom= n-2=2411

At 5% of significance after consulting statistical table we get

Result:- As the value of t is less than the value in the statistical table for given significance we will accept our null hypothesis i.e Actual correlation is zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
r^2 = 0.0006690386

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and population only explains about 0.06% of the variability in population.

Correlation between Life Expectancy and Schooling

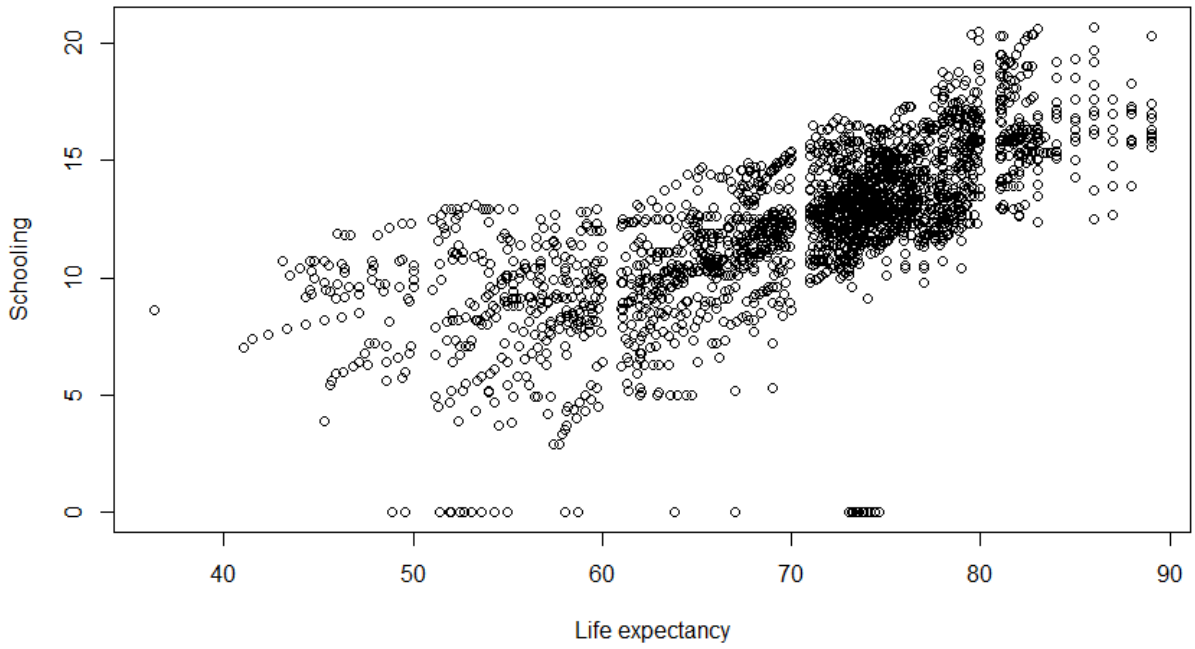
The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Schooling"]
```

Now we have to check for correlation

PLOT

```
 plot(da$Life.expectancy,da$Schooling,xlab='Life
expectancy',ylab='Schooling')
```



After calculating we will get r =0.7089461

To check whether this correlation is significant or not we use t-test
It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=49.35838
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.5026046$

CONCLUSION

Here the value of the coefficient of Determination is moderate and the correlation between life expectancy and schooling explains about 51% of the variability.

Correlation between Life Expectancy and Polio

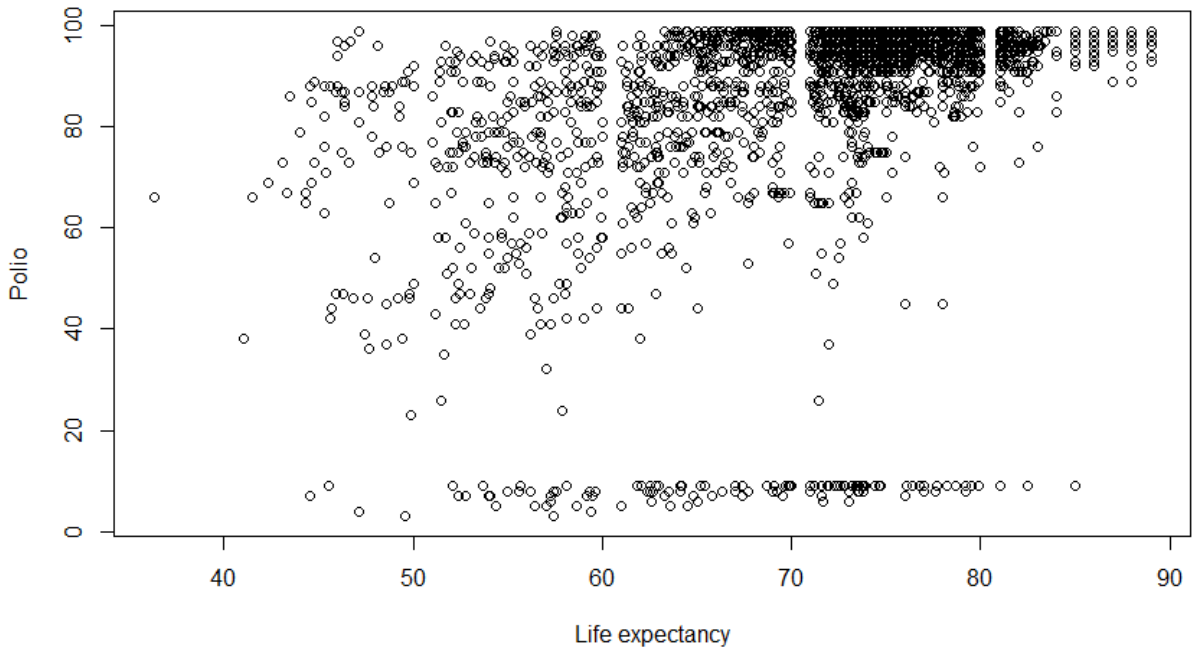
The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Polio"]
```

Now we have to check for correlation

PLOT

plot(da$Life.expectancy,da$Polio,xlab='Life expectancy',ylab='Polio')



After calculating we will get r =0.3767171

To check whether this correlation is significant or not we use t-test
It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=19.969
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In the case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.1419158$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and polio only explains about 14% of the variability.

Correlation between Life Expectancy and under five deaths

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["under.five.deaths"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$under.five.deaths,xlab='Lif
e expectancy',ylab='under five deaths')
```

After calculating we will get r =-0.4250979

As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=-23.06047
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.1807082$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and under five deaths only explains about 18% of the variability in under five deaths.

Correlation between Life Expectancy and HIV AIDS

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["HIV.AIDS"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$HIV.AIDS,xlab='Life
expectancy',ylab='HIV AIDS')
```

After calculating we will get r = -0.5665341
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t= -33.75804
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.3209609$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and HIV AIDS only explains about 32% of the variability in HIV AIDS.

Correlation between Life Expectancy and Hepatitis B

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Hepatitis.B"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Hepatitis.B,xlab='Life
expectancy',ylab='Hepatitis.B')
```

After calculating we will get r =0.1574408
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=7.82828

Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.we conclude that relation is significant.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
r^2 = 0.02478761

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and Hepatitis.B only explains about 2% of the variability in Hepatitis.B.

Correlation between Life Expectancy and Diphteria

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Diphtheria"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Diphtheria,xlab='Life
expectancy',ylab='Diphtheria')
```

After calculating we will get r =0.386158
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=20.55554

Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.we conclude that relation is significant.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.149118$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and Diphtheria only explains about 15% of the variability in Diphtheria.

Correlation between Life Expectancy and thinness 1-19 years

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["thinness..1.19.years"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$thinness..1.19.years,xlab='
Life expectancy',ylab='thinness..1.19.years')
```

After calculating we will get r = -0.5197001
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=-29.86869
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.we conclude that relation is significant.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.2700882$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and thinness..1.19.years only explains about 27% of the variability in thinness..1.19.years.

Correlation between Life Expectancy and thinnes 5-9 years

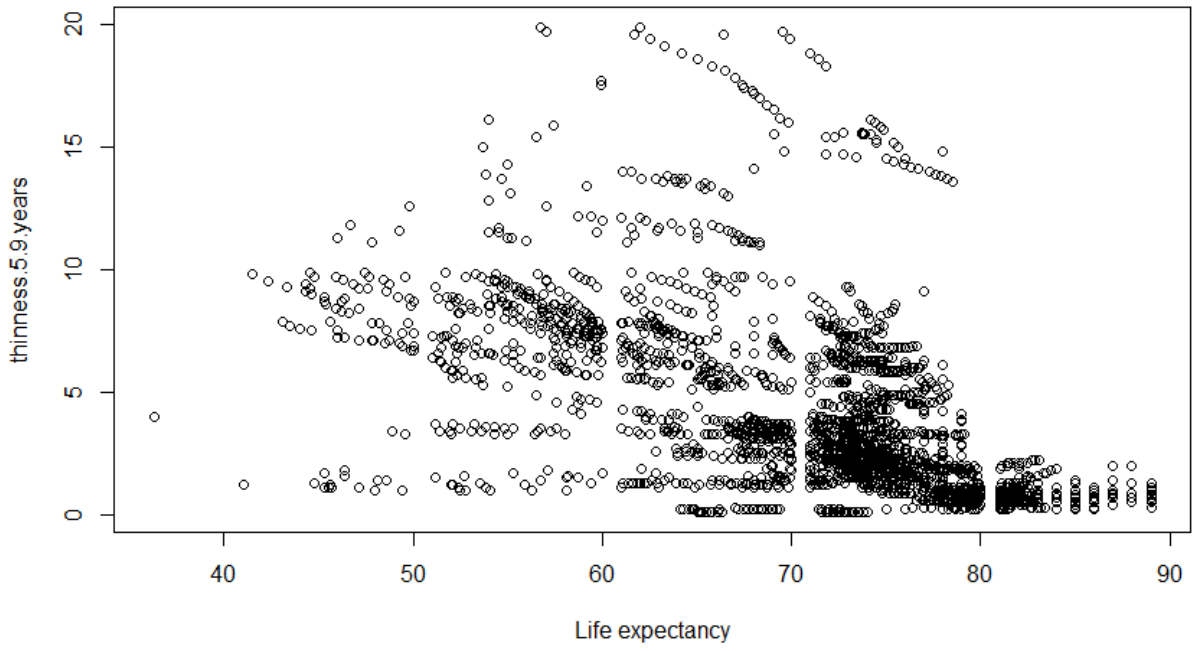
The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["thinness.5.9.years"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$thinness.5.9.years,xlab='Li
fe expectancy',ylab='thinness.5.9.years')
```



After calculating we will get r = -0.5222749
As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero

H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t= -30.07199
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.we conclude that relation is significant.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.2727711$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and thinness.5.9.years only explains about 27% of the variability in thinness.5.9.years.

Correlation between Life Expectancy and Income composition of resources

The same code is used for Calculating r
But we change x and y

```
x=da["Life.expectancy"]
y=da["Income.composition.of.resources"]
```

Now we have to check for correlation

PLOT

```
plot(da$Life.expectancy,da$Income.composition.of.resou
rces,xlab='Life
expectancy',ylab='Income.composition.of.resources')
```

After calculating we will get r =0.6755879

As it is near to zero we are not sure whether the association of variables is strong or not so we have to test for significance of relation
To check whether this correlation is significant or not we use t-test

It's time to set the hypothesis for t-test

H_0: Actual correlation is Zero
H_1: Actual correlation is not equal to zero

We will run the same code for doing t-test

After calculating we get t=44.99335
Degrees of freedom= n-2=2411
At 5% of significance after consulting statistical table we get

RESULT

As the value of t is greater than the value in the statistical table for given significance we will Reject our null hypothesis i.e Actual correlation is not equal to zero.we conclude that relation is significant.

COEFFICIENT OF DETERMINATION

The coefficient of determination or r2 is the proportion of variation in a variable that can be explained by some other variable. In case of the Pearson correlation product, to calculate the coefficient of determination we simply calculate the square of the r value. We get
$r^2 = 0.456419$

CONCLUSION

Since the value of the coefficient of Determination is very low, the correlation between life expectancy and Income.composition.of.resources only explains about 15% of the variability in Income composition of resources.