Day 4
Measures of Association and Modelling

Session 2

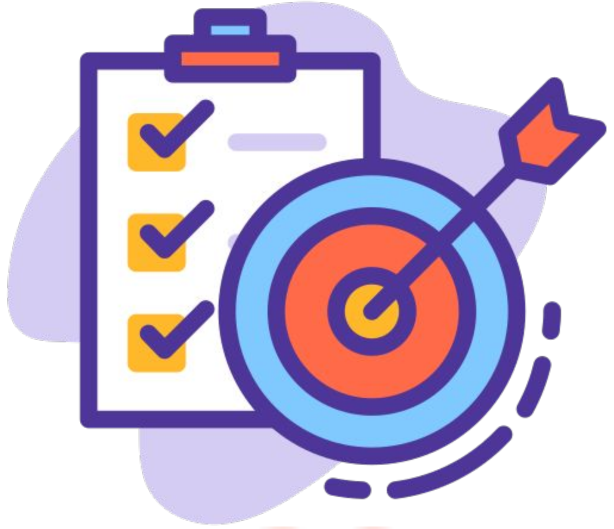
Regression Analysis

Session Outcome

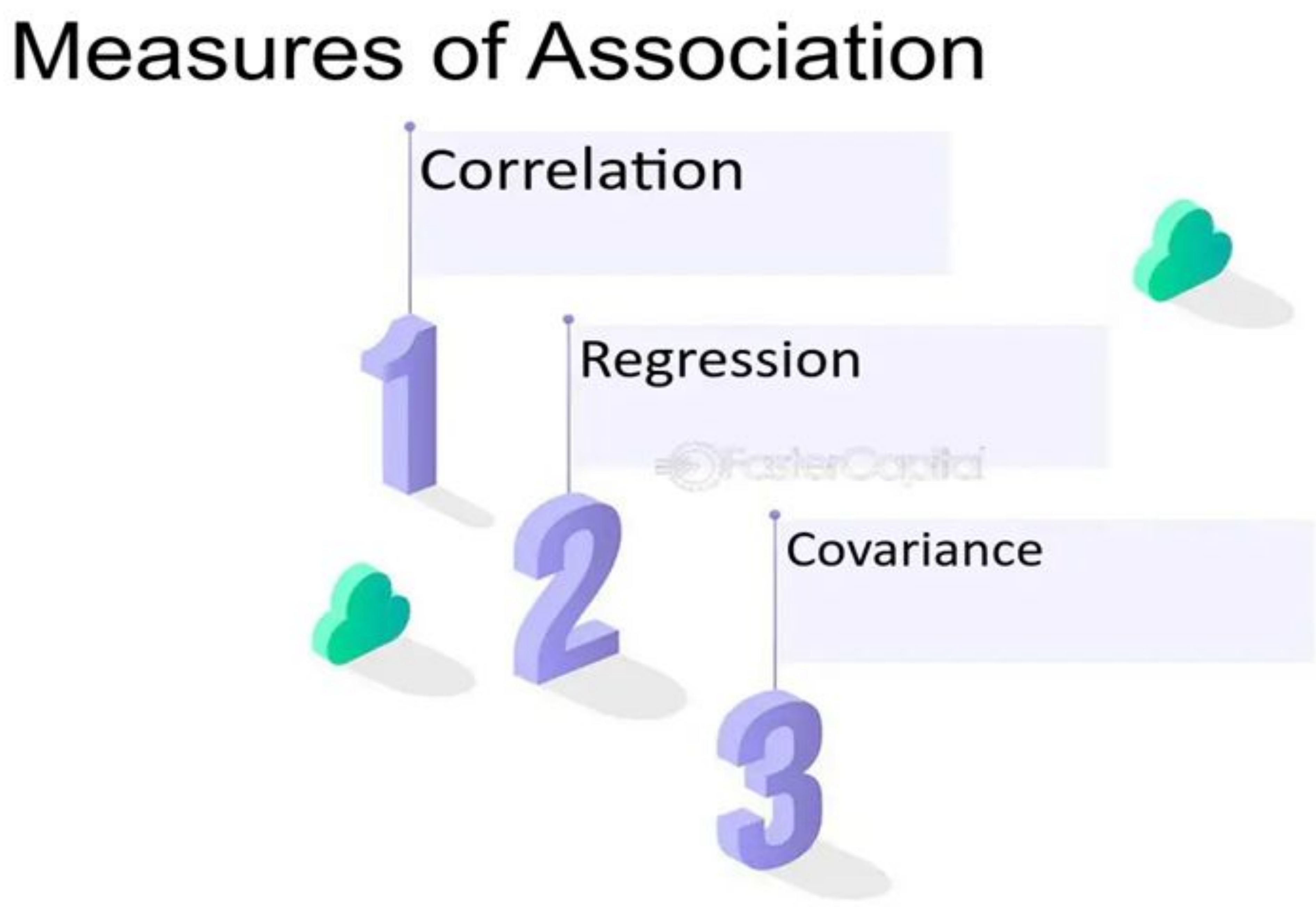
After completing this session, researchers will be able to

- Understand the concept of regression analysis.
- Recognize the crucial role of regression analysis.
- Interpret the parameter from the statistical model.
- Determine the coefficient of determination and its interpretation.

Session Outline



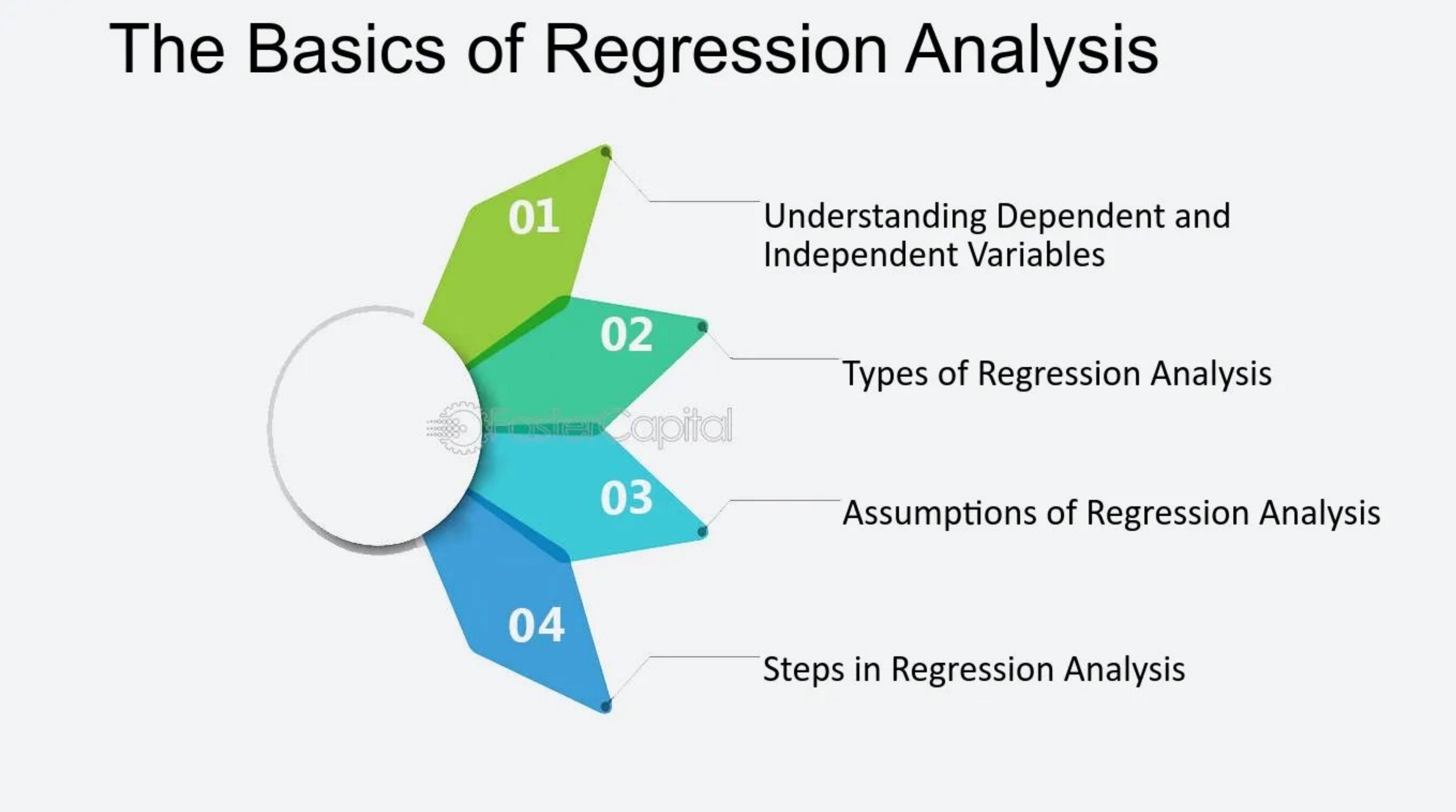
- Simple regression model
- Least square method
- Properties of regression coefficient
- Goodness of fit in regression
- Coefficient of determination



Measures of Association

1. Correlation
2. Regression
3. Covariance

Correlation and Covariance are discussed in the previous session. In this session, we are going to discuss Regression Analysis.



The Basics of Regression Analysis

01 Understanding Dependent and Independent Variables

02 Types of Regression Analysis

03 Assumptions of Regression Analysis

04 Steps in Regression Analysis

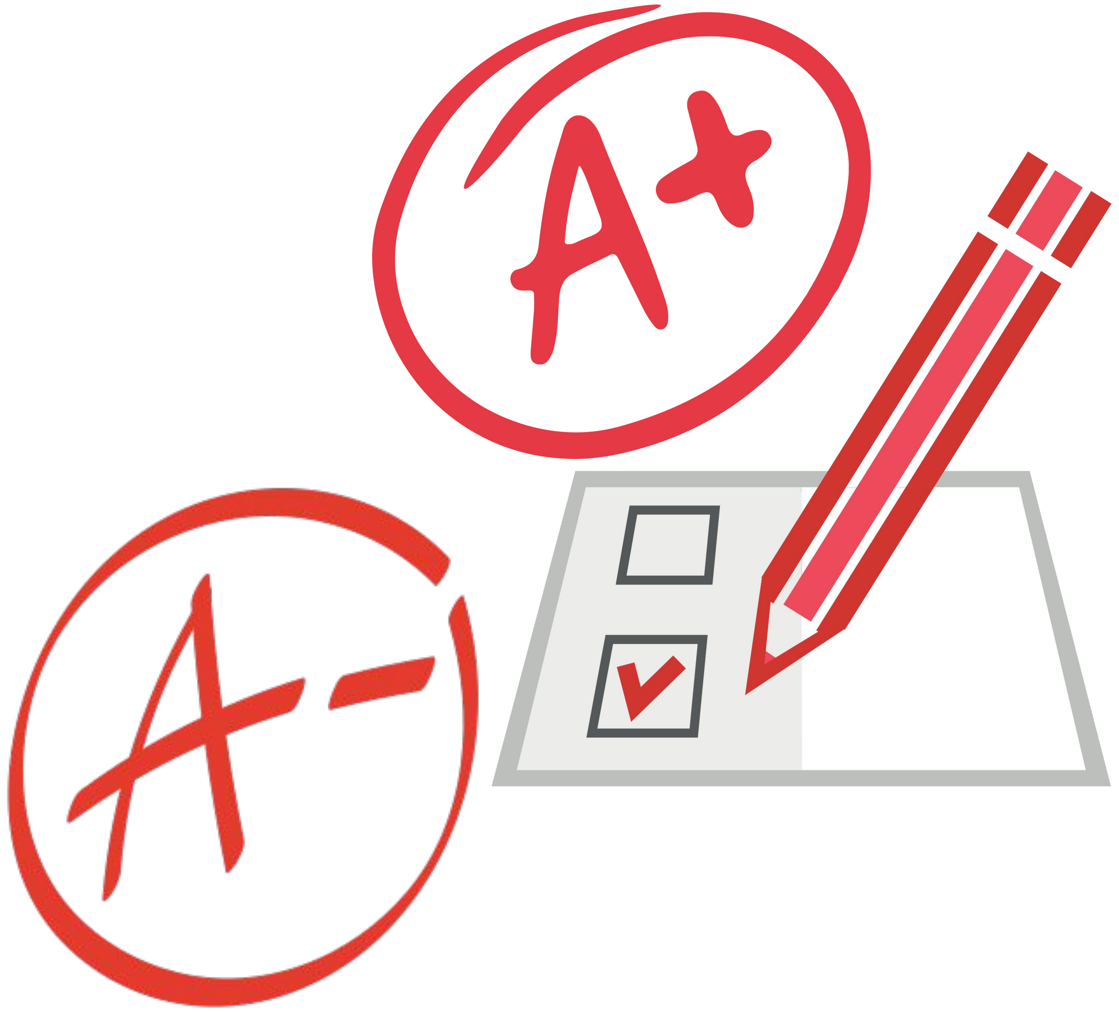
Types of Variables

Independent Variable:

Study time

Dependent variable:

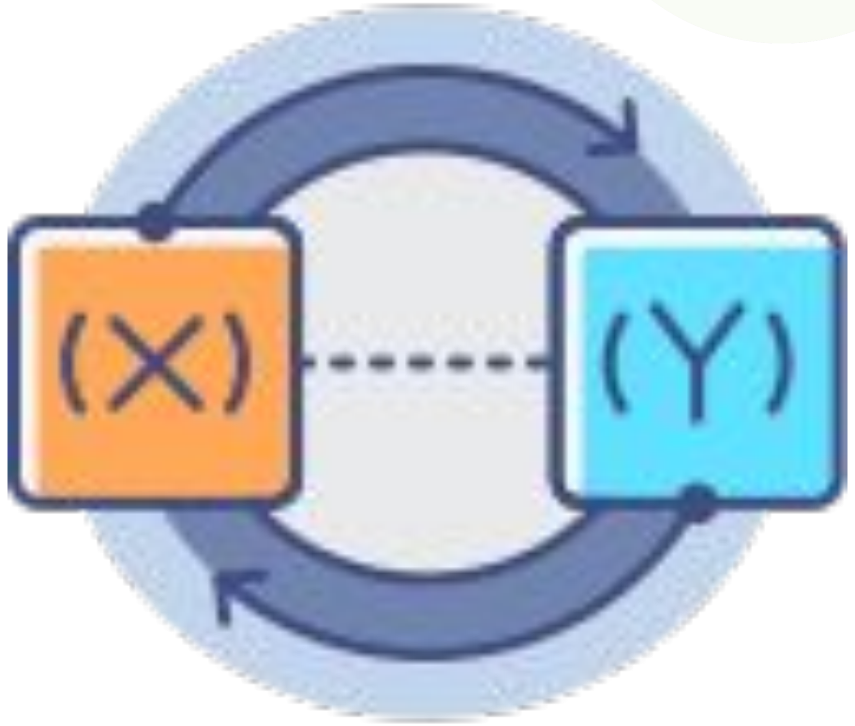
Results / Score



Types of Variables

Independent Variable:

The independent variable is the variable that is manipulated or controlled by the researcher. It is often denoted as X and is hypothesized to cause changes in the dependent variable.

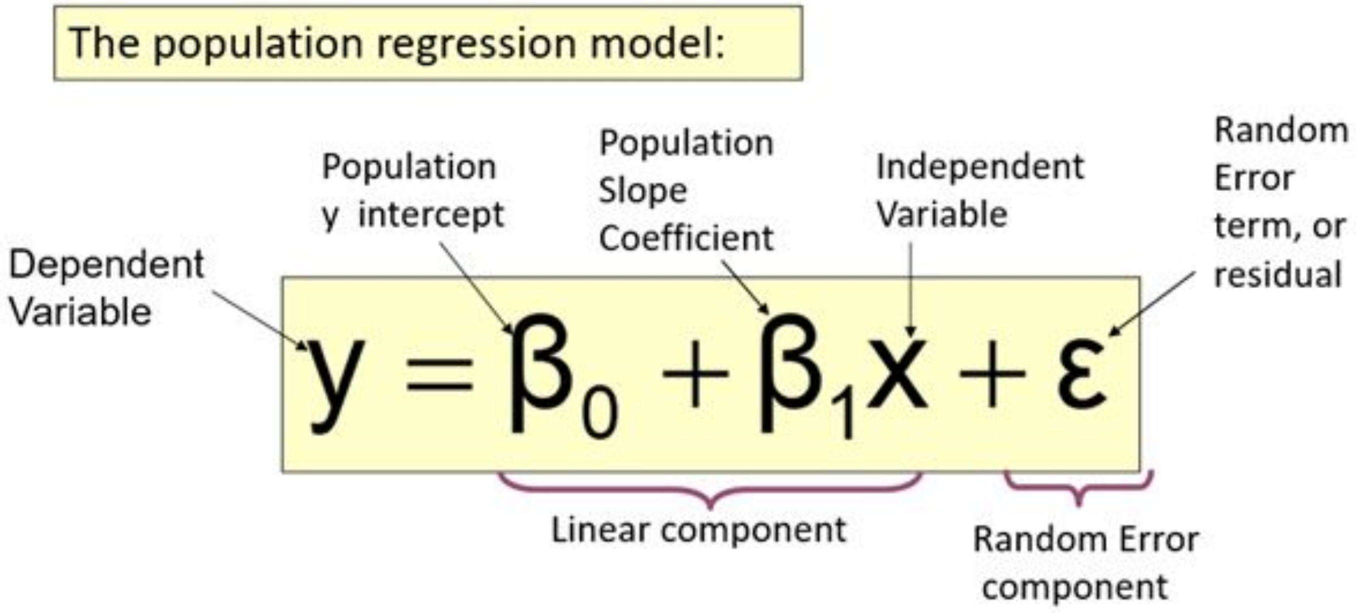


Dependent variable:

The dependent variable is the variable that is observed or measured to assess the effect of the independent variable. It is often denoted as Y and is expected to change in response to variations in the independent variable.

Simple Regression Model

A simple regression model analyzes the relationship between two variables: one predictor (independent) and one outcome (dependent). It predicts or explains changes in the dependent variable based on variations in the independent variable.



The population regression model:

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Linear component}} + \underbrace{\varepsilon}_{\text{Random Error component}}$$

Dependent Variable

Population y intercept

Population Slope Coefficient

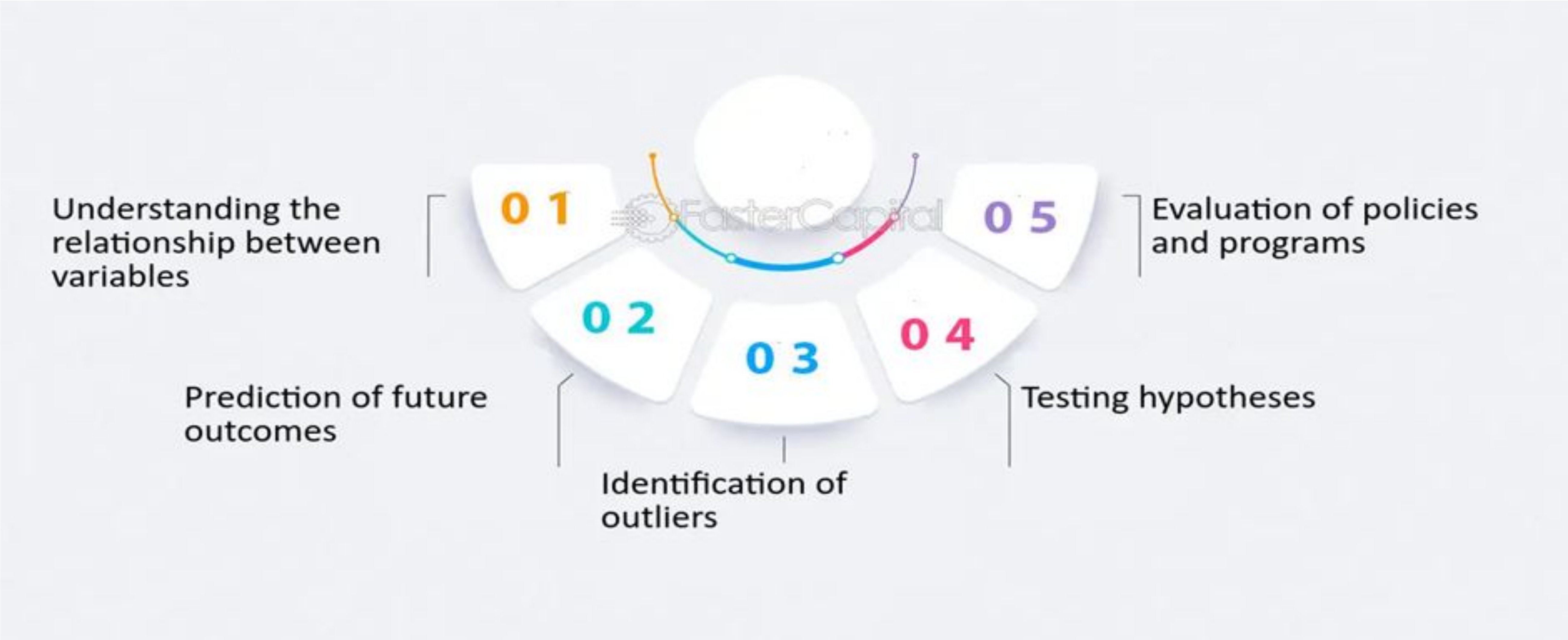
Independent Variable

Random Error term, or residual

Linear component

Random Error component

Importance of ss Regression Analysis



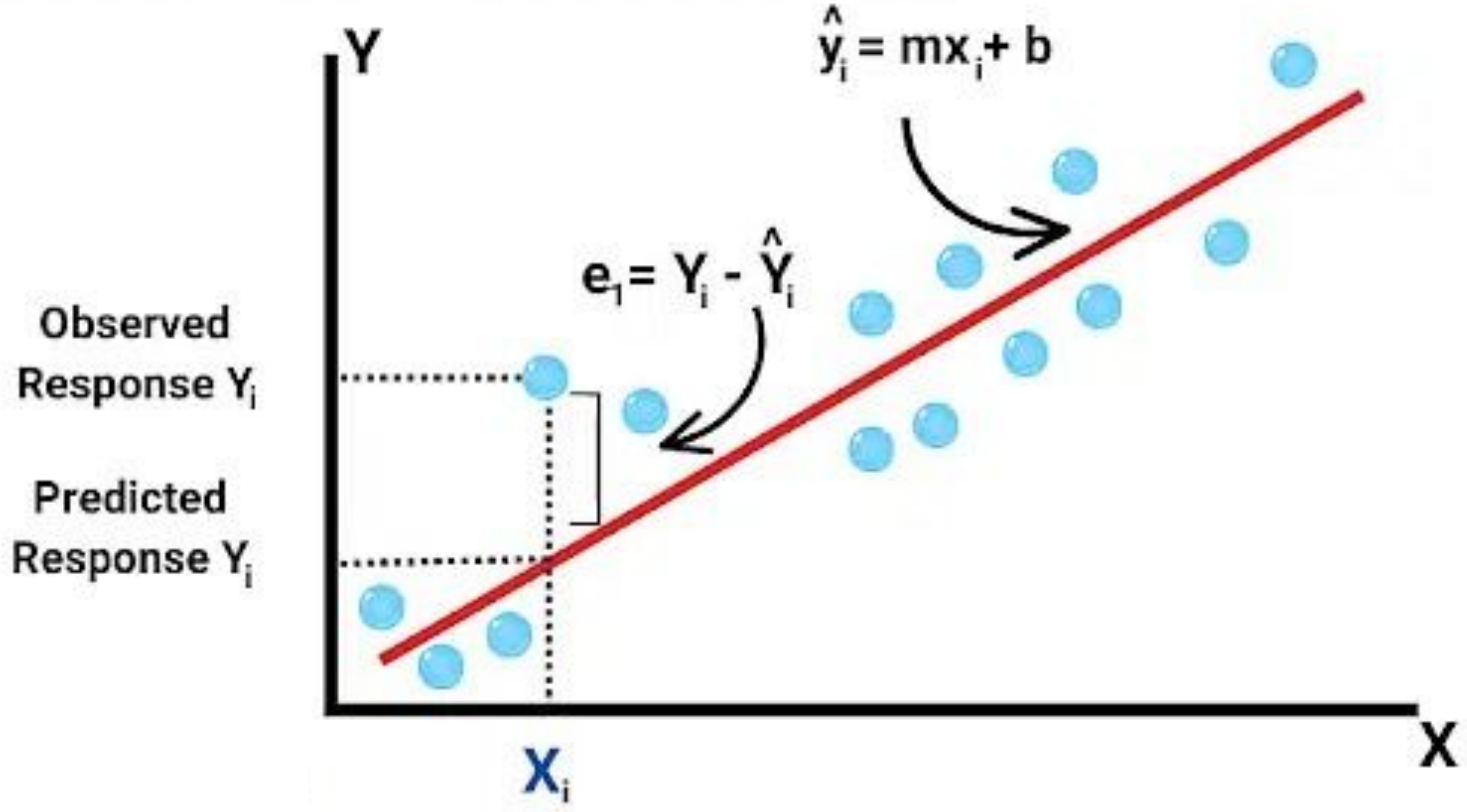
01
Understanding the relationship between variables
02
Prediction of future outcomes
03
Identification of outliers
04
Testing hypotheses
05
Evaluation of policies and programs
FasterCapital

Importance of Regression Analysis

- **Understanding the relationship between variables:** One of the main benefits of regression analysis is that it helps in understanding the relationship between two variables.
- **Prediction of future outcomes:** Regression analysis predicts future outcomes based on the relationship between variables.
- **Identification of outliers:** By identifying outliers, regression analysis can help in improving the accuracy of the analysis and the prediction of future outcomes.
- **Testing hypotheses:** Regression analysis can be used to test hypotheses about the relationship between two variables.
- **Evaluation of policies and programs:** Regression analysis can also be used to evaluate the effectiveness of policies and programs

Least Square Method

The least squares method estimates model parameters by minimizing the sum of squared differences between observed and predicted values, commonly used in regression analysis to find the best-fitting line or curve.



Objectives of Least Square Method

- **Parameter Estimation:** OLS estimates coefficients of a linear regression model that best fit observed data points.
- **Minimization of Residuals:** OLS minimizes the sum of squared differences between observed and predicted values of the dependent variable.
- **Best Linear Unbiased Estimation (BLUE)**: OLS provides unbiased and efficient estimates of regression coefficients.
- **Statistical Inference:** OLS provides estimates of regression coefficients, standard errors, confidence intervals, hypothesis tests, and goodness of fit measures.
- **Prediction:** OLS allows the model to make predictions about the dependent variable for new or unobserved data points.
- **Interpretation of Relationships:** OLS quantifies the effect of changes in independent

Ordinary Least Square Method

Using ordinary least squares estimators to determine optimal parameters

$$\varepsilon = Y_i - \hat{Y}$$

$$\sum \varepsilon^2 = \sum (Y_i - \hat{Y})^2$$

Minimizing sum of the squared differences between observed and predicted values by derivation it with respect to parameter

$$\frac{\partial \sum \varepsilon^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum \varepsilon^2}{\partial \hat{\beta}_1} = 0$$

After simplification,

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i{}^2 - (\sum X_i)^2}$$

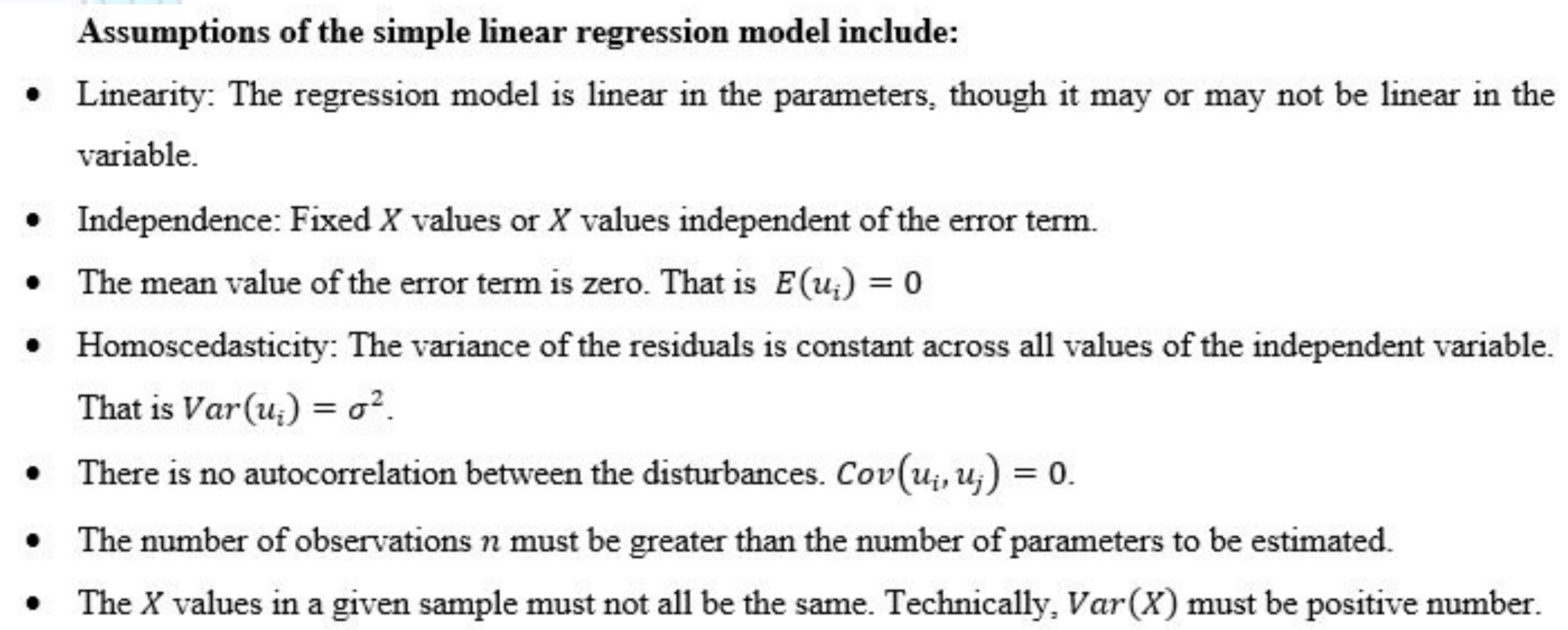
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Interpretation of the parameter:

$\hat{\beta}_0$ represents the value of dependent variable (Y) when the independent variable (X) is zero.

$\hat{\beta}_1$ represents the change of dependent variable (Y) for one unit increase in the independent variable (X).

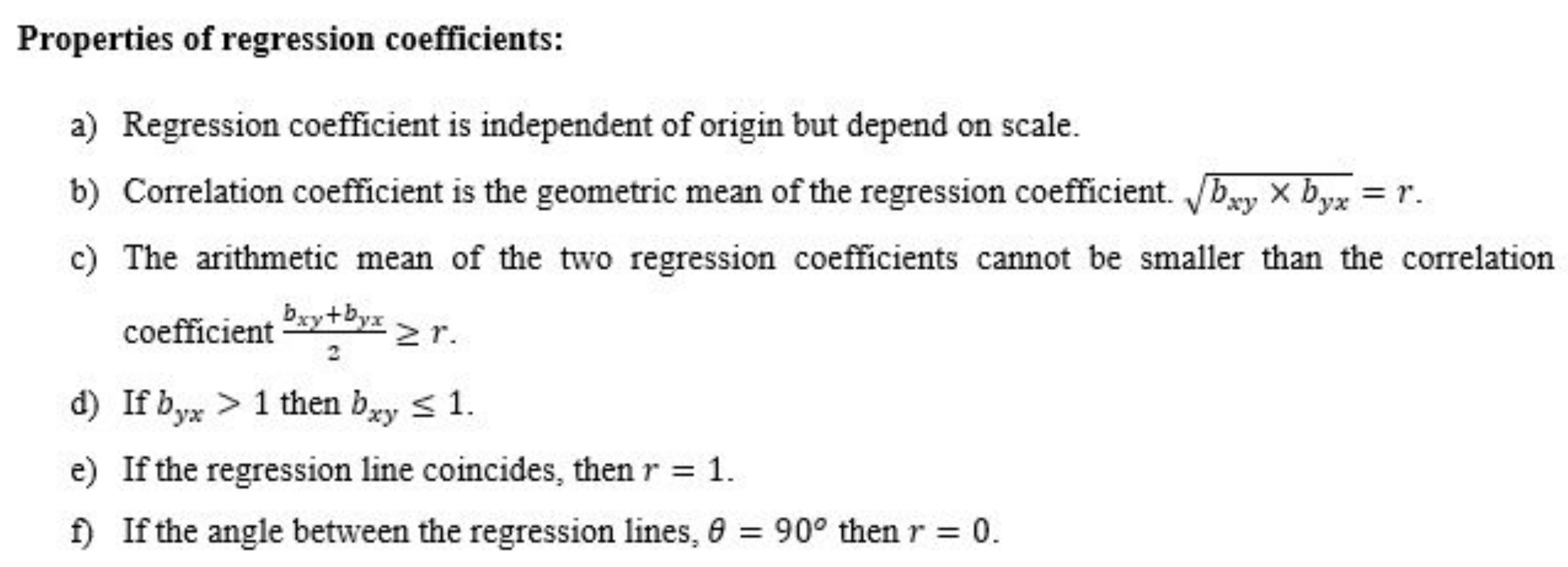
Ordinary Least Square Method



Assumptions of the simple linear regression model include:

- Linearity: The regression model is linear in the parameters, though it may or may not be linear in the variable.
- Independence: Fixed X values or X values independent of the error term.
- The mean value of the error term is zero. That is $E(u_i) = 0$
- Homoscedasticity: The variance of the residuals is constant across all values of the independent variable. That is $Var(u_i) = \sigma^2$.
- There is no autocorrelation between the disturbances. $Cov(u_i, u_j) = 0$.
- The number of observations n must be greater than the number of parameters to be estimated.
- The X values in a given sample must not all be the same. Technically, $Var(X)$ must be positive number.

Ordinary Least Square Method



Properties of regression coefficients:

a) Regression coefficient is independent of origin but depend on scale.

b) Correlation coefficient is the geometric mean of the regression coefficient. $\sqrt{b_{xy} \times b_{yx}} = r$.

c) The arithmetic mean of the two regression coefficients cannot be smaller than the correlation coefficient $\frac{b_{xy}+b_{yx}}{2} \geq r$.

d) If $b_{yx} > 1$ then $b_{xy} \leq 1$.

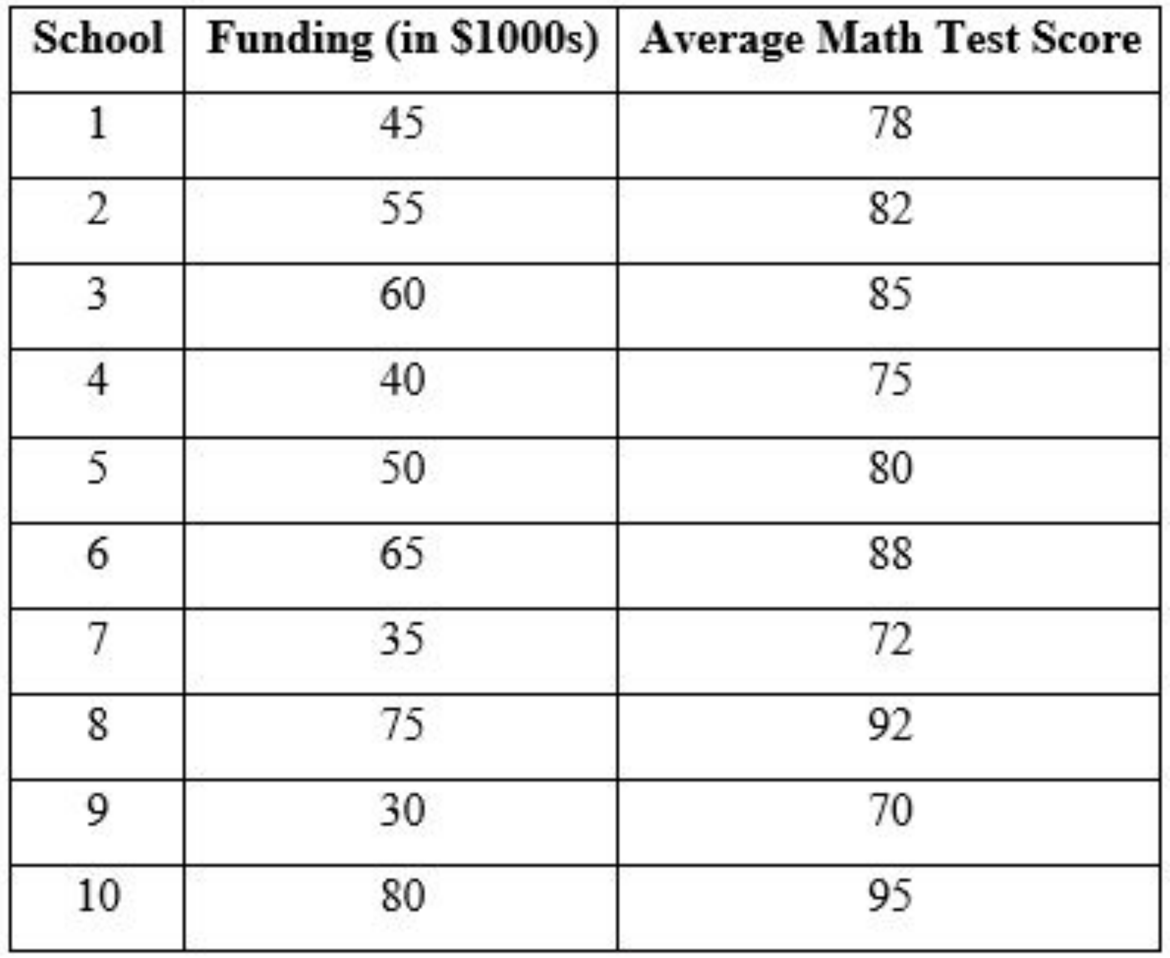
e) If the regression line coincides, then $r = 1$.

f) If the angle between the regression lines, $\theta = 90^o$ then $r = 0$.

Least Square Method

Example

The ministry collects data from 50 primary schools, where they record the annual funding allocated to each school (in thousands of dollars) and the average test scores of students in mathematics.



School	Funding (in $1000s)	Average Math Test Score
1	45	78
2	55	82
3	60	85
4	40	75
5	50	80
6	65	88
7	35	72
8	75	92
9	30	70
10	80	95

To conduct simple regression, we estimate parameters (intercept and slope) using least squares with sample data.

Least Square Method

$$n = 10$$

$$\sum_{i=1}^{n} X_i = 45 + 55 + \cdots + 80 = 545$$

$$\sum_{i=1}^{n} Y_i = 78 + 82 + \cdots + 95 = 797$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{545}{10} = 54.5$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{797}{10} = 79.7$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{10} (X_i - 54.5)^2 = (45 - 54.5)^2 + (55 - 54.5)^2 + \cdots + (80 - 54.5)^2 = 6801$$

$$\sum\nolimits_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum\nolimits_{i=1}^{10} (X_i - 54.5)(Y_i - 79.7) = 161.5$$

Least Square Method

$$\hat{\beta}_1 = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i{}^2 - (\sum X_i)^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{161.5}{6801} = 0.0237$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_2\bar{X} = 79.7 - 0.0237 * 54.5 = 80.9941$$

The estimated model is given by $\hat{Y} = 80.99 + 0.0237X$

For the funding score 30 then the estimated math score is given by 81.701

$\hat{\beta}_0 = 80.994$ represents the value of the average test scores of students when the annual funding is zero.

$\hat{\beta}_1 = 0.0237$ represents the change in the average test scores of students for a one-unit change in the annual funding.

Subset Selection Methods

Subset selection methods in statistical modeling choose a subset of variables from a larger pool, particularly useful in regression analysis with numerous predictors. They aim to identify the most influential variables while omitting less relevant ones, enhancing model interpretability and mitigating overfitting.

Importance of Subset selection method

Improves Model Interpretability: Simplifies complex models by identifying relevant predictor variables.

Reduces Overfitting: Prevents overfitting by choosing important predictor variables.

Enhances Model Performance: Improves prediction accuracy and model conciseness.

Reduces Computational Complexity: Decreases computational complexity in multiple regression models.

Identifies Important Variables: Offers insights into predictor variable importance, aiding in

Subset Selection Methods

Model building strategy for the multiple linear regression model

1. Forward Procedure
2. Backward Procedure
3. Stepwise Selection Procedure

- **Forward Procedure**

First, include the covariates with the smallest p-value.

Next, add each of the covariates separately to the model and include that covariate in the model which leads to the largest increase in the likelihood given that the increase likelihood is significant.

Add the next covariate in the same way that no covariate can be added that leads to a significant increase in likelihood.

Subset Selection Methods

2. Backward Procedure

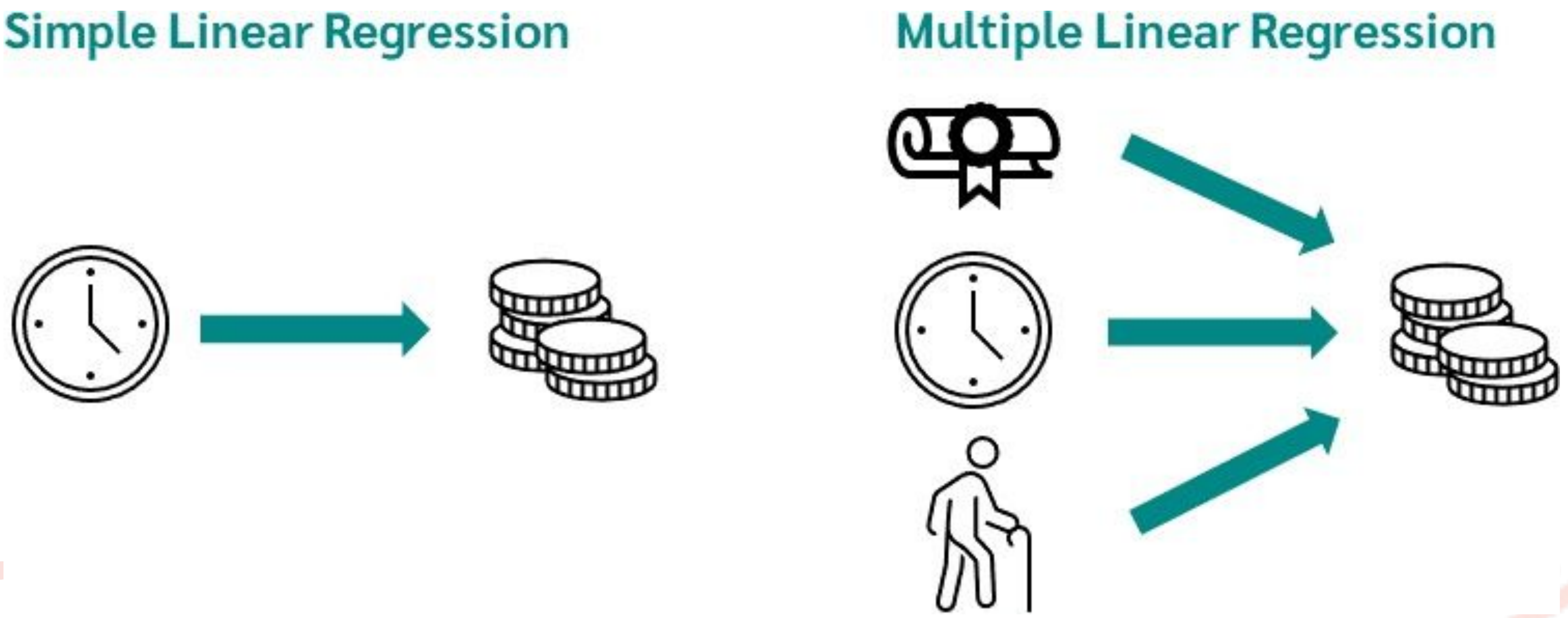
Start from the full model that includes all the covariates.

Continue removing covariates from the model, starting the one which leads to the smallest changes in likelihood, until no covariates can be removed unless it results in a significant decrease in likelihood.

3. Stepwise Procedure

Start as in the forward procedure, but an include covariates could be excluded at latter stage, if it is no longer significant with other covariates in the model.

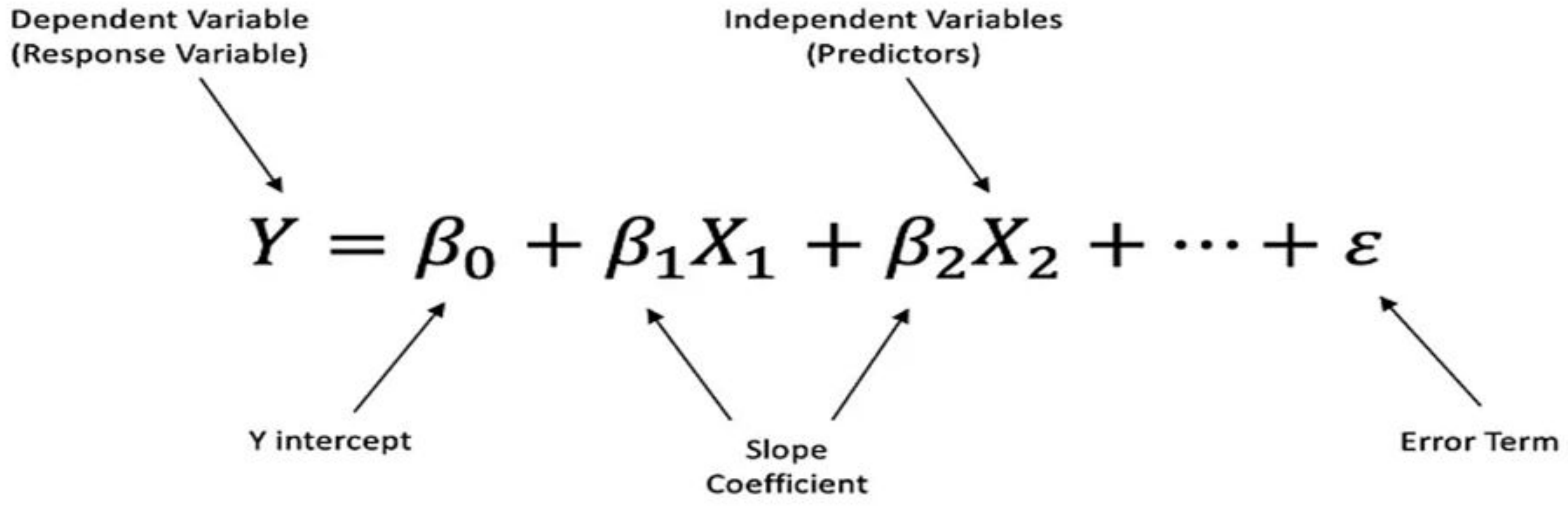
Simple vs Multiple Rgression



Multiple Linear Regression Model

Multiple Linear Regression Model

Multiple Linear Regression analyzes the relationship between multiple independent variables and a single dependent variable, extending from simple linear regression. It accounts for each predictor's effect on the dependent variable while considering other predictors, assuming a linear relationship and aiming to find the best-fitting equation based on observed data.



Dependent Variable (Response Variable)

Independent Variables (Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Y intercept

Slope Coefficient

Error Term

Multiple Linear Regression Model

For three variable regression model is defined by,

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

By using OLS is to be estimated. The slope coefficients are given by,

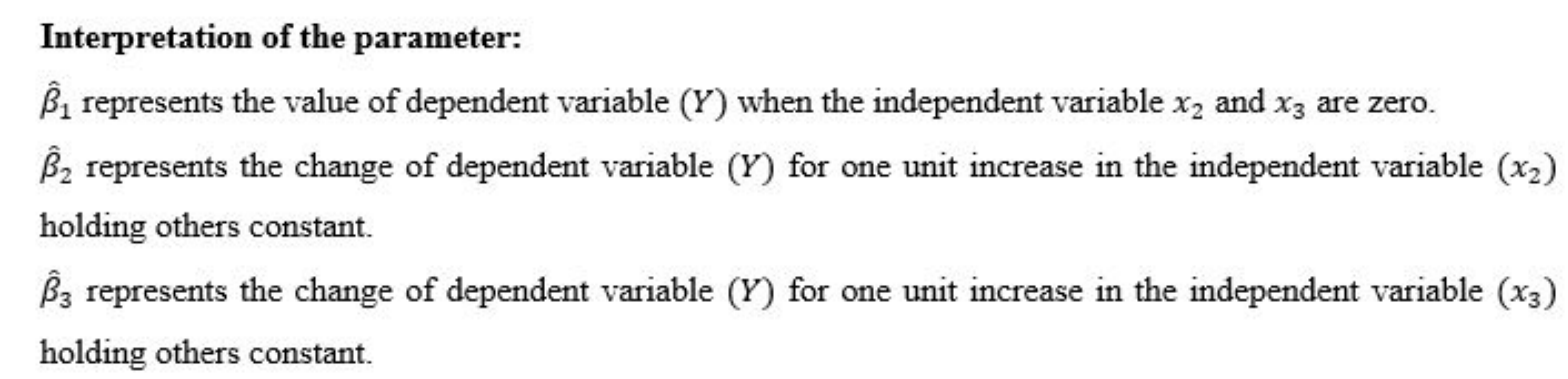
$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{2i} - \bar{x}_1)^2}$$

$$\hat{\beta}_3 = \frac{\sum_{i=1}^{n}(x_{3i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{3i} - \bar{x}_1)^2}$$

And intercept is given by,

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}_2 - \hat{\beta}_3 \bar{x}_3$$

Multiple Linear Regression Model



Interpretation of the parameter:

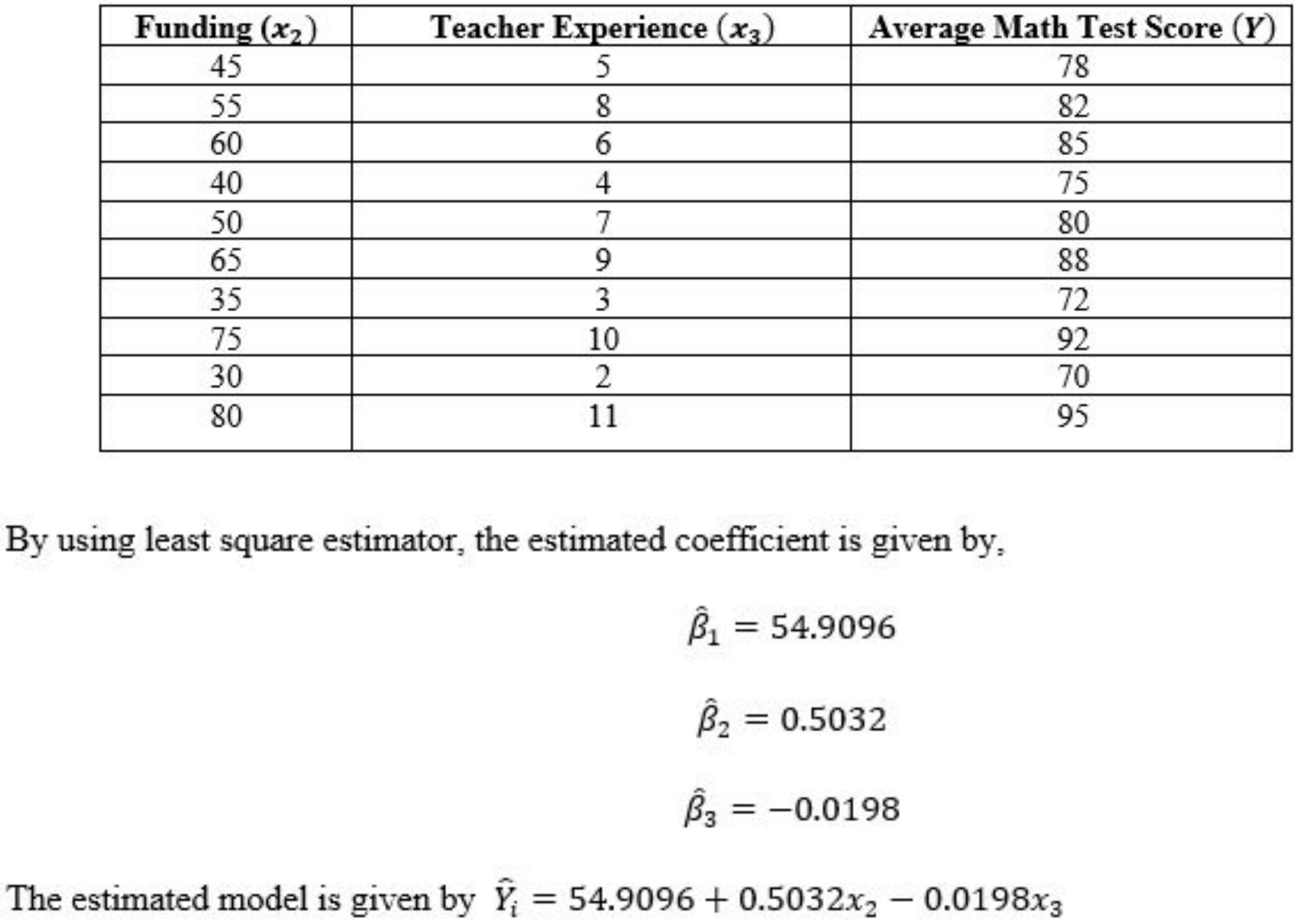
$\hat{\beta}_1$ represents the value of dependent variable (Y) when the independent variable x_2 and x_3 are zero.

$\hat{\beta}_2$ represents the change of dependent variable (Y) for one unit increase in the independent variable (x_2) holding others constant.

$\hat{\beta}_3$ represents the change of dependent variable (Y) for one unit increase in the independent variable (x_3) holding others constant.

Example: A multiple linear regression model where we want to predict an average math test score based on two independent variables: funding and the teacher experience.

Multiple Linear Regression Model



Funding (x_2)	Teacher Experience (x_3)	Average Math Test Score (Y)
45	5	78
55	8	82
60	6	85
40	4	75
50	7	80
65	9	88
35	3	72
75	10	92
30	2	70
80	11	95

By using least square estimator, the estimated coefficient is given by,

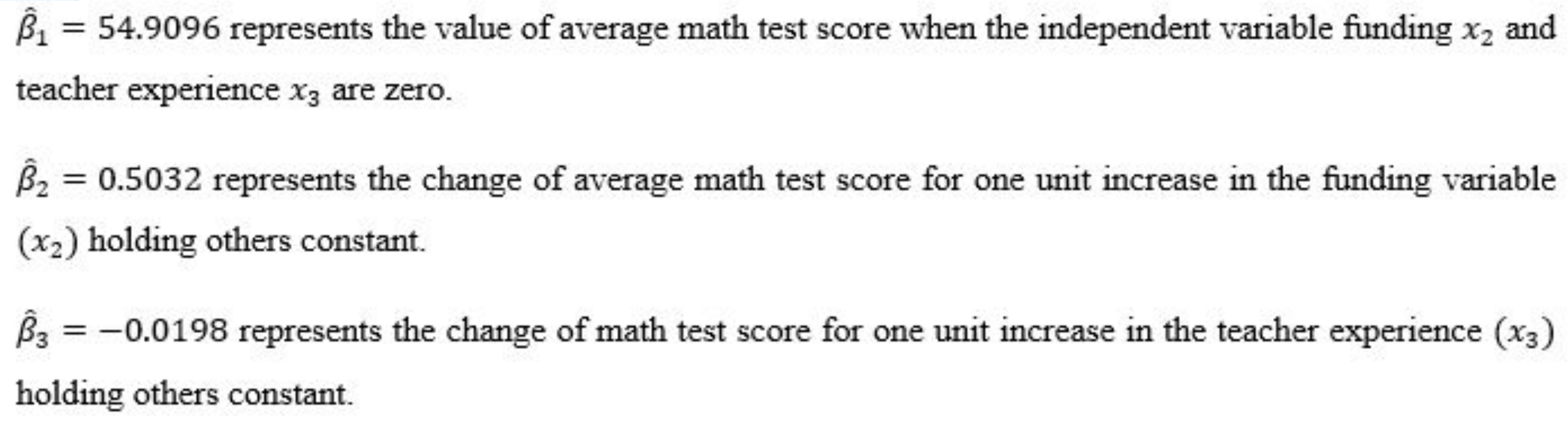
$$\hat{\beta}_1 = 54.9096$$

$$\hat{\beta}_2 = 0.5032$$

$$\hat{\beta}_3 = -0.0198$$

The estimated model is given by $\hat{Y}_i = 54.9096 + 0.5032x_2 - 0.0198x_3$

Multiple Linear Regression Model

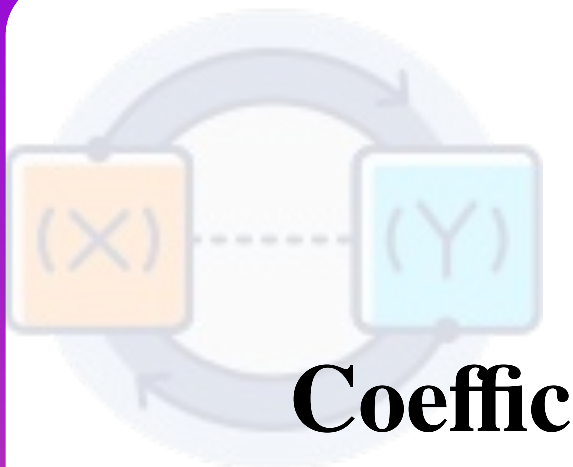


$\hat{\beta}_1 = 54.9096$ represents the value of average math test score when the independent variable funding x_2 and teacher experience x_3 are zero.

$\hat{\beta}_2 = 0.5032$ represents the change of average math test score for one unit increase in the funding variable (x_2) holding others constant.

$\hat{\beta}_3 = -0.0198$ represents the change of math test score for one unit increase in the teacher experience (x_3) holding others constant.

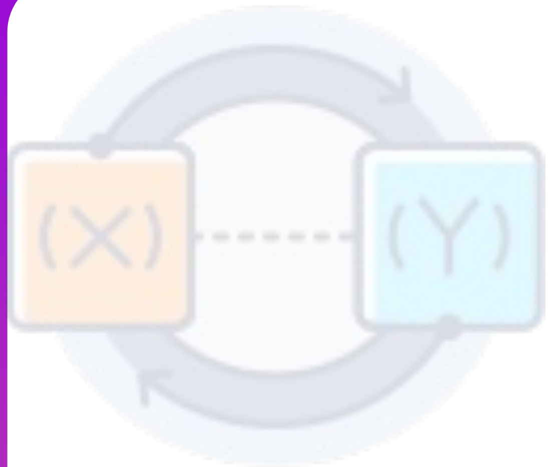
Goodness of Fit



Coefficient of Determination: The coefficient of determination ranges from 0 to 1, indicating the proportion of variance in the dependent variable explained by the independent variable(s).

Adjusted: This version adjusts for predictor count, penalizing unnecessary variables for a more reliable measure of fit, crucial when comparing models with different predictors.

Residual Analysis: Inspecting residuals is vital for assessing fit. Plots should show randomness, constant variance, and normality. Patterns or trends may signal model fit issues.



Goodness of Fit

F-tests: F-tests assess model significance by comparing explained to unexplained variability. A significant result indicates a better fit than a model with no predictors.

Predictive Accuracy: Goodness of fit can be assessed through predictive accuracy, using techniques like cross-validation or comparing predicted values to observed values in a validation dataset.

Coefficient of Determination

It is interpreted as the proportion of the total variability in the dependent variable that is accounted for by the independent variable(s) included in the model.

Here are some key aspects of its importance:

- **Assessment of Model Fit:** It gauges the fit of the regression model to observed data. A higher value signals better fit, indicating more explained variability in the dependent variable by the independent variable(s).
- **Interpretation of Relationships**: It indicates the strength and direction of the relationship between independent and dependent variables. Higher values imply stronger relationships, while lower values signify weaker ones.

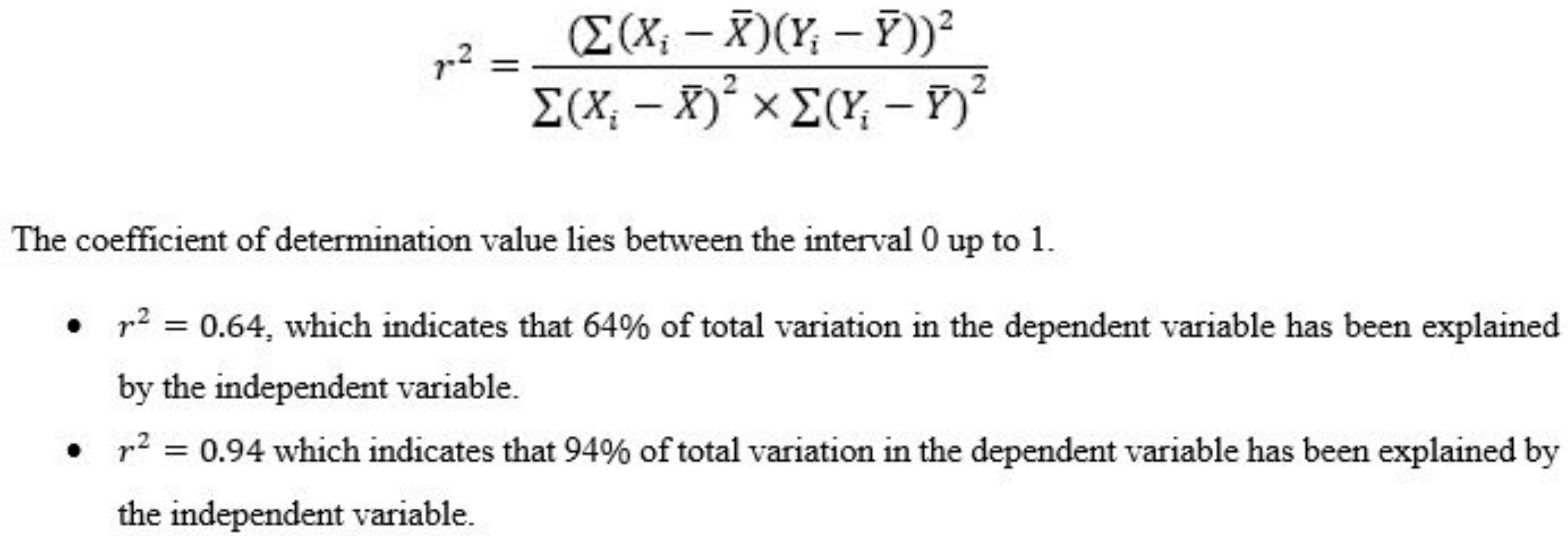
Coefficient of Determination

Comparison of Models: It enables comparison of regression models to identify the one that best explains variability in the dependent variable. Researchers can compare values across models with different independent variables to find the most suitable model.

Decision Making: It aids decision-making in fields like economics, finance, and social sciences. By assessing the impact of factors on outcomes, policymakers and analysts can make informed decisions.

Coefficient of Determination

The coefficient of determination is the square term of the correlation coefficient. The formula for the coefficient of determination is given as follows:



$$r^2 = \frac{(\Sigma(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\Sigma(X_i - \bar{X})^2 \times \Sigma(Y_i - \bar{Y})^2}$$

The coefficient of determination value lies between the interval 0 up to 1.

- $r^2 = 0.64$, which indicates that 64% of total variation in the dependent variable has been explained by the independent variable.
- $r^2 = 0.94$ which indicates that 94% of total variation in the dependent variable has been explained by the independent variable.

Thank You