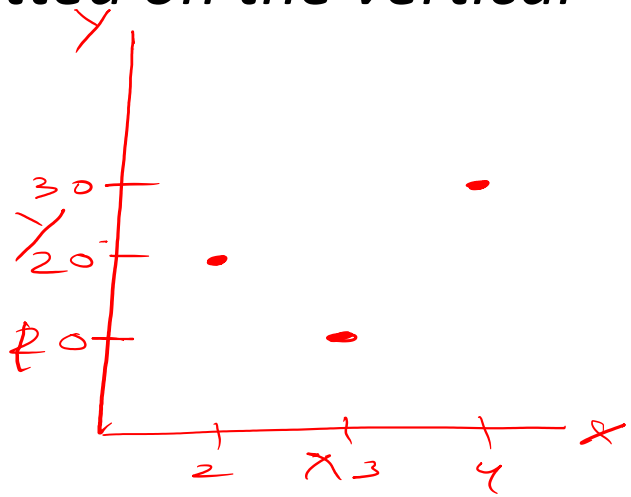
Simple Linear Regression

Scatter Plot

- The independent and dependent variables can be plotted on a graph called a ***scatter plot***. *The independent variable, x, is plotted on the horizontal axis and the dependent* variable, *y, is plotted on the vertical axis.*

I.V → x y ← D.V

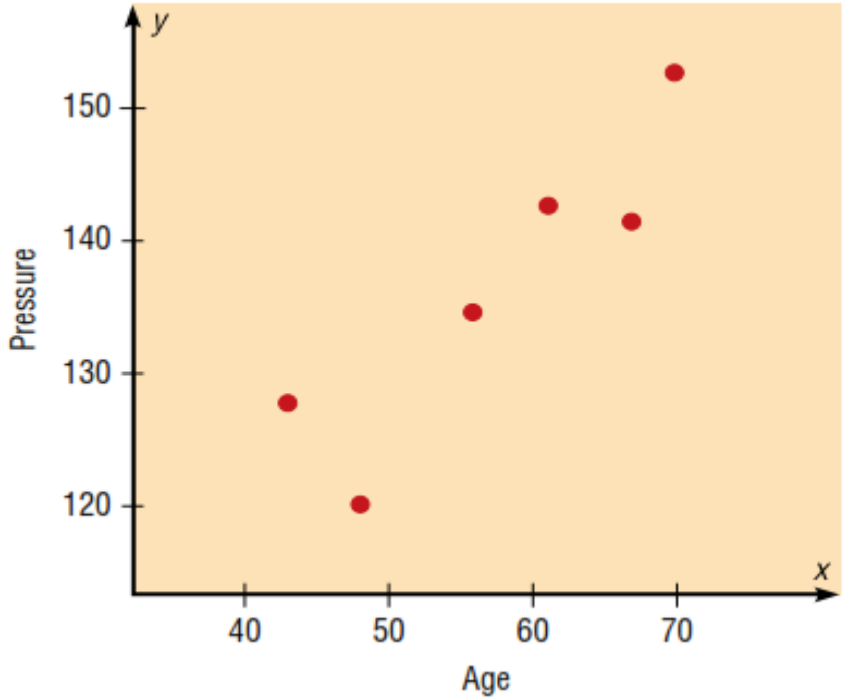
x	y
2	20
3	10
4	30



Scatter plot (Example 01)

- Construct a scatter plot for the data obtained in a study of age and systolic blood pressure of six randomly selected subjects. The data are shown in the following table.

Subject	Age, x	Pressure, y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

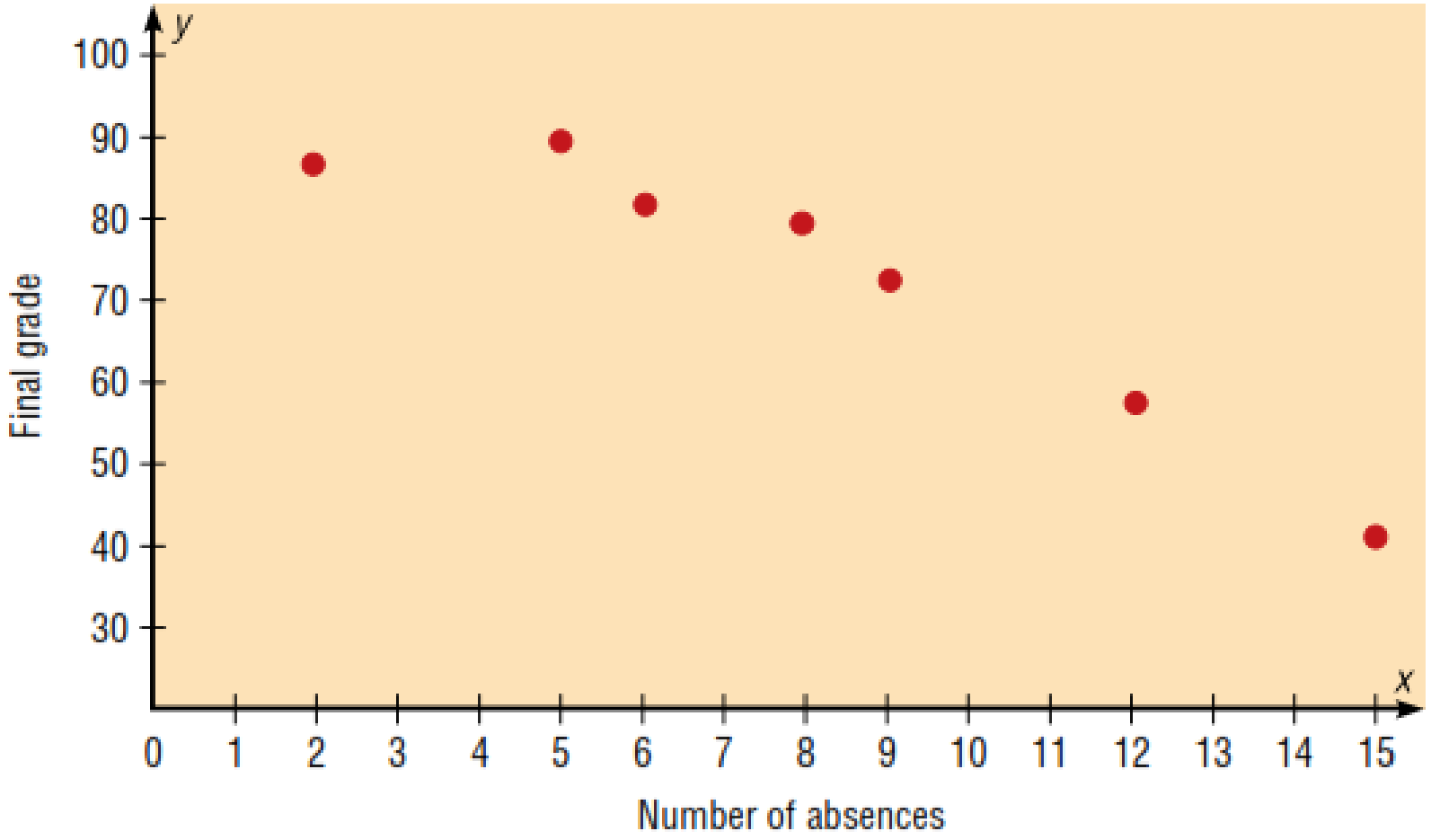


Scatter plot (Example # 02)

- Construct a scatter plot for 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

Scatter plot (Example # 02, Contd.)

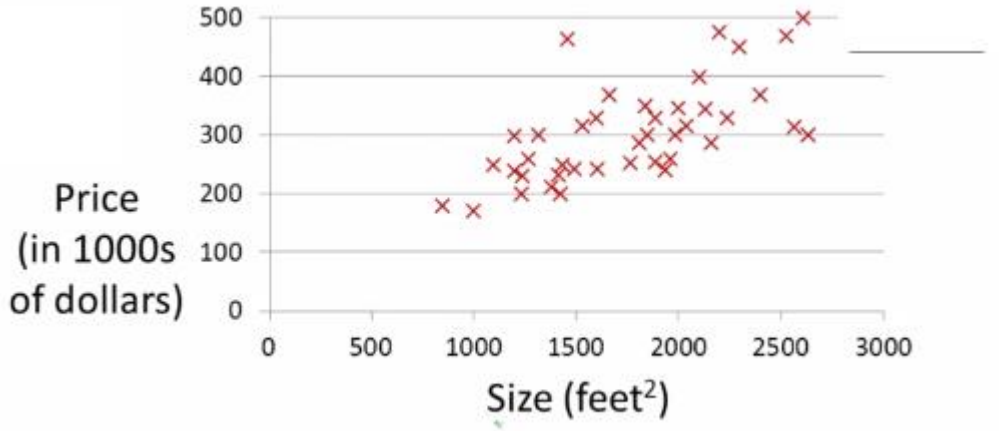


Consider the following example

Consider house prices and area of houses in Karachi. Price of house depends on the area of house and we all would be interested in a method to predict the price by area.

X Y

500



Now, of course, it is highly likely that for many example runs in which the area is the same, say 1500 Sq feet, the price will not be the same.

In house price prediction example

Area of house is the independent variable or regressor
Price of house is the dependent variable

The concept of regression analysis deals with finding the best relationship between Y and x, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressor x.

Deterministic Vs. Statistical Relationship

- In regression analysis we are concerned with what is known as the *statistical, not deterministic* dependence among variables.

- In statistical relationships among variables we essentially deal with **random or stochastic** variables i.e. variables that have probability distributions.

Deterministic Vs. Statistical (Example)

- The dependence of crop yield on temperature, rainfall, sunshine, and fertilizer, for example, is statistical in nature in the sense that the explanatory variables, although certainly important, will not enable the agronomist to predict crop yield exactly.

- In **deterministic phenomena**, on the other hand, we deal with relationships of the type, say, Conversion of temperature from Celsius into kelvin scale

Regression Analysis

Ind var

Analysis of a single regressor is called Simple Regression.

Analysis of more than one regressor is called Multiple Regression.

$y = f(x_1, x_2, x_3)$

In this course we will only consider linear relationship case of Simple regression.

Simple Linear Regression Model

- The statistical model for simple linear regression is given below. The response **Y** is related to the independent variable **x** through the equation:

$$Y = \beta_0 + \beta_1 x + \epsilon.$$ ← epsilon

In the above, β_0 and β_1 are unknown intercept and slope parameters, respectively, and ϵ is a random variable that is assumed to be distributed with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The quantity σ^2 is often called the error variance or residual variance.

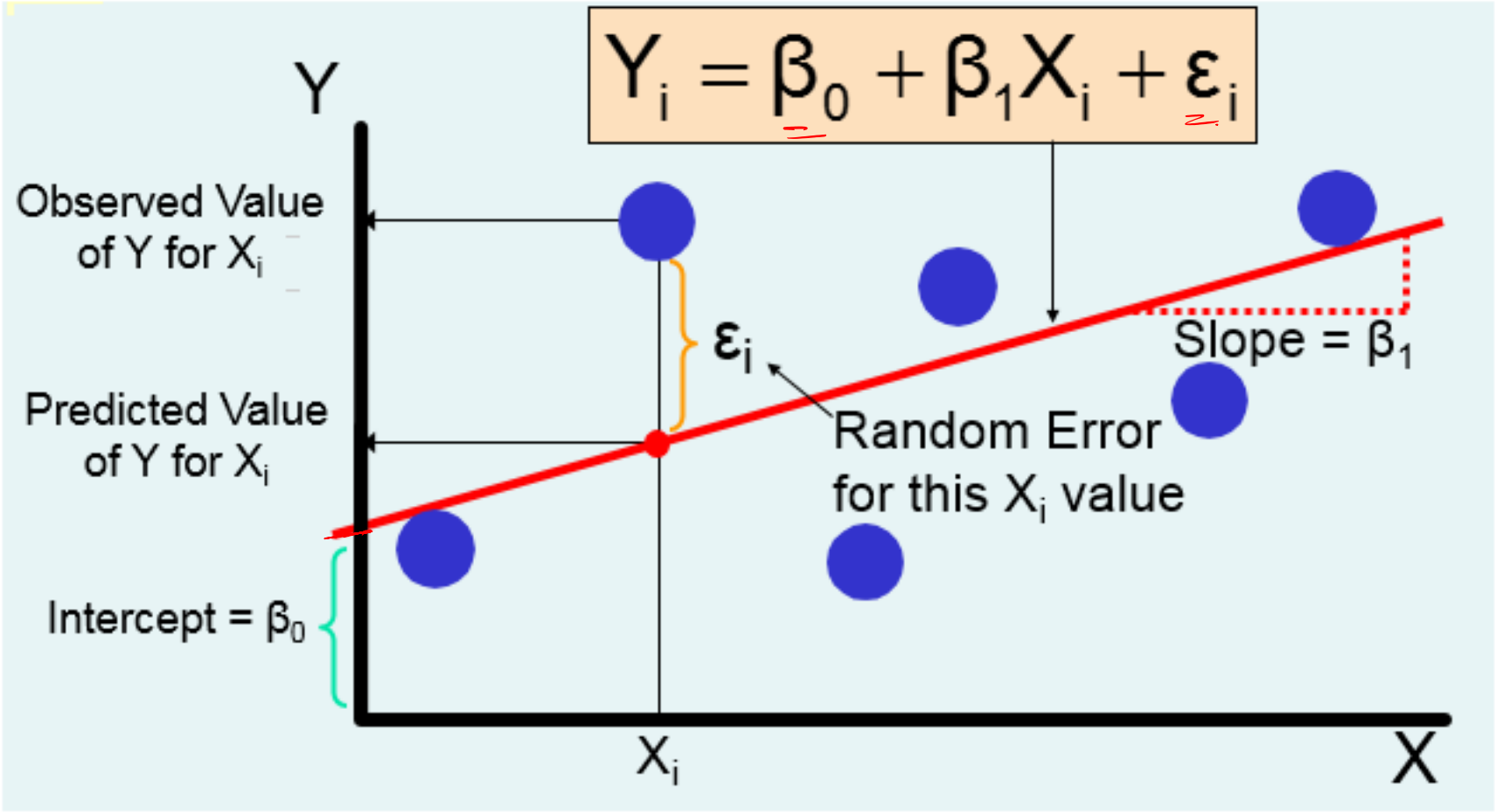
$$Y = \beta_0 + \beta_1 x + \epsilon.$$

- The quantity **Y** is a random variable since is **ϵ** random.
- The value **x** of the regressor variable is not random and, in fact, is measured with negligible error.
- The quantity **ϵ**, often called a **random error** or **random disturbance**, has constant variance (homogeneous variance).
- The presence of this random error, , keeps the model from becoming simply a deterministic equation.

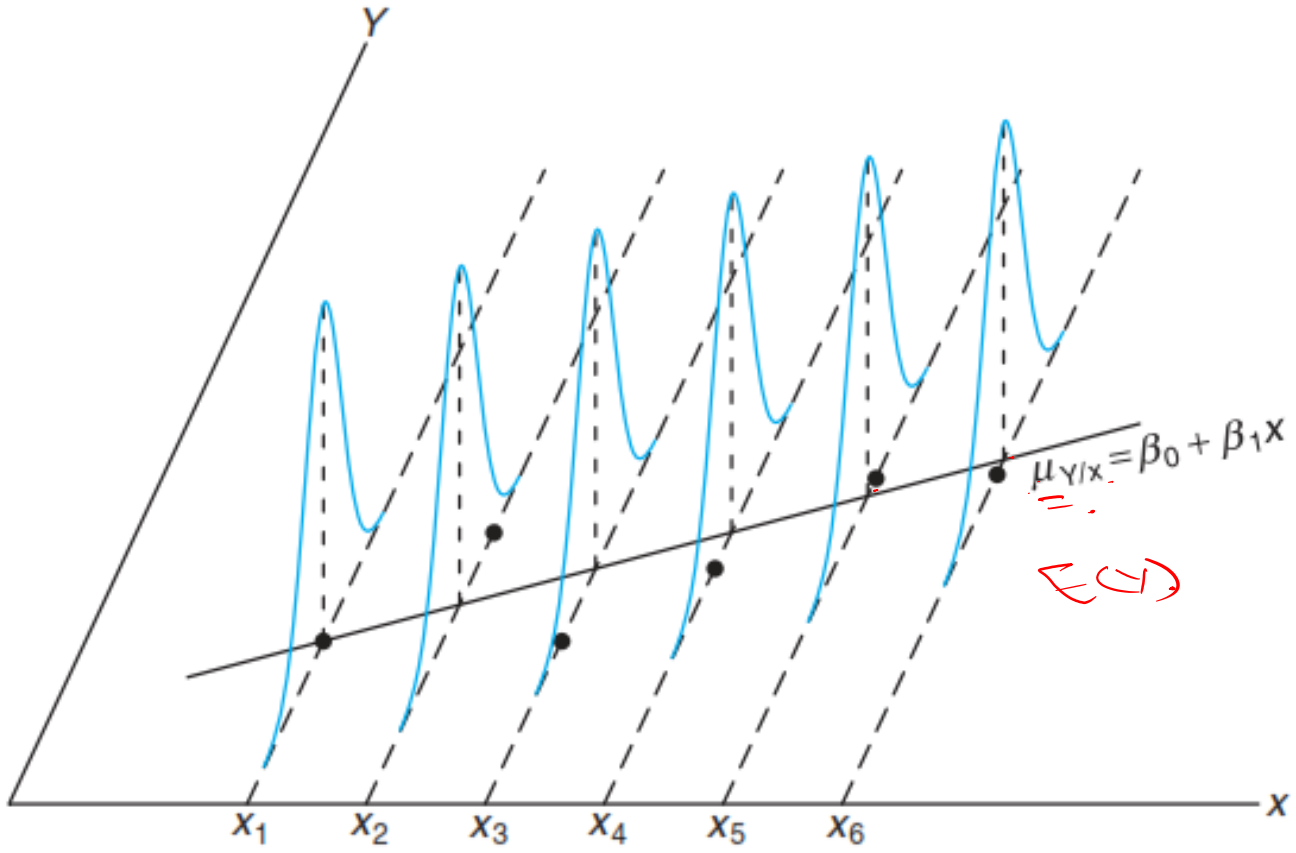
$$Y = \beta_0 + \beta_1 x + \epsilon.$$

- We must keep in mind that:
  - in practice ß₀ and ß₁ are not known and must be estimated from data.
  - we never observe the actual **ϵ** values in practice and thus we can never draw the true regression line
- We can only draw an estimated line.

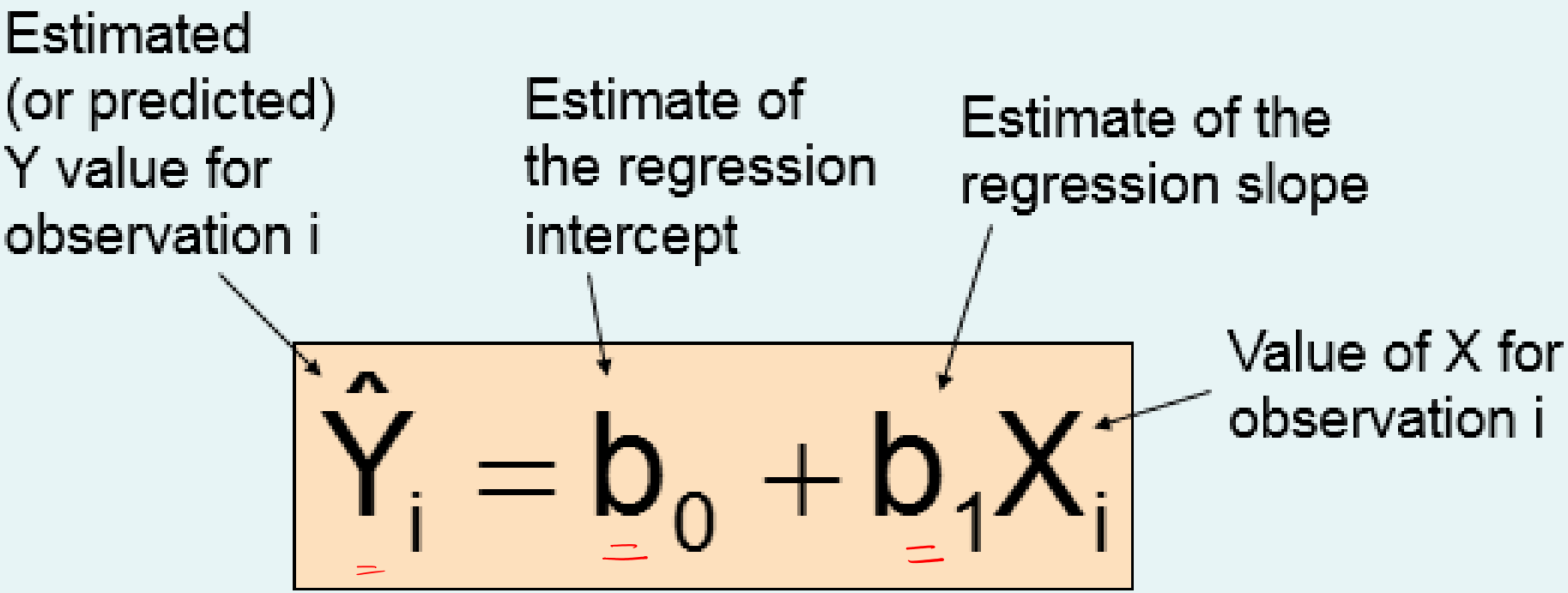
Simple Regression Model (Contd.)



Assumptions of Regression Model



The simple linear regression equation provides an estimate of the population regression line



Determining Regression Equation

- There are several methods for estimating the regression parameters, here we will use **Method of Least Sq.** to estimate the parameters.

$$SSE = e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The Method of Least Squares

- We shall find b_0 and b_1, so that the sum of the squares of the residuals is a minimum.

- The residual sum of squares is often called the sum of squares of the errors about the regression line and is denoted by *SSE*.

- This minimization procedure for estimating the parameters is called the **method of least squares**

→ Gradient Descent ←

The Method of Least Squares (Contd.)

$\hat{y}_i = b_0 + b_1 x_i \rightarrow SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \quad E(y_i - \hat{y}_i)^2$

$\frac{\partial SSE}{\partial b_0} = 0 \qquad \frac{\partial SSE}{\partial b_1} = 0$

$2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-1) = 0 \qquad 2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-x_i) = 0$

Estimating the Regression Coefficients

Given the sample $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

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

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}.$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

11.1 A study was conducted at Virginia Tech to de termine if certain static arm-strength measures hav an influence on the "dynamic lift" characteristics of a individual. Twenty-five individuals were subjected t strength tests and then were asked to perform a weigh lifting test in which weight was dynamically lifted ove head. The data are given here.

Individual	Arm Strength, x	Dynamic Lift, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

(a) Estimate β_0 and β_1 for the linear regression curve $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Find a point estimate of $\mu_{Y|30}$.

(c) Plot the residuals versus the x's (arm strength). Comment.

$\hat{y}_i = b_0 + b_1 x$

xy: 17.3×71.7, 19.3×48.3

x^2: $(17.3)^2$, $(19.3)^2$

$\hat{y}_i$: $64.5 + 0.56(17.3)$, $64.5 + 0.56(19.3)$

b) $\hat{y}_i = 64.5 + 0.56x$

$\hat{y} = 64.5 + 0.56(30)$

$e_i = y_i - \hat{y}_i$

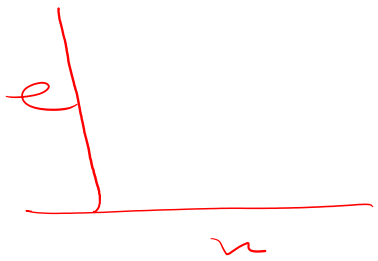
$e_i = 17.3 - 74.2$

Residual: Error in Fit — Given a set of regression data $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ and a fitted model, $\hat{y}_i = b_0 + b_1 x_i$, the ith residual e_i is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$



(a) $\sum_i x_i = 778.7$, $\sum_i y_i = 2050.0$, $\sum_i x_i^2 = 26,591.63$, $\sum_i x_i y_i = 65,164.04$, $n = 25$. Therefore,

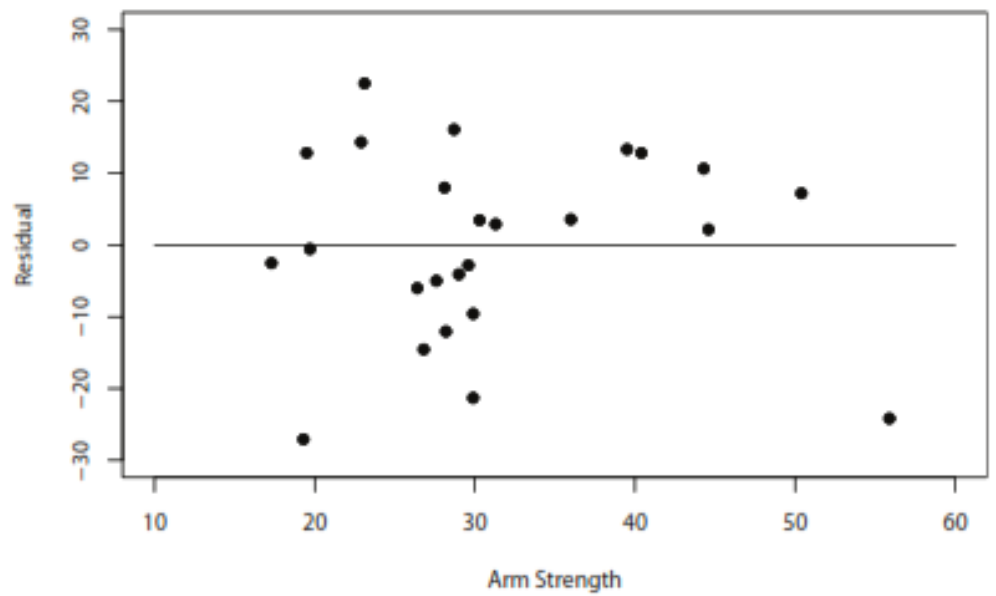
$$b_1 = \frac{(25)(65,164.04) - (778.7)(2050.0)}{(25)(26,591.63) - (778.7)^2} = \underline{0.561},$$

$$b_0 = \frac{2050 - (0.5609)(778.7)}{25} = \underline{64.529}.$$

a) $\hat{y}_i = 64.529 + 0.561 x_i$

(b) Using the equation $\hat{y} = 64.529 + 0.561x$ with $x = 30$, we find $\hat{y} = 64.529 + (0.561)(30) = 81.359$.

(c) Residuals appear to be random as desired.



11.12 Correlation

- to determine the strength of the relationship between two variables. There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**, named after statistician **Karl Pearson,** who pioneered the research in this area.

- The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is **r**. The symbol for the population correlation coefficient is **ρ** (Greek letter rho).

Properties of Correlation Coefficient

- Correlation coefficient is symmetric $r_{xy} = r_{yx}$.
- Correlation coefficient does not depend on units.
- The correlation coefficient lies between – 1 to + 1 i.e. $-1 \leq r \leq +1$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

Correlation Coefficient

The measure ρ of linear association between two variables X and Y is estimated by the **sample correlation coefficient** r, where

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Example # 03

- Compute the value of the correlation coefficient for the data obtained in the study of age and blood pressure:

Subject	Age, x	Pressure, y	xy	x^2	y^2
A	43	128			
B	48	120			
C	56	135			
D	61	143			
E	67	141			
F	70	152			

Example # 03 (contd.)

Subject	Age, x	Pressure, y	xy	x^2	y^2
A	43	128	5,504	1,849	16,384
B	48	120	5,760	2,304	14,400
C	56	135	7,560	3,136	18,225
D	61	143	8,723	3,721	20,449
E	67	141	9,447	4,489	19,881
F	70	152	10,640	4,900	23,104
	$\Sigma x = 345$	$\Sigma y = 819$	$\Sigma xy = 47{,}634$	$\Sigma x^2 = 20{,}399$	$\Sigma y^2 = 112{,}443$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(47{,}634) - (345)(819)}{\sqrt{[(6)(20{,}399) - (345)^2][(6)(112{,}443) - (819)^2]}} = 0.897$$

The correlation coefficient suggests a strong positive relationship between age and blood pressure.

Example # 04

- Compute the value of the correlation coefficient for the data obtained in the study of the number of absences and the final grade of the seven students in the 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

Example # 04 (Contd.)

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
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38{,}993) - (511)^2]}} = -0.944$$

Coefficient of determination

r^2

- The r^2 is a measure of variation of the dependent variable that is explained by the regression line and the independent variable.

$1 - r^2$ $0 \leq r^2 \leq 1$

- Therefore, if r^2 = 0.845 or 84.5% then this result means that 84.5% of the variation in the dependent variable is accounted for by the variations in the independent variable. The rest of the variation, 15.5%, is unexplained.

Practice Questions for Simple Linear Regression

Probability and Statistics for Engineers and Scientists 9th by Walpole,Myers

11.1-11.14

11.43-11.47

$r = 0.97$

r r^2