**Regression**

**Correlation**

* What is Correlation?
* Does it measure causation or association?
* The three different methods of measuring correlation?

**Regression**

*Question: What is the difference between regression and correlation?*

**Intro**

Regression analysis is a very powerful tool in the field of statistical analysis. It helps in predicting the value of one variable, given the value of another variable, when these two variables are related to each other.

Correlation analysis enables us to measure the strength of the association between these variables.

*Question: Can you give some examples of application of regression?*

**Examples**

* Height and weight of individuals
* Hours studied and score in exams
* Advertising and sales volume of a product
* Performance of a student in high school and performance of the student in college

**Variables**

*Questions: In the above example, what are variables – can you identify them?*

* Two variables are there in the above example.
* Value of one variable depends on the value of the other variable.
* Number of hours studied and scores in exams.
* The dependent variable Y is the exam score and the independent variable X is the number of hours studied.
* Weights of a person are correlated to height. So weight is Y (dependent variable) and height is X (dependent variable).
* The general process of predicting the value of one variable Y, on the basis of the known value of the other variable X is known as regression analysis.

**Scope of Regression Analysis**

* Regression analysis can be done for two or more variables.
* When it is done for 2 variables it is called Simple Regression

**Data**

This sort of data is called bi-variate data

We are analysing relationship between height and weight were weight Y depends on height X, the values of these variables are presented as X, Y pairs.

**Example**

|  |  |
| --- | --- |
| Advertising (X) | Profit (Y) |
| 5 | 8 |
| 6 | 7 |
| 7 | 9 |
| 8 | 10 |
| 9 | 13 |
| 10 | 12 |
| 11 | 13 |

Draw a scatter diagram.

**Condition**

We are concerned only with linear or straight-line relationships – that is relationship between the two variables can be adequately described by a straight line.

**Linear Regression Equation**

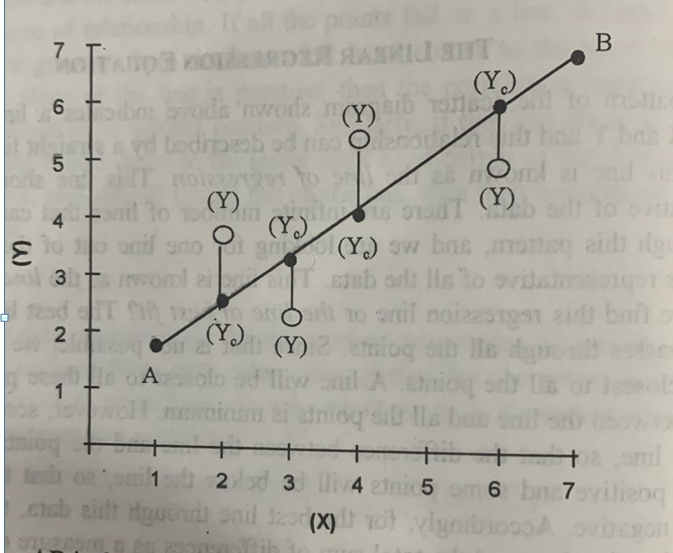
* The scatter plot can be generalised by a straight line through these points
* This line is known as the line of regression.
* This line should be the most representative of the data.
* There are an infinite number of lines that can approximately pass through this pattern – we are looking for one line out of these – that is most suitable as representative of all the data.
* This line is known as the line of best fit.

*Question: How do we find this best fit line? What characteristics?*

**Line of best fit**

* A line that is closes to all the points
* A line will be closest to all the points if the total distance between the line and all the points is minimum.
* Mathematically some points will have +ve distance while others have –ve distance – so sum of differences would cancel out each other and therefore will not be valid.
* The differences must be looked at individually and squared.
* Then a sum of the squared distances must be taken.
* We are looking for a line which has the lowest sum of squared distances (we use squared difference as against absolute difference – magnifies/penalises far away data points).
* Hence this method is called Least Squares Method

*Graphical Representation*

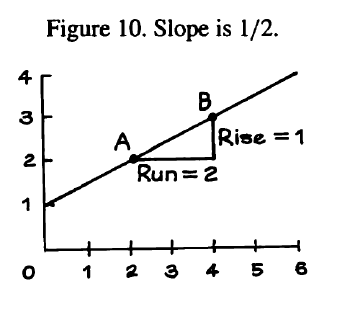


**Algebraic Form**

**Concept of Slope and Intercept**

**Slope**

Illustration: Draw the below line graph

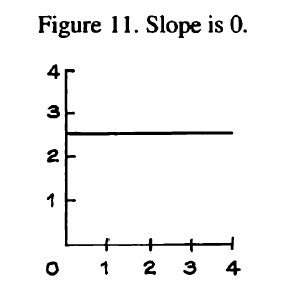


* The above figure shows a line. Take any point on the line – for instance, Point A.
* Now move up the line to any other, for instance point B.
* Your x-coordinate has increased by some amount, called the *run*. In this case, the run was 2.
* At the same time, your y-coordinate has increased by some other amount, called the *rise*.
* In this case, the rise was 1. Rise was half the run.
* The ratio rise/run is called the slope of the line.

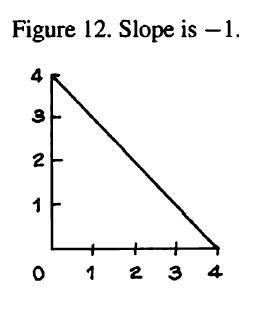
Slope = Rise/Run

* The slope is the rate at which y increases with x, along the line.
* Slope is a measure of steepness.

*Question: What is the slope of the following line?*

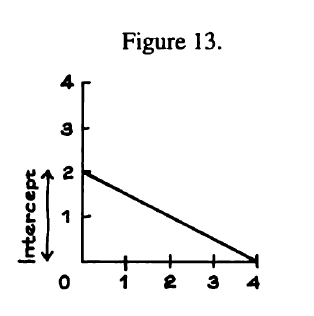


*Question: What about this line?*

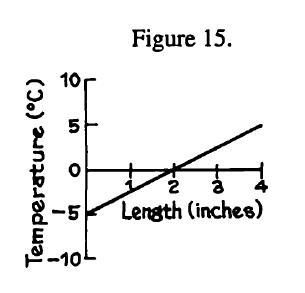


**Intercept**

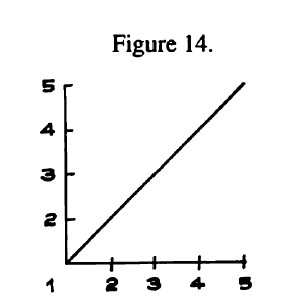
* The intercept of a line is its height at x = 0
* Usually the axes cross at 0.
* Then the intercept is where the line crosses the y-axis.
* Below, the intercept is 2 (refer below figure)



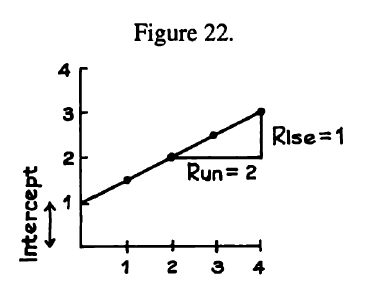
*Question: what is the intercept here?*



*Question: This is one is tricky, what is the intercept in this equation?*



**Standard Equation for a line**



Let’s look at a sample data here

|  |  |
| --- | --- |
| X | Y |
| 1 | 1.5 |
| 2 | 2.0 |
| 3 | 2.5 |
| 4 | 3.0 |

* All the points fall on a line
* We can derive an equation that gives us *y* coordinate for every *x* coordinate
* Y= 1+ 0.5x
* Y = Intercept + slope(X)

**Slope of a regression line has a special feature**

Slope = r x s.d (y) / s.d (x)

For a 1 standard deviation change in X (from X-mean), the value of Y variable changes by r x s.d (y).

**Solved in Class**

**Exercise - 1**

Heights and weights statistics of students are given below.

Average height = 70 inches, SD = 3 inches

Average weight = 180 pounds, SD = 45 pounds, r = 0.4

Estimate the height of a student who is 73 inches

**Exercise – 2**

A university has made a statistical analysis of the relationship between Math SAT scores (ranges from 200 to 800) and first-year GPAs (ranging from 0 to 0.4), for students who complete the first year. The results

Average SAT score = 550, SD = 80

Average first-year GPA = 2.6, SD = 0.6, r = 0.4

The scatter diagram is football-shaped. A student is chosen at random, and has an SAT of 650. Predict the individual’s first-year GPA.

**Exercise - 3**

Education and income for 570 California women age 25-29 summary statistics are shown below;

Average education = 13 years, SD = 3.4 years

Average income = $18,000, SD = $20,000, r = 0.37

1. Find the regression equation for predicting income from education
2. Use the equation to predict the income of a woman whose educational level is 8 years, 12 years and 16 years.

**Equation for the line of best fit**

Yc = B0 + B1X

B1 = [ n(∑(XY)) – ∑X∑Y ] / [ n∑X^2 – (∑X)^2 ]

B0= Y̅ - B1X̅