

Modul 3

# Regression

Data Science Program

# What is Regression?

- Tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables ( $X_1, X_2, \dots, X_n$ )
- Relationship can be linear or non-linear

# Mathematical vs Statistical Relationship

- Mathematical is an exact relationship

$$Y = \beta_0 + \beta_1 X$$

- Statistical is NOT an exact relationship

$$Y = \beta_0 + \beta_1 X + \epsilon$$

# Regression Dictionary

- Dependent variable (*response variable*) measures an outcome of a study (can also be called *outcome variable*)
- Independent variables (*explanatory variables*) explain changes in a response variable
- Given set values of independent variables to see how it affects dependent variables-> predict dependent variables

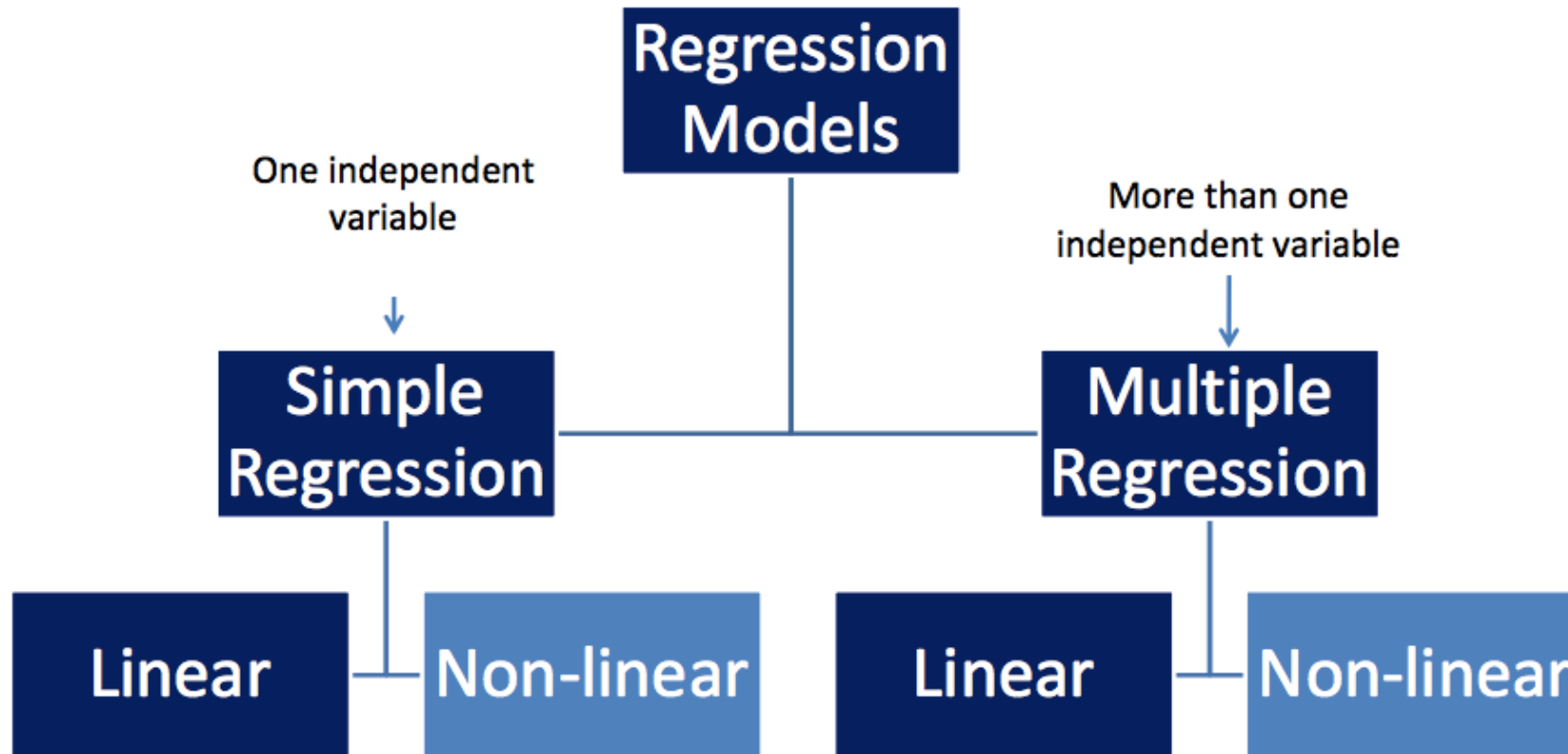
# Dependent and Independent Variables

- Terms dependent and independent does not necessarily imply a causal relationship between two variables
- Regression is NOT to capture causality
- Purpose of regression: predict the value of dependent variable given the values of independent variables

# Why we need Regression?

- Companies would like to know about factors that have significant impact on their key performance indicators.
- Helps to create new hypothesis that assist companies to improve their performance and hence better decision making

# Types of Regression (1)



## Types of Regression (2)

- Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

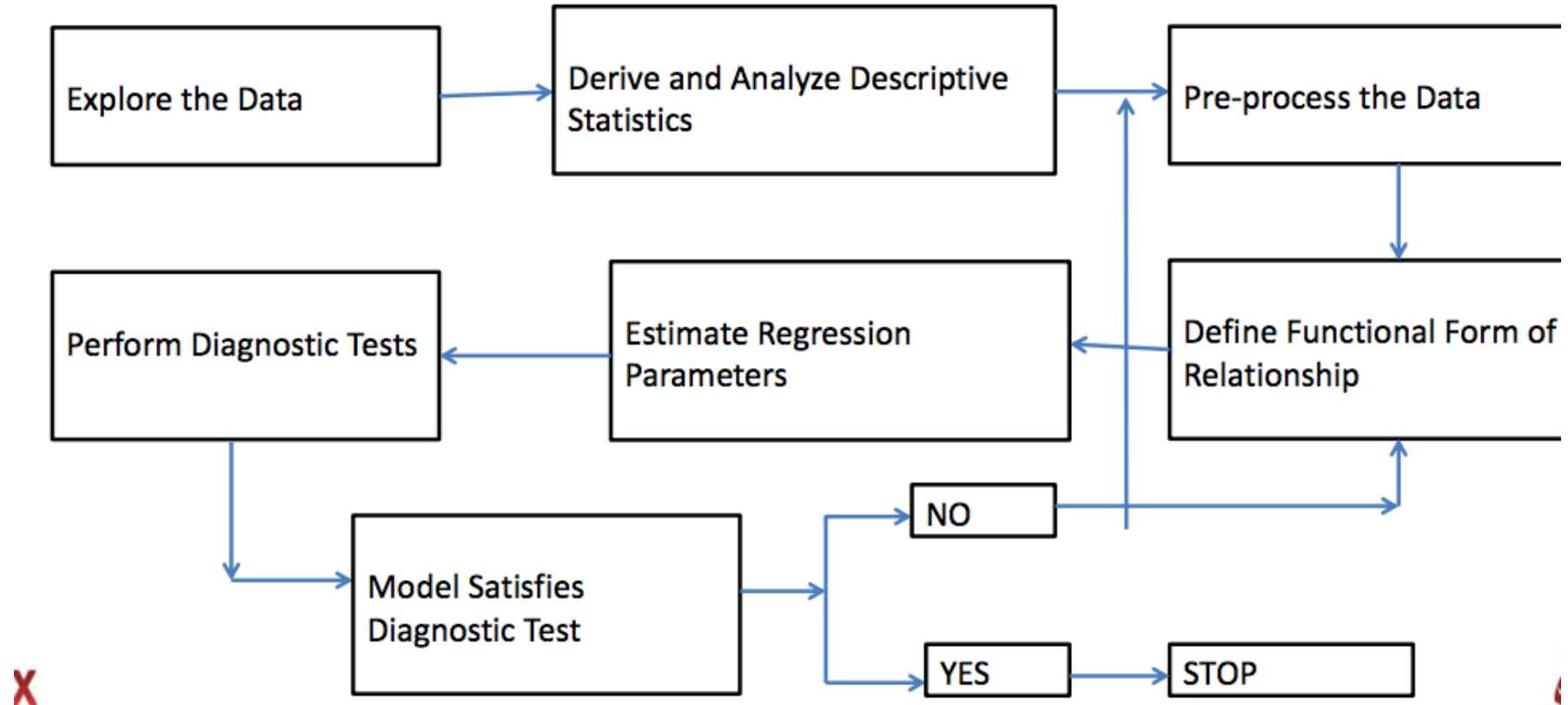
- Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

- Important task in Regression is to estimate beta values



# Regression Model Development



# Model Building

- Identify independent variable
- Specify the nature of relationship between dependent variable and independent variable (intercept and constant)

# Linear Regression Model

- Relationship between variables is a linear function (population)

The diagram illustrates the Linear Regression Model equation:  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ . The equation is centered on a dark blue background. Labels with arrows point to each part of the equation:   
 - "Population Y-Intercept" points to  $\beta_0$ .   
 - "Population Slope" points to  $\beta_1$ .   
 - "Random Error" points to  $\varepsilon_i$ , which is circled in red.   
 - "Dependent (Response)" points to  $Y_i$ .   
 - "Independent (Explanatory)" points to  $X_i$ .

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Population Y-Intercept

Population Slope

Random Error

Dependent (Response)

Independent (Explanatory)

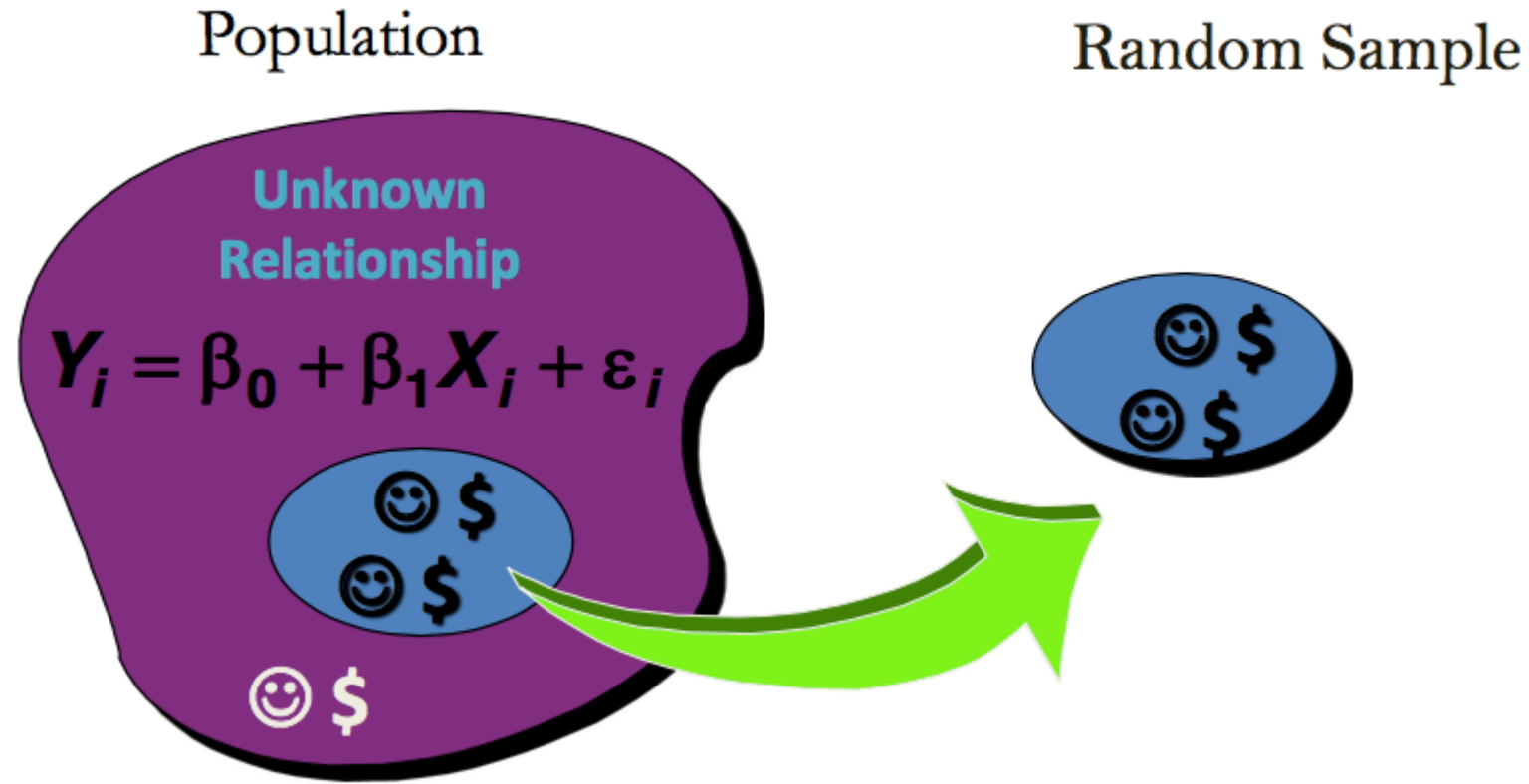
# Linear Regression Model Assumption (Ordinary Least Square(OLS))

Linear Regression is the most common estimation for Linear model, the estimation have some assumption requirements to be fulfilled in order to get the best estimation. The assumption are:

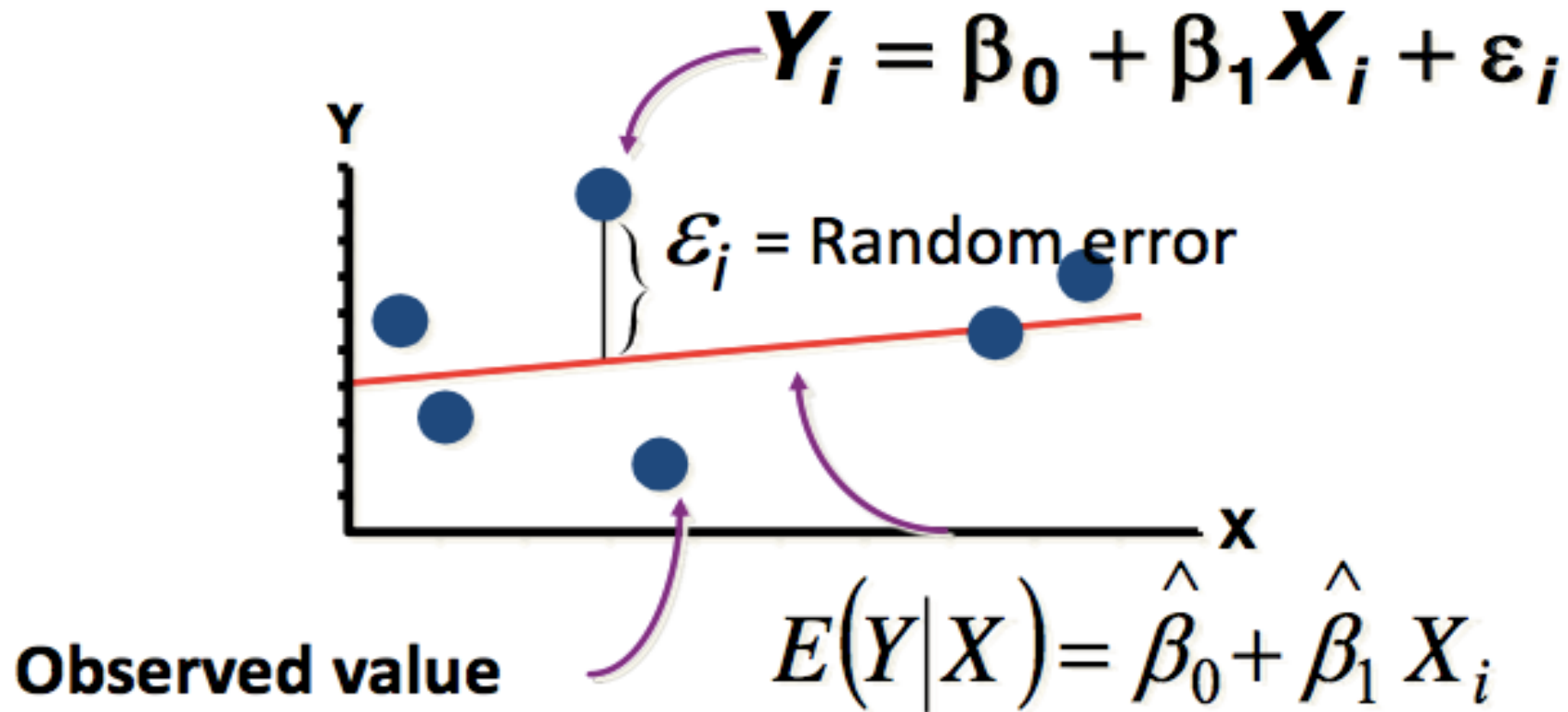
- The regression model is linear in the coefficient and the error term
- The error term has a population mean of zero
- All independent variables are uncorrelated with the error term (Exogeneity)
- Observation of the error terms are uncorrelated with each other
- The error term has constant variance (Homoscedasticity)
- No Independent variable perfectly correlated with the other independent variable (no multicollinearity)
- The error term are normally distributed

When we talk about error term, it is the residual instead of the population error term. We need to use the residual as the population error term is unknown.

# Regression: OLS (Ordinary Least Squares) estimation



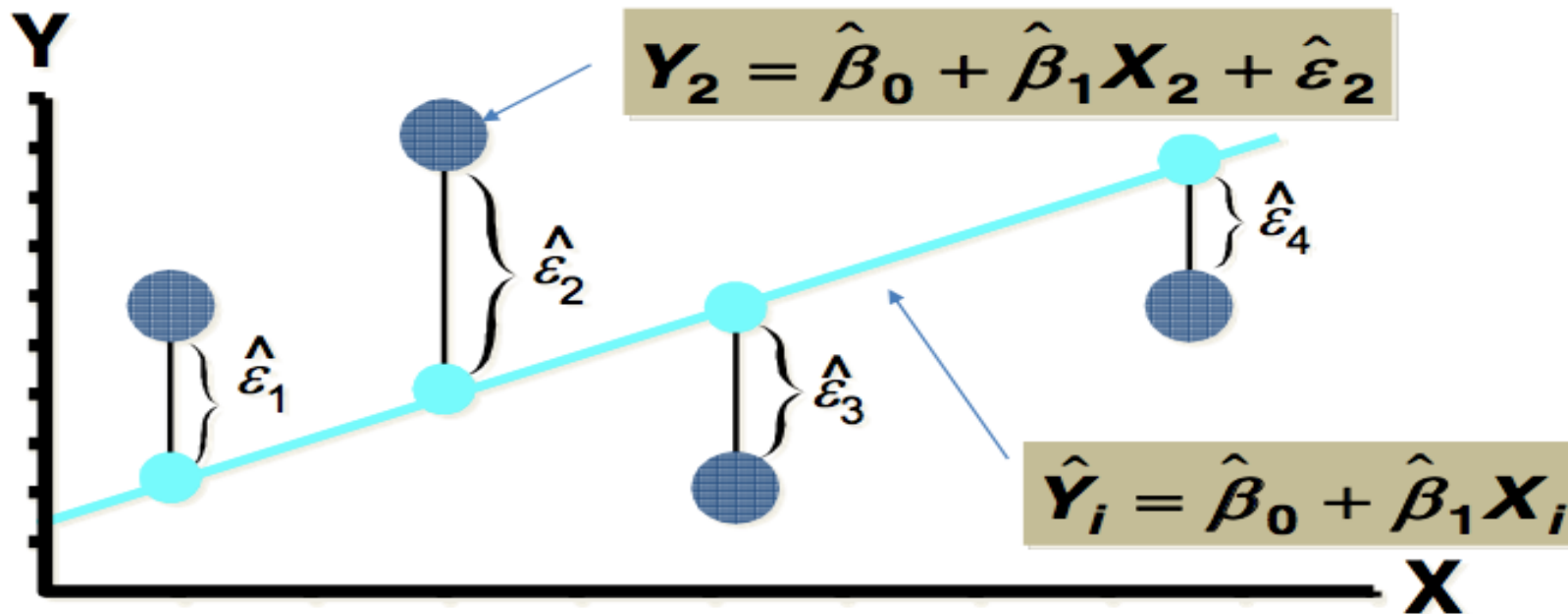
# Population Linear Regression Model



# Least Squares

Least Squares (Sum Square Error (SSE)) is a method that calculate error from the regression model line, and square it. The aim of this method is trying to minimize this term.

$$\text{LS minimizes } \sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \dots + \hat{\varepsilon}_n^2$$



# Estimation of Parameters in Regression

- Least squares function is given by:

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2$$



# Coefficient Equations

- Prediction Equation of Linear Regression:  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$   
Observed independent variable
- Sample Slope: 
$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
  
Observed dependent variable  
Dependent variable mean  
Independent variable mean
- Sample Y-intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Why Least Squares Estimate?

- Remember the OLS assumption? OLS beta estimates provided the error terms are uncorrelated (no auto regression) and have equal variance (homoscedasticity).

$$E\left[\beta - \hat{\beta}\right] = 0$$


- it implies that  $E\{\beta - \hat{\beta}\} = 0$ , where  $\beta$  = population parameter,  $\hat{\beta}$  = parameter that we estimate using sample

# Interpret Beta

- Interpretation depends on the functional form of the relationship between dependent and independent variable.
- Coefficients Interpretation:
  - The intercept,  $\beta_0$  is the mean value of the dependent variable Y, when the independent variable  $X = 0$
  - The slope,  $\beta_1$  is the change in the value of dependent variable Y, for unit change in the independent variable X

# Simple Linear Regression

Variable $x$ and $y$ has <b><i>Linear</i></b> relationship	Assumption of the world
$y = \beta_0 + \beta_1 x + \varepsilon$ , <b><i>Minimize SSE</i></b>	Fitting a model
Is $x$ really related to $y$ ? <b><i>Is <math>\beta_1</math> statistically significant?</i></b>	Validating the model
<b><i>Predict</i></b> $y$ for a given $x$ .	Using a model



# Model validation

- Use of co-efficient of determination to check the goodness of fit of regression
- T-test to validate relationship between dependent and independent variables
- Analysis of Variance (ANOVA) and F-test to check the overall fitness of the regression model
- Residual analysis to check the model adequacies

# Coefficient of Determination

- Measure of how well the regression line fits the data
- Coefficient of determination ( $R^2$  square) lies between 0 and 1
- Percentage of variation that can be explained by the regression model

# Variation in Dependent Variable (Y)

Remember this is our Regression Linear Model

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

$$\text{Variation in } Y_i = \text{Systemic Variation} + \text{Random Variation}$$

or

$$\text{Variation in } Y_i = \text{Explained Variation} + \text{Unexplained Variation}$$

# Variation in Dependent Variable (Y)

Observed dependent variable      Prediction      Dependent variable mean

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

Total variation      Explained variation      Unexplained variation

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

*SST*      *SSR*      *SSE*



# Variation in Dependent Variable (Y)

- SST (Total Sum of Squares):
  - How much error is there in predicting Y without the knowledge of X
- SSE (Sum of Squares Error):
  - How much error is there in predicting Y with the knowledge of X
- SSR (Sum of Squares Regression):
  - Amount of variation explained by the model
- Mathematically,  $SST = SSR + SSE$

# R-square

- What is explained by model over what is total variation

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

# Standard Error of Estimate

- Standard error is the estimate of the standard deviation of the regression errors
- Standard error of estimate( $Se$ ), measures the variability or scatter of the observed values around the regression line. In simpler word, it measure the actual data around the regression line that was created.

# Interpretation of SE Estimate

- Smaller SE of Estimate indicates better fit
- Larger SE of Estimate, the greater the scattering of points around the regression line

# Standard Error of Estimate

$$S_e = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}}$$

If you remember Standard deviation formula, it is similar to that just have more uncertainty. What we want to have is the most minimum score of the Se.

# Standard Error of Estimate for regression coefficient

- We could also measure the amount of sampling error in a regression coefficient by using the Se. Just like before, we want have smaller number of this term.

$$SS_x = \sum_i (X_i - \bar{X})^2$$

$$S(\beta_0) = \frac{S_e \times \sqrt{\sum x^2}}{\sqrt{nSS_x}}$$

$$S(\beta_1) = \frac{S_e}{\sqrt{SS_x}}$$

# T-test

- T-test is a type of hypothesis test which tells how significant the differences between the groups are. In other words, it test if the differences happen because of the random chances or not.
- Beta coefficient is a function of  $Y_i$ . Since  $Y_i$  follows normal distribution,  $\beta_1$  also follows normal distribution
- $Y_i$  follows normal distribution since we assume that the error term follows normal distribution

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

This is the Beta coefficient formula

# T-test

- The T-test point is to check whether the real slope (beta coefficient) is equal to zero or not
- We done Two-tailed hypothesis test for T-test hypothesis

Null hypothesis:

$$H_0: \beta_1 = 0$$

Alternative hypothesis:

$$H_1: \beta_1 \neq 0$$



# Test Statistic

- Errors follow normal distribution; thus test statistic follows t-distribution with  $n-2$  degrees of freedom (df)
- Test statistic:

$$t_{(n-2)} = \frac{\text{Estimate value of parameter} - \text{hypothesis parameter}}{\text{Estimated standard error of estimate}} = \frac{\hat{\beta}_1 - \beta_1}{S_e(\hat{\beta}_1)}$$

$$t_{(n-2)} = \frac{\hat{\beta}_1 \sqrt{SS_x}}{S_e}, \text{ where } \beta = 0$$

# Hypothesis testing decision rule

- From the T-test, we would get the t-score which we could use to get the P-value
- We use the t-distribution table to get where the P-value score is fall

## **P-value**

- Often, we use 95% confidence level. If P-value is less than 0.05, then we reject the null hypothesis and accept alternative hypothesis -> 95% confidence that null hypothesis may highly improbable if it is true

# Multiple Linear Regression

- Several independent variables may influence the change in dependent variable we are trying to study
- Relationship between 1 dependent & 2 or more independent variables is a linear function

The diagram illustrates the Multiple Linear Regression equation: 
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$
 Each term in the equation is labeled with a blue arrow pointing to it: 

- Population Y-intercept** points to  $\beta_0$ .
- Population slopes** points to the  $\beta$  coefficients ( $\beta_1, \beta_2, \dots, \beta_k$ ).
- Random error** points to  $\varepsilon_i$ .
- Dependent (response) variable** points to  $Y_i$ .
- Independent (explanatory) variables** points to the  $X$  variables ( $X_{1i}, X_{2i}, \dots, X_{ki}$ ).

# Multiple Regression Modeling Steps

1. Start with a hypothesis or belief
2. Estimate unknown model parameters (Beta coefficients)
3. Probability distribution of random error term -> assumed to be a normal distribution
4. Check the assumptions of regression (OLS)
5. Evaluate Model
6. Use Model for Prediction and Estimation

# Prediction Model for Multiple Linear Regression

- Prediction equation for the multiple linear regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

k is the number of the independent variable included for the model

# Model Diagnostics

- Test for overall model fitness (R-square and adjusted R-square)
- Test for overall model statistical significance (F test test)
- Test for statistical significance of individual explanatory variables (t test)
- Test for Normality and Homoscedasticity of residuals
- Test for Multi-collinearity and Auto Correlation

# Coefficient of determination in Multiple Regression

- Coefficient of determination increases as the number of independent variables increases.
- In SSR/SST, the SSR increases as the number of independent variables increases, whereas SST remains constant.
- Increase in  $R^2$  can be deceptive, since a greater number of independent variables may overfit the data (overfitting).

# Adjusted R-square

- Inclusion of additional independent variable will increase  $R^2$  value.
- To correct this defect, we adjust the  $R^2$  by taking into account the degrees of freedom (usually number of the samples – 1).



# Adjusted R-Square

$$R_A^2 = 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)}$$

$R_A^2$  = Adjusted R - Square

n = number of observations

k = number of explanatory variables

# Test for overall significance of model – F Test

- F-test is a statistical test that commonly used when comparing statistical model that have been fitted to the data set.
- F-test check for overall significance of multiple regression model.
- F-test checks if there is a statistically significant relationship between Y (dependent variable) and any of the dependent variables

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_A : \text{Not all } \beta \text{ values are zero}$$

# F Statistic

- The null hypothesis is rejected if the  $F$  calculated from the data is greater than the critical value of the  $F$ -distribution (see the  $F$ -table) for some desired significance level
- Relationship between  $F$  and  $R^2$

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

# Testing for Significance of Individual Parameters

- T-test: by rejecting null hypothesis, there is a statistically significant relationship between the dependent variable Y and independent variable  $X_i$ .

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

# Dummy Variable

- Categorical (qualitative) variables in Regression model cannot be integrated as it is, because they are not numerical
- To alleviate this problem, categorical variable in regression are replaced with dummy variables (or indicator variables) in regression model
- A categorical variable with  $n$  levels are replaced with  $(n-1)$  dummy variables. The category for which no dummy variable assigned is known as “Base Category”

# Dummy variable

- When there are more than one categorical variable, it is advisable to use  $(n-1)$  dummy variables for both categorical variables along with the intercept.
- Use of  $n$  dummy variables along with intercept will result in multi-collinearity, known as dummy variable trap.
- This is because  $n-1$  dummy variable already capture the dropped categorical variable during coefficient estimation

# Dummy variables in Regression

- The intercept,  $\beta_0$  is the mean value of the base category.
- The coefficients attached to dummy variables are called differential intercept coefficients -> measure deviation from the base category for that specific dummy variable

# Interaction Variables

- A regression model of type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$



**Interaction  
variable**

- Usually an interaction between a quantitative and qualitative variable



# Interaction Variable Example

- Predict Gender Discrimination
- Consider a regression model with salary as dependent variable Y:

$$Y = \beta_0 + \beta_1 \text{ Gender} + \beta_2 \text{ Work Experience} + \beta_3 \text{ Gender x Work Experience}$$

Let Gender = 1 implies Female:

Then Y for Female is:

$$Y = \beta_0 + \beta_1 + (\beta_2 + \beta_3) \text{ Work Experience}$$

Y for Male is:

$$Y = \beta_0 + \beta_2 \text{ x Work Experience}$$

# Multicollinearity

- High correlation between independent variables is called multi-collinearity.
- Multi-collinearity leads to the unstable coefficients.
- Multi-collinearity is always existing ;it is just the matter of degree(how strong the correlation). Assumption of OLS is that there are no perfect correlation (+1 or -1) between the independent variables

# Properties to check for Multi-collinearity

- Having High  $R^2$  but only few significant t ratios.
- F-test rejects the null hypothesis, but none of the individual t-tests are rejected.
- Correlations between pairs of X variables (independent variables) are more than with Y variables (dependent variables).

# Effects of Multi-collinearity

- The variances of regression coefficient estimators are inflated.
- Magnitudes of regression coefficient estimates may be different
- Adding and removing variables produce large changes in the coefficient estimates.
- Regression coefficient may have opposite sign.

# Identify Multi-collinearity Variance Inflation Factor (VIF)

- The variance inflation factor (VIF) is a relative measure of the increase in the variance in standard error of beta coefficient because of collinearity.
- A VIF greater than 10 indicates that collinearity is very high. A VIF value of more than 4 is not acceptable.

# Variance inflation factor

Variance inflation factor associated with introducing a new variable  $X_j$  is given by:

$$VIF(X_j) = \frac{1}{1 - R_j^2}$$

$R_j^2$  is the coefficient of determination for the regression of  $X_j$  as dependent variable

The standard error of the corresponding Beta is inflated by  $\sqrt{VIF}$

# VIF method

- Take particular X as dependent variable and all other independent variables as independent variables.
- Run a regression between one of those independent variables with remaining independent variables.
- Standard error of estimate is inflated by a quantity which is square root of VIF

# Regression Model Building

- In *Forward selection* method, the entry independent variable is the one with smallest p-value based on F-test
- In *Backward elimination* method, all independent variables are entered into the equation and then sequentially removed starting with the most insignificant variable. At each step, the largest probability of F is removed
- In *Step-wise Regression*, the entry variable is the one with smallest p-value based on F-test. At each step, the independent variable not in the equation that has the smallest probability of F is entered.



# Residual Plot

- Residual plot is a plot of error (or standardized error) against one of the following variables:
  - The dependent variable  $Y$ .
  - The independent variable  $X$ .
  - The standardized independent or dependent variable.

# Residual Analysis

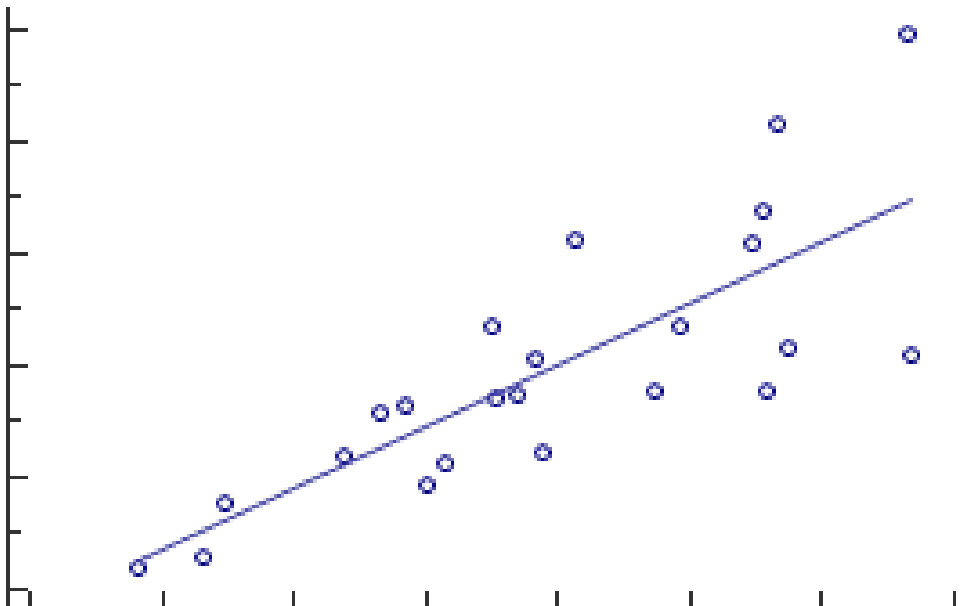
- Analysis of residuals reveal whether the assumption of normally distributed errors hold.
- Residual plots are used to check if there is heteroscedasticity problem (non constant variance for the error term).
- Residual analysis could also indicate if there are any missing variables.
- Residual plot can also reveal if the actual relationship is non-linear.

# Normality of error terms

- Probability plot is a graphical technique for checking whether or not a data set follows a given distribution.
- The data is plotted against a theoretical distribution in such a way that the points should form a straight line.
- In Regression, we create a probability plot of error against normal distribution.
- If residual do not follow normal distribution, t-test and F-test are not valid

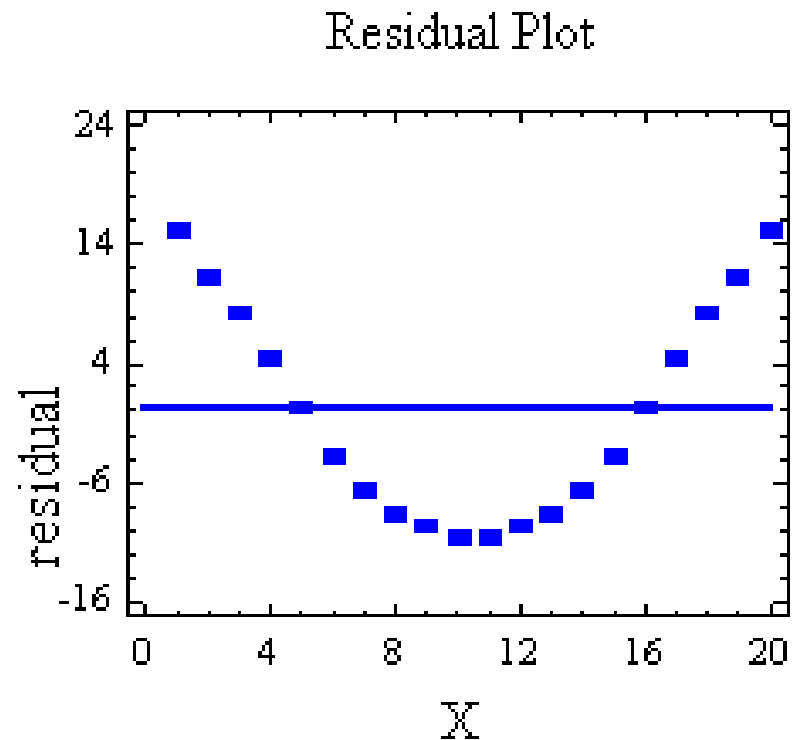
# Check for heteroscedasticity

- A graph of the residuals versus dependent variable Y or independent variable X will reveal whether the variance of the errors are constant.
- If the width of the scatter plot of the residuals either increases or decreases as X (or Y) increases, then the assumption of constant variance is not met.



# Check for non-linearity

- If the residual plot exhibits a curve when plotted, then the actual relationship is non-linear.



# A small bit of review on Linear Regression

*Quiz:*

- *What is linear regression?*
- *What are the assumptions of linear regression?*
- *How can you measure the significance of the model?*