INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: MSc

Subject : Research Methodology

Quantitative Data Analysis

Dr. Satish Takalikar

- Researcher studies the entities, may be subject, object, event, or phenomenon to draw meaningful conclusions about its attributes or characteristics.

- Researcher identifies collection or group of such entities or instances for study.

Variable

- Measurable, quantifiable, countable or classifiable attribute or characteristic which varies from one entity to another in a group is termed as variable.
- Example:
  - Number of units sold in a month
  - Blood pressure of an employee

Types of Variable

- Qualitative
- Quantitative

Qualitative Variable

- Qualitative variable takes values that are names or labels, for categories of like items.
- Examples:
  - Gender of an employee
  - Designation of Faculty in a college

Quantitative Variable

- Quantitative variable takes numerical values These values are measure of how much or how many of something.
- Examples:
  - Amount spent on advertisement
  - Age of an employee
  - Weight of an employee

Types of Quantitative Variables

- Continuous Variables – Variables those take any numerical values

  Example: Weight, Temperature

- Discrete Variables - Variables those take specific numerical values

  (in even or un-even steps)

  Example: Number of units sold

Cause and Effect Relationship

- In quantitative research, researcher often studies effect of one variable on another, to establish cause and effect relationship.

Dependent and Independent Variables

- Dependent Variable – effect, variable being predicted.
  Its value depends upon or is a consequence of the change in value of another variable.

- Independent Variable – cause, variable being used to predict the most likely value of dependent variable.
  Its value is independent of values of other variables.

- Example:

  Researcher is studying effect of BMI – Body Mass Index up on mean systolic blood pressure. Here, BMI is independent variable and mean systolic blood pressure is dependent variable.

Extraneous Variable

- Extraneous Variable - an independent variable that is not related to the purpose of study, but may affect the dependent variable.

Control

- Good research always aims to minimize the influence of extraneous variables on the relationship between independent and dependent variable. When research is done in ‘controlled environment’, effect of extraneous variable is minimized.

Confounded Relationship

- When the dependent variable is not free from the influence of extraneous variables, the relationship between the dependent and independent variables is said to be confounded by an extraneous variable.

Confounding Variable

- When extraneous variables affect the variables under study in spite of stricter controls imposed, then the results do not reflect the actual relationship.
- In such situation, extraneous variables are either corelated with other independent variables or have direct impact on dependent variable.
- They are referred to as confounding variables.

Data

- Data are collection of facts.
- Data is generated when values of variables are obtained and recorded through experiment or process of observing, measuring, and/or counting.
- Entire collection of observations is referred to as Data set and individual observation as Data point.

Types of Data

- Quantitative data – Observed values of quantitative variables.

- Qualitative data – Observed values of quantitative variables. It is also called categorical data.

  Nominal Data - Skilled, semi-skilled, unskilled

  Ranked Data - Seeded players.

Types of Quantitative Data

- Univariate data
- Bivariate data
- Multivariate data

Univariate Data

- When data pertains to only one characteristic or variable of each entity in a category of like items under study, the data is called univariate data.

Examples:

Number of units sold in a month

Number of employees

Bivariate Data

- When two variables are observed simultaneously to study each entity in a category of like items, Data obtained is called Bivariate data.

Examples –

- Number of units sold in a month and amount spent on advertisement
- Body-mass index (BMI) of an employee and Systolic Blood pressure.

Multivariate Data

- When more than two variables are observed simultaneously to study each individual or entity in a certain population, Data obtained is called Multivariate Data.

Data Processing

- Data processing means collection and manipulation of items of data to obtain meaningful information.

- It involves collecting, organizing, presenting, analyzing, and interpreting Data.

Tools for Data Analysis

- Statistical techniques are most important tools used for analyzing the data.

- Types of statistical Techniques on the basis of their nature
  - Descriptive
  - Inferential

- Descriptive techniques help to summarize the characteristics of a data set.

- Inferential techniques help to assess whether your data is generalizable to the broader population.

Analyzing the Univariate Data

Measures in Descriptive Statistics

The measures used to summarize the univariate data -

- The **distribution** concerns the frequency of each value observed.
- The **central tendency** concerns the averages of the values.
- The **variability or dispersion** concerns how spread out the values are.

Techniques of Analysis

- Distribution
  - Tables
  - Charts such as Bar chart, Pie chart or Histogram
- Measures of central tendency
  - Average, Mode, Median
- Variability or Dispersion
  - Range, Variance, Standard deviation
- Skewness

Analyzing the Bivariate Data

- In bi-variate analysis, the frequency and variability of two variables is studied simultaneously to see if they vary together. The central tendency of the two variables can also be compared.

Analyzing the Bivariate Data

Bivariate Data

- When two variables are observed simultaneously to study each entity in a category of like items, Data obtained is called Bivariate data.

Examples –

- Number of units sold in a month and amount spent on advertisement
- Body-mass index (BMI) of an employee and Systolic Blood pressure.

Bivariate analysis involves

- Identifying correlation between two variables if any and then establishing cause and effect relationship between them and describing the nature of it.

Correlation

Correlation is the study that involves

- knowing existence of relationship between two quantitative variables,
- and then knowing its magnitude and direction.

Correlation describes the strength of association between two or more quantitative variables.

Simple and Multiple Correlation

- Simple correlation – Relation between two variables.
- Multiple correlation – Relation between more than two variables.

Correlation and Causal Relationship

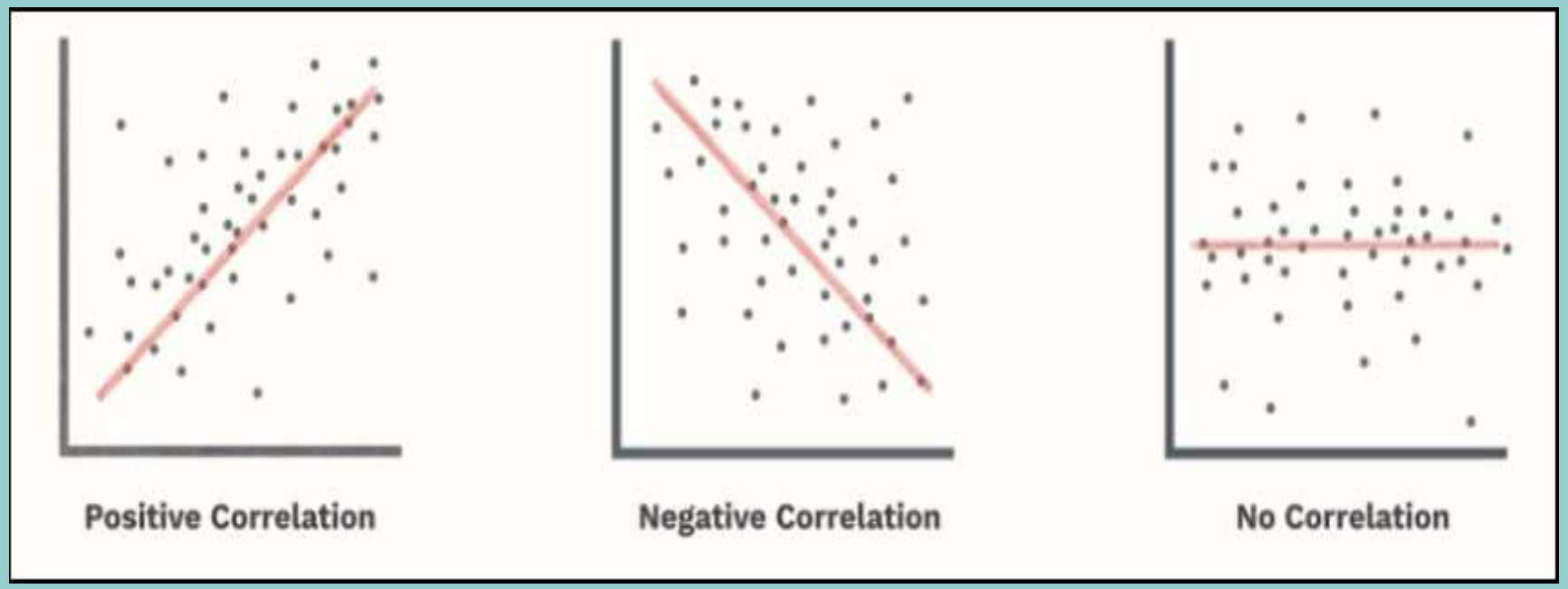
- When two variables are correlated, one of them is cause and other is the effect.

Examples –

- Minimum temperature of town and sale of woolen garments.
- Rain fall and production of paddy

Positive and Negative Correlation

- When one variable increases as the other increases, the correlation is positive.
- When one variable decreases as the other increases, the correlation is negative.


Positive Correlation
Negative Correlation
No Correlation

Spurious Correlation

- Sometimes correlation is observed between two variables but yet there may not be any causal relationship.
- Such a false correlation is called as Spurious correlation.
- It may be due to pure chance OR

  Both variables may be dependent upon third variable.

Linear Correlation

- Correlation is called linear when one variable changes in a fixed amount for a unit change in the other.
- A straight line represents the relationship between these two variables.

Non-Linear Correlation

- The relationship between two variables is not linear but needs some curve to describe it.

Methods of studying Linear Correlation

- Scatter plot
- Covariance
- Coefficient of correlation
- Spearman's Rank correlation coefficient

Scatter plot

- Scatter plot displays the bi-variate data in a graphical form to reveal the relationship between two variables.

- A scatter plot of two variables shows the values of one variable on the X axis (cause) and the values of the other variable on the Y axis (effect) .

Coefficient of Correlation

- It measures the degree of association between two quantitative variables.

Karl Pearson's Coefficient of Correlation

- Most common coefficient of correlation –

$$r = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Interpreting Karl Pearson's Correlation Coefficient

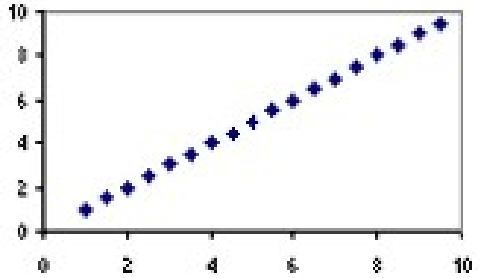
- Value of Karl Pearson's correlation coefficient varies from +1 through 0 to -1.
- The greater the absolute value, the stronger the linear relationship.
- Complete linear correlation between two variables is expressed by either +1 or -1.
- Complete absence of linear correlation is represented by 0.

Weak or No Linear Correlation

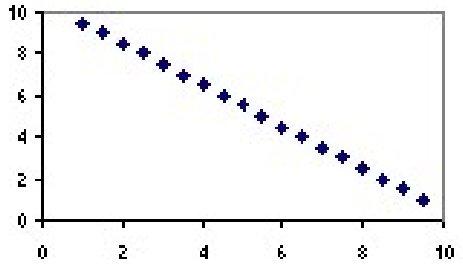
Even if there is weak linear correlation or absence of it between two variables, there may exist strong non-linear correlation between them.

Scatter plot and Karl Pearson's Correlation Coefficient

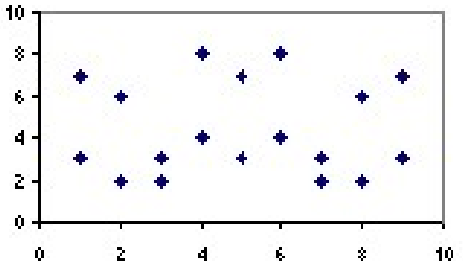
- Maximum positive correlation
  (r = 1.0)



- Maximum negative correlation
  (r = -1.0)

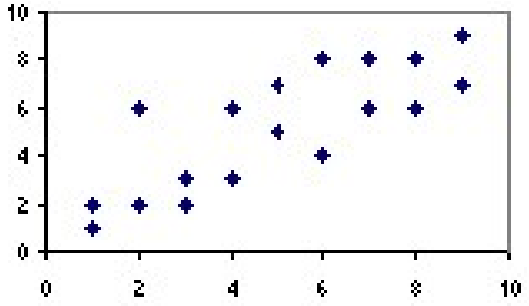


- Zero correlation
  (r = 0)

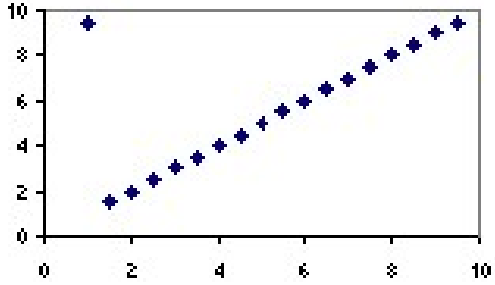


Scatter plot and Karl Pearson's Correlation Coefficient

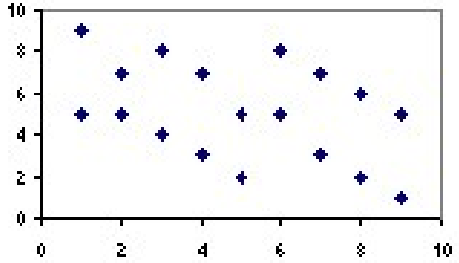
- Strong positive correlation
  (r = 0.8)



- Strong correlation and outlier
  (r = 0.8)



- Moderate negative correlation
  (r = 0.8)



Spearman's Rank Correlation Coefficient

- Sometimes the observations are expressed in comparative terms or ranks.
- In such case, correlation coefficient is given by

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

d – difference between ranks of i^{th} observation

n – number of observations

- It varies from +1 through 0 to -1.

Regression

Regression

- Research inferences and managerial decisions are often based on the nature of relationship between the two or more variables.

  Example – Relation between advertisement expenditure and sales

- Regression is a technique used to develop an equation showing how are the two or more variables are related to each other.

Use of regression equation

- Independent Variable – variable taking observed values (cause).
- Dependent Variable – variable value of which is to be predicted (effect).
- Regression equation expresses independent variable as a function of dependent variables on the basis of data collected.
- Then, it is used to predicts the most likely value of dependent variable for the given value of independent variables.

Note -

- If manager wants to estimate the sales on the basis of advertisement budget provided then

  Dependent variable (Y) is sales and independent variable (X) is advertisement expenditure.

- If manager wants to make provision for advertisement budget to reach certain figure of the sales then

  Dependent variable (Y) is advertisement expenditure and independent variable (X) is sales.

Simple Linear Regression

- Simple linear regression approximates relationship between one dependent and one independent variable by a straight line.

Mathematical Model for Simple Linear Regression

If $\hat{y}$ is the estimated or predicted value of dependent variable and x is the observed value independent variable, then the regression of Y on X relationship is described as follows:

$$\hat{y} = \alpha + \beta x + \varepsilon$$

α and β are population parameters which are not known.

ε is error in estimation i.e. difference in actual value and estimated value of dependent variable.

Note -

- Regression line is a probabilistic model as it enables to develop procedures for making inferences about parameters α and β of the model.

Least square method of fitting Regression line

- The regression equation is the best fit straight line that explains the association between two variables.

- Least square method is used to find the regression equation of y on x that best represents bivariate sample data,

$$\hat{y} = a + b\,x$$

a and b denote sample estimates of α and β.

Calculating Values of a and b

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

where b is the constant in the regression equation representing slope of the line

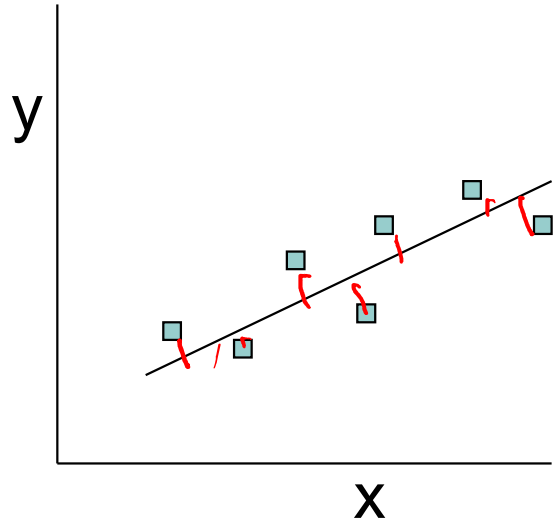
a is the y intercept of regression line.

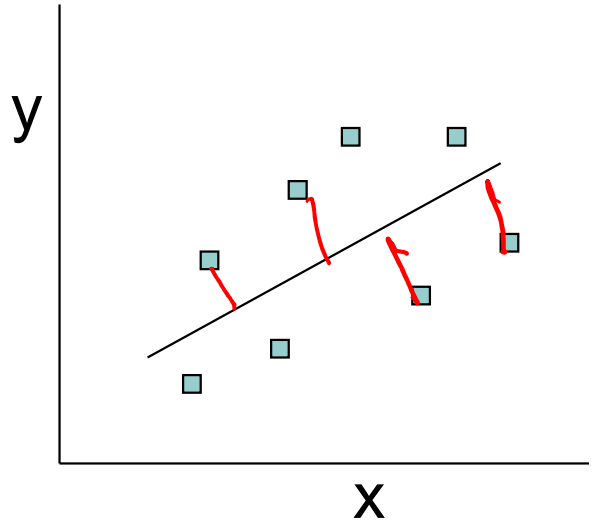
Regression Analysis

For equation $\hat{y} = a + b\,x$, the estimate of y derived from equation may not be equal to the actual observed value of y.

The difference between estimated value and corresponding observed value depends up on the extent of scatter of various points around the line of best fit.

The closer the various paired sample points clustered around the line of best fit, the smaller the difference between the estimated value and observed value.





Scatter differs but the line is same.
Smaller the difference, greater the precision of estimate.

Coefficient of determination

- It is a key output of regression analysis.
- It is equal to square of coefficient of corelation and so denoted as R^2 .
- It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.
- It shows numerical measure about how good is the “fit” between actual observations and predicted value.
- Higher the R^2 value, data points are *less scattered* so it is a good model. Lesser the R^2 value is more scattered the data points.

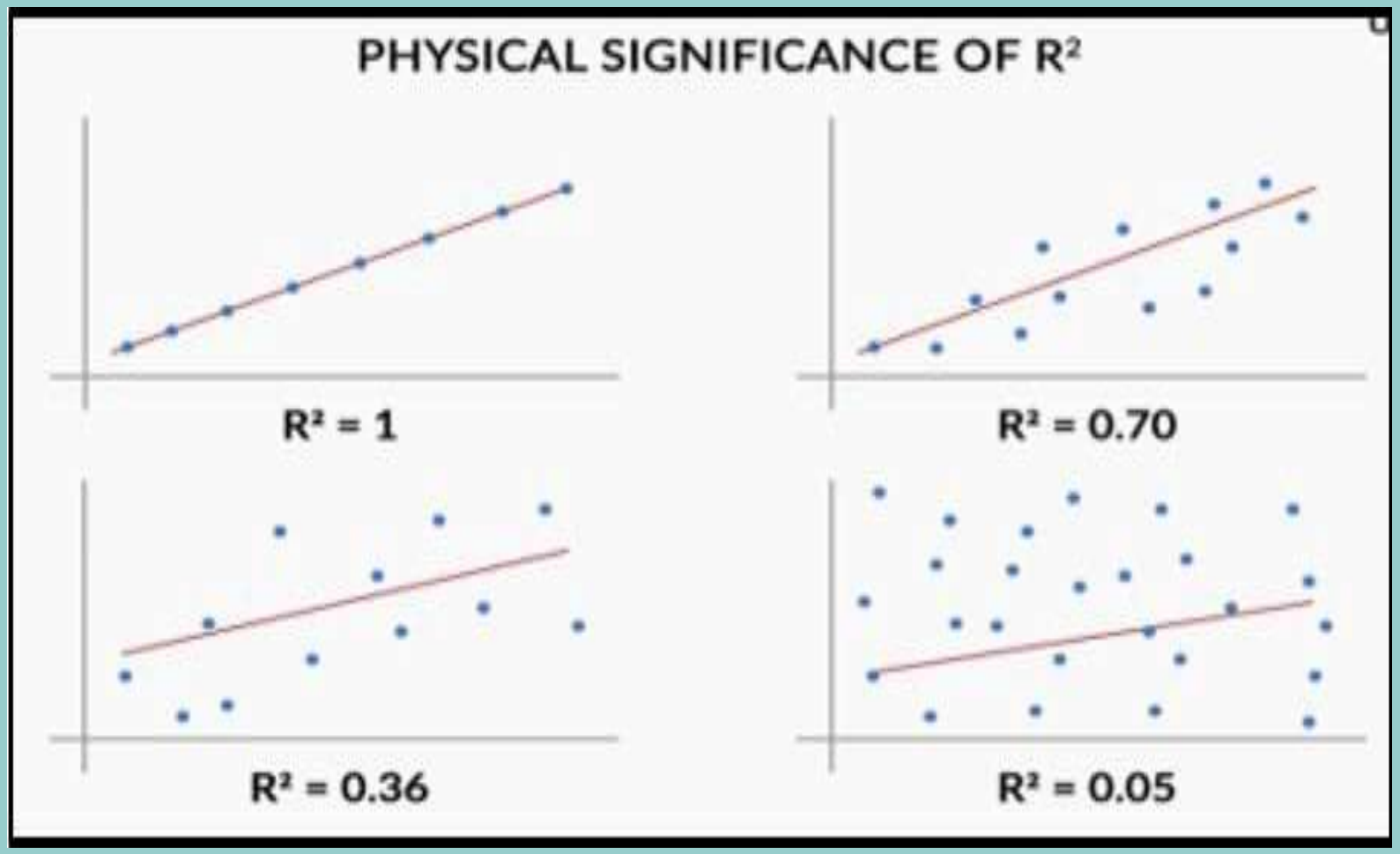
Coefficient of determination

It ranges from 0 to 1 indicating the extent to which the dependent variable is predictable. .

R^2 = 0 means that the dependent variable cannot be predicted from the independent variable.

R^2 = 1 means the dependent variable can be predicted without error from the independent variable.

R^2 = 0.9 means that 90 percent of the variance in *value of dependent variable, y* is explained by *value of independent variable*, *x*. Remaining 10% is unexplained and can due to sampling error or other variables.


PHYSICAL SIGNIFICANCE OF R²
R² = 1
R² = 0.70
R² = 0.36
R² = 0.05

Problem 1 – Following data relates to the number of business bankruptcies in a financial year and the number of firm births (new business start) in immediate previous financial year.

Develop the equation of regression model to find answers to following questions.

- If 6,00,000 new firms are born in FY 2022-23, how many cases of bankruptcies are expected in FY 2023-24.

FY	Bankruptcies (1000s)	Firm Births (10,000s)
16-17	34.3	58.1
17-18	35.0	55.4
18-19	38.5	57.0
19-20	40.1	58.5
20-21	35.5	57.4
21-22	37.9	58.0

Analysis – Dependent variable - number of bankruptcies (y)

Independent variable – number of Firm Births (x)

Regression equation,

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\bar{x} = \Sigma\,x\,/\,n$ $\bar{y} = \Sigma\,y\,/\,n$

Firm Births, x	Bankruptcies, y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x}) * (y - \bar{y})$
58.1	34.3	0.70	-2.58	0.49	-1.81
55.4	35.0	-2.00	-1.88	4.00	3.77
57.0	38.5	-0.40	1.62	0.16	-0.65
58.5	40.1	1.10	3.22	1.21	3.54
57.4	35.5	0.00	-1.38	0.00	0.00
58.0	37.9	0.60	1.02	0.36	0.61

$\Sigma x = 344.4$ $\Sigma y = 221.3$

$\bar{x} = 57.4$ $\bar{y} = 36.88$

$\Sigma(x - \bar{x})^2 = 6.22$ $\Sigma(x - \bar{x})*(y - \bar{y}) = 5.46$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

Σx = 344.4 Σy = 221.3

$\bar{x}$ = 57.4 $\bar{y}$ = 36.88

$\Sigma(x - \bar{x})^2$ = 6.22 $\Sigma(x - \bar{x})^*(y - \bar{y})$ = 5.46

a = - 13.503 b = 0.878

Regression equation, $\hat{y}$ = 0.878 x –13.503

If 6,00,000 new firms are born in FY 2022-23, how many cases of bankruptcies are expected in FY 2023-24.

Number of expected bankruptcies = 39,165

Problem 2 – A specialist in hospital administration stated that the number of FTE (full-time employee) in a hospital can be estimated by counting the beds in the hospital (a common measure of hospital size). A healthcare researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by number of beds. She surveyed 7 hospitals and obtained the following data.

Beds	FTE
46	125
42	126
76	176
64	156
54	178
35	118
78	225

Analysis – Dependent variable - number of FTE (y)
Independent variable – number of beds (x)
Regression equation,

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\bar{x} = \Sigma\,x\,/\,n$ $\bar{y} = \Sigma\,y\,/\,n$

Beds, x	FTE, y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x}) * (y - \bar{y})$
46	125				
42	126				
76	176				
64	156				
54	178				
35	118				
78	225				

$\Sigma x =$ $\Sigma y =$ $\Sigma(x - \bar{x})^2 =$ $\Sigma(x - \bar{x})*(y - \bar{y}) =$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\bar{x} = \Sigma\,x\,/\,n$ $\bar{y} = \Sigma\,y\,/\,n$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x}) * (y - \bar{y})$
46	125	-10.43	-32.71	108.76	341.16
42	126	-14.43	-31.71	208.18	457.59
76	176	19.57	18.29	383.04	357.88
64	156	7.57	-1.71	57.33	-12.98
54	178	-2.43	20.29	5.90	-49.27
35	118	-21.43	-39.71	459.18	851.02
78	225	21.57	67.29	465.33	1451.45

$\Sigma x = 395$ $\Sigma y = 1104$

$\bar{x} = 56.49$ $\bar{y} = 157.71$

$\Sigma(x - \bar{x})^2 = 1687.71$ $\Sigma(x - \bar{x})*(y - \bar{y}) = 3396.86$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\Sigma x = 395$ $\Sigma y = 1104$ $\bar{x} = 56.49$ $\bar{y} = 157.71$

$\Sigma(x - \bar{x})^2 = 1687.71$ $\Sigma(x - \bar{x})^*(y - \bar{y}) = 3396.86$

a = 44.14 b = 2.013

Regression equation, $\hat{y} = 44.14\,x + 2.01$

How many FTE are needed for 100 bed hospital?

FTE needed = 245

Because of certain constraints, policy decision is taken to limit number of FTE to 200. How many indoor patients should admitted to cater effective health care service?

Here, Dependent variable - number of beds (y)
Independent variable –number of FTE (x)
Regression equation,

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\bar{x} = \Sigma\,x\,/\,n$ $\bar{y} = \Sigma\,y\,/\,n$

FTE, x	Beds, y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x}) * (y - \bar{y})$
125	46				
126	42				
176	76				
156	64				
178	54				
118	35				
225	78				

$\Sigma x =$ $\Sigma y =$ $\Sigma(x - \bar{x})^2 =$ $\Sigma(x - \bar{x})*(y - \bar{y}) =$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\bar{x} = \Sigma\,x\,/\,n$ $\bar{y} = \Sigma\,y\,/\,n$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x}) * (y - \bar{y})$
125	46	-32.71	-10.43	1070.22	341.16
126	42	-31.71	-14.43	1005.80	457.59
176	76	18.29	19.57	334.37	357.88
156	64	-1.71	7.57	2.94	-12.98
178	54	20.29	-2.43	411.51	-49.27
118	35	-39.71	-21.43	1577.22	851.02
225	78	67.29	21.57	4527.37	1451.45

$\Sigma x = 1104$ $\Sigma y = 395$

$\bar{x} = 157.71$ $\bar{y} = 56.49$

$\Sigma(x - \bar{x})^2 = 8929.43$ $\Sigma(x - \bar{x})*(y - \bar{y}) = 3396.86$

Analysis – Calculating values of coefficients a and b.

$$\hat{y} = a + b\,x$$

$$b = \frac{\Sigma\,[(x_i - \bar{x})\,(y_i - \bar{y})]}{\Sigma[(x_i - \bar{x})^2]}$$

$$a = \bar{y} - b\,\bar{x}$$

$\Sigma x = 1104$ $\Sigma y = 395$

$\bar{x} = 157.71$ $\bar{y} = 56.49$

$\Sigma(x - \bar{x})^2 = 8929.43$ $\Sigma(x - \bar{x})^*(y - \bar{y}) = 3396.86$

a = - 3.57 b = 0.38

Regression equation, $\hat{y} = 0.38\,x - 3.57$

For FTE = 200, How many indoor patients should admitted to cater effective health care service

Beds or number of patients = 73

THANKS!!!

THANKS!!!