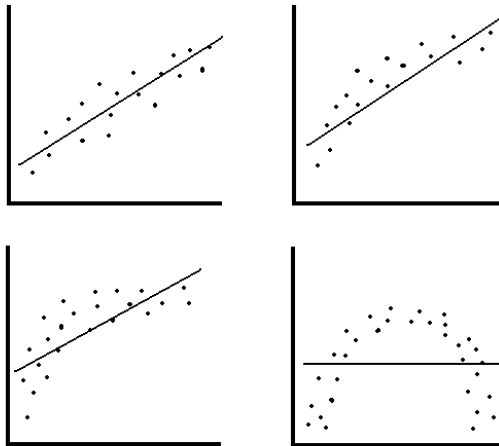


Correlation and Regression

Correlation and regression are very much related. They both involve the relationship between two variables. Correlation is primarily concerned with finding out whether or not a relationship exists between two variables. It provides information on the magnitude and direction of the relationship. Regression, on the other hand, is primarily concerned with using the relationship for predictive purposes.

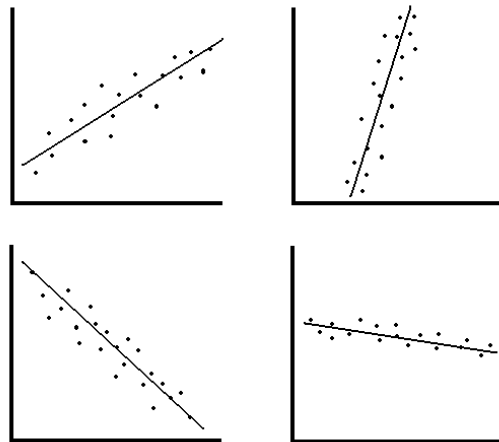
I. Linear and Curvilinear Relationships

- a. A linear relationship between two variables is one in which the relationship can be most accurately represented by a straight line.
- b. Not all relationships are linear. Some relationships are curvilinear. In these cases, when a scatter plot of two variables is drawn, a curved line fits the points better than a straight line.



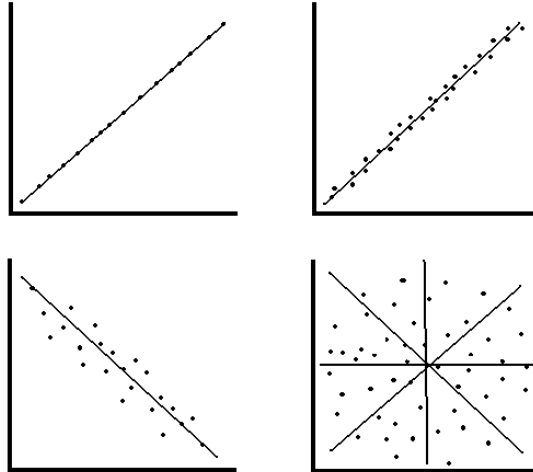
II. Positive and Negative Relationships

- a. A positive relationship indicates that there is a direct relationship between two variables.
- b. A negative relationship exists when there is an inverse relationship between two variables



III. Perfect and Imperfect Relationships

- a. A perfect relationship is one in which a positive or negative relationship exists and all of the points fall on the line.
- b. An imperfect relationship is one in which a relationship exists, but all of the points do not fall on the line.



IV. Correlation

- a. Correlation is a topic that focuses on the direction and magnitude of the relationship between two variables. The direction refers to whether the relationship is positive or negative. The magnitude of the relationship refers to the degree of the relationship, which can vary from nonexistent to perfect.
- b. A correlation coefficient expresses quantitatively the magnitude and direction of the relationship between two variables.
 - i. A correlation coefficient can vary from +1 to -1. The sign of the coefficient tells us whether the relationship is positive or negative, and the numerical portion of the coefficient describes the magnitude of the correlation. Because 1 is the highest number possible, it represents a perfect correlation; -1 also represents a perfect correlation. When the relationship is nonexistent, the correlation coefficient equals 0.
- c. Pearson correlation coefficient
 - i. The Pearson correlation coefficient (r) is a widely used way of quantifying the correlation between two mensural variables. A Pearson correlation is a measure of the extent to which paired scores occupy the same or opposite positions within their own distributions. Pearson r can also be interpreted in terms of the variability of one variable accounted for by the other variable.
 - ii. The equation for calculating Pearson r is:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N} \right] \left[\sum Y^2 - \frac{(\sum Y)^2}{N} \right]}}$$

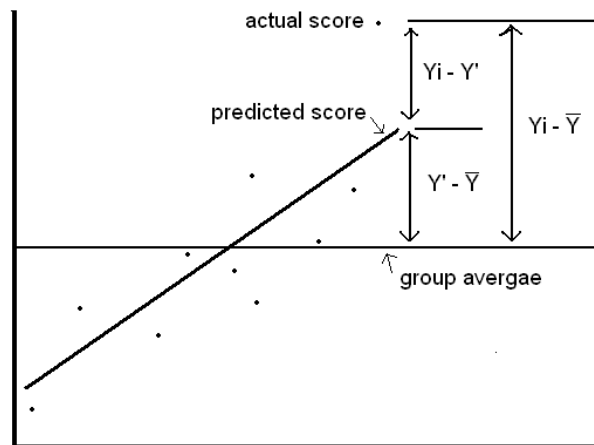
- d. Spearman rank order correlation coefficient
 - i. The Spearman rank order correlation coefficient (*rho*) is used when one or both of the variables are of ordinal scale. Spearman *rho* is really the linear correlation coefficient Pearson *r* applied to data that meet the requirements of ordinal scaling.
 - ii. The easiest way of calculating rho is:

$$rho = 1 - \frac{6 \sum [R(X_i) - R(Y_i)]^2}{N^3 - N}$$

- e. Things to watch out for
 - i. Effects of range
 - ii. Effects of outliers
 - iii. Correlation does not mean causation
 - 1. When two variables are correlated, it is tempting to conclude that one of them is the cause of the other; however, to do so without further experiments would be a serious error because whenever two variable are correlated, there are four possible explanations of eth correlation
 - a. The correlation between X and Y is spurious
 - b. X is the cause of Y
 - c. Y is the cause of X
 - d. A third variable is the cause of the correlation between X and Y

V. Regression

- a. Regression is a topic that considers using the relationship between two mensural variables for prediction. We can have three main ways in which to conduct regression analyses. We can have the regression of Y on X, the regression or X on Y, or major-axis regression. In each of these cases, the least-squares regression line can be remarkable different. In addition to the regression equation, regression analysis typically gives coefficient of determination (r^2).
- b. Least-squares regression line: $\sum (Y_i - \bar{Y})^2 = \sum (Y_i - Y')^2 + \sum (Y' - \bar{Y})^2$



c. Regression of Y on X

- i. The least-squares regression line is the prediction line that minimizes $\sum(Y - Y')^2$. The equation for the least-squares regression line for predicting Y given X is $Y' = b_Y X + a_Y$; where Y' is the predicted value of Y, b_Y is the slope of the line for minimizing errors in predicting Y, and a_Y is the Y axis intercept for minimizing errors in predicting Y.

- ii. The b_Y regression constant is equal to
$$b_Y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$
.

- iii. The a_Y regression constant is given by $a_Y = \bar{Y} - b_Y \bar{X}$.
- iv. Because we need the b_Y constant to determine the a_Y constant, the procedure is to first find b_Y then a_Y . Once they are both found, they are substituted into the regression equation.

d. Regression of X on Y

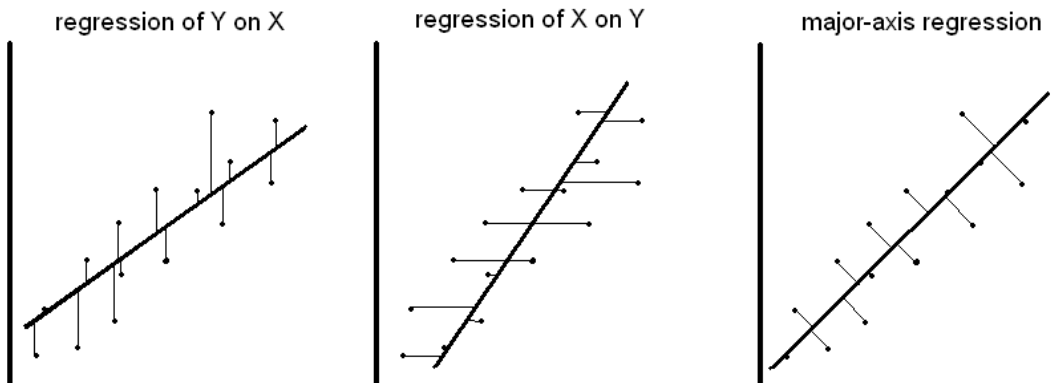
- i. The least-squares regression line is the prediction line that minimizes $\sum(X - X')^2$. The equation for the least-squares regression line for predicting X given Y is $X' = b_X Y + a_X$; where X' is the predicted value of X, b_X is the slope of the line for minimizing errors in predicting X, and a_X is the X axis intercept for minimizing errors in predicting X.

- ii. The b_X regression constant is equal to
$$b_X = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}}$$
.

- iii. The a_X regression constant is given by $a_X = \bar{X} - b_X \bar{Y}$.
- iv. Because we need the b_X constant to determine the a_X constant, the procedure is to first find b_X then a_X . Once they are both found, they are substituted into the regression equation.

e. Major-axis regression

- i. Used when both variables have error around them.



f. Coefficient of Determination

- a. The coefficient of determination is the amount of the variation in y that is explained by the regression line (accounted for by variation in x)
- b. $r^2 = \frac{\text{explained variation}}{\text{total variation}}$