Predictive Modeling
Correlation and Linear Regression

Objectives

01 Understanding Correlation and its use

02 From correlation to regression

03 Simple Linear Regression

04 Least Squares Method

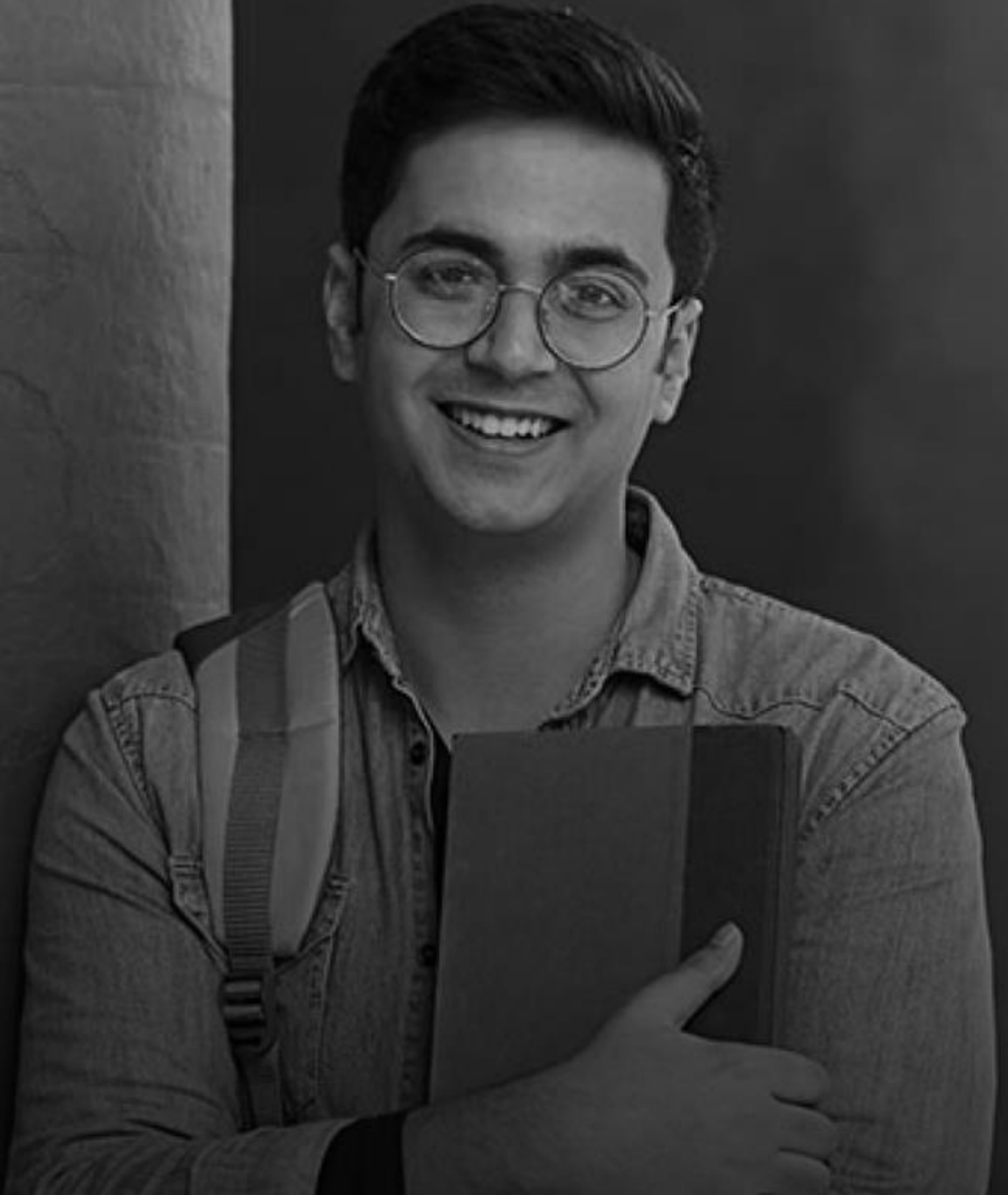
Objectives

05 | Goodness of fit

06 | Assumptions of SLR

07 | Model Diagnostics

Correlation



What is need of correlation?

- Is there any association between hours of study and grades?
- Is there a relationship between the price of a Big Mac and the net hourly wages of workers around the world? If so, how strong is
- the relationship?
- What happens to sweater sales with increase in temperature? What is the strength of association between them?
- How to quantify this association?
- Which of the above examples has very strong association?

Correlation coefficient

- It is a measure of linear association between two variables.

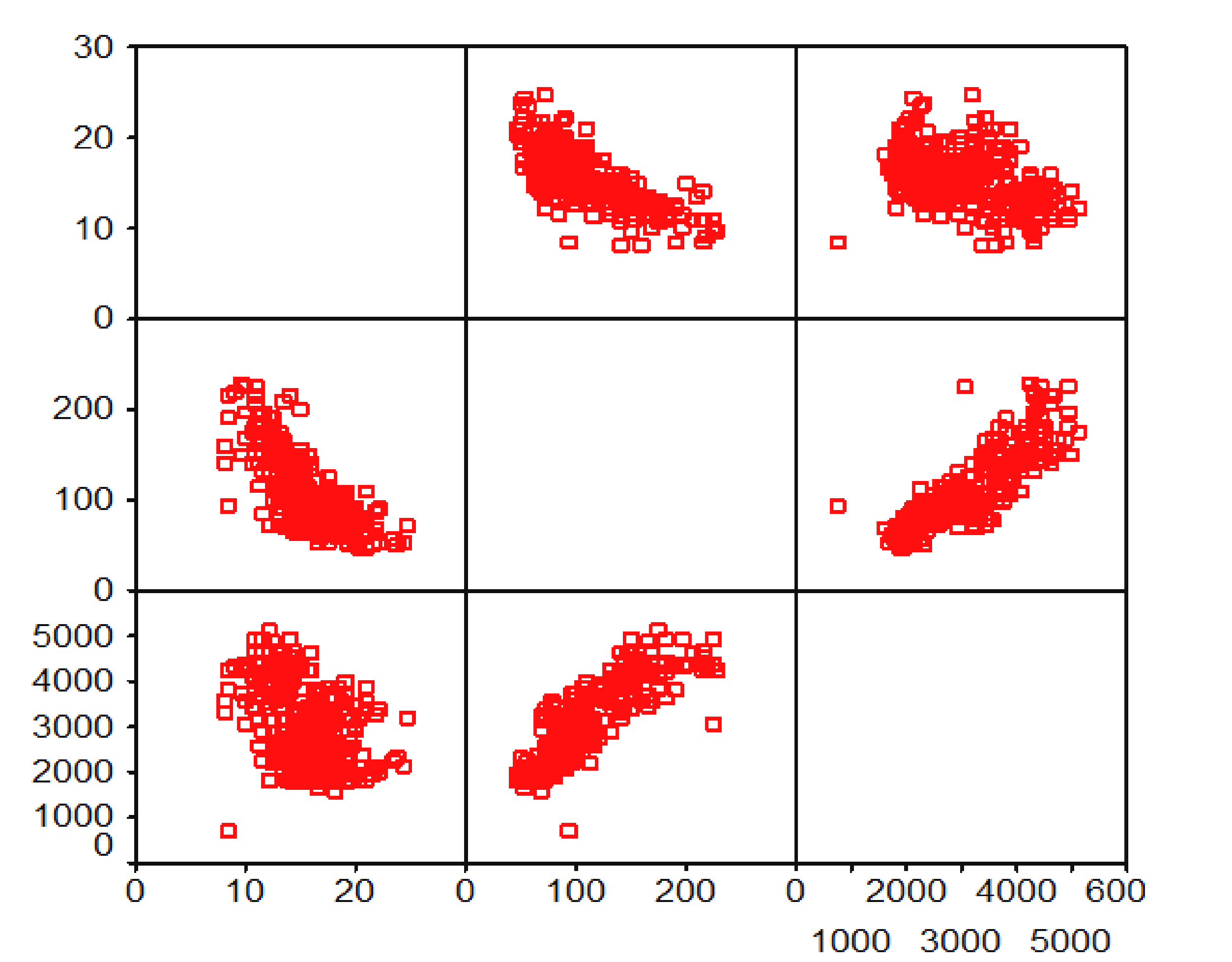
- Methods of studying correlation:
    - Scatter Diagram Method
    - Karl Pearson’s Correlation Coefficient
    - Rank Correlation Method

Correlation coefficient

- Correlation coefficient varies between -1 to +1
- Correlation 0 : No linear relationship
- Correlation 0 to 0.25 : Negligible positive relationship
- Correlation 0.25-0.5 : Weak positive relationship
- Correlation 0.5-0.75 : Moderate positive relationship
- Correlation > 0.75 : Very Strong positive relationship

- Similarly, we can conclude about negative correlation coefficients.

Method of Studying Correlation – Scatter Diagram Method



Business Case:

Suppose you work for an Airline company and your boss asks you, is it necessary to spend money on Promotional Budget to attract Passengers?

From Correlation to Regression

- In the above example, Promotional Budget and number of Passengers are highly correlated.
- Can we estimate number of Passengers given the Promotional Budget?

From Correlation to Regression

- Correlation is just a measure of linear association. It can’t be used for prediction.
- Given the predictor variable, we can’t estimate the dependent variable.
- In the Air Passengers example, given the Promotional Budget, we can’t get an estimated value of Passengers.
- We need a model, an equation, a fit for the data: that is known as regression line.

What is Regression

What is Regression?

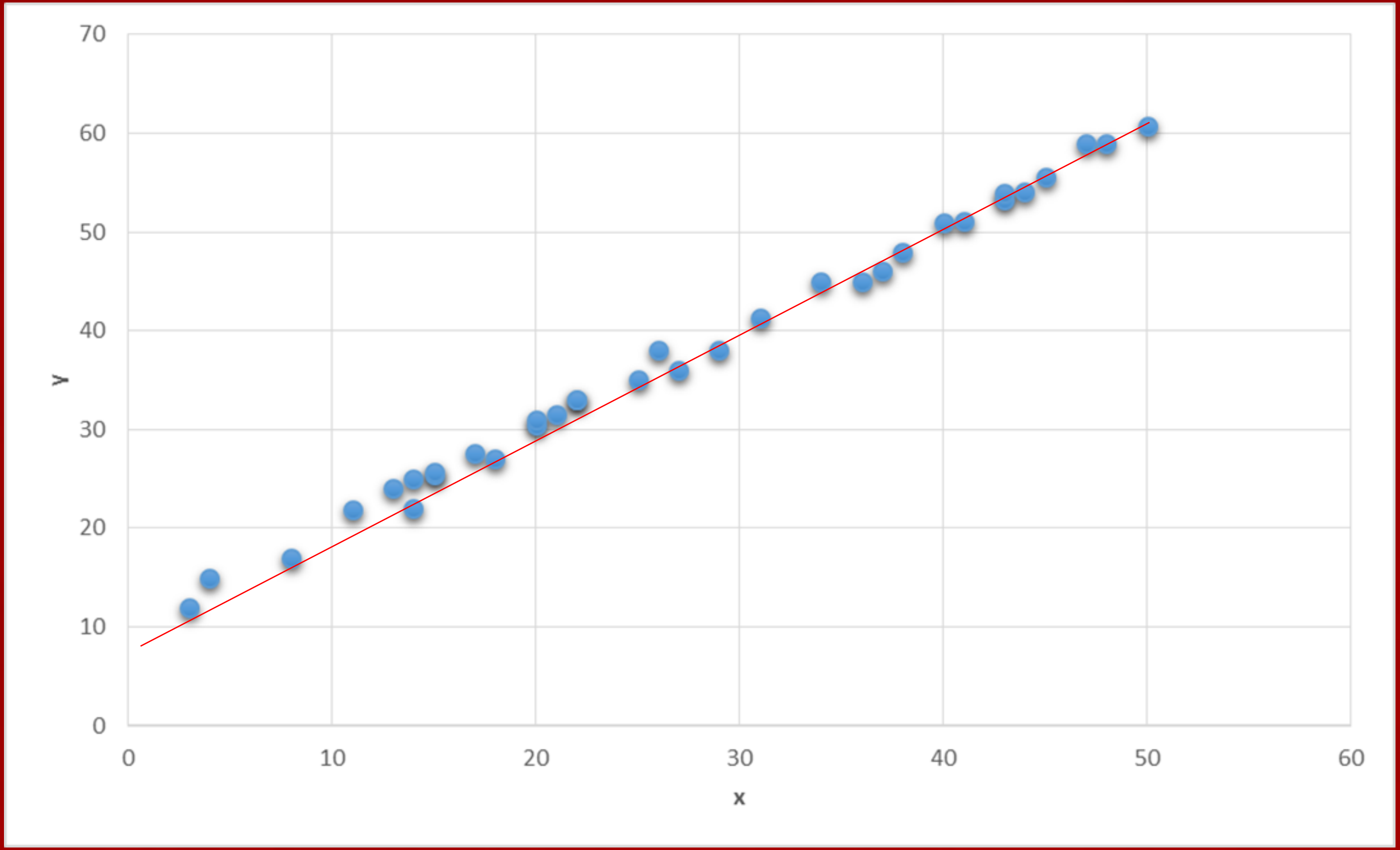
- A regression line is a mathematical formula that quantifies the general relationship between a predictor/independent (or known) variable x and the target/dependent (or the unknown) variable y.

- Below is the regression line. If we have the data of x and y then we can build a model to generalize their relation.

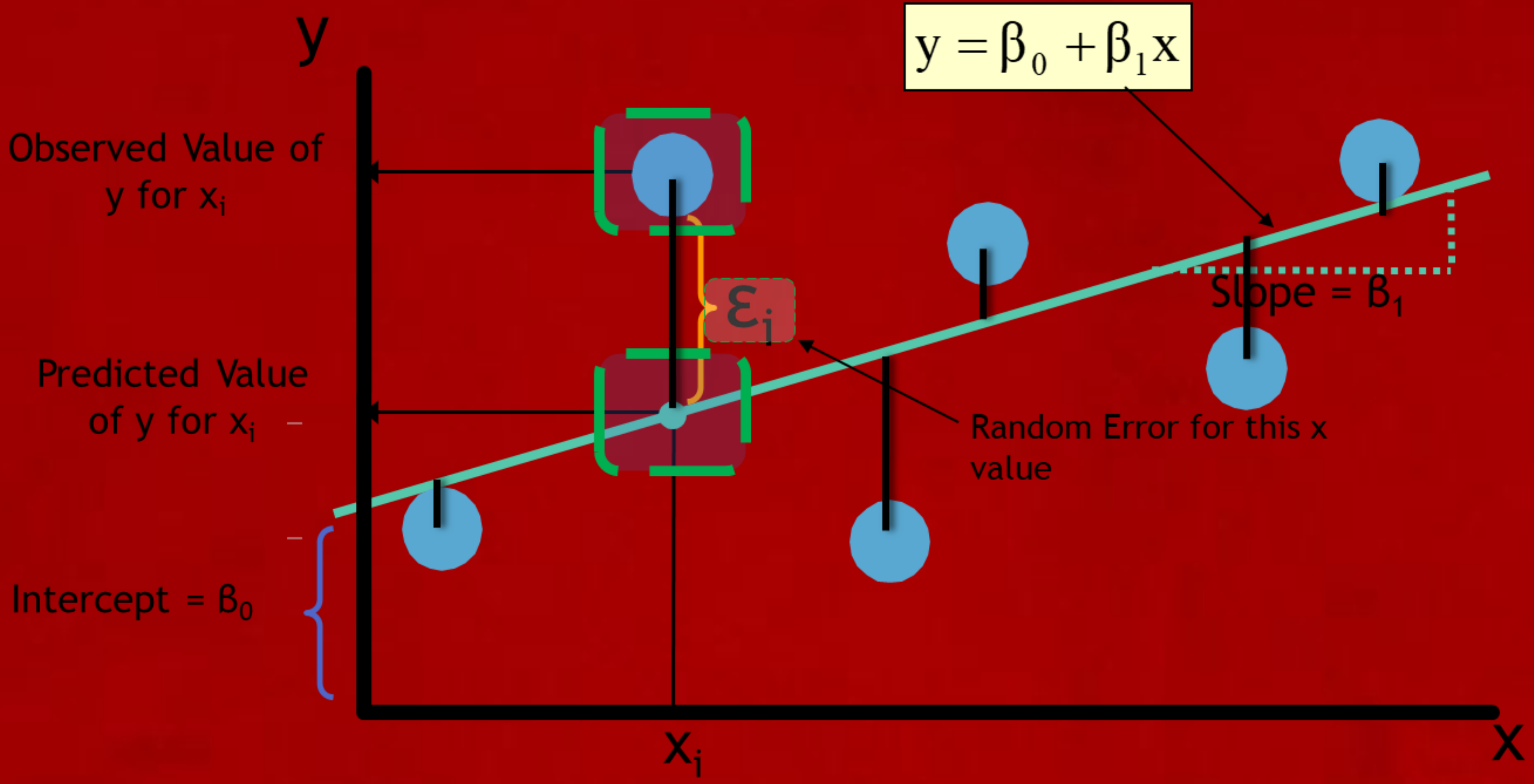
$$\boxed{y = \beta_0 + \beta_1 x}$$

- When it contains a Single predictor variable, it is called as Simple Linear Regression.
- And when it contains more than one predictor variable, it is called as Multiple Linear Regression.

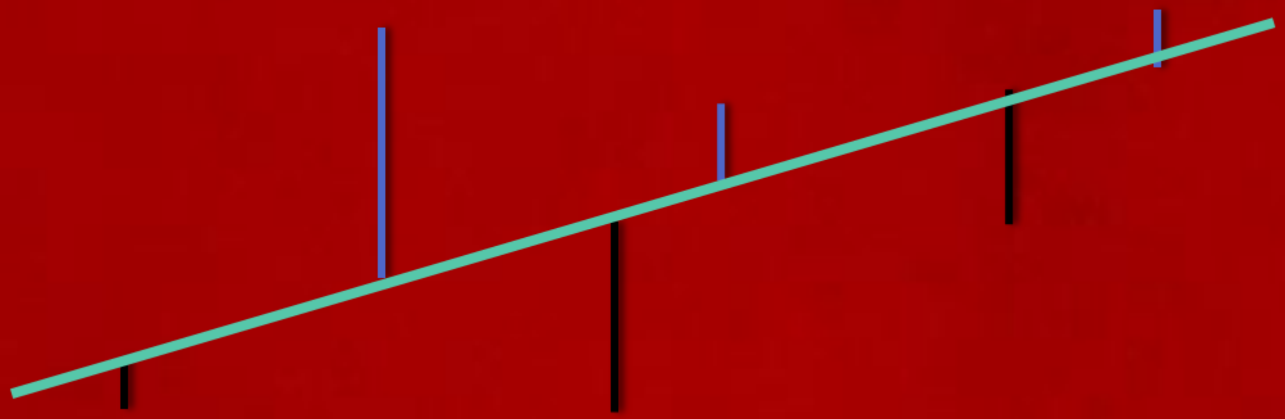
Regression



Regression Line Fitting

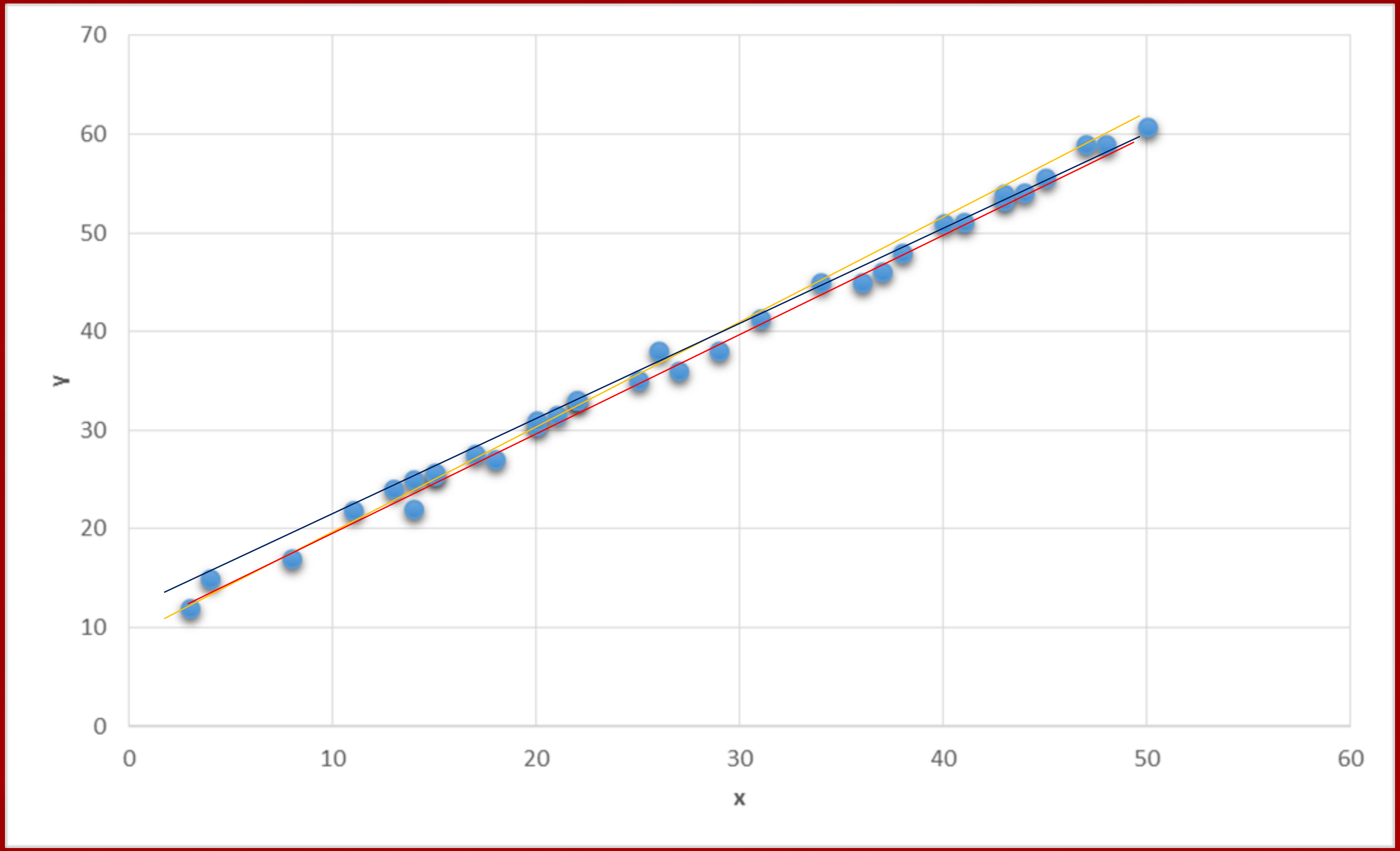

y
$y = \beta_0 + \beta_1 x$
Observed Value of y for x_i
ε_i
Slope = B_1
Predicted Value of y for x_i
Random Error for this x value
Intercept = B_0
x_i
X

Regression Line Fitting

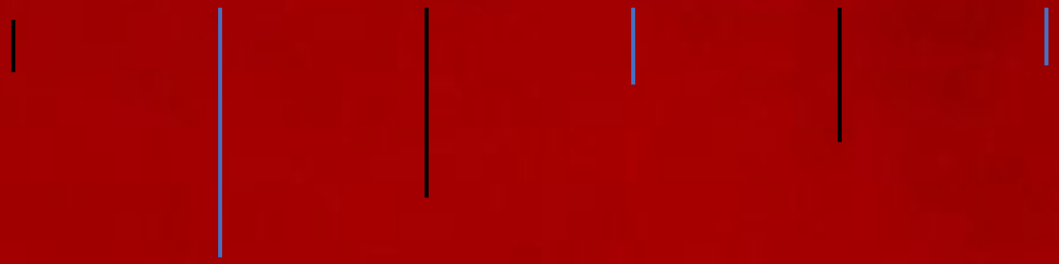


Regression Line Fitting

What is the best fit for our data ?



Minimizing the Error



- The best line will have the minimum error.
- Some errors are positive, and some errors are negative. Taking their sum is not a good idea
- We can either minimize the squared sum of errors Or we can minimize the absolute sum of errors.
- Squared sum of errors is mathematically convenient to minimize.
- The method of minimizing squared sum of errors is called Least Squared Method of Regression.

Least Squares Estimation

- X: x1, x2, x3, x4, x5, x6, x7,……..
- Y:y1, y2, y3, y4, y5, y6, y7…….
- Imagine a line through all the points
- Deviation from each point (residual or error)
- Square of the deviation
- Minimizing sum of squares of deviation

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (\beta_0 + \beta_1 x))^2\end{aligned}$$

beta$_0$ and beta$_1$ are obtained by minimizing the sum of the squared residuals

Assumption of Simple Linear Regression

- There should be linear relationship between Target and Predictor variable
- The error term must have constant variance
- There should be no correlation between the residual (error) terms
- The error term must be normally distributed (with mean zero)

How good is my regression line?

- Take an (x,y) point from data.
- Imagine that we submitted x in the regression line, we got a prediction as y_{pred}
- If the regression line is a good fit, then the we expect y_{pred} = y or $(y-y_{pred})$ = 0
- At every point of x, if we repeat the same, then we will get multiple error values $(y-y_{pred})$ values
- Some of them might be positive, some of them may be negative, so we can take the square of all such errors

$$\boxed{SSE = \sum (y - \hat{y})^2}$$

How good is my regression line?

- For a good model we need SSE to be zero or near to zero

- Standalone SSE will not make any sense. For example, SSE= 100, is very less when y is varying in terms of 1000’s. Same value is very high when y is varying in terms of decimals.

- We must consider variance of y while calculating the regression line accuracy

How good is my regression line?

- Error Sum of squares (SSE: Sum of Squares of error)
  - $SSE = \sum(y - \hat{y})^2$

- Total Variance in Y (SST: Sum of Squares of Total)
  - $SST = \sum(y - \overline{y})^2$
  - $SST = \sum(y - \hat{y} + \hat{y} - \overline{y})^2$
  - $SST = \sum(y - \hat{y})^2 + \sum(\hat{y} - \bar{y})^2$ {using assumptions of linear regression}
  - $\mathrm{SST} = \mathrm{SSE} + \sum(\hat{y} - \bar{y})^2$
  - $\mathrm{SST} = \mathrm{SSE} + SSR$

How good is my regression line?

- Total variance in Y is divided into two parts,
  - Variance that can’t be explained by x (error)
  - Variance that can be explained by x, using regression

$$\mathrm{SST} = \mathrm{SSE} + SSR$$

How good is my regression line?

So, total variance in Y is divided into two parts,

- Variance that can be explained by x, using regression
- Variance that can't be explained by x

$$SST = SSE + SSR$$

- Total sum of Squares

Sum of Squares Error

Sum of Squares Regression

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

$$SSR = \sum (\hat{y} - \bar{y})^2$$

R-Squared

R-Squared

A good fit will have

- SSE (Minimum or Maximum?)
- SSR (Minimum or Maximum?)
- And we know SST= SSE + SSR
- SSE/SST(Minimum or Maximum?)
- SSR/SST(Minimum or Maximum?)

R-Squared

A good fit will have

- SSE (Minimum or Maximum?)
- SSR (Minimum or Maximum?)
- And we know SST= SSE + SSR
- SSE/SST(Minimum or Maximum?)
- SSR/SST(Minimum or Maximum?)

The Coefficient of Determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

The coefficient of determination is also called R-squared and is denoted as R^2

where,

$$\boxed{R^2 = \frac{SSR}{SST}}$$

$$\boxed{0 \leq R^2 \leq 1}$$

R-Squared

Here are some basic characteristics of the measure:

- Since *R-squared* is a proportion, it is always a number between 0 and 1.
- If *R-squared* = 1, all the data points fall perfectly on the regression line. The predictor *x* accounts for *all* the variation in *y*!
- If *R-squared* = 0, the estimated regression line is perfectly horizontal. The predictor *x* accounts for *none* of the variation in *y*!

Model Diagnostics

Model Diagnostics

- After fitting a regression model, it is important to determine whether all the necessary model assumptions are valid before performing inference.
- If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions. Therefore, it is crucial to perform appropriate model diagnostics.
- In constructing our regression model, we assumed that the response y to the explanatory variables were linear in the parameters and that the errors were independent and identically distributed (i.i.d) normal random variables with mean 0 and constant variance.
- Model diagnostic procedures involve both graphical methods and formal statistical tests. These procedures allow us to explore whether the assumptions of the regression model are valid and decide whether we can trust subsequent inference results.

Model Diagnostics

Studying the Variables:

- There are several graphical methods appropriate for studying the behavior of the explanatory variables. We are primarily concerned with determining the range and concentration of the values of the X variables and whether there exist any outliers.
- Graphical procedures include histograms, boxplots and sequence plots.

Model Diagnostics

Residual Analysis:

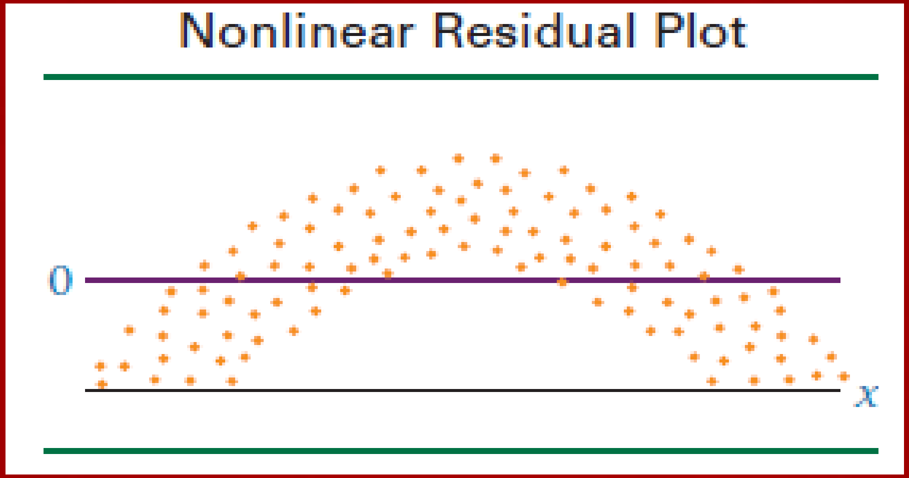
It uses Residual Plot in which Residuals are usually plotted against the *x*-variable (for SLR), which reveals a view of the residuals as *x* increases.

We can use residuals to study whether:

- The regression function is nonlinear.
- The error terms have non constant variance.
- The error terms are not independent.
- There are outliers.
- The error terms are not normally distributed.

Model Diagnostics

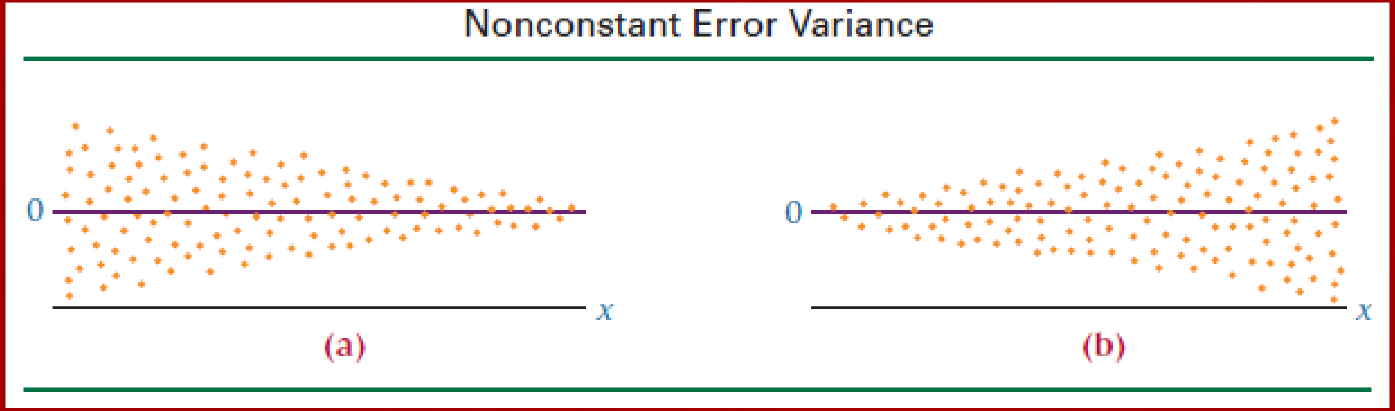
- If a residual plot such as the one in Figure attached appears, the assumption that the model is linear does not hold.



- Note that the residuals are negative for low and high values of *x* and are positive for middle values of *x*. The graph of these residuals is parabolic, not linear.
- Any significant deviation from an approximately linear residual plot may mean that a nonlinear relationship exists between the two variables.

Model Diagnostics

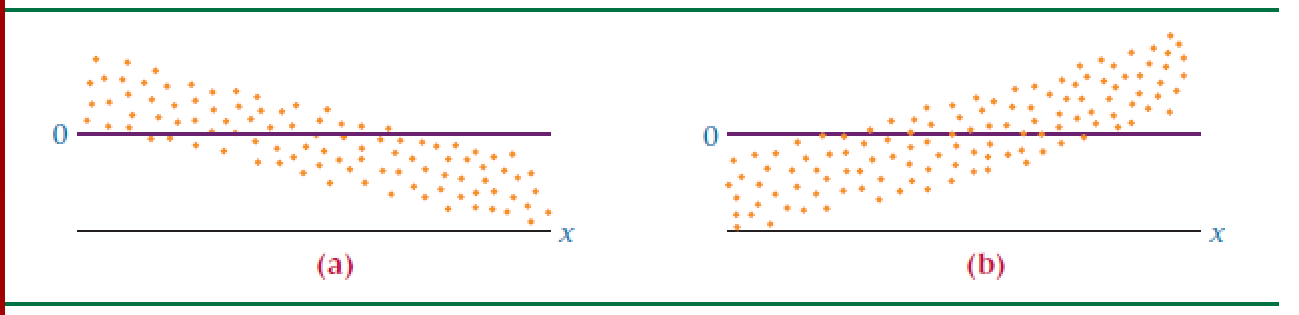
- The assumption of *constant error variance* sometimes is called **homoscedasticity**.
- If *the error variances are not constant* (called **heteroscedasticity**), the residual plots might look like one of the two plots in Figure below.
- Note in Figure (a) that the error variance is greater for small values of *x* and smaller for large values of *x*. The situation is reversed in Figure (b).



Nonconstant Error Variance

Model Diagnostics

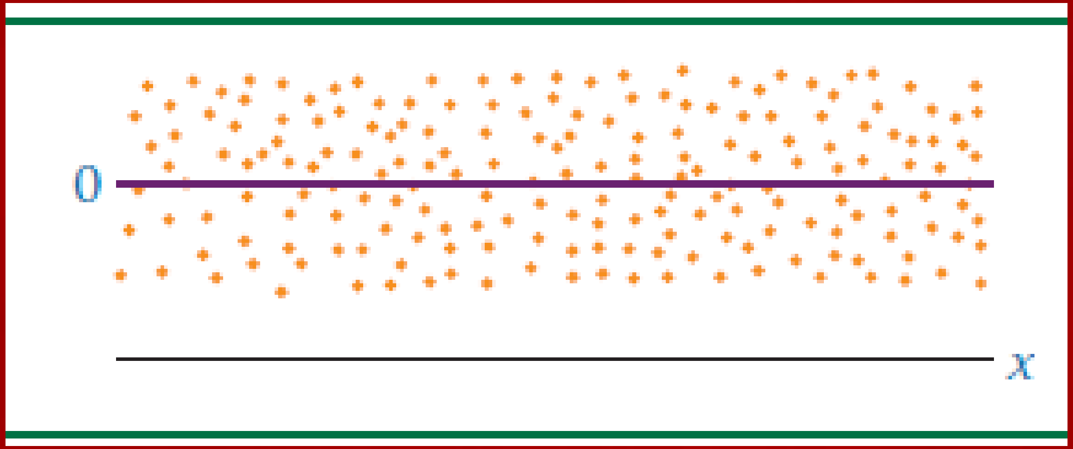
- If the error terms are not independent, the residual plots could look like one of the graphs given below.



- According to these graphs, instead of each error term being independent of the one next to it, the value of the residual is a function of the residual value next to it. For example, a large positive residual is next to a large positive residual and a small negative residual is next to a small negative residual

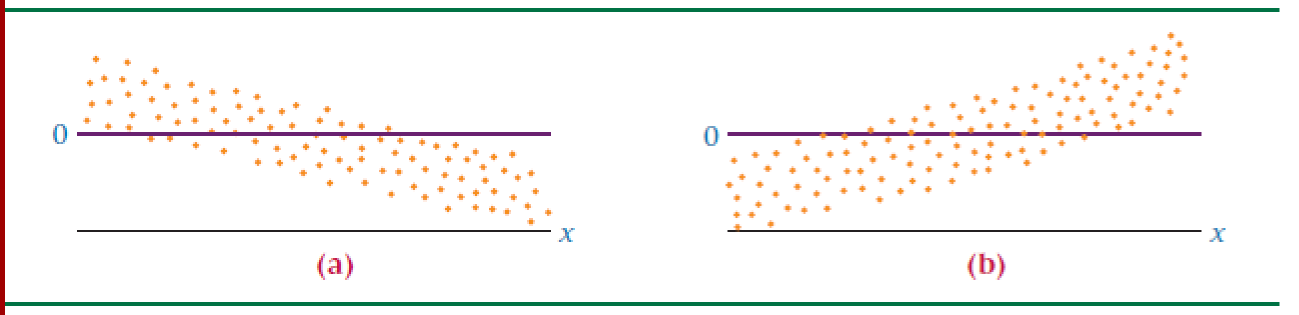
Model Diagnostics

- The graph of the residuals from a regression analysis that meets the assumptions—a *healthy residual graph*—might look like the graph as given in figure below.
- The plot is relatively linear; the variances of the errors are about equal for each value of *x*, and the error terms do not appear to be related to adjacent terms.



Model Diagnostics

- If the error terms are not independent, the residual plots could look like one of the graphs given below.



- According to these graphs, instead of each error term being independent of the one next to it, the value of the residual is a function of the residual value next to it. For example, a large positive residual is next to a large positive residual and a small negative residual is next to a small negative residual

Model Diagnostics

For validating normality condition:

we can use Q-Q plot and/or any of the following non-parametric tests:

- Kolmogorov Smirnov Test
- Shapiro Wilk Test
- Anderson Darling Test

Model Diagnostics

For validating normality condition:

we can use Q-Q plot and/or any of the following non-parametric tests:

- Kolmogorov Smirnov Test
- Shapiro Wilk Test
- Anderson Darling Test

Thank You

Hero
Vired