Advanced Statistics

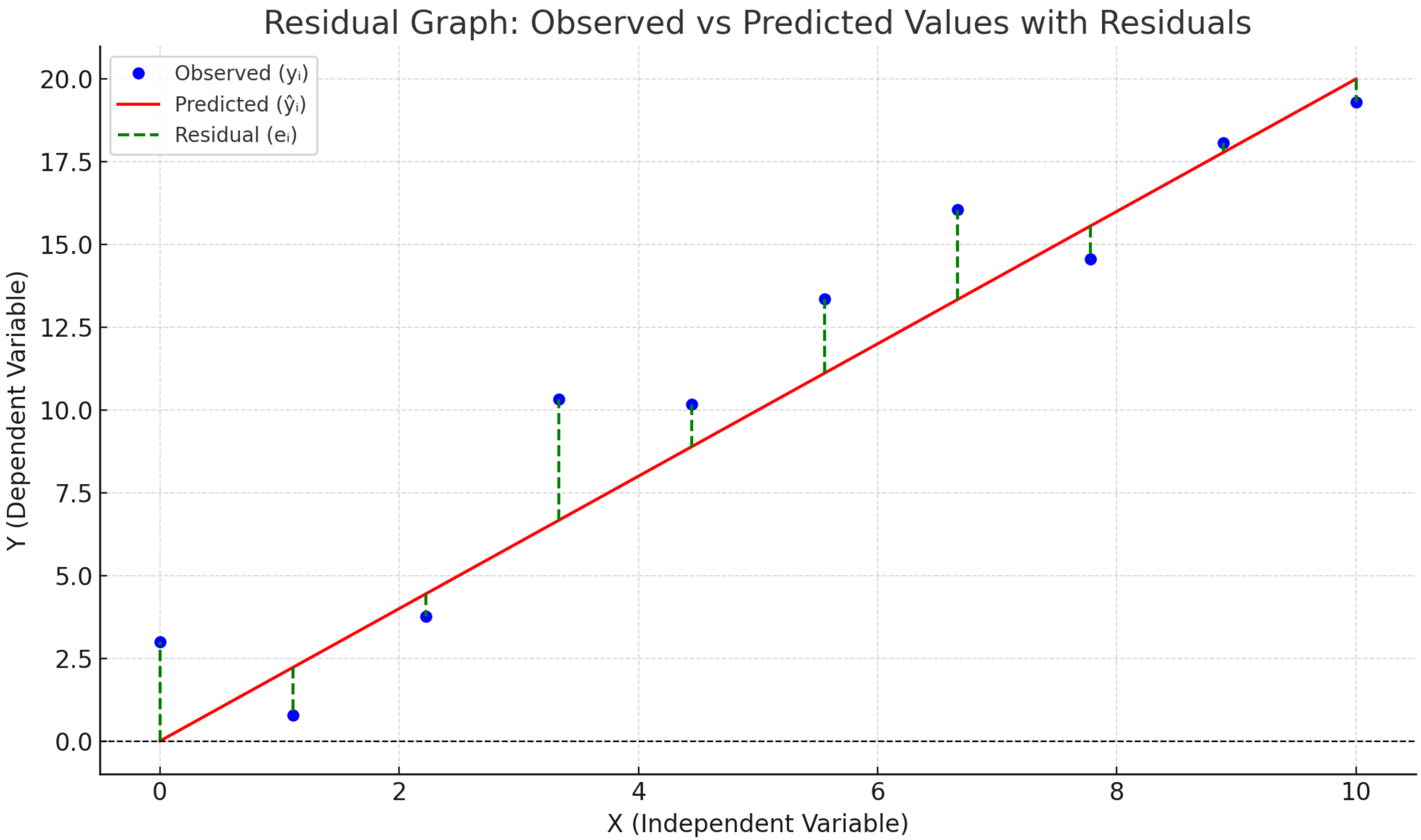
Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics**, 13th Edition, Mario F. Triola


Residual Graph: Observed vs Predicted Values with Residuals
Observed (y_i)
Predicted ($\hat{y}_i$)
Residual (e_i)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Y (Dependent Variable)
0
2
4
6
8
10
X (Independent Variable)

Cost Function

The **cost function** in least squares regression is the **Sum of Squared Errors (SSE)**:

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

where:

- $\mathbf{y_i}$ are the **observed values**
- $\hat{\mathbf{y}}_{\mathbf{i}} = b_0 + b_1 x_i$ are the **predicted values**
- $\mathbf{b_0}$ is the **y-intercept**
- $\mathbf{b_1}$ is the **slope**.

Steps to Derive the Cost Function

Step1: Start with the **residuals**:

$e_i = y_i - \hat{y}_i$

Step 2: Square the residuals to avoid cancellation of positive and negative values:

$e_i^2 = (y_i - (b_0 + b_1 x_i))^2$

Step 3: Sum up the squared residuals across all data points:

$\text{SSE} = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$

Importance of Minimizing SSE

- **Minimizing the SSE ensures the best-fit line:**
  - Reduces the overall discrepancy between **observed** and **predicted values**.
  - Provides the most **accurate regression line** to model the relationship between variables.
- **The smaller the SSE, the better the model fits the data.**

Objective Function

To minimize the **Sum of Squared Errors (SSE)**:

SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $\because \hat{y}_i = b_0 + b_1 x_i$

SSE = $\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$ ------------------------(1)

Differentiating Eq(1) with respect to $\boldsymbol{b_0}$, we get

$$\frac{\partial(SSE)}{\partial b_0} = 2(y_i - b_0 - b_1 x_i)^{2-1} \times \frac{\partial(y_i - b_0 - b_1 x_i)}{\partial b_0}$$

$$\Longrightarrow \frac{\partial(\text{SSE})}{\partial \text{b}_\text{o}} = 2(y_i - b_0 - b_1 x_i) \times (-1)$$

$$\Longrightarrow \frac{\partial(\text{SSE})}{\partial \text{b}_\text{o}} = -2(y_i - b_0 - b_1 x_i) \text{ ------------------------------(2)}$$

Objective Function

SSE = $\sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$ ------------------------(1)

Differentiating SSE (1) with respect to b_1

$$\frac{\partial(\text{SSE})}{\partial b_1} = 2(y_i - b_0 - b_1 x_i))^{2-1} \times \frac{\partial(y_i - b_0 - b_1 x_i)}{\partial b_1}$$

$$\frac{\partial(\text{SSE})}{\partial b_1} = 2(y_i - b_0 - b_1 x_i) \times (-x_i)$$

$$\frac{\partial(\text{SSE})}{\partial b_1} = -2x_i(y_i - b_0 - b_1 x_i) \text{------------------------(3)}$$

Setting the partial derivatives to zero and rearranging equation (2) to get the first normal equation

$$\frac{\partial(\mathrm{SSE})}{\partial b_o} = -2(y_i - b_0 - b_1 x_i) = 0$$

$$y_i = b_0 + b_1 x_i$$

Apply summation, we get

$$\sum_{i=1}^{n} y_i = \mathrm{n}b_0 + b_1 \sum_{i=1}^{n} x_i \text{ ----------------(4)}$$

Setting the partial derivatives to zero and rearranging equation (3) to get the second normal equation

$$\frac{\partial(\text{SSE})}{\partial b_1} = -2x_i(y_i - b_0 - b_1 x_i) = 0$$

$$x_i y_i = b_0 x_i + b_1 x_i^2$$

Apply summation, we get

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \quad \text{----------------(5)}$$

$\sum_{i=1}^{n} y_i$ = nb_0 + $b_1 \sum_{i=1}^{n} x_i$------------------------------(4)

$\sum_{i=1}^{n} x_i y_i$ = $b_0 \sum_{i=1}^{n} x_i$ + $b_1 \sum_{i=1}^{n} x_i^2$ ---------------(5)

(4) × ∑ $\mathbf{x_i}$ - (5) × n:

$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$ = n$b_0 \sum_{i=1}^{n} x_i$ + $b_1 \left(\sum_{i=1}^{n} x_i\right)^2$

n$\sum_{i=1}^{n} x_i y_i$ = $nb_0 \sum_{i=1}^{n} x_i$ + $nb_1 \sum_{i=1}^{n} x_i^2$

- - -

$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$ - n$\sum_{i=1}^{n} x_i y_i$= $b_1 \left(\sum_{i=1}^{n} x_i\right)^2$ - $nb_1 \sum_{i=1}^{n} x_i^2$

$$\Rightarrow \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n\sum_{i=1}^{n} x_i y_i = b_1 \{(\sum_{i=1}^{n} x_i)^2 - n\sum_{i=1}^{n} x_i^2\}$$

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i)^2 - n \sum_{i=1}^{n} x_i^2}$$

or

$$\Rightarrow \mathbf{b_1} = \frac{\mathbf{n} \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\mathbf{n} \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

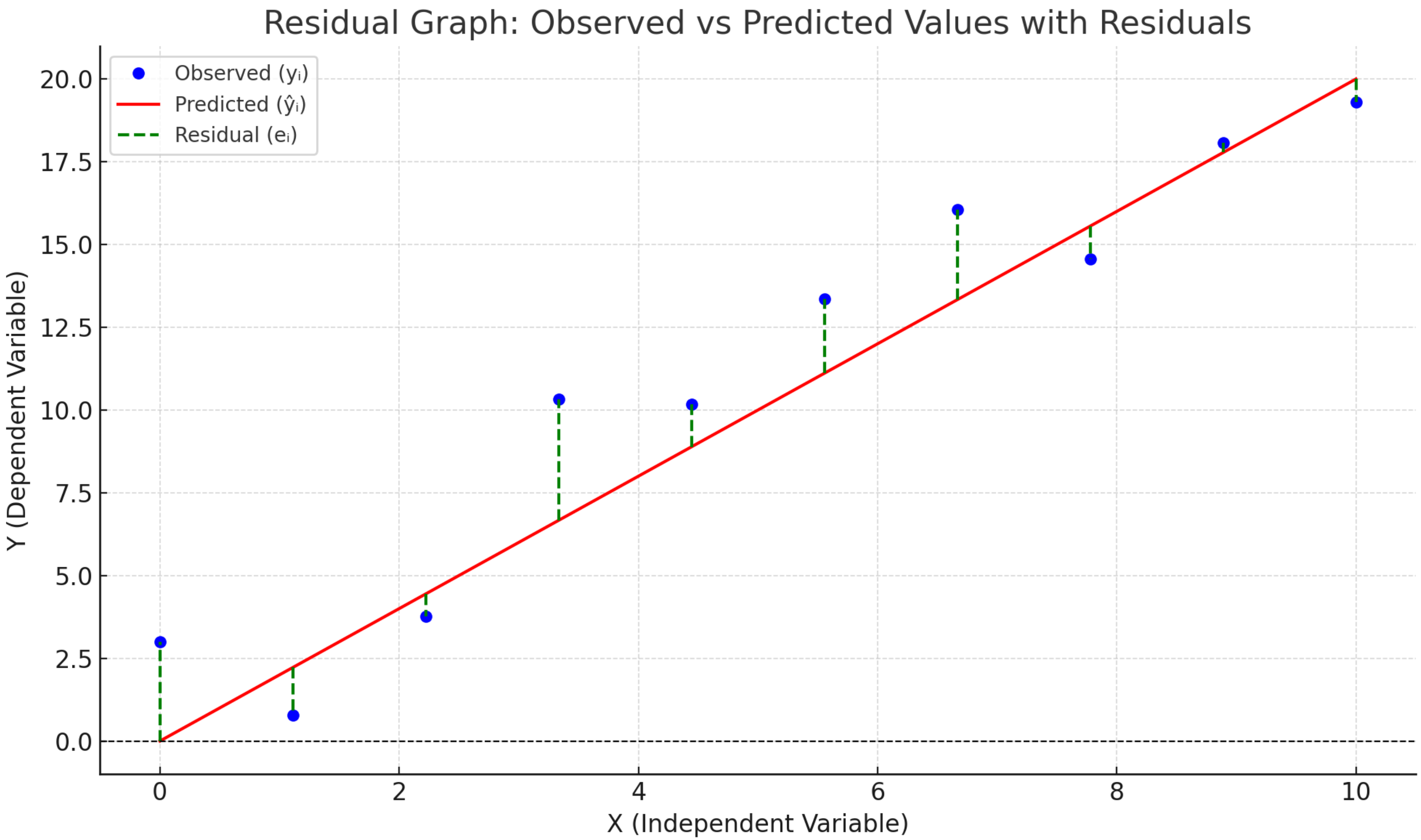
Using equation 4, we get

$$\sum_{i=1}^{n} y_i = \text{n}b_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\Rightarrow \text{n}b_0 = \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i$$

$$\Rightarrow b_0 = \frac{\sum \text{y}_\text{i}}{n} - \frac{b_1 \sum \text{x}_\text{i}}{n}$$

$$\boldsymbol{\Rightarrow b_0 = \overline{y} - b_1 \overline{x}}$$


Residual Graph: Observed vs Predicted Values with Residuals
Observed (y_i)
Predicted ($\hat{y}_i$)
Residual (e_i)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
2
4
6
8
10
Y (Dependent Variable)
X (Independent Variable)

1. Minimizes the Sum of Squared Errors (SSE)

The least squares regression line minimizes:

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

where:

y_i are the observed values

$\hat{y}_i$ are the predicted values

2. Passes Through the Mean of the Data

The regression line passes through the mean:

$(\bar{x}, \bar{y})$, where:

$\bar{x}$ is the mean of the independent variable

$\bar{y}$ is the mean of the dependent variable

3. Residuals Sum to Zero

The sum of the residuals (errors) is zero:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

4. Uncorrelated Residuals and Predicted Values

The residuals (e_i) and the predicted values ($\hat{y}_i$) are uncorrelated:

$\mathrm{Cov}(e, \hat{y}) = 0.$

5. Minimizes Variance of Residuals

The least squares regression line minimizes the variance of residuals compared to any other possible line.

6. Unique Solution

The regression line has a unique solution for the slope ($\mathbf{b_1}$) and intercept ($\mathbf{b_0}$)

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Or

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Example: A real estate agent wants to predict housing prices (y) in USD based on the square footage (x) of houses (in square meters). Using the given dataset, derive the least-squares regression line and predict the house price for a property with a square footage of 2,000 square meters:

House Index	Square Footage (x)	Price (y)
1	800	150,000
2	1000	180,000
3	1200	200,000
4	1500	240,000
5	1800	300,000

Square Footage (x)	Price (y)	x^2	xy
800	150000	640000	120000000
1000	180000	1000000	180000000
1200	200000	1440000	240000000
1500	240000	2250000	360000000
1800	300000	3240000	540000000
$\sum x_i = 6300$	$\sum y_i = 1070000$	$\sum x_i^2 = 8570000$	$\sum x_i y_i = 1440000000$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Substitute values:

$$b_1 = \frac{(5)(1440000000) - (6300)(1070000)}{5(8570000) - (6300)^2}$$

$b_1 \approx 145.25$

$b_0 = \bar{y} - b_1 \bar{x}$

$\bar{y} = \frac{1070000}{5} = 214000$

$\bar{x} = \frac{6300}{5} = 1260$

Substitute values:

$b_o = 214000 - 145.25(1260)$

$b_o \approx 30981.01$

Step 4: Regression Line and Prediction

Regression Line:

$\hat{y} = 30981.01 + 145.25x$

Prediction:

For a house with 2,000 square feet:

$\hat{y} = 30981.01 + 145.25(2000)$

$\hat{y} \approx 321{,}487.34$

Predicted price: 321,487.34 USD

Problem: A computer scientist wants to predict the execution time (y) of an algorithm based on the size of the input data (x).

Input Size (x)	Execution Time (y) (ms)
100	20
200	50
300	70
400	100
500	150

Task:

1. Derive the least-squares regression line.
2. Predict the execution time for an input size of 350.

x	y	x^2	xy
100	20	10000	2000
200	50	40000	10000
300	70	90000	21000
400	100	160000	40000
500	150	250000	75000
$\sum_{i=1}^{n} x_i$ = 1500	$\sum_{i=1}^{n} y_i$ = 390	$\sum_{i=1}^{n} x_i^2$ = 550000	$\sum_{i=1}^{n} x_i y_i$ = 148000

$b_1 = 0.31$

$b_0 = -15.00$

Regression Line:

$\hat{y} = -15.00 + 0.31x$

Predict Execution Time for x = 350

$\hat{y} = -15.00 + 0.31(350) = 93.50$

Predicted execution time = 93.50 ms