Bivariate Analysis

Regression and Correlation

Department of Mathematical Sciences

March 28, 2019

1 Intorduction

Regression analysis is a conceptually simple method for investigating functional relationships among variables.

In research laboratories, experiments are being performed daily. These are usually small, carefully planned studies and result in sets of data of modest size. The objective is often a quick yet accurate analysis, enabling the experimenter to move on to "better" experimental conditions, which will produce a product with desirable characteristics. Additional data can easily be obtained if needed, however, if the decision is initially unclear.

In any system in which variable quantities change, it is of interest to examine the effects that some variables exert (or appear to exert) on others. There may in fact be a simple functional relationship between variables; in most physical processes this is the exception rather than the rule. Often there exists a functional relationship that is too complicated to grasp or to describe in simple terms. In this case we may wish to approximate to this functional relationship by some simple mathematical function, such as a polynomial, which contains the appropriate variables and which approximates to the true function over some limited ranges of the variables involved.

By examining such approximating function we may be able to learn more about the underlying true relationship and to appreciate the separate and joint effects produced by changes in certain important variables.

Even where no sensible physical relationship exists between variables, we may wish to relate them by some sort of mathematical equation. While the equation might be physically meaningless, it may nevertheless be extremely valuable for predicting the values of some variables from knowledge of other variables, perhaps under certain stated restrictions.[**?**]

2 Variables in Regression

We can distinguish two main types of variable at this stage. We shall usually call these predictor variables and response variables.

By predictor variables we shall usually mean variables that can either be set to a desired value (e.g., input temperature or catalyst feed rate) or else take values that can be observed but not controlled (e.g., the outdoor humidity).

As a result of changes that are deliberately made, or simply take place in the predictor variables, an effect is transmitted to other variables, the response variables (e.g., the final color or the purity of a chemical product).

In general, we shall be interested in finding out how changes in the predictor variables affect the values of the response variables. Other names frequently seen are the following:

Predictor variables = input variables = inputs
= X-variables = regressors
= independent variables.
Response variables = output variables = outputs
= Y-variables
= dependent variables.

We shall concerned with relationships of the form:
Response variable = Model function + Random error.
The model function will usually be "known" and of specified form and will involve the predictor variables as well as parameters to be estimated from data. If We denote the response variable by Y and the set of predictor variables by $X_1, X_2, \ldots, X_p$, where p denotes the number of predictor variables. The true relationship between Y and $X_1, X_2, \ldots, X_p$ can be approximated by the regression model

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon \tag{2.1}$$

where ε is assumed to be a random error representing the discrepancy in the approximation.It accounts for the failure of the model to fit the data exactly.

An example is the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \ldots, + \beta_p X_p + \varepsilon \tag{2.2}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ called the regression parameters or coefficients, are unknown constants to be determined (estimated) from the data.

We need to select the form of the function $f(X_1, X_2, \ldots, X_p)$ in (2.1). This function can be classified into two types: linear and nonlinear.

An example of a linear function is

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{2.3}$$

while a nonlinear function is

$$Y = \beta_0 + e^{\beta_1 X_1} + \varepsilon \tag{2.4}$$

Note that the term linear (nonlinear) here does not describe the relationship between Y and $X_1, X_2, \ldots, X_p$. It is related to the fact that the regression parameters enter the equation linearly (nonlinearly). Each of the following models are linear

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 \ln X + \varepsilon$$

If the two models are re-expressed, respectively, as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

In first equation we have $X_1 = X$ and $X_2 = X^2$ and in second equation we have $X_1 = \ln X$.The variables here are re-expressed or transformed.

A regression equation containing only one predictor variable is called a simple regression equation. An equation containing more than one predictor variable is called a multiple regression equation.

When we deal only with one response variable, regression analysis is called univariate regression and in cases where we have two or more response variables, the regression is called multivariate regression.

3 Simple Linear Regression

We start with the simple case of studying the relationship between a response variable Y and a predictor variable X_1 . Since we have only one predictor variable, we shall drop the subscript in X_1 and use X for simplicity.

3.1 Covariation and Correlation Coefficient

Suppose we have observations on n subjects consisting of a dependent or response variable Y and an explanatory variable X . The observations are usually recorded as in Table (3.1). We wish to measure both the direction and the strength of the relationship between Y and X . Two related measures, known as the covariance and the correlation coefficient, are developed below.

Table 3.1: Notation for the Data Used in Simple Regression and Correlation

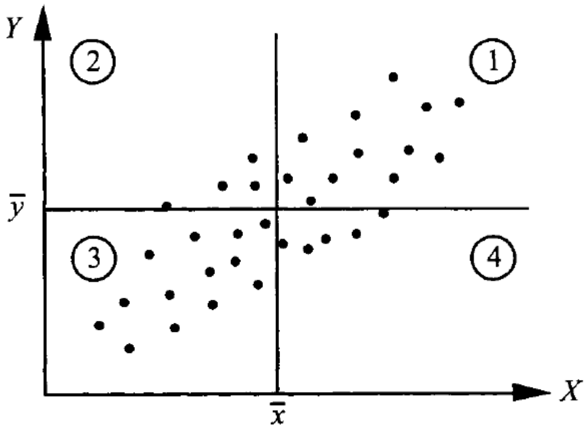
Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

On the scatter plot of Y versus X , let us draw a vertical line at $\bar{x}$ and a horizontal line at $\bar{y}$, as shown in Figure (1), where

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

are the sample mean of Y and X, respectively. The two lines divide the graph into four quadrants.

Figure 1: A graphical illustration of the correlation coefficient



For each point i in the graph, compute the following quantities:

- $y_i - \bar{y}$, the deviation of each observation y_i from the mean of the response variable,

- $x_i - \bar{x}$, the deviation of each observation x_i from the mean of the predictor variable, and
- the product of the above two quantities, $(y_i - \bar{y})(x_i - \bar{x})$.

It is clear from the graph that the quantity $(y_i - \bar{y})$ is positive for every point in the first and second quadrants, and is negative for every point in the third and fourth quadrants. Similarly, the quantity $(x_i - \bar{x})$ is positive for every point in the first and fourth quadrants, and is negative for every point in the second and third quadrants.

Table 3.2: Algebraic Signs of the Quantities $(y_i - \bar{y})$ and $(x_i - \bar{x})$

Quadrant	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x_i - \bar{x})$
1	+	+	+
2	+	-	-
3	-	-	+
4	-	+	-

If the linear relationship between Y and X is positive (as X increases Y also increases), then there are more points in the first and third quadrants than in the second and fourth quadrants. In this case, the sum of the last column in Table (3.2) is likely to be positive because there are more positive than negative quantities. Conversely, if the relationship between Y and X is negative (as X increases Y decreases), then there are more points in the second and fourth quadrants than in the first and third quadrants. Hence the sum of the last column in Table (3.2) is likely to be negative. Therefore, the sign of the quantity

$$Cov(Y, X) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} \tag{3.2}$$

which is known as the covariance between Y and X, indicates the direction of the linear relationship between Y and X

If $Cov(Y, X) > 0$, then there is a positive relationship between Y and X, but if $Cov(Y, X) < 0$, then the relationship is negative. Unfortunately, $Cov(Y, X)$ does not tell us much about the strength of such a relationship because it is affected by changes in the units of measurement.To avoid this disadvantage of the covariance, we standardize the data before computing the covariance.

We compute

$$z_i = \frac{y_i - \bar{y}}{s_y} \tag{3.3}$$

where

$$s_y = \sqrt{\frac{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \tag{3.4}$$

is the sample standard deviation of Y.It can be shown that the standardized variable z_i in (3.3) has mean zero and standard deviation one.We standardize X in a similar way by subtracting the mean $\bar{x}$ from each observation x_i then divide by the standard deviation s_x.

The covariance between the standardized X and Y data is known as the correlation coefficient between Y and X and is given by

$$Cor(Y, X) = \frac{\sum\limits_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s_y}\right)\left(\frac{x_i - \bar{x}}{s_x}\right)}{n-1} \tag{3.5}$$

Equivalent formulas for the correlation coefficient are:

$$Cor(Y,X) = \frac{Cov(Y,X)}{s_x s_y} \tag{3.6}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{3.7}$$

Thus, $Cor(Y,X)$ can be interpreted either as the covariance between the standardized variables or the ratio of the covariance to the standard deviations of the two Variables. From (3.7), it can be seen that the correlation coefficient is symmetric,that is, $Cor(Y,X) = Cor(X,Y)$.

Unlike $Cov(Y,X)$, $Cor(Y,X)$ is scale invariant, that is, it does not change if we change the units of measurements. Furthermore, $Cor(Y,X)$ satisfies

$$-1 \leq Cor(Y,X) \leq 1 \tag{3.8}$$

These properties make the $Cor(Y,X)$ a useful quantity for measuring both the direction and the strength of the relationship between Y and X . The magnitude of $Cor(Y,X)$ measures the strength of the linear relationship between Y and X . The closer $Cor(Y,X)$ is to 1 or -1, the stronger is the relationship between Y and X . The sign of $Cor(Y,X)$ indicates the direction of the relationship between Y and X . That is, $Cor(Y,X) > 0$ implies that Y and X are positively related. Conversely, $Cor(Y,X) < 0$, implies that Y and X are negatively related. $Cor(Y,X) = 0$,implies that Y and X are not linearly related because the correlation coefficient measures only linear relationships.

Example 3.1. *Standard aqueous solutions of fluorescein are examined in a flu-*

orescence spectrometer, and yield the following fluorescence intensities (in arbitrary units):

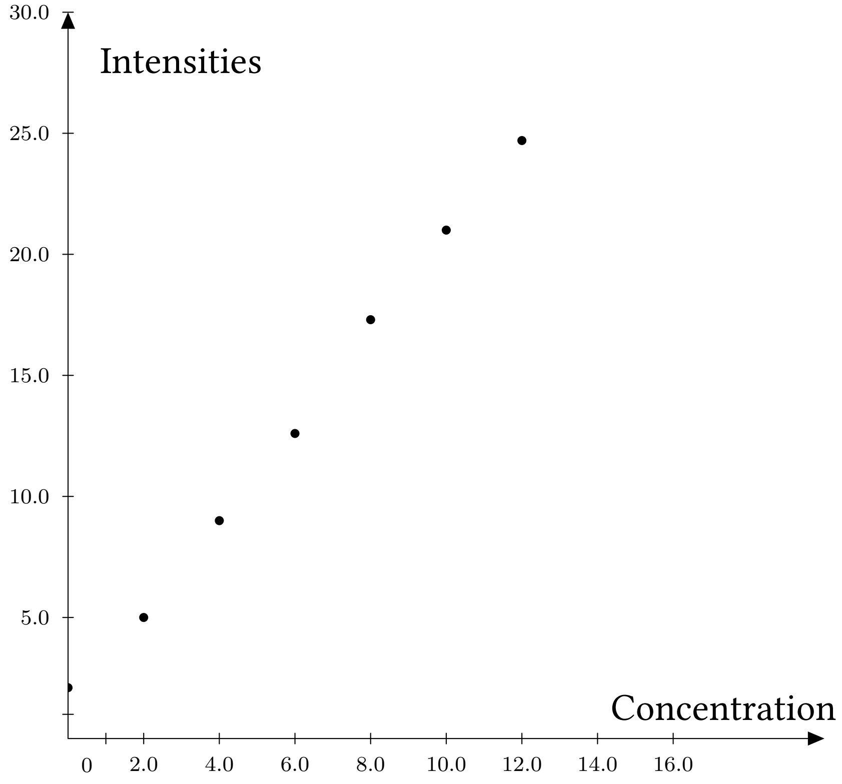
Table 3.3: fluorescence intensities (in arbitrary units)

Fluorescence intensities:	2.1	5.0	9.0	12.6	17.3	21.0	24.7
Concentration, pg ml^{-1}:	0	2	4	6	8	10	12

Determine the correlation coefficient.

Solution 3.1. *Let x denotes Concentration (pg ml^{-1})and y denotes Fluorescence intensities.*

Figure 2: Scatter Plot



From table 3.4

$$\begin{aligned}
\bar{y} &= \frac{\sum_{i=1}^{n} y_i}{n} \\
&= \frac{91.7}{7} = 13.1 \\
\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{42}{7} = 6
\end{aligned}$$

Table 3.4: Computing Correlation

Observation Number	x_i	y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	0	2.1	-6	36	-11.0	121.00	66.0
2	2	5.0	-4	16	-8.1	65.61	32.4
3	4	9.0	-2	4	-4.1	16.81	8.2
4	6	12.6	0	0	-0.5	0.25	0.0
5	8	17.3	2	4	4.2	17.64	8.4
6	10	21.0	4	16	7.9	62.41	31.6
7	12	24.7	6	36	11.6	134.56	69.6
Total	42	91.7	0	112	0	418.28	216.2

$$
\begin{aligned}
Cor(Y,X) &= \frac{\sum\limits_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}} \\
&= \frac{216.2}{\sqrt{418.28}\sqrt{112}} \\
&= 0.9989
\end{aligned}
$$

4 Simple Linear Regression Model

The relationship between a response variable Y and a predictor variable X is linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{4.1}$$

where β_0 and β_1, are constants called the model regression coefficients or parameters, and ε is a random disturbance or error. That is Y is approximately a linear function of X , and ε measures the discrepancy in that approximation.

In particular ε contains no systematic information for determining Y that is not already captured in X. The coefficient β_1, called the slope, may be interpreted as the change in Y for unit change in X. The coefficient β_0, called the constant coefficient or intercept, is the predicted value of Y when $X = 0$. According to (4.1), each observation in Table 3.1 can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \ldots, n \tag{4.2}$$

where y_i represents the i^{th} value of the response variable Y , x_i represents the i^{th} value of the predictor variable X, and ε_i represents the error in the approximation of y_i.

Regression analysis differs in an important way from correlation analysis. The correlation coefficient is symmetric in the sense that $Cor(Y, X)$ is the same as $Cor(X, Y)$. The variables X and Y are of equal importance. In regression analysis the response variable Y is of primary importance. The importance of the predictor X lies on its ability to account for the variability of the response variable Y and not in itself per se. Hence Y is of primary importance.

5 Parameter Estimation

Based on the available data, we wish to estimate the parameters β_0 and β_1. This is equivalent to finding the straight line that gives the best fit (representation) of the points in the scatter plot of the response versus the predictor variable (See Figure 2). We estimate the parameters using the popular least squares method, which gives the line that minimizes the sum of squares of the vertical distances from each point to the line. The vertical distances represent the errors in the response variable. These errors can be obtained by rewriting (4.2) as

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, i = 1, 2, \ldots, n \tag{5.1}$$

The sum of squares of these distances can then be written as

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2, i = 1, 2, \ldots, n \tag{5.2}$$

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes $S(\beta_0, \beta_1)$ is given by

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.3}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{5.4}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimates of β_0 and β_1 because they are the solution to the least squares method, the intercept and the slope of the line that has the smallest possible sum of squares of the vertical distances from each point to the line. For this reason, the line is called the least squares regression line. The least squares regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{5.5}$$

For each observation in our data we can compute

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \ldots, n \tag{5.6}$$

These are called the fitted values. Thus, the i^{th} fitted value, $\hat{y_i}$, is the point on the least squares regression line (5.5) corresponding to x_i. The vertical distance corresponding to the i^{th} observation is

$$e_i = y_i - \hat{y_i} \tag{5.7}$$

These vertical distances are called the ordinary least squares residuals. One properties of the residuals in (5.7) is that their sum is zero.This means that the sum of the distances above the line is equal to the sum of the distances below the line.

Using data in Table 3.4 we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{216.2}{112} = 1.93$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 13.1 - (1.93)(6) = 1.52$$

Then the equation of the least squares regression line is

$$\hat{Y} = 1.52 + 1.93X \tag{5.8}$$

Figure 3: least squares regression line $\hat{Y} = 1.52 + 1.93X$

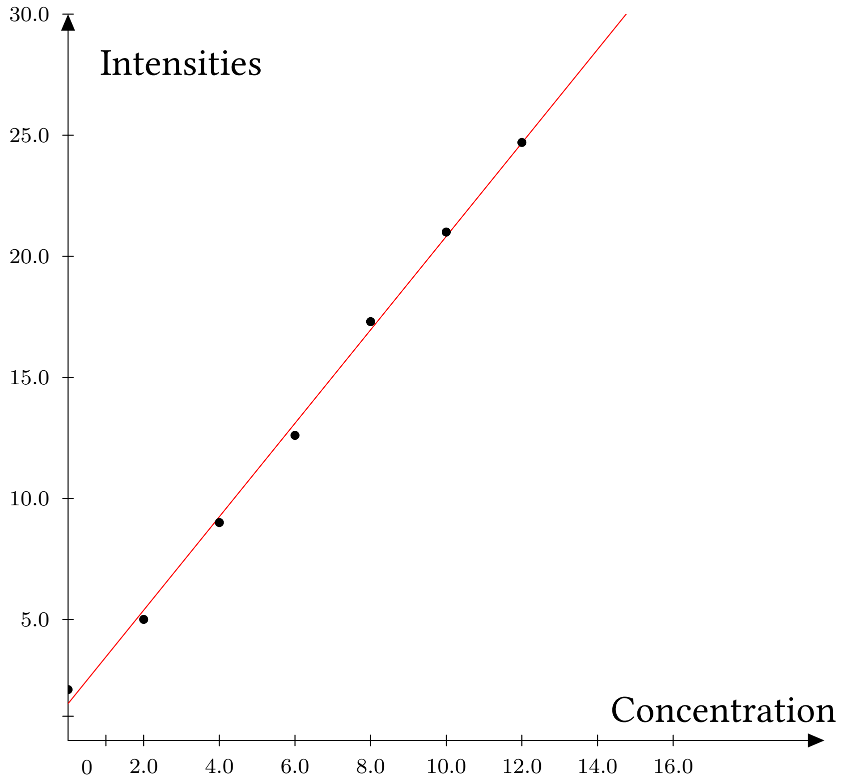


Table 5.1: The Fitted Values, y_i, and the Ordinary Least Squares Residuals, e_i

Observation Number	x_i	y_i	$\hat{y}_i$	e_i
1	0	2.1	1.52	0.58
2	2	5.0	5.38	-0.38
3	4	9.0	9.24	-0.24
4	6	12.6	13.10	-0.50
5	8	17.3	16.96	0.34
6	10	21.0	20.82	0.18
7	12	24.7	24.68	0.02
Total	42	91.7	91.70	0.00

We should note here that by comparing (3.2), (3.7), and (5.3), an alternative formula for $\hat{\beta}_1$ can be expressed as

$$\hat{\beta}_1 = \frac{Cov(Y,X)}{\text{Var}(X)} = Cor(Y,X)\frac{s_y}{s_x} \tag{5.9}$$

from which it can be seen that ,$\hat{\beta}_1$, $Cov(Y,X)$, and $Cor(Y,X)$ have the same sign. This makes intuitive sense because positive (negative) slope means positive (negative) correlation.

6 Exercise

Example 6.1. *The chemical of compound y, which were dissolved in 100 grams in water at various temperatures, x were recorded as follows.*

Table 6.1: Data for Regression

$x(^oc)$	15	15	30	30	45	45	60	60
y (grams)	12	10	25	21	31	33	44	39

Obtain the line of regression of y on x and estimate the amount of chemical that will dissolve in 100 grams of water at 50 ^0c

Example 6.2. *People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months. Table 6.7 gives data on both HbA and FPG for 7 diabetics five months after they had completed a diabetes education class.*

Table 6.2: Two measures of glucose level in diabetics

Subject	1	2	3	4	5	6	7
HbA(%)	7.1	7.5	7.7	7.9	8.7	9.4	10.4
FPG (mg/ml)	95	96	78	148	172	200	271

(a) Make a scatterplot of the data.

(b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?

Example 6.3. *The data come from a study that investigated optimal conditions for the extraction of tobacco alkaloids using ultrasonic and microwave extraction methods. The first row shows the extraction temperature ($^{\circ}C$), while the second row gives the percentage nicotine extracted from tobacco.*

Table 6.3: The nicotine study data

Temperature °C	41	41	74	74	30	85	57.5	57.5
% Nicotine	3.279	3.401	3.973	4.319	3.145	4.595	3.945	4.243

(a) Identify dependent and independent variables from the data (Table 6.3)

(b) Assume a linear relationship,Predict the % Nicotine at Temperature $50°C$ from regression line.

Example 6.4. *The following are the ages (years) and systolic blood pressures of 10 apparently healthy adults:*

Table 6.4: Data for Correlation

Age (X)	46	53	70	20	63	43	26	19	31	23
BP(Y)	128	136	146	124	143	130	124	121	126	123

(a) Construct the scatter diagram of variables in Table 6.4

(b) Use the transformation $u = \dfrac{X-40}{5}$ and $v = \dfrac{Y-130}{5}$ and compute the simple correlation coefficient r_{uv} from the data (Table 6.4)

Example 6.5. *Consider the data in Table (6.5).Construct the scatter diagram and obtain a line of regression of Y on X.*

Table 6.5: Data for regression

x	1	2	3	4	5	6	7
y	3	2	5	8	7	10	15

Example 6.6. *In medicine manufacturing process, mainly two processes viz. process -X and process -Y, are performed in a company which has cause and*

effect relationship. Following data is of time (in minutes) required to complete process X and process Y.

Table 6.6: time in minutes

X	65	66	67	67	68	69	70	72
Y	67	68	65	72	72	72	69	71

(i) Obtain the regression line of Y on X.

(ii) What will be the time required for completing the process Y if process X consumes 75 minute?

Example 6.7. *Consider the data in Example 6.6 again*

(i) Construct the scatter plot of variables in table (6.6)

(ii) Find simple linear correlation r between variables in table (6.6) and write your conclusion about value of r.

Example 6.8. *People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months. Table 6.7 gives data on both HbA and FPG for 18 diabetics five months after they had completed a diabetes education class.*

Table 6.7: Two measures of glucose level in diabetics

	HbA	FPG		HbA	FPG		HbA	FPG
Subject	(%)	(mg/ml)	Subject	(%)	(mg/ml)	Subject	(%)	(mg/ml)
1	6.1	141	7	7.5	96	13	10.6	103
2	6.3	158	8	7.7	78	14	10.7	172
3	6.4	112	9	7.9	148	15	10.7	359
4	6.8	153	10	8.7	172	16	11.2	145
5	7.0	134	11	9.4	200	17	13.7	147
6	7.1	95	12	10.4	271	18	19.3	255

(i) *Make a scatterplot of the data.*

(ii) *Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?*

Example 6.9. *In a stability testing programme samples of a drug product were stored at 25°C and 60% relative humidity. Determinations of the potency of the drug product were made at seven time points: zero, 3, 6, 9, 12, 18 and 24 months after manufacture. The reported results are 'percentage of label claim' shown in Table (6.8)*

Table 6.8: The stability data

Month	0	3	6	9	12	18	24
Potency (%)	102.6	100.4	98.4	99.4	99.8	97.8	97.6

(i) *Construct the scatter diagram from the data and write interpretation.*

(ii) *Obtain regression line to predict average % potency of drug product two years after manufacture.*

(iii) Determine the time when potency of drug product will have fallen to 95$\%$ of its label claim using regression line obtained in question (ii)

Example 6.10. *Consider, the data from biomedical analysis in Table (6.9). They represent Cu, Mn and Zn concentrations determined in 8 different structures of the human brain.*

Table 6.9: Concentration of Cu, Zn and Mn in different brain structures

Brain Structure	1	2	3	4	5	6	7	8
Cu	25.8	24.2	27.3	32.8	27.3	17.9	14.0	13.3
Mn	1.0	1.0	1.1	1.5	1.8	1.2	1.1	1.0
Zn	78.0	81.8	69.4	76.1	62.5	60.1	34.2	35.5

(i) Construct scatter diagram of variables Cu and Zn and write the interpretation.

(ii) Compute simple correlation between the variables Cu and Zn

Example 6.11. *Investigators at a sports health center are interested in the relationship between oxygen consumption and exercise time in athletes recovering from injury. Appropriate mechanics for exercising and measuring oxygen consumption are set up, and the results are presented in table (6.10) below.*

Table 6.10: Data on Exercise time and Oxygen Consumption

x exercise time (min)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y oxygen consumption	620	630	800	840	840	870	1010	940	950	1130

(i) Construct the scatter plot of the data.

(ii) Determine a regression line that estimate oxygen consumption.

(iii) Estimate oxygen consumption for exercise time 7 minutes based on regression line obtained in (ii)

Example 6.12. *Following data show chest circumference and Birth Weight of 10 babies.*

Table 6.11: Chest Circumference and Birth Weight

x chest circum (cm)	22.4	27.5	28.5	28.5	29.4	29.4	30.5	32	31.4	32.5
y Birth Weight (kg)	2.00	2.25	2.10	2.35	2.45	2.50	2.80	2.80	2.55	3.00

(i) Construct the scatter plot of the data in table (6.11).

(ii) Determine correlation coefficient r.

(iii) Write your interpretation from the plot and numerical value of r.