INNOMATICS RESEARCH LABS

Simple Linear Regression

Contents

- Measures of Association
  - Covariance
  - Correlation
- Correlation Properties
- Regression Analysis
- Ordinary Least Squares method (OLS)
- Assumptions of Regression Analysis
- Output interpretation

Covariance

- **Covariance** is a measure of association between two random variables.
- Measures the linear relationship between two variables.
- Covariance can be negative or positive or zero.
- Formula for Covariance

$$\mathrm{Cov}\ (X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n} \qquad \mathrm{Cov}\ (X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n\ - 1}$$

Covariance

- A positive Covariance value → the two variables tend to vary in the same direction (i.e. if one increases, then the other one increases too).
- A negative value → they vary in opposite directions (i.e. if one increases, then the other one decreases).
- Zero means that they don't vary together.

- **Limitations**
  - measures the directional relationship between two variables.
  - does not show the strength of the relationship between them.
  - Covariance values are not standardized.

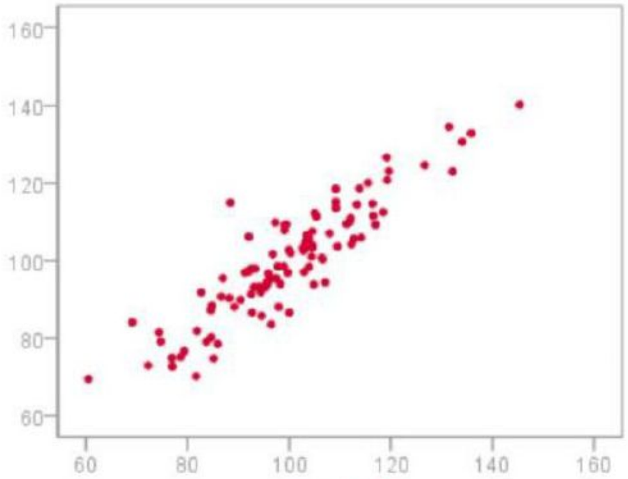
Correlation

- A correlation coefficient measures the extent to which two variables tend to change together. The coefficient describes both the **strength** and the **direction** of the relationship.
- It is considered to be the normalised version of the Covariance.

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

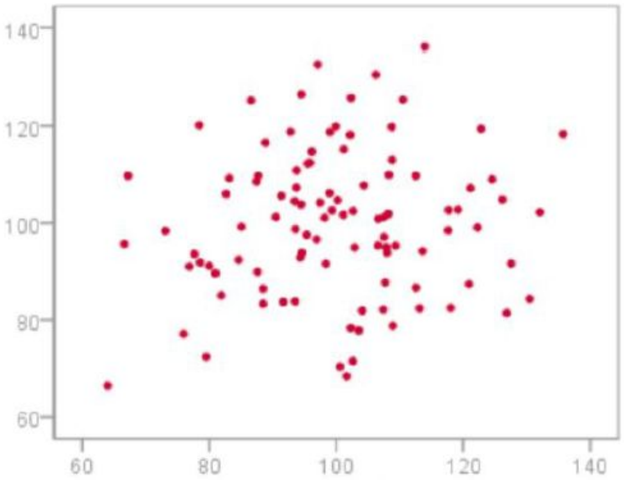
- The correlation is bounded between **-1 and 1**.

Correlation



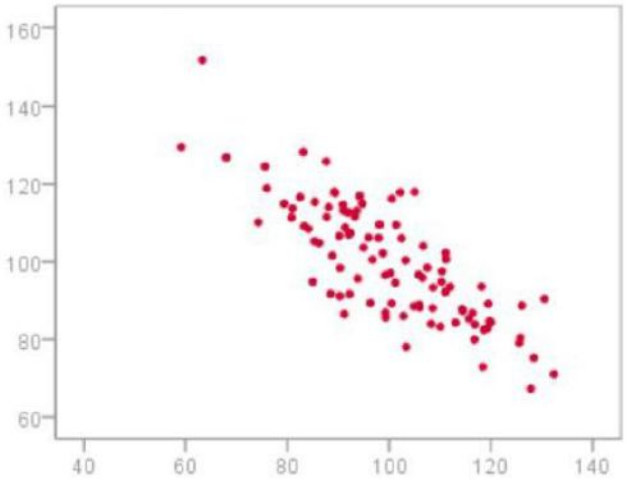
r = 0.9

Positve Correlation



r = 0.1

No Correlation



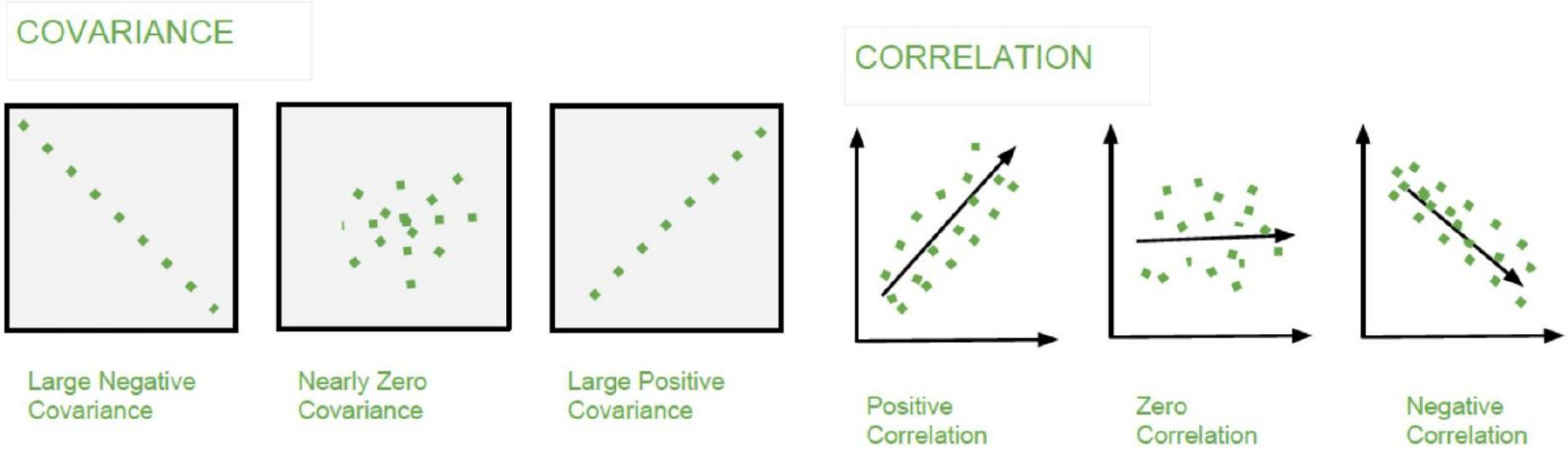
r = -0.8

Negative Correlation

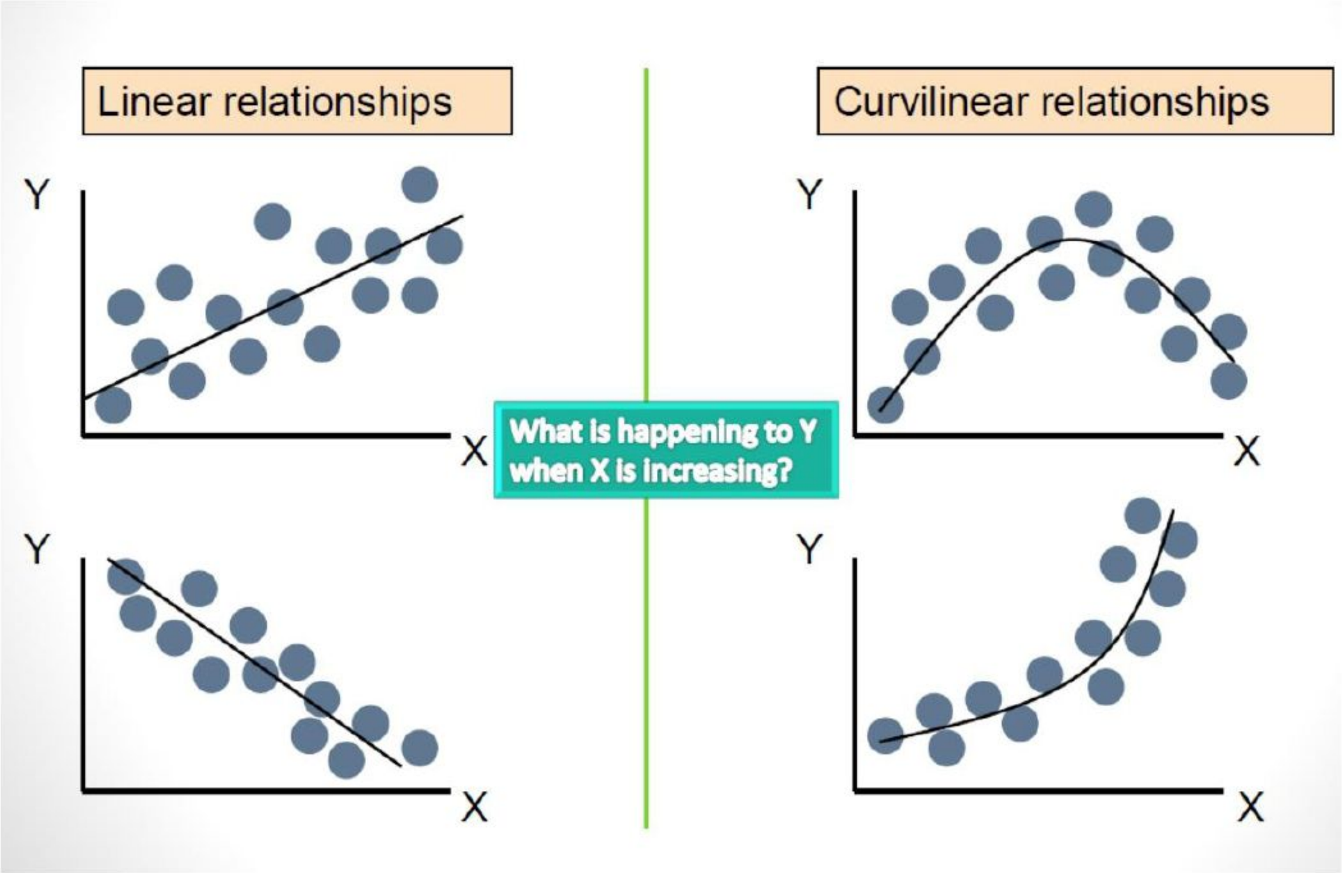
Limitation

Correlation is not and cannot be taken to imply causation. Even if there is a very strong association between two variables we cannot assume that one causes the other.

Covariance vs Correlation



Types of Relationships

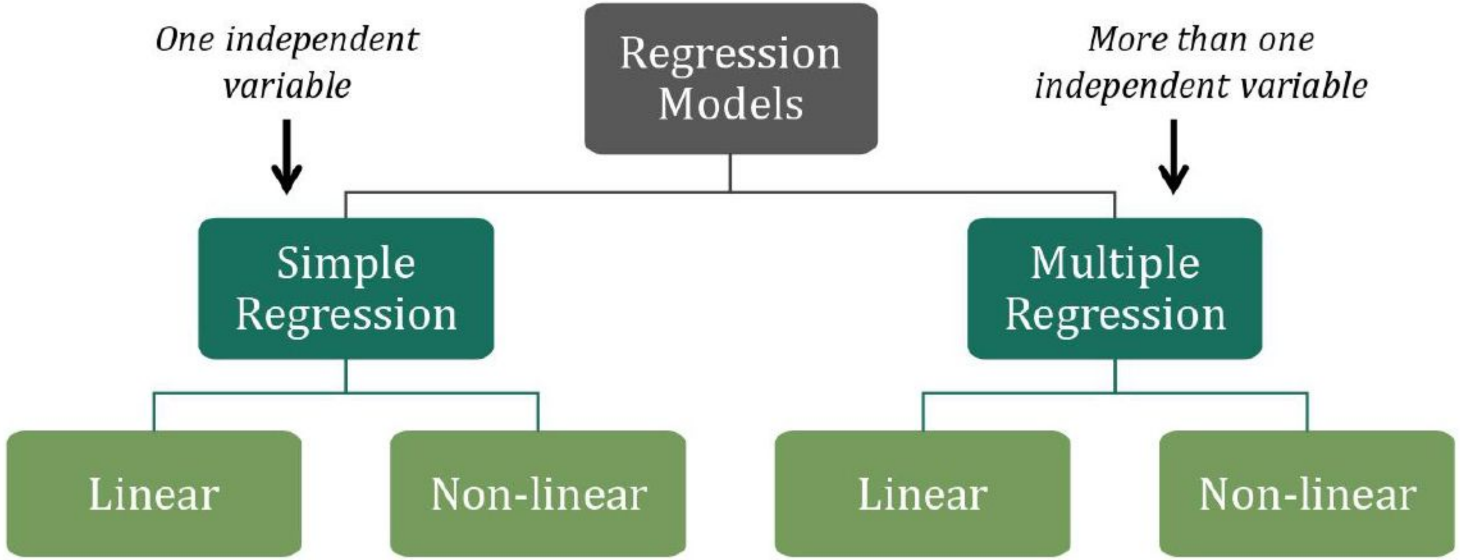


Regression Analysis

- Linear regression is a very simple approach for supervised learning.
- Useful technique for predicting a quantitative response.
- Aim of regression analysis is to find a best fit line that passes through the points.
- Describes the relationship between two variables x and y can be expressed by the following equation:

$$\boldsymbol{Y = c + mx +} \ \varepsilon$$

Regression Analysis

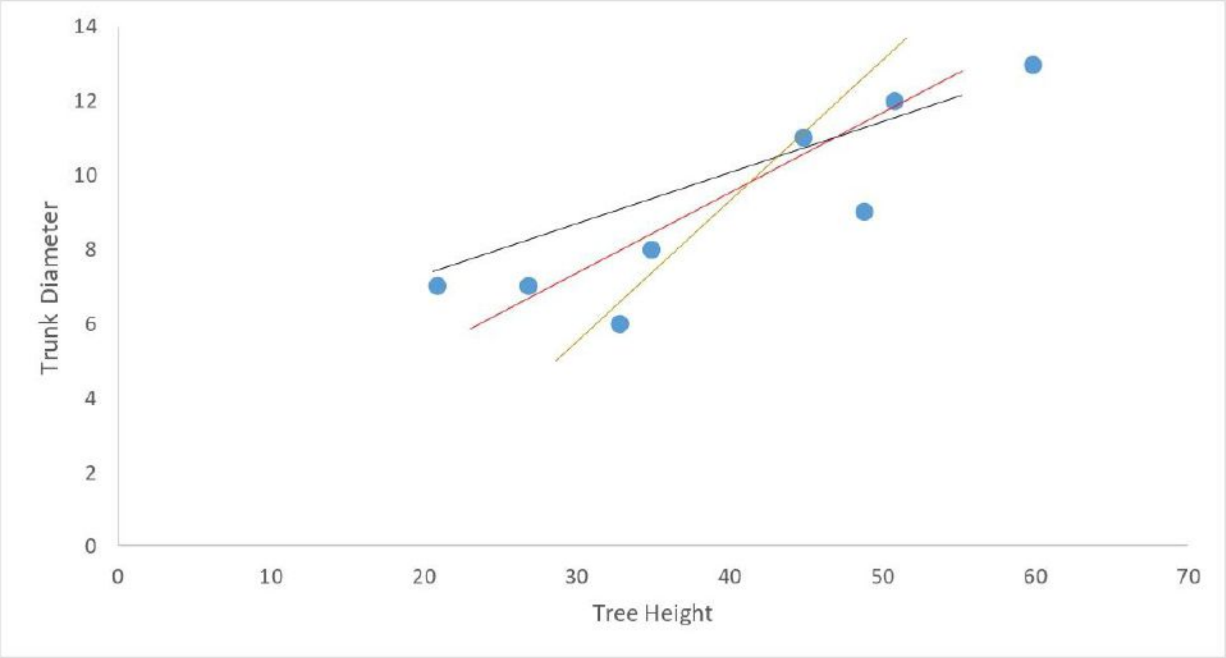


Ordinary Least Squares method (OLS)

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

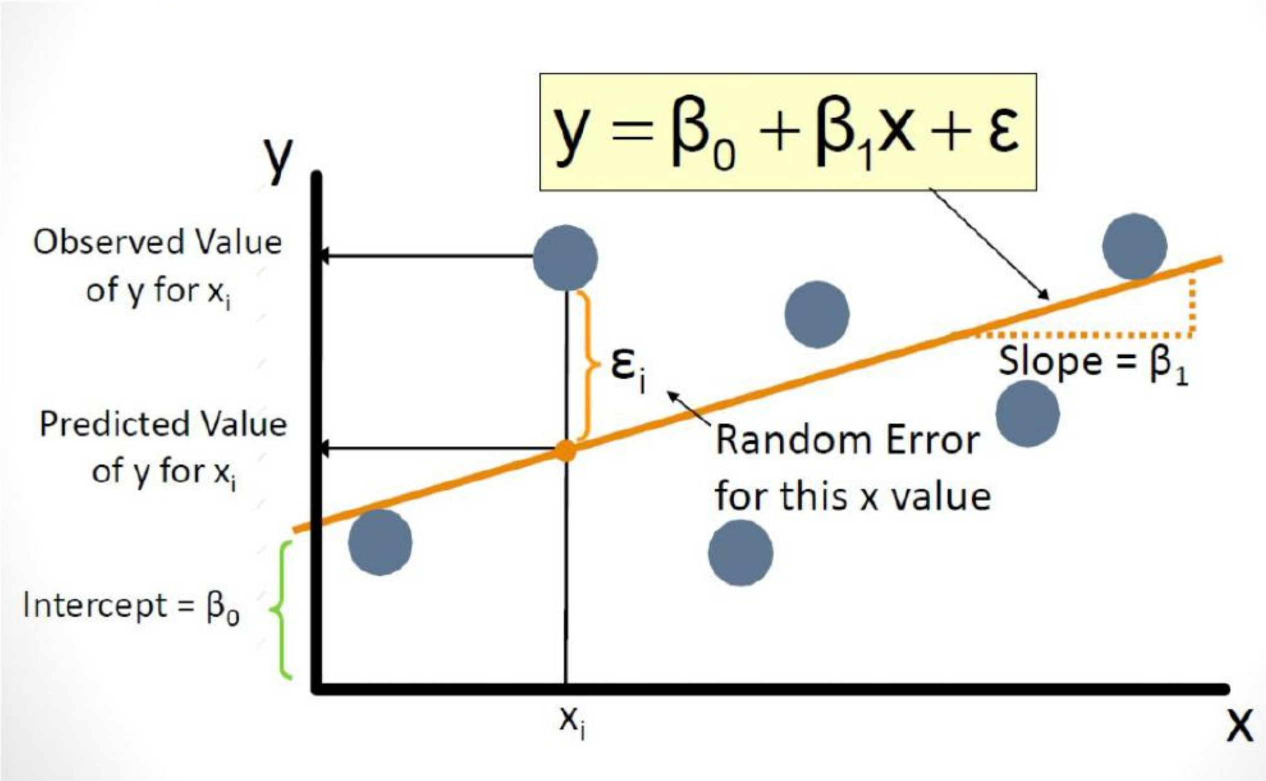
- Y - predicted value of the dependent variable, Dependent/Response variable.
- β_0 - the intercept (the predicted value of Y when all the predictor variables equal zero).
- β_1 - the regression coefficient (Slope) for the predictor – X.
- x - predictor value, Independent/Explanatory variable (Input).
- ε - Random error/Noise.

OLS – What is the best fit?



OLS – Least Squares

Min. Sum of Squares $$\sum_{i=0}^{6} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2$$



OLS – Least Squares

- Let $\varepsilon_\mathrm{i} = (y_i - \widehat{y}_i)$ be the prediction error for observation i.
- Sum of Squares of Errors, $SSE = \sum_{i=1}^{n} \varepsilon_i^2$
- For good fit, SSE should be minimum, that is "Least Squares".

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Least Squares Regression Properties

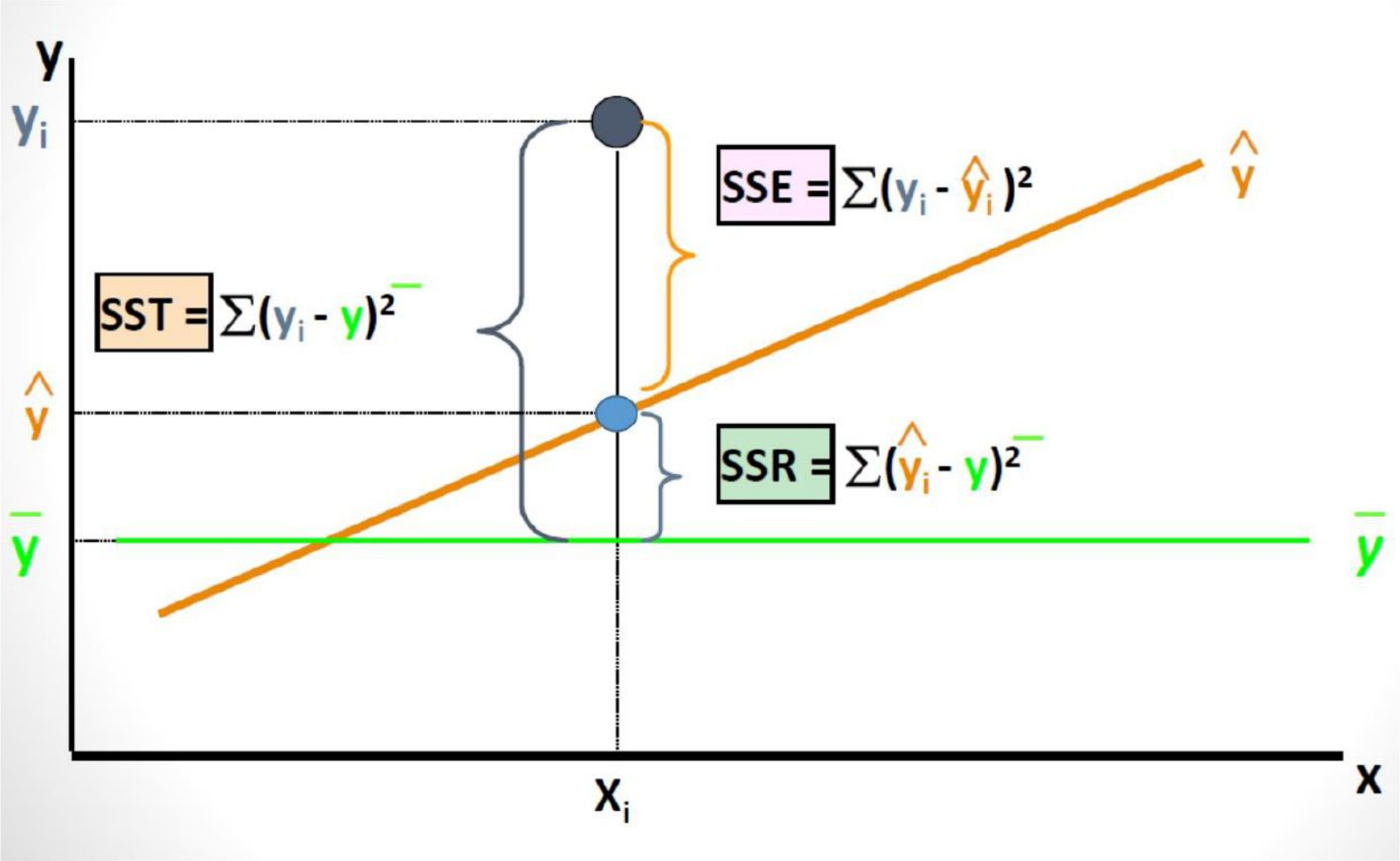
The sum of the residuals from the least squares regression line is 0.
($\sum(y-\hat{y})=0$)

The sum of the squared residuals is a minimum .
(minimized $\sum(y-\hat{y})^2$)

The simple regression line always passes through the mean of the y variable and the mean of the x variable.

The least squares coefficients are unbiased estimates of β_0 and β_1.

Explained and Unexplained variation



Explained and Unexplained variation

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum (y - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$

Where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

Explained and Unexplained variation

SST = Total sum of squares

- Measures the variation of the y_i values around their mean y

SSE = Error sum of squares

- Variation attributable to factors other than the relationship between x and y (Unexplained)

SSR = Regression sum of squares

- Explained variation attributable to the relationship between x and y

Coefficient of determination (R^2)

- How to judge a good fit line -
  - SSE (Minimum or Maximum?)
  - SSR (Minimum or Maximum?)
  - SSR/SSE(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.
- R-squared provides an estimate of the strength of the relationship between your model and the response variable.
- The coefficient of determination is also called R-squared and is denoted as R^2.

$$R^2 = \frac{SSR}{SST} \qquad where \qquad 0 \leq R^2 \leq 1$$