Covariance

- **Covariance** : measure the linear relationship between 2 random variables
  \[ \text{Cov}_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y})}{n-1} \]

Correlation coefficient

- **Correlation coefficient** : strength of the linear relationship
  \[ r_{XY} = \frac{\text{Cov}_{XY}}{\sigma_X \sigma_Y} \]
  \[-1 \leq r_{XY} \leq 1 \]

Scatter plot

- **Scatter plot** : collection of points on the graph, each represents the value of 2 variables (X and Y)
  - Upward scatter plot : positive correlation
  - Downward scatter plot : negative correlation

Limitation to correlation analysis

1. **Outliers** : extreme values for sample observations
   → statistical evidence that significant relationship exist when there is none, or
   → no relationship when there is
2. **Spurious Correlation** : may appear to have a relationship when there is none
3. **Correlation only measure linear relationship, but not non-linear relationship**

Test the correlation between 2 variables

- **Methodology** : use t-test to test whether the correlation between 2 variables = 0
  \[ H_0: \rho = 0 \; ; \; H_1: \rho \neq 0 \]
  \[ t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}} \]
  \[ df = n-2 \]
  \[ \text{Reject } H_0 \text{ if } t > +t_{\text{critical}} \text{ or } t < -t_{\text{critical}} \]

Dependent / Independent variables

- **Dependent variables** : variable whose variation is explained by independent variables
- **Independent variables** : variable is used to explain the variation of dependent variables

Assumptions of linear regression

1. Linear regression exists between dependent and independent variables
2. Independent variable : uncorrelated with residuals
3. Expected value of residual term \( \mathbb{E}(\varepsilon_i) = 0 \)
4. Variance of the residual term is constant for all observations \( \mathbb{E}(\varepsilon_i^2) = \sigma^2 \)
5. Residual term is independently distributed (residual for observation A is not correlated with residual for observation B)
6. Residual term is normally distributed

Linear regression model

- \( Y = b_0 + b_1 \times X_i + \varepsilon_i \)
  - In which :
    - \( Y \) = dependent variable
    - \( X \) = independent variable
    - \( b_0 \) = regression intercept
    - \( b_1 \) = regression slope
    - \( \varepsilon \) = error term

Linear equation for regression line

- \( \hat{Y} = \hat{b}_0 + \hat{b}_1 \times \bar{X} \)
  - In which :
    - \( \hat{b}_0 \) = estimated slope coefficient \( = \frac{\text{Cov}_{XY}}{\sigma_X^2} \)
    - \( \hat{b}_1 \) = estimated intercept term \( = \bar{Y} - \bar{b}_1 \times \bar{X} \)
    - \( \bar{Y} \) = mean of \( Y \)
    - \( \bar{X} \) = mean of \( X \)

Confidence interval for regression slope coefficient (Range of \( b_1 \))

- **Confidence interval of regression slope coefficient** :
  \[ \hat{b}_1 \pm t_{\alpha/2} \times s_{b_1} \]
  or
  \[ \hat{b}_1 - (t_{\alpha/2} \times s_{b_1}) < b_1 < \hat{b}_1 + (t_{\alpha/2} \times s_{b_1}) \]
  - In which:
    - \( t_{\alpha/2} \) = critical 2-tailed t value for the selected confidence level, with \( df = n-2 \)
    - \( s_{b_1} \) = standard of error

Test hypothesis that slope coefficient = hypothesized value

- From the confidence interval for slope coefficient \( \rightarrow \) use t-test to test the hypothesis that slope coefficient = hypothesized value
  \[ t_{\alpha/2} = \frac{\hat{b}_1 - b_0}{s_{b_1}} \]
  \[ \text{Reject } H_0 \text{ if } t > +t_{\text{critical}} \text{ or } t < -t_{\text{critical}} \]