STATISTICS IS THE GRAMMAR OF SCIENCE

PROBABILITY AND STATISTICS

LECTURE – 22

REGRESSION ANALYSIS

SIMPLE LINEAR REGRESSION MODEL

PREPARED BY
HAZBER SAMSON
FAST NUCES ISLAMABAD

REGRESSION ANALYSIS

INTRODUCION: At the beginning of the nineteenth century, **Legendre and Gauss** published papers on the *method of least squares*, which implemented the earliest form of what is now known as *linear regression*. The term *regression* was first introduced by **Francis Galton**. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" toward the average height in the population as a whole.1 In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton's *law of universal regression* was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups.2 He found that the average height of sons of a group of tall fathers was less than their fathers' height and the average height of sons of a group of short fathers was greater than their fathers' height, thus "regressing" tall and short sons alike toward the average height of all men. In the words of Galton, this was "regression to mediocrity. The modern interpretations of regression is, as follows however,

DEFINITION: Regression analysis is a statistical technique for investigating and modeling the relationship between variables.

Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

Usually, Researchers want more than just finding the best linear approximation of one variable given a set of others. They want valid statistical relationships. They want to draw conclusions about what happens if one of the variables actually changes. That is: they want to say something about things that are not observed (yet). In this case, they want to find a fundamental relationship. To do this it is assumed that there is a general relationship that is valid for all possible observations from a well-defined population. Restricting attention to linear relationships, we specify a statistical model as linear regression model. Some examples of analyses using regression models include the following:

• Estimating weight gain by the addition to children's diet of different amounts of various dietary supplements
• Predicting scholastic success (grade point ratio) based on students' scores on an aptitude or entrance test
• Estimating amounts of sales associated with levels of expenditures for various types of advertising
• Predicting fuel consumption for home heating based on daily temperatures and other weather factors
• Estimating changes in interest rates associated with the amount of deficit spending

Regression analysis is used to predict the value of one variable on the basis of the other variables. Regression analysis is generally classified into two kinds, **simple** and **multiple**. The regression will be **simple** if it involves only one independent variable and it will be **multiple** if it involves more than one independent variables.

SIMPLE LINEAR REGRESSION MODEL (SLRM)

By linear regression model we mean an approximate linear association between dependent and independent variables. There could be two types of linearity
1. Linearity in Parameters
2. Linearity in Variables

In case of SLRM both linearity's must hold.

DEFINITION If there is only one independent variable and one dependent variable than regression is called Simple regression. In this case regression equation forms a straight line so it is also called SLRM. It is also called the *two-variable linear regression model* or *bivariate linear regression model* because it relates the two variables *x* and *y*.

It is defined as

$$Y = \alpha + \beta X + \varepsilon$$

Here

$Y = dependent\ variable$

$X = independent\ variable$

$\alpha = y - intercept$

$\beta = Slope$

$\varepsilon = error term \ \ or \ \ residual\ term$

y	*x*
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

In SLRM $\alpha \ and \ \beta$ are also called parameters of the Model.

Different Names of dependent and independent variables are given in the table.

PROPERTIES OF REGRESSION LINE

1. Regression line always passes through the sample means $\overline{X}$ *and* $\overline{Y}$.
2. Mean value of the estimated values of Y is equal to the mean value of actual Y ie $\overline{\hat{Y}_i} = \overline{Y_i}$
3. The mean value of the residuals is zero ie $\overline{\varepsilon_i} = 0$
4. The residuals are uncorrelated with predicted values.
5. The residuals are uncorrelated with independent variables.

ESTIMATION OF REGRESSION LINE

We will discuss three methods for the estimation of regression model
The Method of Least Squares
The Method of Moments
The Method of Maximum Likelihood

ORDINARY LEAST SQUARES (OLS)

OLS is technique for fitting the best straight line to the sample x, y observations. It involves minimizing the sum of squares of residuals. i.e. $Min \sum \left(Y - \hat{Y}\right)^2 = Min \sum e_i^2$

Note that Y: Observed Values, $\hat{Y}$: estimated values and $Y - \hat{Y}_i = e_i$

COEFFICIENT OF DETERMINATION

The coefficient of determination is statistic used to determine how well a regression is fit. It is the ratio of the explained variation to the total variation. It shows proportion of variation in y variable due to explanatory variables of the model. We can say that how much of the variability in y variable can be explained by the linear relationship of y to the independent variable x. It is denoted by $R^2 \text{ or } r^2$ and is given by

$$R^2 = \frac{Explained\,Variation}{Total\,Variation} = \frac{SSR}{SST} = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$

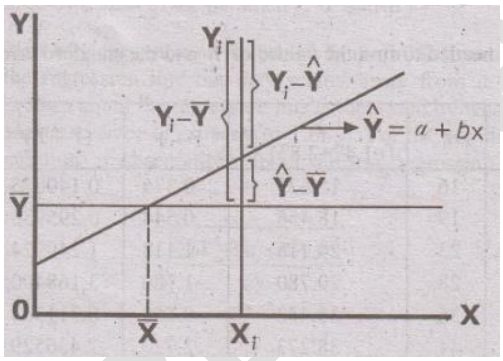
or

$$R^2 = 1 - \frac{Un\exp lained\,Variation}{Total\,Variation} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$

Note that Total Variation = Explained Variation + Unexplained Variation

$$i.e.\ SST = SSR + SSE \ \ or \ \ \sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum \left(Y - \hat{Y}\right)^2$$

Graphically



Also Note that

1. $Total\ Varation = SST = Sum\ of\ \ Squares\ Total = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$
2. $SSR = Sum\ of\ Squares\ \text{Re}\ gression = \sum (\hat{Y} - \bar{Y})^2 = \sum (Y - \bar{Y})^2 - \sum \left(Y - \hat{Y}\right)^2$
3. $SSE = Sum\ of\ \ Squares\,Errors = \sum \left(Y - \hat{Y}\right)^2 = \sum Y^2 - a\sum Y - b\sum XY$

PROPERTIES OF COEFFICIENT OF DETERMINATION

1. It is a non-negative measure.
2. Its value lies between o and 1 i.e. $0 \le R^2 \le 1$
3. Coefficient of determination is just the square of Correlation coefficient. $R^2 = r^2$

INTERPRETATION OF COEFFICIENT OF DETERMINATION

If $R^2 = 0.90$ than it means that 90% of the variation in dependent variable is due to the linear relationship of the independent variables where the remaining 10% of the variation is due other factors.

ORDINARY LEAST SQUARES (OLS)

$$Y = \alpha + \beta X + \varepsilon$$

NORMAL EQUATIONS

- $\sum Y = n\alpha + \beta \sum X$
- $\sum XY = \alpha \sum X + \beta \sum X^2$

ESTIMATORS

Now

- $\hat{\beta} = \dfrac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$
- $\hat{\beta} = \dfrac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sum \left(X - \overline{X}\right)^2}$
- $\hat{\beta} = \dfrac{S_{XY}}{S_{XX}}$

Also

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

STANDARD ERROR OF ESTIMATE or RESIDUAL STANDARD ERROR

RSE is a measure of the lack of fit of the model to the data

$$RSE = S_{y.x} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum \left(Y - \hat{Y}\right)^2}{n-2}} = \sqrt{\frac{\sum Y^2 - \alpha \sum Y - \beta \sum XY}{n-2}}$$

CO-EFFICIENT OF DETERMINATION

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST} = \frac{\sum \left(\hat{Y} - \overline{Y}\right)^2}{\sum \left(Y - \overline{Y}\right)^2}$$

or

$$R^2 = \frac{Unexplained\ Variation}{Total\ Variation} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum \left(Y - \hat{Y}\right)^2}{\sum \left(Y - \overline{Y}\right)^2}$$

EXAMPLES OF SIMPLE LINEAR REGRESSION MODEL

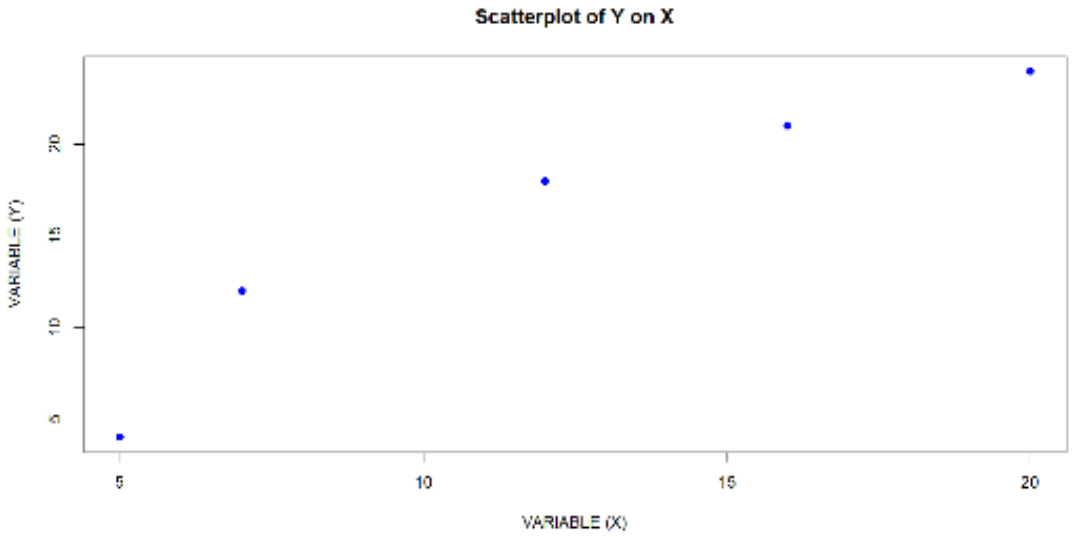
EXAMPLE-1 Consider the data given below

X	5	7	12	16	20
Y	4	12	18	21	24

(a) Construct a Scatter Plot for the data.
(b) Find a regression equation of Y on X and interpret the Results.
(c) Find the value of Y when X = 25 using Fitted regression Model.
(d) Find Correlation Coefficient and Interpret it.
(e) Find Coefficient of Determination and Interpret it.

SOLUTION First of all we shall construct the scatter plot and analyze.

(A) *SCATTER PLOT*



Scatter Plot shows that there is a positive relation between X and Y.

(B) *FITTING REGRESSION THE MODEL* $Y = \alpha + \beta X + \varepsilon$

For calculations we shall consider only $Y = \alpha + \beta X$

$$We\ \ know\ that\ \ \hat{\beta} = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$and \qquad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

Let's make table for the values used in formulas

X	Y	XY	X^2	Y^2	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$Y-\overline{Y}$	$(Y-\overline{Y})^2$
5	4	20	25	16	7.26	-3.26	10.63	-11.8	139.24
7	12	84	49	144	9.7	2.3	5.29	-3.8	14.44
12	18	216	144	324	15.8	2.2	4.84	2.2	4.84
16	21	336	256	441	20.68	0.32	0.10	5.2	27.04
20	24	480	400	576	25.56	-1.56	2.43	8.2	67.24
60	**79**	**1136**	**874**	**1501**	**79**	**0**	**23.29**	**0**	**252.8**

Now $\hat{\beta}=\dfrac{n\left(\sum XY\right)-\left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right)-\left(\sum X\right)^2}=\dfrac{(5)(1136)-(60)(79)}{(5)(874)-(60)^2}=\dfrac{940}{770}=1.22$

Also $\overline{X}=\dfrac{\sum X}{n}=\dfrac{60}{5}=12$ and $\overline{Y}=\dfrac{\sum Y}{n}=\dfrac{79}{5}=15.8$

So $\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{X}=15.8-(1.22)(12)=1.16$

So the required regression equation will be given as $\hat{Y}=1.16+1.22\,X$

INTERPRETATION OF RESULTS OF THE MODEL

Interpretation of $\hat{\alpha}=1.16$ **:** If X = 0 then Y will be 1.16.

Interpretation of $\hat{\beta}=1.22$: If X is increased by 1 unit Y will be increased by 1.22 units on average in the long run keeping other factors as constant.

(C) ***VALUE OF Y FOR GIVEN X*** Fitted regression equation is $\hat{Y}=1.16+1.22\,X$

At $X=25$ we have $Y=1.16+1.22(25)=31.66$

(D) ***COEFFICIENT OF CORRELATION***

Coefficient of Correlation is given by

$$r=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{[n\sum x^2-\left(\sum x\right)^2][n\sum y^2-\left(\sum y\right)^2]}}=\frac{(5)(1136)-(60)(79)}{\sqrt{\left[(5)(874)-(60)^2\right]\left[(5)(1501)-(79)^2\right]}}=0.9528$$

Interpretation: $r=0.9528$ shows that there is strong positive relation between X and Y.

(E) ***COEFFICIENT OF DETERMINATION***

Coefficient of Determination is given by

$$R^2=1-\frac{SSE}{SST}=1-\frac{\sum\left(Y-\hat{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2}=1-\frac{23.29}{252.8}=0.91 \quad \text{or} \quad R^2=r^2=(0.9528)^2=0.91$$

Interpretation: $R^2=0.91$ shows that 91% variation in Y is explained due to X and remaining 9% is due to other factors.

EXAMPLE-2 Consider the data shows for car rental companies in the United States for a recent year.

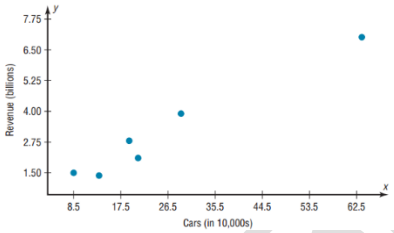
Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Source: *Auto Rental News.*

(a) Construct a Scatter Plot for the data.
(b) Find the equation of the regression line for the data and interpret the results.
(c) Graph the regression line on the scatter plot of the data.
(d) Find the Coefficient of Correlation and Coefficient of Determination and Interpret it.
(e) Use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

SOLUTION First of all we shall construct the scatter plot and analyze.

(A) *SCATTER PLOT*



Clearly scatter plot shows a positive relation between number of cars and Revenue.

(B) *FITTING THE REGRESSION MODEL* $Y = \alpha + \beta X + \varepsilon$

For Calculations we shall consider only $Y = \alpha + \beta X$

Find the values of xy, x^2, and y^2 and place these values in the corresponding columns of the table.

Company	Cars x (in 10,000s)	Revenue y (in billions)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Now $\hat{\beta} = \dfrac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2} = \dfrac{(6)(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$

Also $\overline{X} = \frac{\sum X}{n} = \frac{153.8}{6} = 25.633$ *and* $\overline{Y} = \frac{\sum Y}{n} = \frac{18.7}{6} = 3.116$

So $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 3.116 - (0.106)(25.63) = 0.396$

So the required regression equation will be given as $\hat{Y} = 0.396 + 0.106\,X$

INTERPRETATION OF RESULTS OF THE MODEL

Interpretation of $\hat{\alpha} = 1.16$ **:** If X = 0 then Y will be 0.396, ie if number of cars are zero even then revenue will be 0.396 billion.

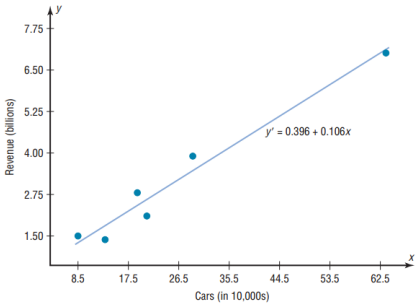
Interpretation of $\hat{\beta} = 0.106$: If number of cars is increased by 10000 then revenue will be increased by 1.22 billion on average in the long run keeping other factors as constant.

(C) SCATTER PLOT WITH REGRESSION LINE

To graph a line, select any two points for and find the corresponding values of y.

If $x = 15$ *then* $y = 1.986$ *and If* $x = 40$ *then* $y = 4.636$

So two points are $A(15,1.99)$ & $B(40, 4.64)$



(D) COEFFICIENT OF CORRELATION AND DETERMINATION

Coefficient of Correlation is given by

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{\left[(6)(5859.26) - (153.8)^2\right]\left[(6)(80.67) - (18.7)^2\right]}} = 0.9819$$

Interpretation: $r = 0.9819$ shows that there is strong positive relation between X and Y.

Coefficient of Determination is given by

$$R^2 = r^2 = (0.9819)^2 = 0.96$$

Interpretation: $R^2 = 0.96$ shows that 96% variation in Y is explained due to X and remaining 9% is due to other factors.

(E) VALUE OF Y FOR GIVEN X *Here* $X = 200000/10000 = 20$ *Since X is in Thousands*

Fitted regression equation is $\hat{Y} = 0.396 + 0.106\,X$

At $X = 20$ we have $\hat{Y} = 0.396 + 0.106(20) = 2.516$

Hence when a rental agency has 200000 cars, its revenue will be app. $2.516 billion.

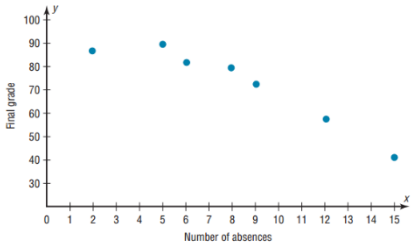
EXAMPLE-3 Consider the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

(a) Construct a Scatter Plot for the data.
(b) Find the equation of the regression line for the data and interpret the results.
(c) Graph the regression line on the scatter plot of the data.
(d) Find the Coefficient of Correlation and Coefficient of Determination and Interpret it.
(e) Use the equation of the regression line to predict the grade of a student having 10 absences.

SOLUTION First of all we shall construct the scatter plot and analyze.

(A) *SCATTER PLOT*



Scatter plot shows a negative relation between Number of absences and Final grade.

(B) *FITTING THE REGRESSION MODEL* $Y = \alpha + \beta X + \varepsilon$

For Calculations we shall consider only $Y = \alpha + \beta X$

Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the table.

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

Now $\hat{\beta} = \dfrac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2} = \dfrac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} = -3.622$

Also $\overline{X} = \frac{\sum X}{n} = \frac{57}{7} = 8.1428$ *and* $\overline{Y} = \frac{\sum Y}{n} = \frac{511}{7} = 73$

So $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 73 - (-3.622)(8.1428) = 102.493$

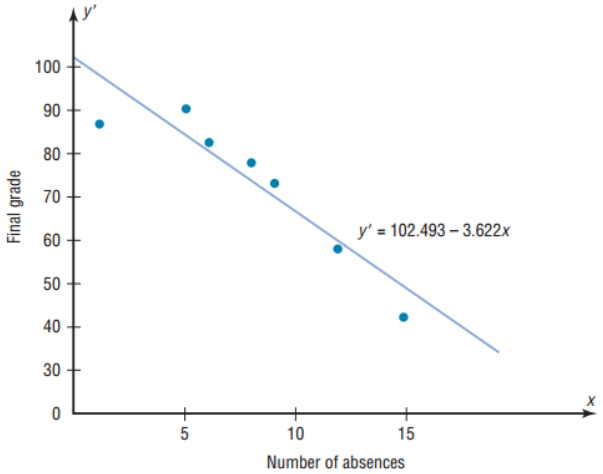
So the required regression equation will be given as $\hat{Y} = 102.493 - 3.622X$

(C) *SCATTER PLOT WITH REGRESSION LINE*

To graph a line, select any two points for and find the corresponding values of y.

If $x = 0$ *then* $y = 102.5$ *and If* $y = 0$ *then* $x = 28.3$

So two points are $A(0,102.5)$ & $B(28.3, 0)$



(D) *COEFFICIENT OF CORRELATION AND DETERMINATION*

Coefficient of Correlation is given by

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38993) - (511)^2]}} = -0.944$$

Interpretation: $r = -0.944$ shows that there is strong negative relation between X and Y.

Coefficient of Determination is given by

$$R^2 = r^2 = (-0.944)^2 = 0.89$$

Interpretation: $R^2 = 0.89$ shows that 89% variation in final grades is due to Number of absences and remaining 11% is due to other factors.

(E) *VALUE OF Y FOR GIVEN X* *Here* $X = 10$ *Since X is number of absences.*

Fitted regression equation is $\hat{Y} = 102.493 - 3.622X$

At $X = 10$ we have $\hat{Y} = 102.493 - 3.622(10) = 66.273$

Hence if the number of absences are 10 then final grade will be 66.27approximately.

EXERCISE – 5.2

SIMPLE LINEAR REGRESSION MODEL

1. The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Fit a linear regression model using the above data set also interpret the results.
(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.
(c) Find Coefficient of Determination and interpret it.
$(Ans: (a)\hat{y} = 12.06 + 0.78x + 12.16x_2 \quad (b)\hat{y} = 78 \quad (c) R^2 = \quad)$

2. A study was made by a retail merchant to determine the relation between weekly advertising expenditures and sales.

Advertising Costs ($)	Sales ($)
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

(a) Fit a linear regression model using the above data set also interpret the results.
(b) Estimate the weekly sales when advertising costs are $35.
(c) Find Coefficient of Determination and interpret it.
$(Ans: (a)\hat{y} = 343.71 + 3.22x \quad (b)\hat{y} = \$456.43 \quad (c) R^2 = \quad)$

3. A professor in the School of Business in a university polled a dozen colleagues about the number of professional meetings they attended in the past five years (x) and the number of papers they submitted to refereed journals (y) during the same period. The summary data are given as follows:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12,$$
$$\sum_{i=1}^{n} x_i^2 = 232, \quad \sum_{i=1}^{n} x_i y_i = 318.$$

(a) Fit a linear regression model using the above data set also interpret the results.
(b) Comment on whether attending more professional meetings would result in publishing more papers.
(c) Find Coefficient of Determination and interpret it.

$(Ans: (a)\hat{y} = 37.8 - 6.45x \qquad (c) R^2 =$
(b) *It appears that attending professional meetings would not result in publishing more papers.*$)$

4. Raw material used in the production of a synthetic fiber is stored in a place that has no humidity control. Measurements of the relative humidity and the moisture content of samples of the raw material (both in percentages) on 12 days yielded the following results:

Humidity	*Moisture content*
46	12
53	14
37	11
42	13
34	10
29	8
60	17
44	12
41	10
48	15
33	9
40	13

(a) Fit a linear regression model using the above data set also interpret the results.
(b) Predict the moisture content when the relative humidity is 38 percent.
(c) Find Coefficient of Determination and interpret it.

$(Ans:(a)\hat{y}=0.49+0.27x\,(b)\,\hat{y}=10.75\;\;(c)\,R^2=0.8464)$

5. A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature, x	**Converted Sugar, y**
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

(a) Fit a linear regression model using the above data set also interpret the results.
(b) Predict the mean amount of converted sugar produced when the coded temperature is 1.75.
(c) Find Coefficient of Determination and interpret it.

$(Ans:(a)\hat{y}=6.41+1.81x\;(b)\,\hat{y}=9.580\;\;(c)\,R^2=\quad)$