

**CORRELATION & REGRESSION**

1. Relationship between two variables

- Are two variables associated each other?
- To what degree (strength) are they associated?
- In which directions is the relationship?
  - Positive or negative
- Change in dependent variable that corresponds to change in independent variable.
  - Prediction

<b>Correlation</b>	<b>Regression</b>
<ul style="list-style-type: none"> <li>- presence of association</li> <li>- strength (degree) of association</li> <li>- direction of association</li> </ul>	<ul style="list-style-type: none"> <li>- prediction</li> </ul>

**CORRELATION**

1. Is a measure of relationship between two numerical variables
  - E.g. the relationship between height and weight, the relationship between cholesterol and blood pressure.
2. Pattern:
  - Elliptical pattern – degree of elongation of the ellipses – proportional to the correlation coefficient.

- Elliptical pattern – indicative of normally distributed variables

3. Correlation coefficient (r)

- X increase                      Y increase                      r = 1 (perfect positive)
- X increase                      Y decrease                      r = -1 (perfect negative)
- No linear relationship                      r = 0
- r
  - o < 0.25                      → poor
  - o 0.26 – 0.50                      → fair
  - o 0.51 – 0.75                      → good
  - o 0.76 – 1.00                      → excellent
- r does not imply a cause and effect relationship
- Correlation should be assessed mathematically, not visually.
- r for statistical sample, ρ (rho) for parameter of population.
- Correlation coefficient:

<b>Pearson’s Correlation coefficient</b>	<b>Spearman’e Ranked Correlation coefficient</b>
<ul style="list-style-type: none"> <li>- A measure of degree of straight line relationship between two numerical variable</li> <li>- At least one variable have a normal distribution</li> </ul>	<ul style="list-style-type: none"> <li>- Correlation coefficient calculated on the ranks of the observation of two variables</li> <li>- Rank correlation and Spearman’s correlation – similar</li> <li>- Different when the scatter plot deviates from an elliptical shape</li> </ul>

4. Example: Relationship between height and weight

- Step 1: state the null and alternative hypothesis
  - $H_0$ : There is no correlation between height and weight
  - $H_A$ : There is correlation between height and weight (2-tailed)
- Step 2: set significance level
  - $\alpha = 0.05$
- Step 3: check the assumption
  - Both numerical variable
  - One of the variables has normal distribution
    - Histogram
    - Box and Whisker plot
- Step 4: statistical test
  - Pearson correlation (if assumption is met)
  - Spearman's correlation (if assumption is not met)
- Step 5: Interpretation

**Correlations**

		height Height	weight Weight
height Height	Pearson Correlation	1	.878(**)
	Sig. (2-tailed)	.	.000
	N	100	100
weight Weight	Pearson Correlation	.878(**)	1
	Sig. (2-tailed)	.000	.
	N	100	100

\*\* Correlation is significant at the 0.01 level (2-tailed).

- p-value = <0.001
  - reject  $H_0$
- step 6: conclusion
  - There is a significant, positive and excellence correlation between height and weight ( $r = 0.88, p < 0.001$ )
- Checklist for reporting correlation (Figure 1)



## **SIMPLE LINEAR REGRESSION (SLR)**

### 1. Regression Analysis

- Regression analysis is a statistical tool that utilizes the relation between variables so that one variable can be predicted from the other or others
- Linear regression
  - Simple (one independent variable (factor) and one outcome)
  - Multiple (more than one factor and one outcome)
- Logistic Regression (dichotomous dependent variables)

### 2. Simple Linear Regression

- Example of research questions
  - Does a relationship exist between oral contraceptive and the incidence of thromboembolism?
  - What is the relationship of a mother's weight to her baby's birth weight?
  - Relationship between an animal's pulse rate and the amount of particular drug administered?
- Simple because only one independent variable
- Linear means the relationship between y (dependent/outcome) and x (independent/factor) variables can be represented by a straight line
- Analysed linear relationship between two quantitative (numerical) variables
- Involves estimating the equation of a straight line that defines the relationship between a dependent variable using a given data set
- The method involved is called method of least squares
- We choose a line such that the sum of squares of vertical distances of all points from the line is minimized ( $Q = \sum e_i^2$ )

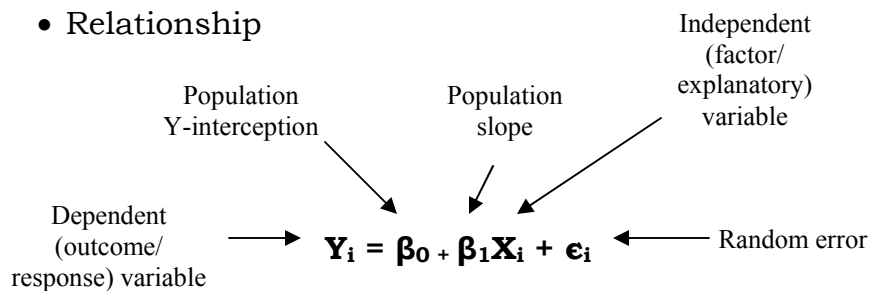
- These vertical distances between  $y$  values and their corresponding estimated values on the line are called residuals ( $e_i = y_i - \hat{y}_i$ )
- The line thus obtained is called the regression line or the least-squares line of best fit

### 3. Regression line (least squares line of best fit)

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ 
  - $Y_i$  is the value of dependent variable when the value of the independent variable is  $X_i$
  - $\beta_0$  is Y-interception and is constant
  - $\beta_1$  is the slope of the regression line. It is the change in  $Y_i$  when  $X_i$  is increased by one unit
  - $\beta_0$  and  $\beta_1$  are called regression coefficients
  - $\epsilon_i$  is random error terms, normally distributed, independent, with zero mean, and constant variance  $\sigma^2$

### 4. Linear Regression Model

- Relationship



### 5. True Regression line

- The random error term  $\epsilon_i$  the regression equation accounts for the scattering of the data points about the regression line
- As the mean of the  $\epsilon_i$ s is zero, the mean of  $Y_i$  (at  $X_i$ ) is:
  - $E(Y_i) = \beta_0 + \beta_1 X_i$
  - The notation  $E(Y_i)$  means 'expected value' of  $Y_i$  and represents the mean of  $Y_i$

- Not that the mean of y and on x and the relationship is represented by a straight line
- This equation represents the true regression line

#### 6. Least square estimate

- Time regression line is unknown
- Estimated regression line:
  - $\hat{Y} = \beta^{\wedge}_0 + \beta^{\wedge}_1 X$  → least square estimate
    - $\hat{Y}$  = is estimated mean
    - $\beta^{\wedge}_0$  is y-intercept and is constant
      - if  $x = 0$ ,  $\beta^{\wedge}_0$  is the estimated mean value of Y
    - $\beta^{\wedge}_1$  is the slope of the regression line. It is the change in Y when X is increased by one unit.

#### 7. Least Squares (LS)

- “Best Fit” means difference between actual Y values and predicted Y values are minimum.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$$

- LS minimizes the Sum of the Squared Differences (SSE)

#### 8. Interpretation of Coefficient

- Slope ( $\beta^{\wedge}_1$ )
  - The change in the estimated mean value of Y when X is increased by 1 unit

- If  $\beta^{\wedge}_1 = 0.05$ , then the estimated mean cholesterol level (Y) changes by 0.05 mmol/dl when the age is (X) increased by 1 year.
- Y-intercept ( $\beta^{\wedge}_0$ )
  - Average value of Y when  $X = 0$ 
    - If  $\beta^{\wedge}_0 = 3.3$ , then the mean cholesterol level (Y) is expected to 3.3, when the age (X) is 0 (???)

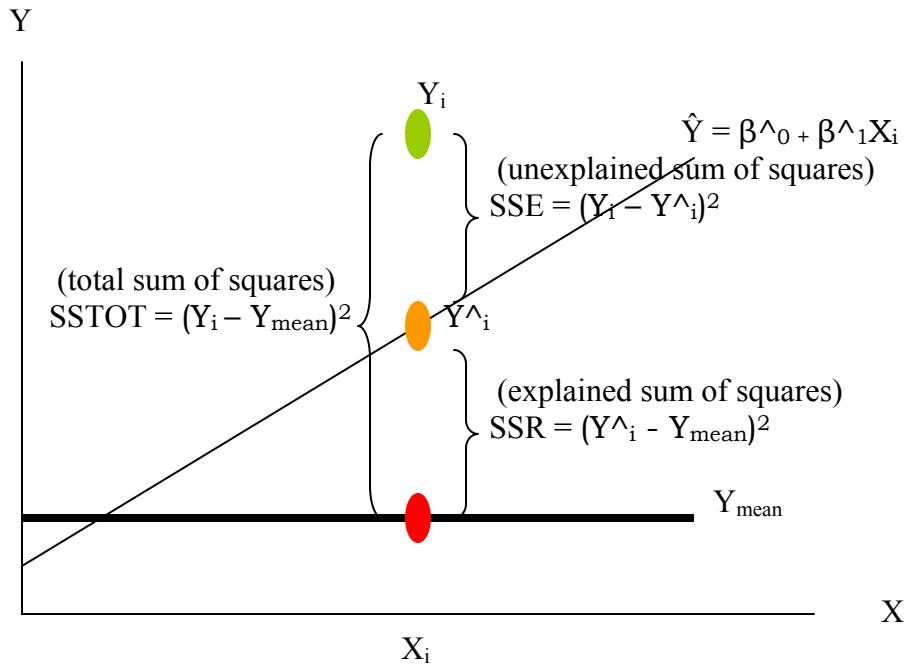
#### 8. Measures of variation in Regression

- Total variation (Total Sum of Squares (SSTOT))
  - Measures variation of observed  $Y_i$  around the mean  $Y_{\text{mean}}$
- Explained variation (Squared Sum of Regression (SSR))
  - Variation due to relationship between X & Y
- Unexplained variation (Square Sum of Error (SSE))
  - Variation due to other factor

#### 9. Sum of squares

- Total sum of square (SSTOT)
  - Measure of total variation in dependent variable Y
  - $SSTOT = \sum (Y_i - Y_{\text{mean}})^2 = SSR + SSE$
- Regression Sum of square (SSR)
  - Measure the variation 'explained' by the regression line
  - $SSR = \sum (Y^{\wedge}_i - Y_{\text{mean}})^2$
- Error Sum of squares (SSE)
  - Measures of the 'unexplained' variation in Y or the scatter around the regression line
  - $SSE = \sum (Y_i - Y^{\wedge}_i)^2$

**Measure of variation**



Notes:

X, Y and slope:

- Positive slope, Y increases with increase in X
- Negative slope, Y decreases with increase in X

10. Hypothesis Testing:

- For Simple Linear Regression
  - $H_0: \beta_1 = 0$  (no linear relationship)
  - $H_A: \beta_1 \neq 0$  (there is linear relationship)
  - Test statistics: t-distribution
  - Rejection rule:
    - Reject  $H_0$  if p-value less than 0.05 (assumed  $\alpha$ )
- For Multi Linear Regression
  - $H_0: \beta_1 = 0$  (no linear relationship)
  - $H_A: \beta_1 \neq 0$  (there is linear relationship)
  - Test statistics: F-test for ANOVA table:

- $F = MSR/MSE$
- $MSR = SSR/df_{Reg}$
- $MSE = SSE/df_{Error}$
- Rejection rule:
  - Reject  $H_0$  if p-value for the F-test less than 0.05 (assumed  $\alpha$ )
- Assumption
  - The errors are normally distributed
  - They are independent
  - The mean of random error term is equal to zero
  - The variance of random error,  $\sigma^2$  (sigma square), is constant.

11. How to analyse

- Exploration of the data
  - Descriptive
  - Scatter plot between two variables
    - Check for distribution, relationship and outliers
- Fit the square least line (regression line)
  - Using least square method
  - It is the best fitting straight line through the data points in a scatter plot
  - It represents the least square equation and estimates the constant (a) and slope (b) for  $\alpha$  and  $\beta \rightarrow Y^{\wedge} = a + bx$
  - It is constructed by using the method of least square – minimizes the sum of squared deviations of each point from the mean (regression line)
- Evaluation of model by  $R^2$  (R square)

**Model Summary**

model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	Std Error of the estimate
1	.592 <sup>a</sup>	.350	.338	.9043

a. Predictors: (constant), Time

- $R^2 = 0.35$ , meaning that 35% of the total variation in GPA is explained by the study time
- $R^2$  measures the closeness of fit of the sample regression equation to the observed values of Y
- It ranges from 0 to 1
- Is called **coefficient of determination**
- Evaluation b
  - Evaluation of  $\beta$  using t-statistics

**Coefficients**

Model	Unstandardized Coefficient		t	Sig.	95% CI for	
	B	Std error			Lower	Upper
1 (constant)	1.461	.315	4.639	.000	.829	2.093
Time	.389	.073	5.342	.000	.243	.534

a- dependent variable for GPA

- $H_0: \beta_1 = 0$  (no linear relationship)
- $H_A: \beta_1 \neq 0$  (there is linear relationship)
- As p value < 0.001, we reject  $H_0$  at 5% significance level and have sufficient evidence to conclude that there is linear relationship between study time and GPA.
- Positive  $\beta$  means direct relationship
- Estimated Least Square (LS) equation
  - $GPA = \text{constant} + b (\text{study time})$
  - $GPA = 1.461 + 0.389 (\text{study time})$
- Diagnostic checking for assumption
  - The assumptions:
    - The errors are normally distributed
    - They are independent
    - The mean of random error term is equal to zero (linearity)
    - The variance of random error,  $\sigma^2$  (sigma square), is constant or equal

- Model adequacy checks
  - After obtaining the least square line or fit
  - Linear model appropriate? ... $R^2$
  - Investigate model assumption
  - Diagnostic procedures carried out through examination of Residuals (difference between the observed value Y and the fitted or the predicted value Y at a given value X)
  - Normality
    - Histogram of unstandardized residuals
  - Linearity
    - Plot of unstandardised residuals against unstandardised predicted values
    - Creating residual: go to analyse → regression bivariate → save → unstandardised residual and predicted values
  - Let say the assumption is met.
- Interpretation and conclusion
  - 35% of the variation in GPA is explained by study time
  - There is significant linear association between GPA and study time
  - For each 1 hour increase in study, the GPA of a student increase by 0.39
  - We are 95% confident that for each 1 hour change in the study time, the GPA increase will lie between 0.24 to 0.53