Lab-5

(Challenging Experiment 3)

Find the linear correlation co-efficient and fit the regression line for future prediction. Applying simple linear regression model to real dataset; computing and interpreting the coefficient of determination	Conceptual understanding of Model fitting and investigate relationships between two variables within a regression framework	Learn to do future prediction with two variable

Correlation and Linear regression

Aim: Model fitting and investigating relationships between two variables within a regression framework.

Correlation Definition:-

Correlation refers to the relationship between two or more variables. Simple correlation studies the relationship between two variables. Correlation analysis attempts to determine the degree of relationship between variables.

Measures of Correlation:

Scatter Diagram:

Scatter diagram is the simplest way of graphic representation of a bivariate data, where the given set of 'n' pairs of observations on two variables X and Y say (X_1, Y_2), (X_2, Y_2) …(X_n, Y_n) may be plotted as dots by considering X-values on X-axis and Y-values on Y-axis. By scatter diagram, we can get some idea about the correlation between X and Y.

Problem:-

AGE GROUP	REPRESENTATIVE AGE	HOURS SPEND IN THE LOCAL LIBRARY
10-19	15	302.38
20-29	25	193.63
30-39	35	185.46
40-49	45	198.49

AGE GROUP	REPRESENTATIVE AGE	HOURS SPEND IN THE LOCAL LIBRARY
50-59	55	224.30
60-69	65	288.71

illustrate the relationship between the average age versus the time spent in the library, by using scatterplot.

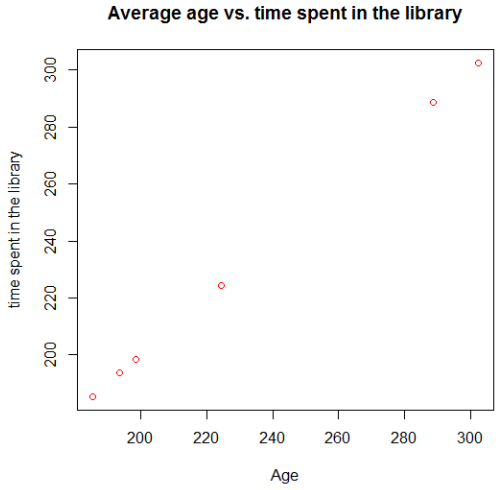
R code:-

>x <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71)

> y <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71)

>plot(x,y, main="Average age vs. time spent in the library", xlab="Age", ylab="time spent in the library",col="red")

OUTPUT:-



Karl Pearson's Coefficient of Correlation

It is defined as the ratio of covariance between x and y say Cov (X,Y) to the product of the standard deviations of X and Y, say $\sigma_X \sigma_Y$

$$i.e \qquad r_{XY} = \frac{Cov(XY)}{\sigma_X \sigma_Y}$$

Consider a set of 'n' pairs of observations (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n) on two variables X and Y. Then we have, Covariance between X and Y

R code:-

```
> x=c(23,27,28,28,29,30,31,33,35,36)
> y=c(18,20,22,27,21,29,27,29,28,29)
>var(x)
[1] 15.33333
>var(y)
[1] 18.22222
>var(x,y)
[1] 13.66667
> r=var(x,y)/sqrt(var(x)*var(y))
> r
[1] 0.8176052
Or
>cor(x,y)
[1] 0.8176052
Or
>cor.test(x,y) Or
>cor.test(x,y,method="pearson")
Pearson's product-moment correlation
data:  x and y
t = 4.0164, df = 8, p-value = 0.003861
alternative hypothesis: true correlation is not equal to 0
```

95 percent confidence interval:

0.3874142 0.9554034

sample estimates:

cor

0.8176052

There is a Positive correlation between X and Y

SPEARMAN'S RANK CORRELATION COEFFICIENT

Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient ρ is given by

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right]$$ [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'

SPEARMAN'S RANK CORRELATION COFFICIENT FOR A DATA WITH AND WITHOUT TIED OBSERVATIONS:

Problem : Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below :

Recruit	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Selection Test Score	*44*	*49*	*52*	*54*	*47*	*76*	*65*	*60*	*63*	*58*	*50*	*67*
Proficiency Test Scrore	*48*	*55*	*45*	*60*	*43*	*80*	*58*	*50*	*77*	*46*	*47*	*65*

Calculate rank correlation coefficient and comment on your result.

Solution:-

> selection =c(44,49,52,54,47,76,65,60,63,58,50,67)
> proficiency =c(48,55,45,60,43,80,58,50,77,46,47,65)
>cor.test(selection,proficiency,method="spearman")

Spearman's rank correlation rho

data: selection and proficielncy
S = 80, p-value = 0.01102

alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.7202797

There is a positive correlation between selection and Proficiency

KENDALL'S COEFFICIENT OF CONCURRENT DEVIATIONS

The Kendall's coefficient of concurrent deviations is denoted by r_c and defined as

$$r_c = \pm\sqrt{\pm\left[\frac{2C - n}{n}\right]}$$

Where, C = Number of concurrent deviations or position signs of (D_X, D_Y);

n = Number of pairs of deviations

Problem: The following data gives the marks obtained by 12 students in statistics and computer science :

Students		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Marks	*Statistics*	*55*	*40*	*70*	*60*	*62*	*73*	*65*	*65*	*20*	*35*	*46*	*50*
	Computer Science	*35*	*32*	*65*	*50*	*63*	*45*	*50*	*65*	*70*	*72*	*72*	*40*

Compute the coefficient of correlation by the method of concurrent deviations.
R code:

```
> statistics=c(55,40,70,60,62,73,65,65,20,35,46,50)
> mathematics=c(35,32,65,50,63,45,50,65,70,72,72,40)
>cor.test(statistics,mathematics,method="kendall")
Kendall's rank correlation tau
data:  statistics and mathematics
z = -0.27688, p-value = 0.7819
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
```

-0.06250763

There is a negative correlation between mathematics and statistics

R^2 *(Coefficient of determination):-*

Code:

```
examdata=read.csv("C:\\Users\\aadmin\\Desktop\\mokesh\\examdata.csv")
examdata2 <- examData[, c("Exam", "Anxiety", "revise")]
cor(examdata2)
```

OUTPUT:-

	Exam	Anxiety	revise
Exam	1.0000000	-0.6381787	0.6281441
Anxiety	-0.6381787	1.0000000	-0.8190752
revise	0.6281441	-0.8190752	1.0000000

Interpretation:-

Provides a matrix of the correlation coefficients for the three variables. Each variable is perfectly correlated with itself (obviously) and so r = 1 along the diagonal of the table. Exam performance is negatively related to exam anxiety with a Pearson correlation coefficient of r = −.441. This is a reasonably big effect. Exam performance is positively related to the amount of time spent revising, with a coefficient of r = .397, which is also a reasonably big effect. Finally, exam anxiety appears to be negatively related to the time spent revising, r = −.709, which is a substantial effect size. In psychological terms, this all means that as anxiety about an exam increases, the percentage mark obtained in that exam decreases. Conversely, as the amount of time revising increases, the percentage obtained in the exam increases. Finally, as revision time increases, the student's anxiety about

the exam decreases. So there is a complex interrelationship between the three variables.

$R^{2:-}$

```
> examdata=read.csv("C:\\Users\\aadmin\\Desktop\\examdata.csv")
> examdata2 <- examdata[, c("Exam", "Anxiety", "revise")]
> cor(examdata2)^2     #coefficient of determination
             Exam   Anxiety    revise
Exam    1.0000000 0.4072769 0.3945650
Anxiety 0.4072769 1.0000000 0.6708793
revise  0.3945650 0.6708793 1.0000000
```

Interpretation:-

*Coefficient a step further by squaring it. The correlation coefficient squared (known as the **coefficient of determination**, R^2) is a measure of the amount of variability in one variable that is shared by the other. From the above we may look at the relationship between exam anxiety and exam performance. Exam performances vary from person to person because of any number of factors (different ability, different levels of preparation and so on). then we would have an estimate of how much variability exists in exam performances. We can then use* R^2 *to tell us how much of this variability is shared by exam anxiety. These two variables had a correlation of -0.6381787 and so the value of R^2 will be (-0.6381787) ^2 = 0.4072721.This value tells us how much of the variability in exam performance is shared by exam anxiety.*

If we convert this value into a percentage (multiply by 100) we can say that exam anxiety shares 40.7%1of the variability in exam performance. So, although exam anxiety was highly correlated with exam performance, it can account for only 40.7% of variation in exam scores. To put this value into perspective, this leaves 59.3 % of the variability still to be accounted for by other variables

Linear Regression Model

To draw conclusions about a population based on a regression analysis done on a sample, several assumptions must be true (see Berry, 1993):

***Variable types:** All predictor variables must be quantitative or categorical (with two categories), and the outcome variable must be quantitative, continuous and*

unbounded. By 'quantitative' I mean that they should be measured at the interval level and by 'unbounded' I mean that there should be no constraints on the variability of the outcome. If the outcome is a measure ranging from 1 to 10 yet the data collected vary between 3 and 7, then these data are constrained.

Non-zero variance: *The predictors should have some variation in value (i.e., they do not have variances of 0).*

No perfect multicollinearity: *There should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly (see section 7.7.2.4).*

Predictors are uncorrelated with 'external variables': *External variables are variables that haven't been included in the regression model which influence the outcome variable.9 These variables can be thought of as similar to the 'third variable' that was discussed with reference to correlation. This assumption means that there should be no external variables that correlate with any of the variables included in the regression model. Obviously, if external variables do correlate with the predictors, then the conclusions we draw from the model become unreliable (because other variables exist that can predict the outcome just as well).*

Homoscedasticity: *At each level of the predictor variable(s), the variance of the residual terms should be constant. This just means that the residuals at each level of the predictor(s) should have the same variance (homoscedasticity); when the variances are very unequal there is said to be heteroscedasticity (see section 5.7 as well).*

Independent errors: *For any two observations the residual terms should be uncorrelated (or independent). This eventuality is sometimes described as a lack of autocorrelation. This assumption can be tested with the Durbin–Watson test, which tests for serial correlations between errors. Specifically, it tests whether adjacent residuals are correlated. The test statistic can vary between 0 and 4, with a value of 2 meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value less than 2 indicates a positive correlation. The size of the Durbin–Watson statistic depends upon the number of predictors in the model and the number of observations. As a very conservative rule of thumb, values less than 1 or greater than 3 are definitely cause for concern; however, values closer to 2 may still be problematic depending on your sample and model. R also provides a p-value of the autocorrelation. Be very careful with the Durbin–Watson test, though, as it depends on the order of the data: if you reorder your data, you'll get a different value.*

Normally distributed errors: *It is assumed that the residuals in the model are random, normally distributed variables with a mean of 0. This assumption simply means that the differences between the model and the observed data are most frequently zero or very close to zero, and that differences much greater than zero happen only occasionally. Some people confuse this assumption with the idea that predictors have to be normally distributed. Predictors do not need to be normally distributed .*

Independence: *It is assumed that all of the values of the outcome variable are independent (in other words, each value of the outcome variable comes from a separate entity).*

Linearity: *The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line. In plain English this means that it is assumed that the relationship we are modelling is a linear one. If we model a non-linear relationship using a linear model then this obviously limits the generalizability of the findings.*

Problem:-

The body weight and the BMI of 12 school going children are given in the following table

Weight	***15***	***26***	***27***	***25***	***25.5***	***27***	***32***	***18***	***22***	***20***	***26***	***24***
BMI	***13.35***	***16.12***	***16.74***	***16.00***	***13.59***	***15.73***	***15.65***	***13.85***	***16.07***	***12.8***	***13.65***	***14.42***

Let us fit a simple regression model BMI on weight and examine the results.

Answer:-

```
> weight=c(15,26,27,25,25.5,27,32,18,22,20,26,24)
> bmi=c(13.35,16.12,16.74,16.00,13.59,15.73,15.65,13.85,16.07,12.8,13.65,14.42)
> cor(weight,bmi)
[1] 0.5790235
> model<-lm(bmi~weight)
> summary.lm(model)

Call:
lm(formula = bmi ~ weight)

Residuals:
     Min       1Q   Median       3Q      Max
-1.52988 -0.75527  0.04426  0.95286  1.57397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.73487    1.85405   5.790 0.000175 ***
weight       0.17096    0.07612   2.246 0.048524 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.155 on 10 degrees of freedom
Multiple R-squared:  0.3353,    Adjusted R-squared:  0.2688
F-statistic: 5.044 on 1 and 10 DF,  p-value: 0.04852
```

Interpretation :

Correlation r=0.5790,which is the correlation coefficient between the body 'weight' and BMI.There is a positive correlation between these two variables.The Value of R^2 is 0.3353,which means that about 33.53% variation in BMI can be explained by 'weight'through this linear model.This is apparently low because more than 67% of variation remains unexplained.There could be several reasons for this and one of them is that there might be some other influencing variables that have not been included in the present model.

The F value shown in the above output gives the statistics for the variance ratio test of the regression model.The significance of F, which is given as 0.0485,is the p value of the F-test carried out in ANOVA. If this value is less than 0.05 we say that the regression is statistical significant at 5% level of significance .Here

regression is significant which means that the relationship is not an occurrence by chance

In the above output we find b_0 is the intercept which value of 10.73487 and b_1 is the regression coefficient due to weight with a value of 0.1710.The regression coefficient is positive ,which shows that the BMI is positively related to weight,

The regression output can be written as mathematical equation

$$BMI = 10.7349 + 0.1710 * weight$$

Suppose body weight of one student is known as 25 kg.Using the above equation,the estimated BMI is 15.01.since this is only an estimate we have to intercept it as the average BMI corresponding to the given weight assuming that other parameters are unchanged.

Obtain a linear relationship between weight (kg) and height (cm) of 10 subjects.

Height	***175***	***168***	***170***	***171***	***169***	***165***	***165***	***160***	***180***	***186***
Weight	***80***	***68***	***72***	***75***	***70***	***65***	***62***	***60***	***85***	***90***

CODE:-

```
> height = c(175, 168, 170, 171, 169, 165, 165, 160, 180, 186)
> weight = c(80, 68, 72, 75, 70, 65, 62, 60, 85, 90)
> cor(height,weight)
[1] 0.9849472
> model = lm(formula = height ~ weight)
> model

Call:
lm(formula = height ~ weight)

Coefficients:
(Intercept)       weight
   115.2002       0.7662

> summary(model)

Call:
lm(formula = height ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6622 -0.9683 -0.1622  0.5679  2.2979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.20021    3.48450   33.06 7.64e-10 ***
weight        0.76616    0.04754   16.12 2.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.405 on 8 degrees of freedom
Multiple R-squared:  0.9701,    Adjusted R-squared:  0.9664
F-statistic: 259.7 on 1 and 8 DF,  p-value: 2.206e-07
```

Interpretation:-

The regression equation is

Height=115.2002+0.7662 weight. Since the p- value of the test is 16.12 is greater than 0.05 we reject the hypothesis. Therefore the model we found is significant. The Multiple R-squared is the coefficient of determination. It provides a measure of how well future outcomes are likely to be predicted by the model. In this case the R suare value is 0.9701. Therefore 97.01% of data is well predicted.

Practice Problem:

1. ***The following data refers to the daily sales of tomatoes (in kg) at different prices(in Rupess) observed on different days in a market***

Price	***4.5***	***5.5***	***4.5***	***4.5***	***4.0***	***5.5***	***5.5***	***6.5***	***5.0***	***5.5***	***6.0***	***4.5***
Quantity Sold	***125***	***115***	***140***	***140***	***150***	***150***	***130***	***120***	***130***	***100***	***105***	***150***

Let us carry out linear regression analysis for this data.

2. ***The success of a shopping center can be represented as a function of the distance (inmiles) from the center of the population and the number of clients (in hundreds of people) who will visit. The data is given in the table below:***

No. Customer(x)	***8***	***7***	***6***	***4***	***2***	***1***
Distance(y)	***15***	***19***	***25***	***23***	***34***	***40***

   a) ***Calculate the linear correlation coefficient***
   b) ***If the mall is located 2 miles from the center of the population, how many customers should the shopping center except?***
   c) ***To receive 500 customers, at what distance from the center of the population should the shopping centre be located?***

3. ***Find the correlation between Experience and Income. Also fit a regression equation and interpret the result.***