Statistics for Computing

(CSC 502 0.0)

MSc in Computer Science

Prof. (Dr.) R.M. KAPILA RATHNAYAKA

B.Sc. Special (Math. & Stat.) (Ruhuna), M.Sc. (Industrial Mathematics) (USJ), M.Sc. (Stat.) (WHUT, China),

Ph.D. (Applied Statistics, WHUT)

Introduction to Correlation and Regression Analysis

Chapter 05

The Regression Analysis

- Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest.

- A regression analysis generates an equation to describe the statistical relationship between one or more predictors and the response variable and to predict new observations.

- It is a statistical tool used to determine the probable change in one variable for the given amount of change in another. This means, the value of the unknown variable can be estimated from the known value of another variable

Regression Equation

- The **Regression Equation** is the algebraic expression of the regression lines.

$$Y = a + b X$$

- X- independent variable
- Y - dependent variable
- a - intercept on Y axis
- b - slope of the line

- Dependent Variable: This is the main factor that you're trying to understand or predict.

- Independent Variables: These are the factors that you hypothesize have an impact on your dependent variable.

- There are two methods of obtaining regression line

  1) The scatter diagram method

  2) Method of least square

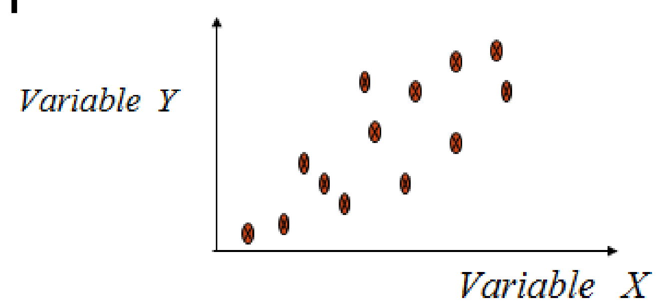
The scatter diagram method

- Scatter diagram is the simplest method for representing data.

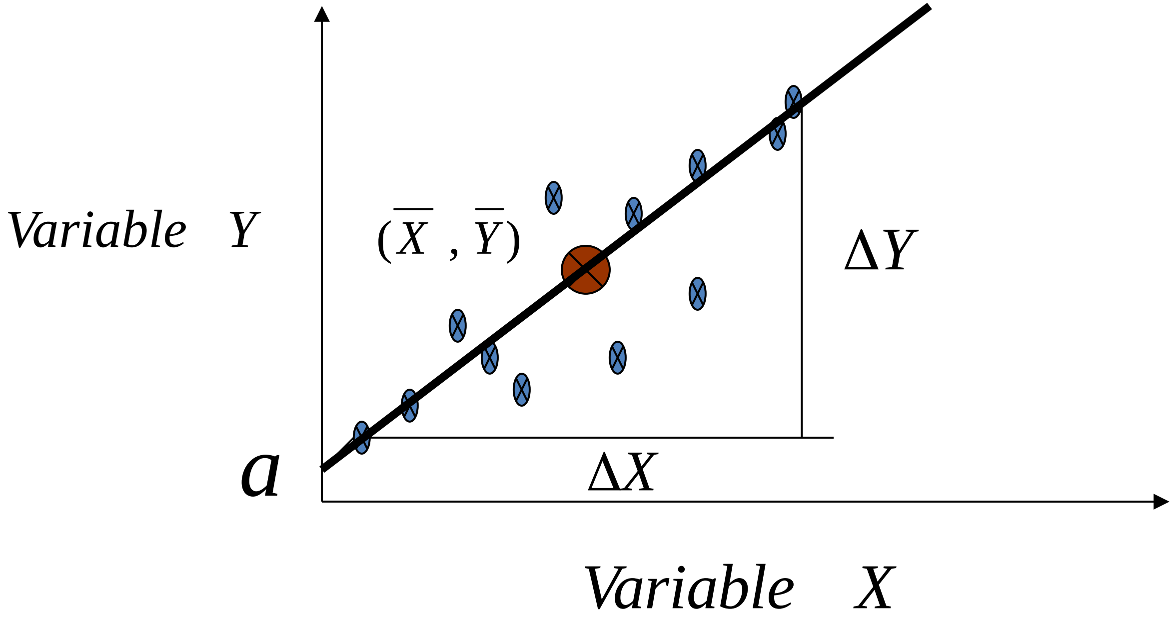
- Suppose the two variables are X and Y and there are ‘n’ pairs of values $(x_1, y_1), (x_2, y_2), \dots\dots\dots, (x_n, y_n)$

- Generally independent variable is plotted along the horizontal (X) axis and depend variable plotted along the vertical (Y) axis.

- Plotting your data is the first step in figuring out if there is a relationship between your independent and dependent variable



- Calculate $(\bar{X}, \bar{Y})$ values.
- The paired observations are plotted.
- Then draw the line through the mean point.



$$\boxed{b = \frac{\Delta Y}{\Delta X}}$$

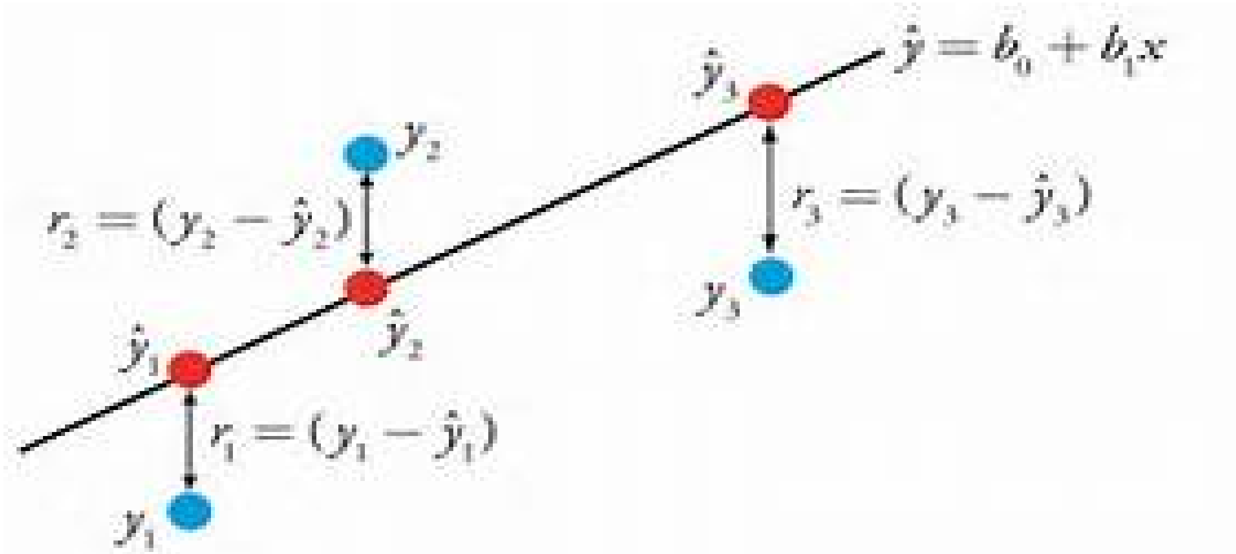
$$\boxed{Y = a + b\,X}$$

Why should your organization use regression analysis?

- Regression analysis is helpful statistical method that can be leveraged across an organization to determine the degree to which particular independent variables are influencing dependent variables.

The Method of Ordinary Least Squares

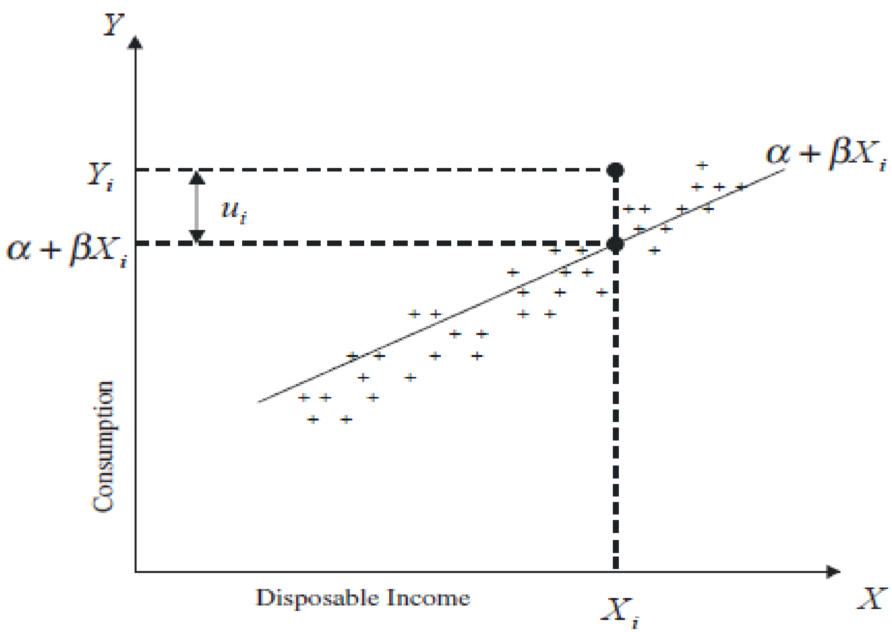
- In ordinary least squares (OLS) regression, the estimated equation is calculated by determining the equation that minimizes the sum of the squared distances between the sample's data points and the values predicted by the equation.



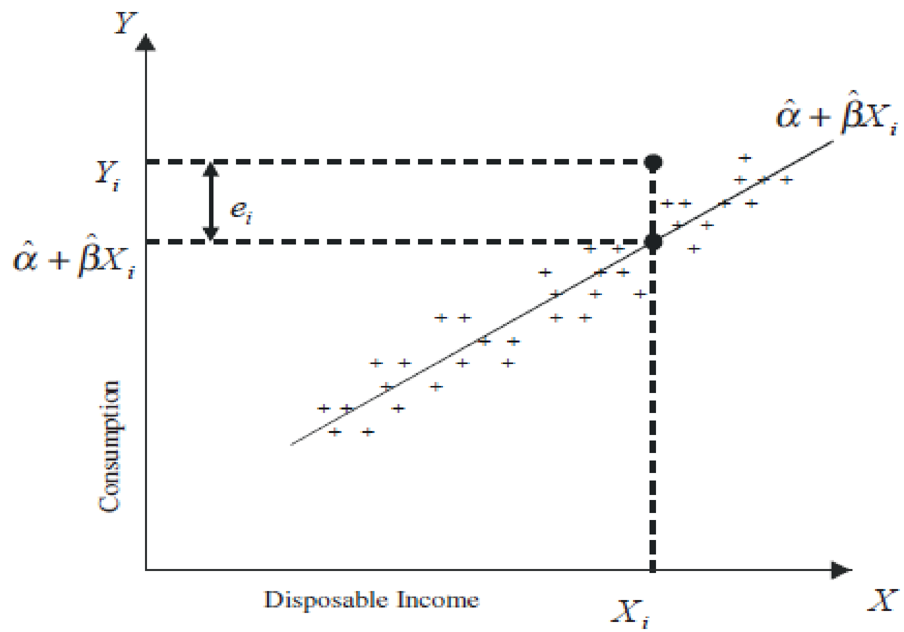
The Classical Assumptions

Assumption 1: The disturbances have zero mean, i.e., $\mathrm{E}(u_i) = 0$ for every $i = 1, 2, \ldots, n$.

- This assumption is needed to insure that on the average we are on the true line.



'True' Consumption Function



Estimated Consumption Function

Assumption 2: The disturbances have a constant variance, i.e., $\mathrm{Var}(u_i) = \sigma^2$ for every $i = 1,2,\dots,n$. This insures that every observation is equally reliable.

Assumption 3: The disturbances are not correlated, i.e., $\mathrm{E}(u_i u_j) = 0 \text{ for } i \neq j, \quad i = 1,2,\dots,n$

Assumption 4: The explanatory variable X is non-stochastic, i.e., fixed in repeated samples, and hence, not correlated with the disturbances. Also, $\sum_{i=1}^{n} x_i^2 / n \neq 0$ and has a finite limit as n tends to infinity.

Least squares Estimation

- Least squares minimizes the residual sum of squares where the residuals are given by

$$e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i \quad i = 1, 2, \ldots, n$$

and $\hat{\alpha}$ and $\hat{\beta}$ denote guesses on the regression parameters α and β, respectively.

- The residual sum of squares denoted by

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

is minimized by the two first-order conditions:

$$\frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_0} = \frac{\partial \left\{ \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \right\}}{\partial \hat{\beta}_0} = -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

$$= -2\left(\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right) = 0$$

and

$$\frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_1} = \frac{\partial \left\{ \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \right\}}{\partial \hat{\beta}_1} = -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]x_i$$

$$= -2\left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \right) = 0.$$

- The equations $\frac{\partial RSS}{\partial \alpha} = 0$ and $\frac{\partial SSE}{\partial \beta} = 0$ are called the least-squares equations for estimating the parameters of a line.

- The equations $\frac{\partial RSS}{\partial \alpha} = 0$ and $\frac{\partial SSE}{\partial \beta} = 0$ are called the least-squares equations for estimating the parameters of a line.

- The least-squares equations are linear in $\hat{\alpha}$ and $\hat{\beta}$and hence can be solved simultaneously. The solutions are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2},$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Case Study

- The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

B.P (y)	Age (x)	B.P (y)	Age (x)
128	46	120	20
136	53	128	43
146	60	141	63
124	20	126	26
143	63	134	53
130	43	128	31
124	26	136	58
121	19	132	46
126	31	140	58
123	23	144	70

1. **Find the correlation between age and blood pressure using simple and Spearman's correlation coefficients, and comment.**

2. **Find the regression equation?**

3. **What is the predicted blood pressure for a man aging 25 years?**

Serial	x	y	xy	x2
1	20	120	2400	400
2	43	128	5504	1849
3	63	141	8883	3969
4	26	126	3276	676
5	53	134	7102	2809
6	31	128	3968	961
7	58	136	7888	3364
8	46	132	6072	2116
9	58	140	8120	3364
10	70	144	10080	4900

Serial	x	y	xy	x2
11	46	128	5888	2116
12	53	136	7208	2809
13	60	146	8760	3600
14	20	124	2480	400
15	63	143	9009	3969
16	43	130	5590	1849
17	26	124	3224	676
18	19	121	2299	361
19	31	126	3906	961
20	23	123	2829	529
Total	852	2630	114486	41678

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{114486 - \frac{852 \times 2630}{20}}{41678 - \frac{852^2}{20}} = 0.4547$$

$\hat{y}$ =112.13 + 0.4547 x

for age 25

B.P = 112.13 + 0.4547 * 25=123.49 = 123.5 mm hg

Regression validation

- Model validation is possibly the most important step in the model building sequence.

- There are many statistical tools for model validation can be seen in the literature.

- But the primary tool for most process modeling applications is **graphical residual analysis**.

Residual Plots

- A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.

- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

x	y	ŷ	e
		65.411	4.589
		71.849	-6.849
		78.288	-8.288
		81.507	13.493
		87.945	-2.945

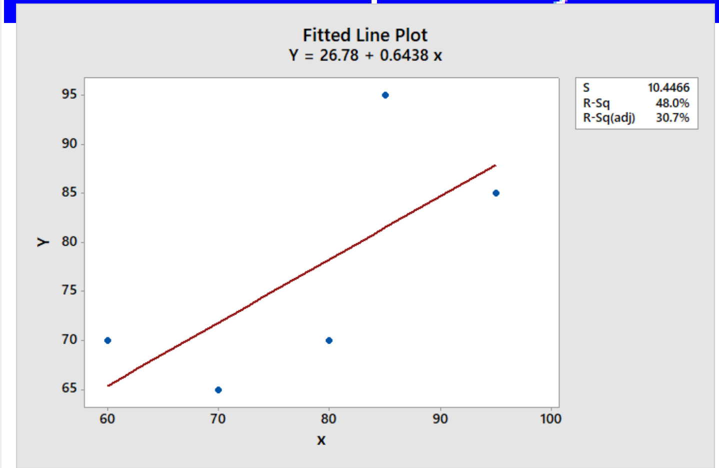
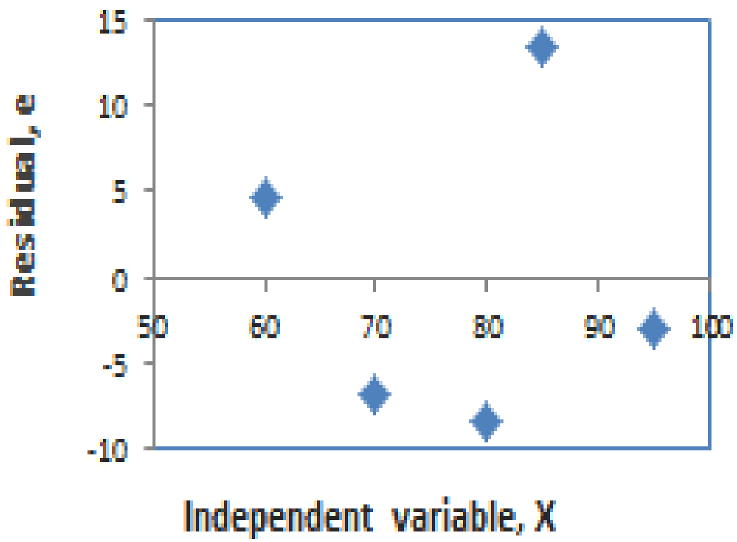


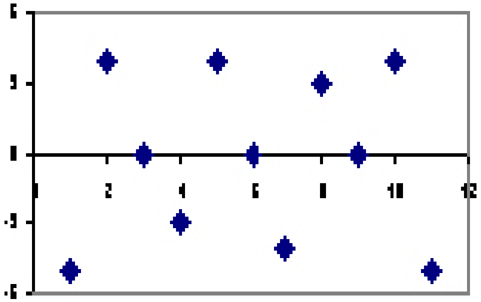
Chart displays the residual (e) and independent variable (X) as a residual plot.

- The residual plot shows a fairly random pattern
  - The first residual is positive,
  - the next two are negative,
  - the fourth is positive,
  - and the last residual is negative.
- <u>This random pattern indicates that a lin</u>
  <u>model provides a decent fit to the data</u>.

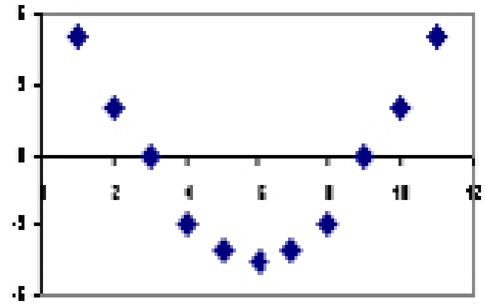


Residual Plots

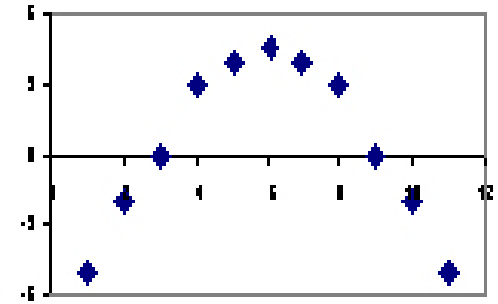
- The residual plots show three typical patterns.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

- The first plot shows a random pattern, indicating a good fit for a linear model.

- The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

What Is R-squared?

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

Total Variation = $\sum_{i=1}^{n} Y^2 - \frac{(\sum_{i=1}^{n} Y)^2}{n}$

Explained Variation= $\widehat{\beta_0} \sum_{i=1}^{n} Y + \widehat{\beta_1} \sum_{i=1}^{n} XY - \frac{(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{Y})^2}{\mathbf{n}}$

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

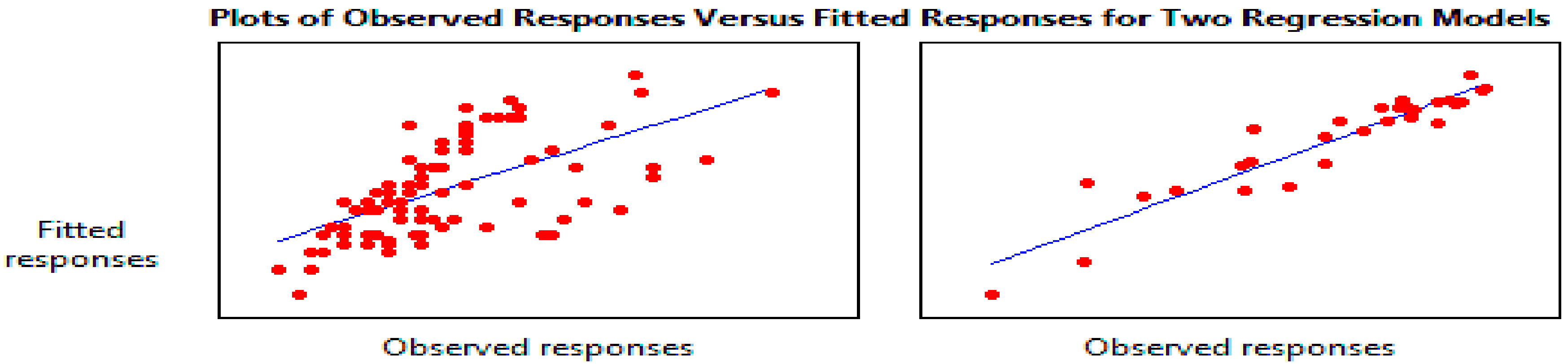
What Is R-squared?

- R-squared is always between 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.

- 100% indicates that the model explains all the variability of the response data around its mean.

- In general, the higher the R-squared, the better the model fits your data.



- The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%.

- The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

- Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.

Exercise

- The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

B.P (y)	Age (x)	B.P (y)	Age (x)
128	46	120	20
136	53	128	43
146	60	141	63
124	20	126	26
143	63	134	53
130	43	128	31
124	26	136	58
121	19	132	46
126	31	140	58
123	23	144	70

1. Find the correlation between age and blood pressure using simple or Spearman's correlation coefficients, and comment.

2. Find the regression equation?

3. Calculate R- Square ; comment.

4. What is the predicted blood pressure for a man aging 25 years?

Serial	x	y	xy	x2
1	20	120	2400	400
2	43	128	5504	1849
3	63	141	8883	3969
4	26	126	3276	676
5	53	134	7102	2809
6	31	128	3968	961
7	58	136	7888	3364
8	46	132	6072	2116
9	58	140	8120	3364
10	70	144	10080	4900

Serial	x	y	xy	x2
11	46	128	5888	2116
12	53	136	7208	2809
13	60	146	8760	3600
14	20	124	2480	400
15	63	143	9009	3969
16	43	130	5590	1849
17	26	124	3224	676
18	19	121	2299	361
19	31	126	3906	961
20	23	123	2829	529
Total	852	2630	114486	41678

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{114486 - \frac{852 \times 2630}{20}}{41678 - \frac{852^2}{20}} = 0.4547$$

$\hat{y}$ =112.13 + 0.4547 x

for age 25

B.P = 112.13 + 0.4547 * 25=123.49 = 123.5 mm hg