

Unit 3 *Correlation and Regression*

Introduction

Learning objectives

The previous two units in this strand deal with measures of central tendency and variation; here we consider correlation and regression between two variable. After studying this unit you should

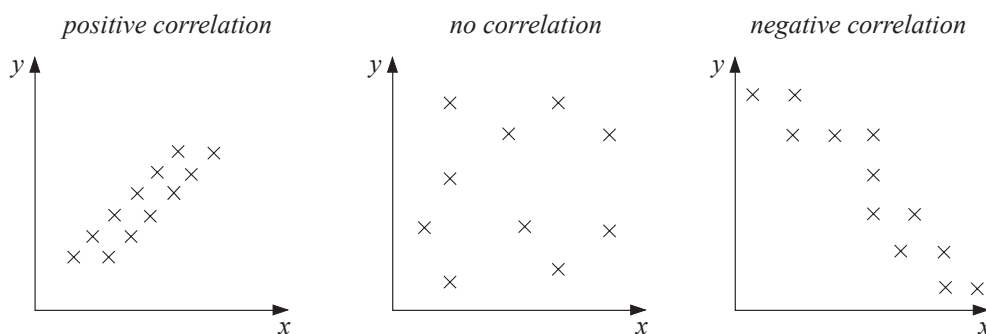
- understand the concept and measuring of variation
- use Spearman's Rank Correlation Coefficient for data that is ranked
- use and interpret the product Moment Correlation Coefficient
- understand the concept of regression
- use the line of regression, based on the method of least squares.

Notes

Correlation is a measure of how two sets of data are related whereas regression shows well data fits a straight line. Regression should only be used when the data is strongly correlated (positively or negatively).

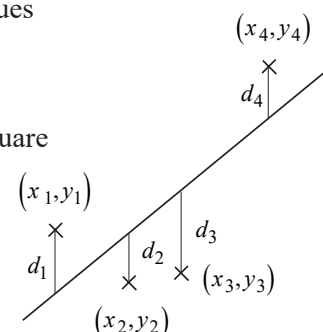
Key points

- Measures of correlation, r , are based on $-1 \leq r \leq 1$ and $r = 1$ for perfect positive correlation and $r = 0$ for no correlation and $r = -1$ for perfect negative correlation



- Spearman' Rank Correlation Coefficient is based on ranked data
- Product Moment Correlation Coefficient is based on actual data values
- Regression lines are essentially lines of best fit based on given data
- The equation of the line of regression is based on minimising the square of the distances of the data points from the regression line,

i.e. minimising $\sum_{i=1}^n d_i^2$



Unit 3 *Correlation and Regression*

Introduction

Facts to remember

Spearman's rank correlation coefficient is given by

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Product moment correlation coefficient (sometimes known as Pearson's PMCC) is given by

$$r = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{s_x s_y}$$

where

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}$$

$$s_y = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}$$

The 'y on x' line of regression is given by

$$y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$$

when $S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$

and $S_{xx} = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$

Glossary of terms

Bivariate data: data that has two variables

Correlation: a measure of how bivariate data, that is, points (x_i, y_i) for $i = 1, 2, \dots, n$ are related

Regression: gives a straight line approximation for strongly correlated bivariate data