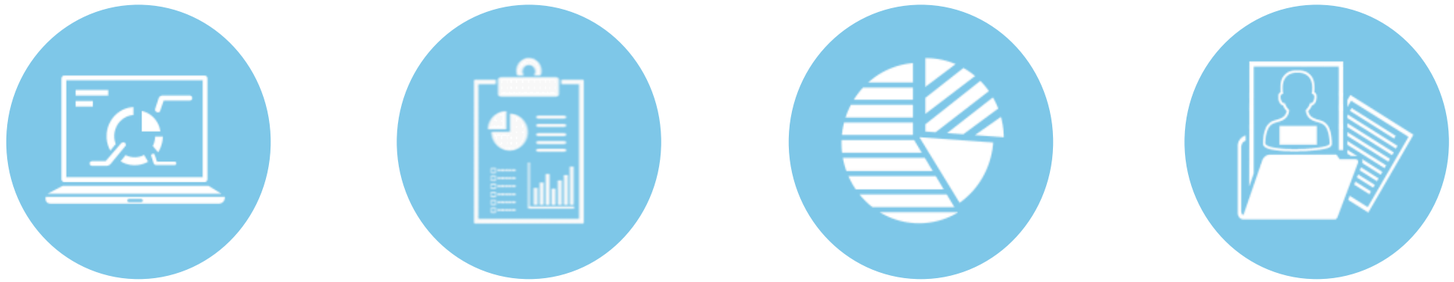
Data Science with R

Lesson 7— Regression Analysis



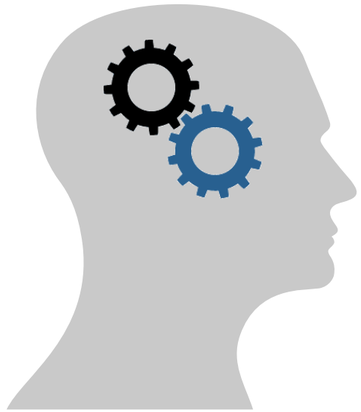
Learning Objectives

- Explain the meaning and uses of regression analysis
- Describe the different types of regression analysis models
- List the functions to convert non-linear models to linear models
- Discuss R squared and adjusted R squared models
- Explain Principal Component Analysis and Factor Analysis of Dimensionality Reduction

Regression Analysis

Topic 1— Introduction to Regression Analysis

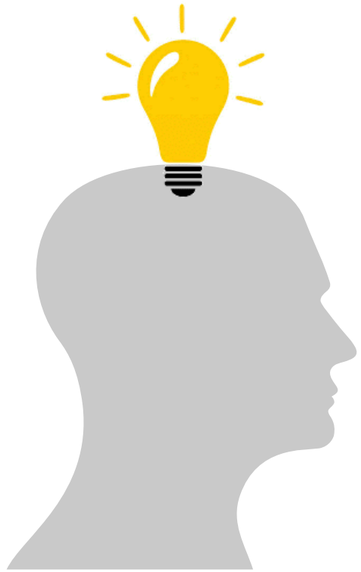
Introduction



Consider a scenario where an apparel manufacturing company needs to find out the increase in sales for a particular month based on its TV ad.

The analysis involves finding the relation between the promotion campaign ad and the sales.

Introduction



In such cases (where a relation between two variables needs to be derived), regression is used. **Regression analysis** is used to make effective business prediction.

What Is Regression Analysis?

Regression Analysis is a technique used to estimate the relationship between variables and predict the value of one variable (dependent variable) on the basis of other variables (independent variables).

Regression Analysis

Topic 2—Types of Regression Analysis Models

Types of Regression Analysis Models

Simple Regression

Multiple Regression

Types of Regression Analysis Models

Simple Regression

Multiple Regression

- It depicts the relationship between a dependent variable and an independent variable.
- It considers one quantitative and independent variable X to predict the other quantitative, but dependent, variable Y.
- A straight line is fit to the data.

Types of Regression Analysis Models

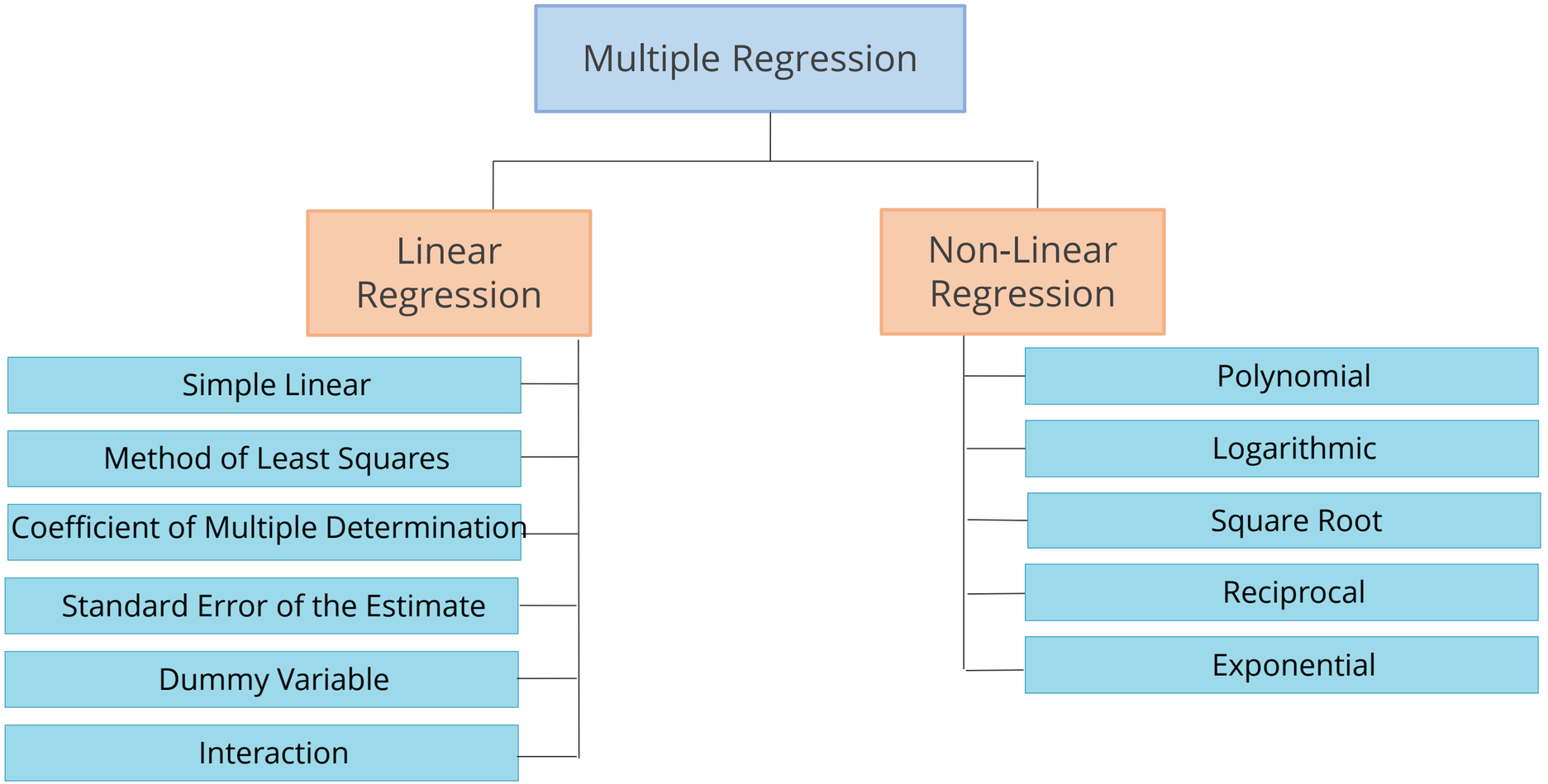
Simple Regression

Multiple Regression

- It predicts the value of a variable based on the value of two or more other variables.
- It considers more than one quantitative and qualitative variable ($X_1 \ldots X_N$) to predict a quantitative and dependent variable Y.

Types of Regression Analysis Models

- Simple Regression
- Multiple Regression



Regression Analysis

Topic 3—Linear Regression

What Is Linear Regression?

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

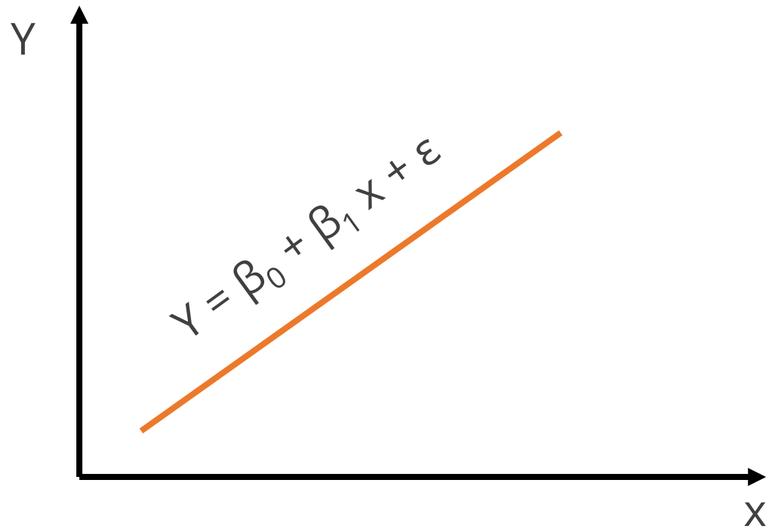
It depicts the relationship between one dependent and two or more independent variables.

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

EXAMPLE

Consider the following graph with the equation of the line as shown:



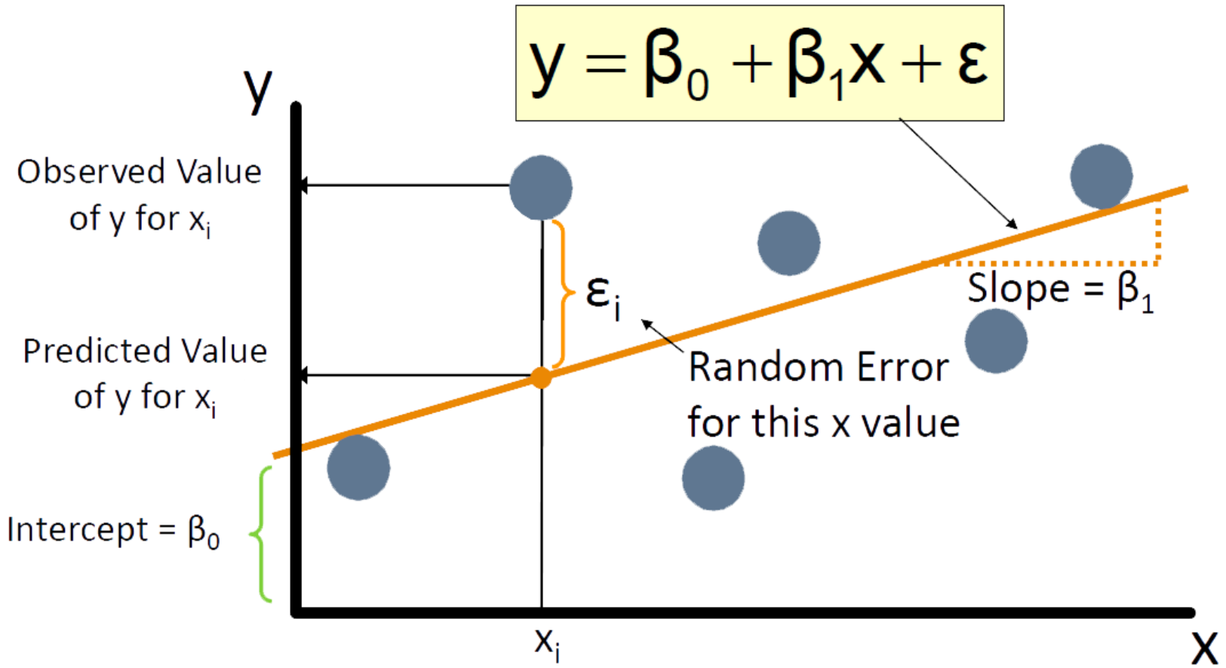
- β_1 represents the slope.
- β_1 represents the estimated change in the average value of y as a result of a one-unit change in x.
- β_0 represents the estimated average value of y when the value of x is zero.

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

EXAMPLE

The values are labeled as shown:



Demo

Perform Simple Linear Regression

Given the class dataset with 5 variables, i.e., name, sex, age, height, and weight representing the information for a class of students, predict the weight based on height.

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

Method of least squares identifies the line with the lowest total sum of squared prediction errors (Sum of Squares of Error or SSE).

Mathematically,

SSR (Sum of squared residuals) = $\sum (\hat{y} - \overline{y})^2$ (measure of an explained variation)

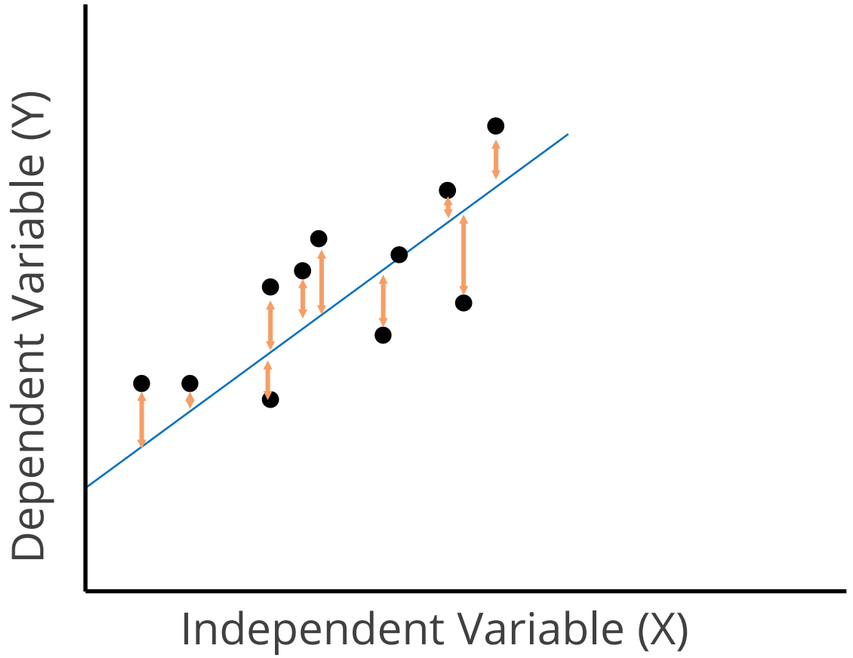
SSE (Sum of squared errors) = $\sum (y - \hat{y})$ (measure of an unexplained variation)

SST (Total sum of squares) = SSR + SSE = $\sum (y - \overline{y})^2$ (measure of the total variation in y)

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

EXAMPLE



Here, the observations (black) are assumed to be the result of random deviations (orange) from an underlying relationship (blue) between the dependent variable (*y*) and independent variable (*x*).

Types of Linear Regression

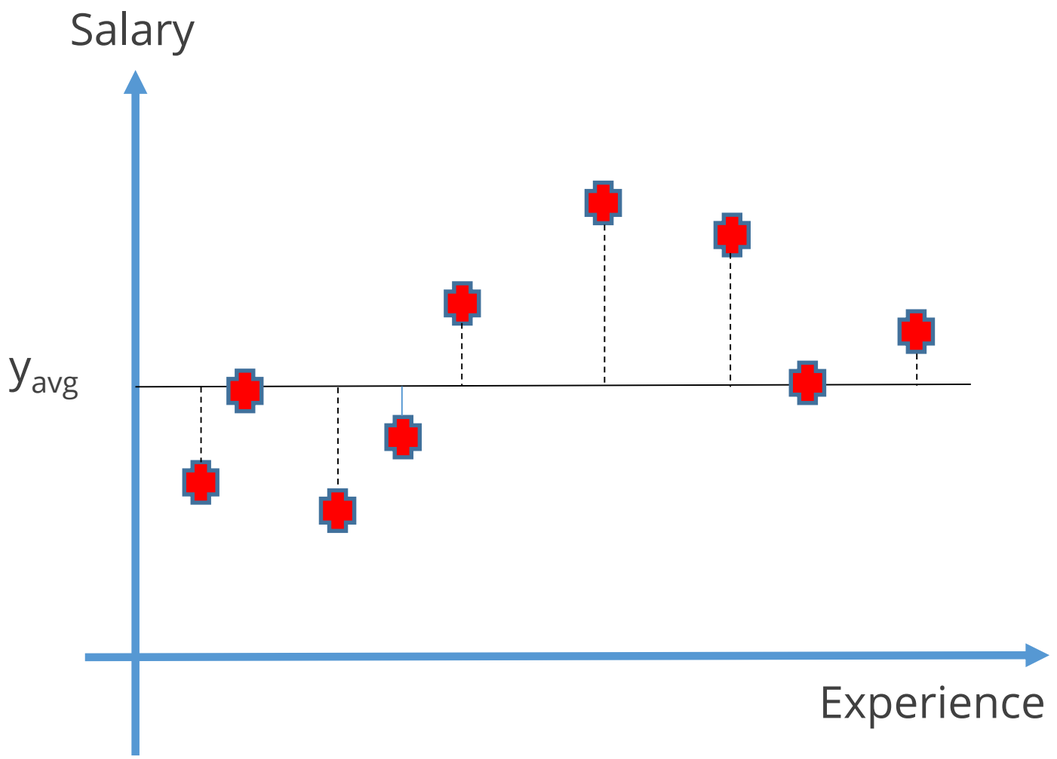
- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

Coefficient of multiple determination or R^2 is 'Goodness of Fit' of a regression model that is used to determine the total variance with the help of independent variables.

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

EXAMPLE



Here, the

$SSR = SUM(y_i - y_i(hat))^2$

$SST = SUM(y_i - y_{avg})^2$

$R^2 = 1 - SS_{res}/SS_{tot}$

Since, SS_{res} > Min ⇒ The value of R^2 will never decrease

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

ADJUSTED R^2

The R^2 of a model increases with an increase in the number of independent variables.

The adjusted R^2 penalizes the model, thereby increasing the variables that do not represent the model correctly.

$$R^2 = 1 - SSR/SST$$

$$\text{Adj } R^2 = 1 - (1-R^2)(n-1/n-p-1)$$

Where,
p = number of regressors
N = sample size

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

Standard error of a regression (Sy.x) is the measure of variability around the line of regression and can be used the same way as standard deviation (SD).

Standard Error = $\sqrt{(SSE/(n-k))}$
where,
n = Number of observations in the sample
k = Total number of variables in the model

Y ± 2 standard errors provide approximately 95% accuracy. However, 3 standard errors provide a 99% confidence interval.

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

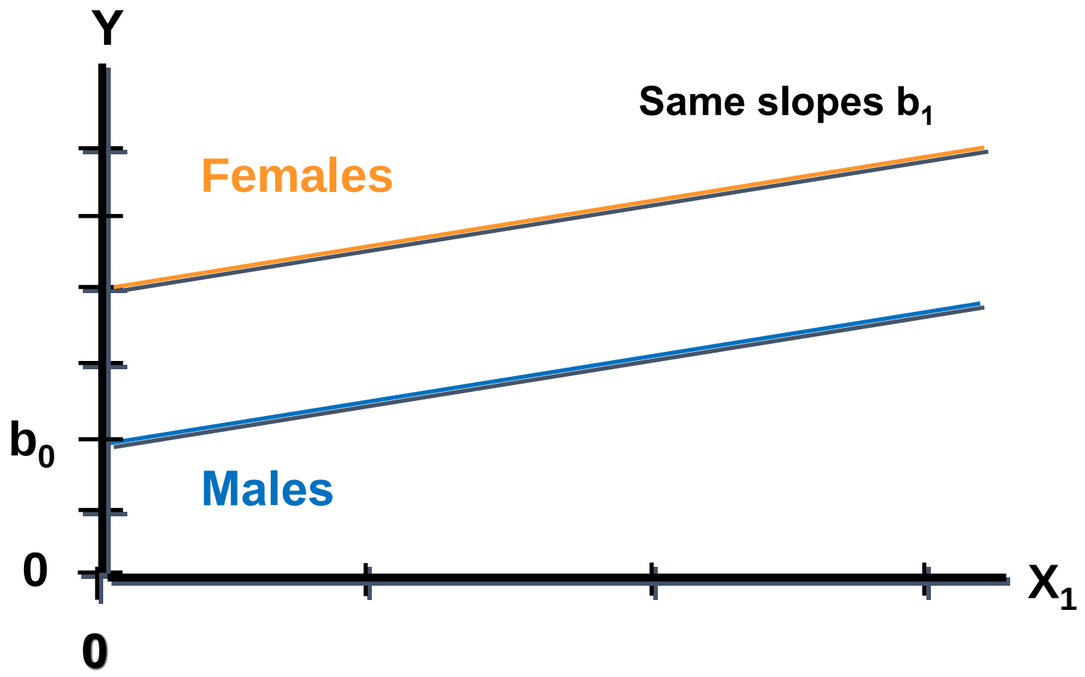
A dummy variable includes variables, each with two levels coded 0 and 1.

- Assumes that only the intercept is different
- Includes slopes that are constant across categories
- Permits the use of qualitative data
- Incorporates large residuals and influences measures

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

EXAMPLE



Types of Linear Regression

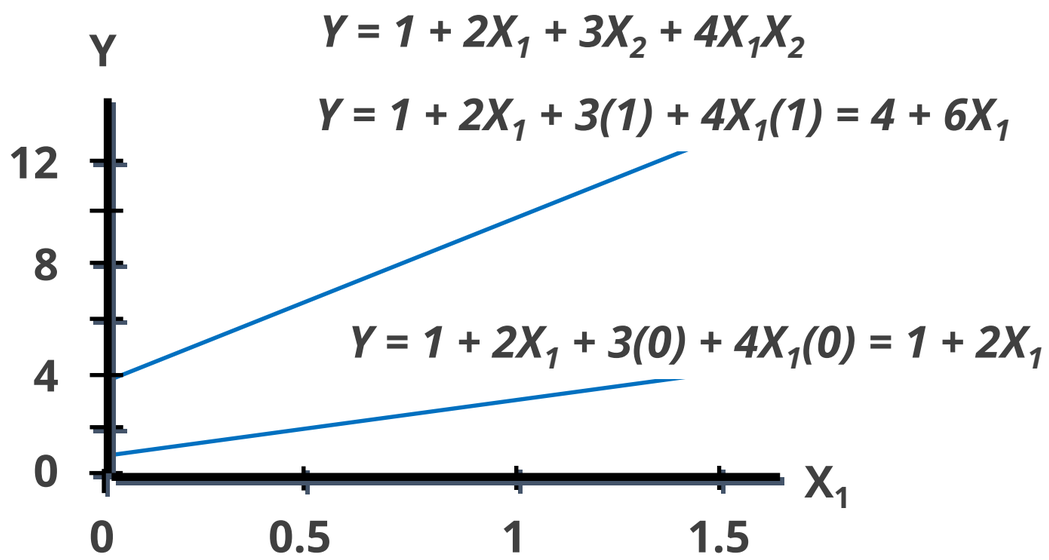
- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

- Assumes the interaction between pairs of X variables
- Includes two-way cross-product terms
- Can be combined with other models, such as a dummy variable regression model

Types of Linear Regression

- Simple Linear
- Method of Least Squares
- Coefficient of Multiple Determination
- Standard Error of the Estimate
- Dummy Variable
- Interaction

EXAMPLE



In $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$ without and with the interaction term, the effect of X_1 on Y is measured by 1 and $\beta_1 + \beta_3 X_2$, respectively.

Regression Analysis

Topic 4—Non-Linear Regression

What Is Non-Linear Regression?

Nonlinear regression is a form of regression analysis in which observational data is modeled by a function that is a nonlinear combination of the model parameters and depends on one or more independent variables.

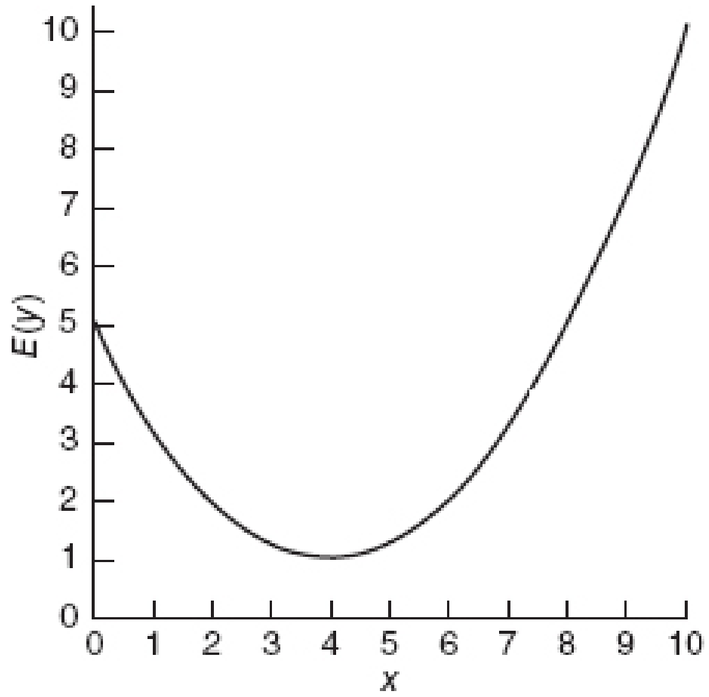
It is used when the number of predictors is large. Prior to fitting a regression model with all the predictors, stepwise or best subsets model-selection methods are used. This is done to screen out predictors that are not associated with the responses.

Types of Non-Linear Regression

- Polynomial
- Logarithmic
- Square Root
- Reciprocal
- Exponential

A Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an n^{th} degree polynomial in x.

Equation: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$

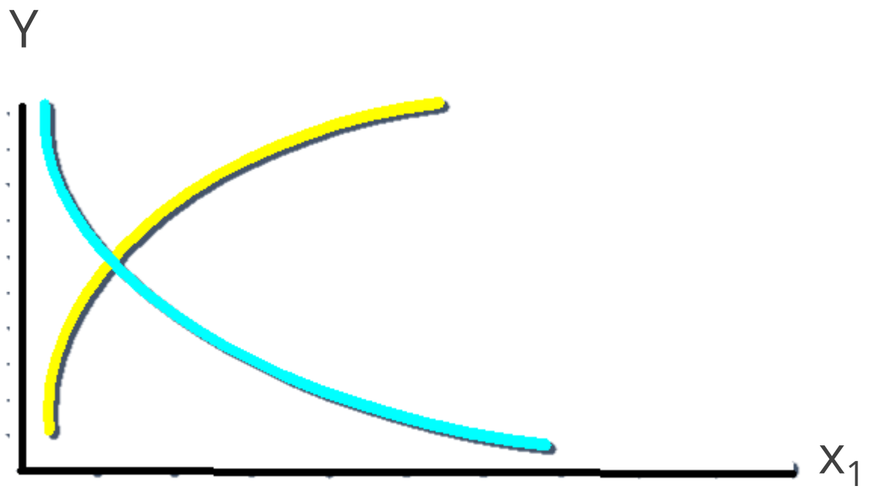


Types of Non-Linear Regression

- Polynomial
- Logarithmic
- Square Root
- Reciprocal
- Exponential

Logarithmic regression is used to model situations where growth or decay accelerates rapidly at first and then slows over time.

Equation: $Y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$

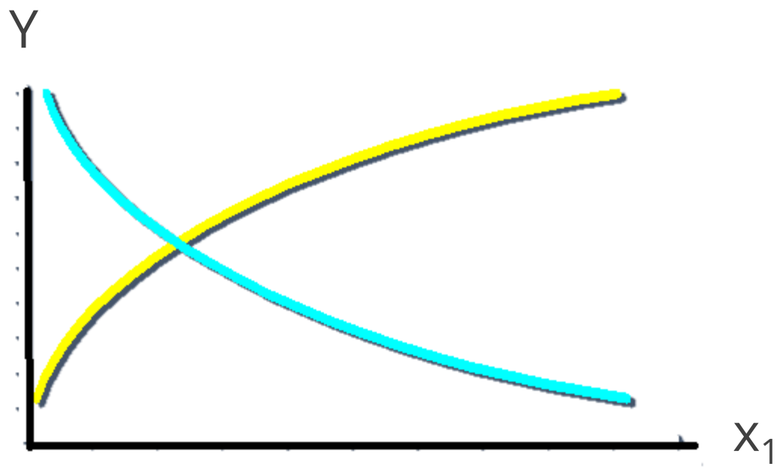


Types of Non-Linear Regression

- Polynomial
- Logarithmic
- Square Root
- Reciprocal
- Exponential

Square root regression is used to model situations to improve the distribution of the dependent variable.

Equation: $Y_i = \beta_0 + \beta_1\sqrt{X_{1i}} + \beta_2\sqrt{X_{2i}} + \varepsilon_i$

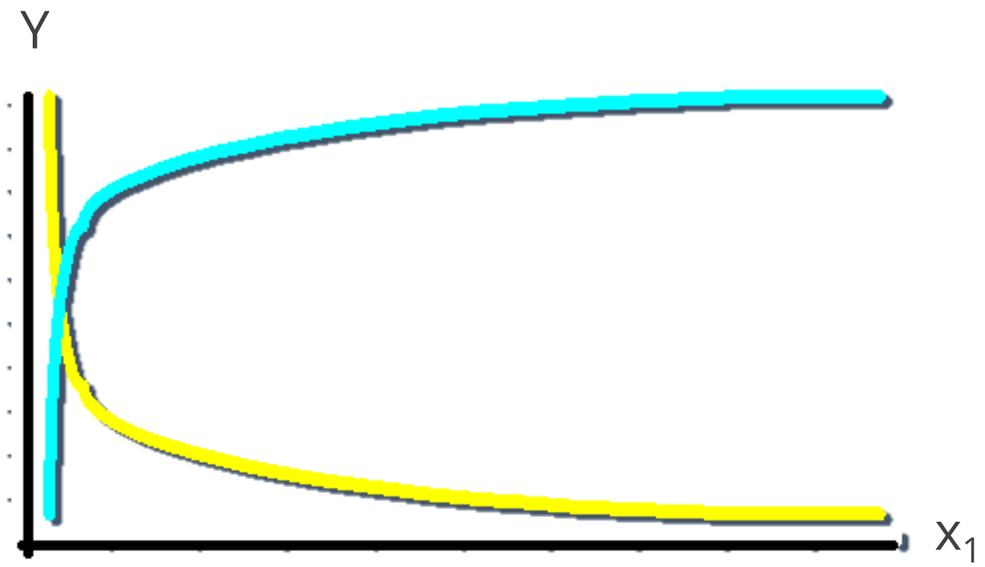


Types of Non-Linear Regression

- Polynomial
- Logarithmic
- Square Root
- **Reciprocal**
- Exponential

Reciprocal regression helps in making good predictions for a curved relationship.

Equation: $Y_i = \beta_0 + \beta_1 \frac{1}{X_{1i}} + \beta_2 \frac{1}{X_{2i}} + \varepsilon_i$

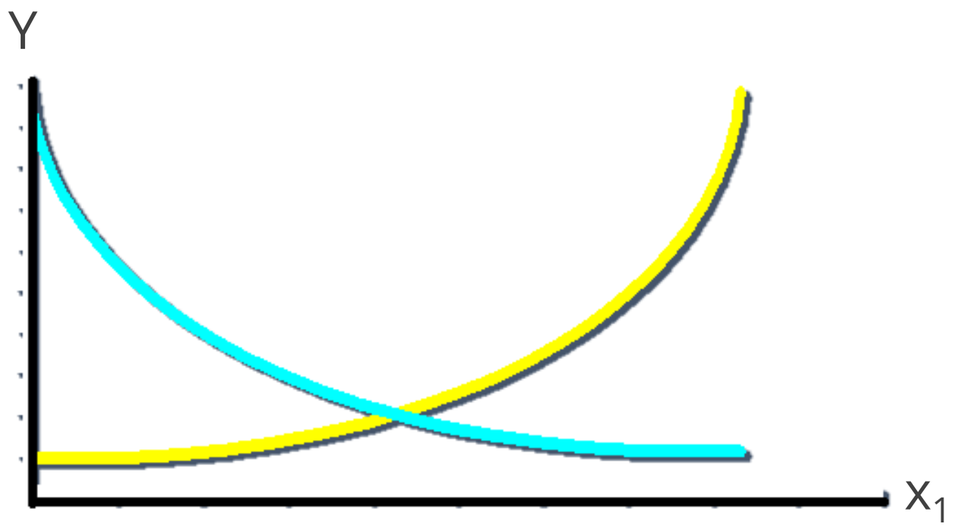


Types of Non-Linear Regression

- Polynomial
- Logarithmic
- Square Root
- Reciprocal
- Exponential

An exponential regression refers to the process of deriving the equation of the exponential function that fits best for a set of data.

Equation: $$Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i$$



Demo

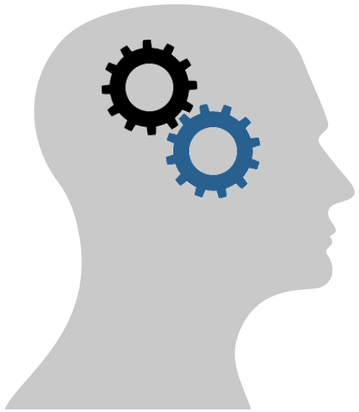
Perform Regression Analysis with Multiple Variables

The buyers of cars are interested in determining the factors that influence the car mileage

Boston dataset consists of variables to predict price of house. Analyze the factors that influence price of house.

Linear Regression Model

LIMITATIONS

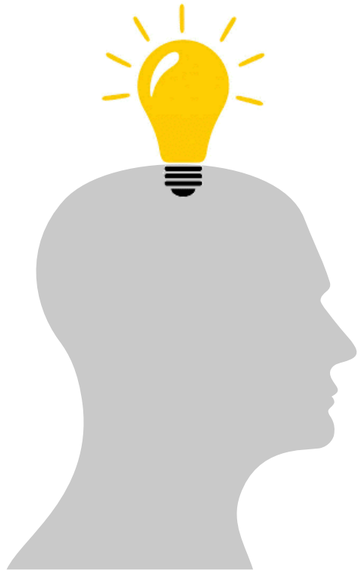


While building up the linear regression equation, the following challenges may occur:

- The sample of data analyzed for the regression model can be biased.
- The model developed may be a good fit for an existing dataset but not for new data.
- All data used to train may result in no data for testing.

Linear Regression Model

LIMITATIONS



To overcome such challenges, the concept of **cross-validation** is used.

Regression Analysis

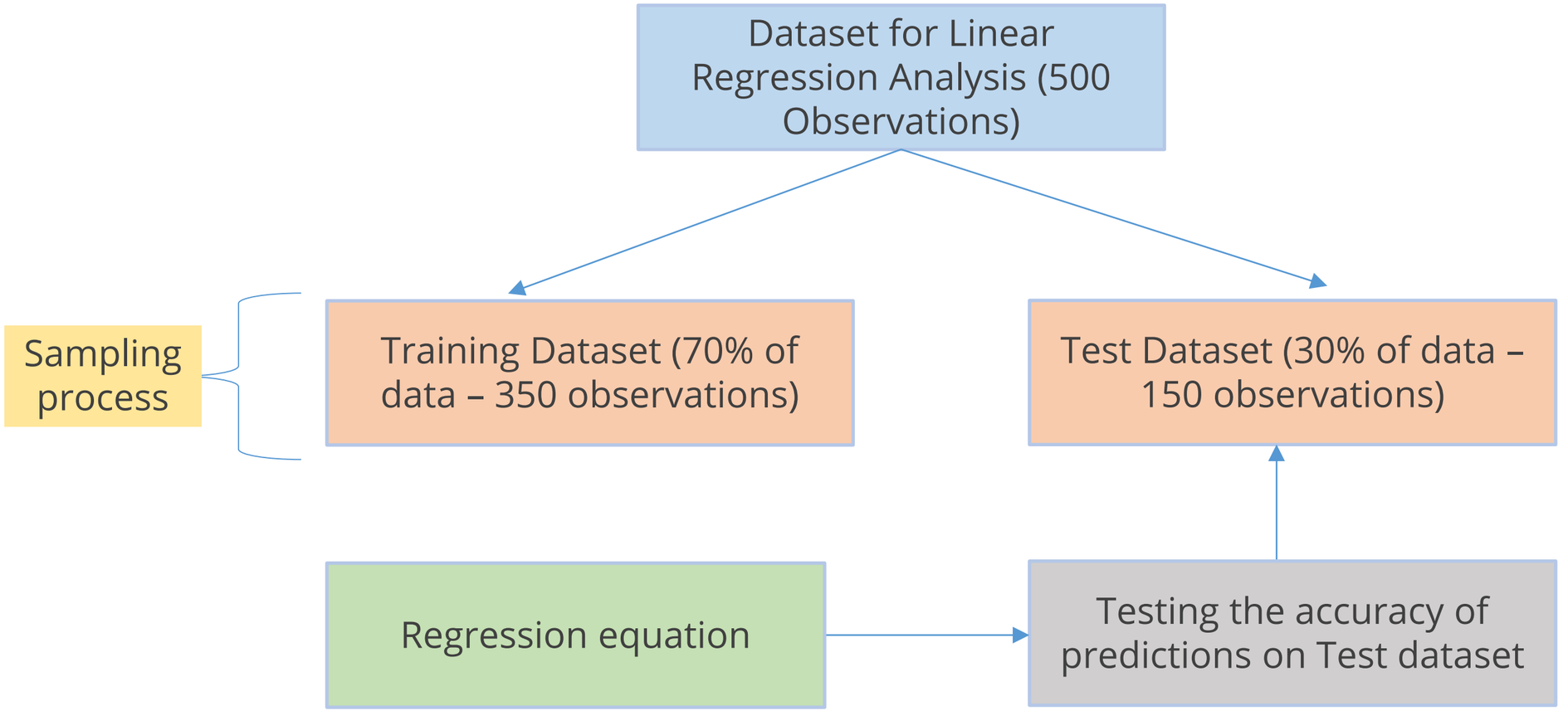
Topic 5—Cross Validation

What Is Cross Validation?

Cross Validation is a technique used to determine the accuracy in predicting models.

In cross validation, the dataset is split into 70:30 (Training data : Testing data) ratio.

Cross Validation: Example



K-fold cross validation is the most used cross validation. You will learn it in detail in the next lesson.

Regression Analysis

Topic 6—Non-Linear to Linear Models

Non- Linear Models to Linear Models

Non-linear models can be converted to linear models by applying the following functions:

- Exponential
- Logarithmic
- Trigonometric
- Power

Measures of Regression Models

Absolute Percentage Error (APE)	mae<-function(a,b){round(abs(a-b)),2)}
Mean Absolute Error (MAE)	mae<-function(a,b){round(mean(abs(a-b)),2)}
Mean Absolute Percentage Error/Deviation (MAPE/MAPD)	mape<-function(a,b){round(mean(abs(a-b)/a),4)}
Root Mean Squared Error (RMSE)	rmse<-function(c,d){round(sqrt(mean((c-d)^2)),2)}

Regression Analysis

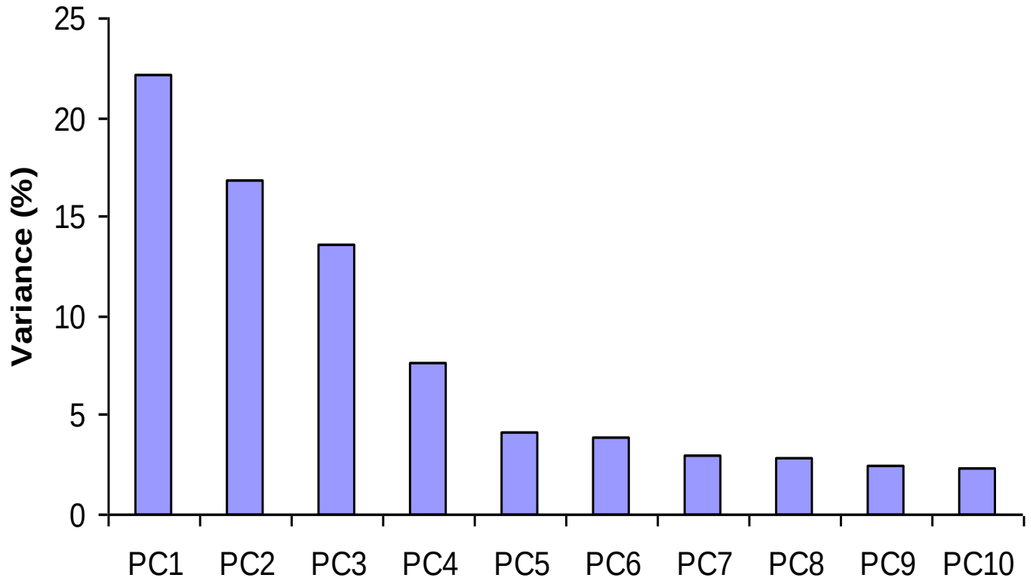
Topic 7—Principal Component Analysis (PCA)

Principal Components

Principal components are linear components of the original variables. They tend to capture as much variance as possible in a dataset.

Facts about principal components:

- Are summary variables
- Are formed by the linear combination of the original variables
- Are not correlated to each other
- Try to capture maximum variance within themselves



Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a process of extracting variables from a dataset to explain maximum variance in the dataset.

From “n” independent variables in a dataset, PCA extracts “k” new variables that explain the most variance in the dataset.

Principal Component Analysis (PCA)

USES

- It is used to eliminate the duplicate variables in cases where many variables are present in the dataset, to avoid redundancy.
- Since dependent variable is not considered, this model can be categorized as an unsupervised model.

Regression Analysis

Topic 8—Factor Analysis of Dimensionality Reduction

Factor Analysis

Factor analysis is a commonly used technique to find latent variables or factors in a model.

It is also considered a dimensionality reduction technique.

Demo

PCA and Factor Analysis for Dimensionality Reduction

The basics of PCA and factor analysis in R.

Key Takeaways

- Regression analysis is used to estimate the relationship between variables.
- Simple regression considers one quantitative and independent variable X to predict the other quantitative, but dependent, variable Y.
- Multiple regression considers more than one quantitative and qualitative variable ($X_1 \ldots X_N$) to predict a quantitative and dependent variable Y.
- R squared and adjusted R squared are important measures of a regression model and explain the variance with the help of independent variables.
- Factor analysis and principal component analysis are the methods used to decrease the number of variables or factors in a model.
- Multiple regression has two types of models; linear and non-linear

Quiz

QUIZ

1

What is the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) called?

a. Explanatory variable

b. Coefficient

c. Residual

d. Target variable

QUIZ

1

What is the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) called?

a. Explanatory variable

b. Coefficient

c. Residual

d. Target variable

The correct answer is **c.**

The difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the residual.

QUIZ

2

Which of the following statements is true about the Method of Least Squares Regression Model? Select all that apply.

a. It selects the line with the mean total sum of squared prediction errors

b. It is a linear regression model.

c. It selects the line with the lowest total sum of squared prediction errors.

d. It determines the relation between X and Y, when you reject H0.

QUIZ

2

Which of the following statements is true about the Method of Least Squares Regression Model? Select all that apply.

a. It selects the line with the mean total sum of squared prediction errors

b. It is a linear regression model.

c. It selects the line with the lowest total sum of squared prediction errors.

d. It determines the relation between X and Y, when you reject H0.

The correct answer is **b and c.**

Method of Least Squares Regression Model is a linear regression model that selects the line with the lowest total sum of squared prediction errors.

This concludes “Regression Analysis.”

The next lesson is “Classification.”