# 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

# Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

# References

- Elementary Statistics, 14th Edition, Mario F. Triola

These notes contain material from the above resources.

# The Least Squares Criterion

- **Objective:** To provide a fitted line that ensures "closeness" between the line and the plotted data points.
- **Method:** Minimizes the **sum of squared residuals** to achieve this closeness.

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

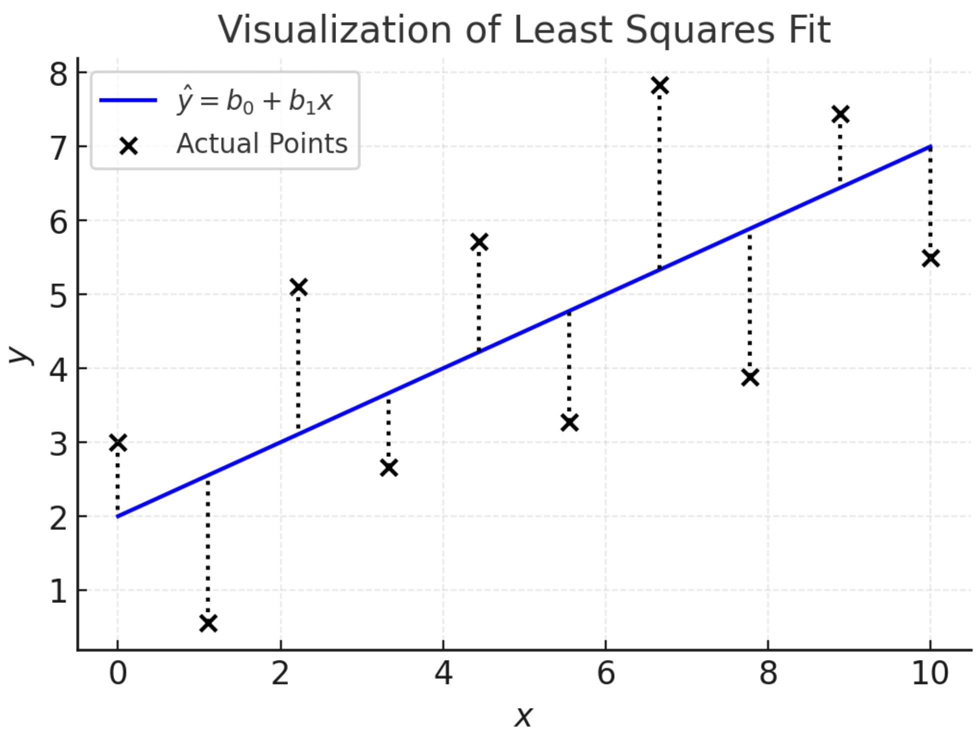
# Residuals and Fitted Line in Least Squares

**Predicted Values:** Points on the fitted line represent predicted values based on the model.

**Residuals:** Vertical deviations from the observed data points to the line.

**Key Insight:**

The least squares procedure generates a line that **minimizes the sum of squares of these vertical deviations**.



Visualization of Least Squares Fit
$\hat{y} = b_0 + b_1 x$
Actual Points
y
x
0
2
4
6
8
10
1
2
3
4
5
6
7
8

# Closeness and Alternative Measures

There are various ways to measure closeness between the line and data points:

- **Minimizing the sum of absolute residuals:**

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Minimizing the sum of residuals raised to a power:**

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|^{1.5}$$

These approaches, like least squares, aim to make residuals **"small".**

# Benefits of Least Squares

- Provides a systematic and consistent method to fit a line.

- Ensures **residuals** are minimized in a **squared sense**, reducing the **impact of larger errors**.

- Widely used due to simplicity and statistical properties, such as **unbiased estimators**.

# Definitions

Given a collection of paired sample data, the **regression equation**

$\hat{y} = b_0 + b_1 x$

algebraically describes the relationship **between the two variables**. The graph of the **regression equation** is called the **regression line** (or ***line of best fit***, or ***least-squares line***).

# Notation for Regression Equation

| | Population Parameter | Sample Statistic |
|---|---|---|
| y-intercept of regression equation | $\beta_0$ | $b_0$ |
| Slope of regression equation | $\beta_1$ | $b_1$ |
| Equation of the regression line | $Y = \beta_0 + \beta_1 x$ | $\hat{y} = b_0 + b_1 x$ |

**Finding the slope $b_1$ and $y$-intercept $b_0$ in the regression equation $\hat{y} = b_0 + b_1 x$**

| | |
|---|---|
| **Slope** | $\mathbf{b_1} = \dfrac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$ |
| ***y*-intercept:** | $\mathbf{b_0} = \bar{y} - \mathbf{b_1}\bar{x}$ <br> **or** <br> $\mathbf{b_0} = \dfrac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$ |

**Finding the slope $b_1$ and *y*-intercept $b_0$ in the regression equation $\hat{y} = \mathbf{b_0 + b_1 x}$**

**Slope:** $b_1 = r\frac{s_y}{s_x}$

where $r$ is the linear correlation coefficient, $s_y$ **is the standard deviation** of the *y* values, and $s_x$ **is the standard deviation** of the *x* values.

***y*-intercept:** $\mathbf{b_0} = \bar{y} - \mathbf{b_1}\bar{x}$

**Example:** Table 1 is reproduced here. (Jackpot amounts are in millions of dollars and numbers of tickets sold are in millions.) Find the equation of the **regression line in which the explanatory variable (or *x* variable)** is the amount of the lottery jackpot and the response variable (or *y* variable) is the corresponding number of lottery tickets sold.

**Table 1 Powerball Tickets Sold and Jackpot Amounts**

| Jackpot | 334 | 127 | 300 | 227 | 202 | 180 | 164 | 145 | 255 |
|---|---|---|---|---|---|---|---|---|---|
| Tickets | 54 | 16 | 41 | 27 | 23 | 18 | 18 | 16 | 26 |

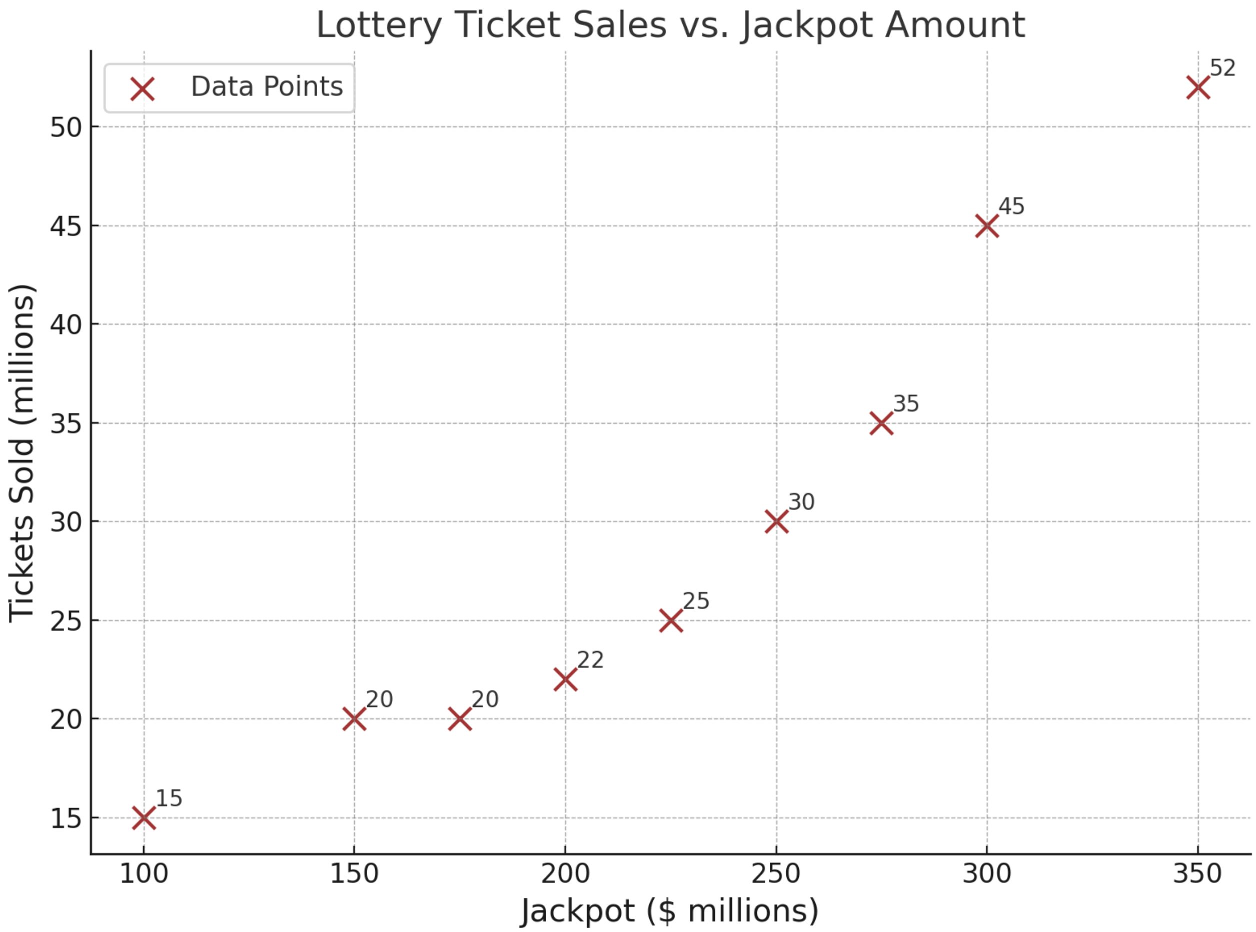



**FIGURE 1 Scatterplot from Table 1**

1. The data are a simple random sample.

2. The scatterplot in Figure 1 on previous slide shows that the pattern of points is reasonably close to a **straight-line pattern**.

3. The scatterplot also shows that there are **no outliers**.

The requirements are satisfied.

| x(Jackpot) | y(Tickets) | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 334 | 54 | 111,556 | 2916 | 18,036 |
| 127 | 16 | 16,129 | 256 | 2032 |
| 300 | 41 | 90,000 | 1681 | 12,300 |
| 227 | 27 | 51,529 | 729 | 6129 |
| 202 | 23 | 40,804 | 529 | 4646 |
| 180 | 18 | 32,400 | 324 | 3240 |
| 164 | 18 | 26,896 | 324 | 2952 |
| 145 | 16 | 21,025 | 256 | 2320 |
| 255 | 26 | 65,025 | 676 | 6630 |
| $\sum x$ = 1934 | $\sum y$ = 239 | $\sum x^2$= 455,364 | $\sum y^2$= 7691 | $\sum xy$ = 58,285 |

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{9(158{,}2852) - (1934)(239)}{\sqrt{9(455{,}364) - (1943)^2}\sqrt{9(7651) - (239)^2}}$$

$$r = \frac{62{,}339}{\sqrt{357{,}920}\sqrt{12{,}098}} = 0.947$$

$$\boldsymbol{b_1} = \boldsymbol{r}\frac{\boldsymbol{s_y}}{\boldsymbol{s_x}}$$

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} \mathrm{y}^2{}_i - (\sum_{i=1}^{n} y_i)^2\}}$$

$$s_y = \sqrt{\frac{1}{9(9-1)}\{9(7691) - (239)^2\}}$$

$$s_y = 70.50611$$

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2\}}$$

$$s_x = \sqrt{\frac{1}{9(9-1)}\{9(455{,}364) - (1934)^2\}}$$

$$= 12.96255$$

$$\boldsymbol{b_1} = \boldsymbol{r}\frac{\boldsymbol{s_y}}{\boldsymbol{s_x}}$$

$$= 0.947 \times \frac{12.9625}{70.5061}$$

$$= 0.1742$$

$$\bar{x} = \frac{1934}{9} = 214.8889$$

$$\bar{y} = \frac{239}{9} = 26.5556$$

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

$$b_0 = 26.5556 - (0.1742)(214.8889)$$

$$\mathbf{b_0} = -10.8716$$

$$\hat{y} = \mathbf{b_0} + \mathbf{b_1 x}$$

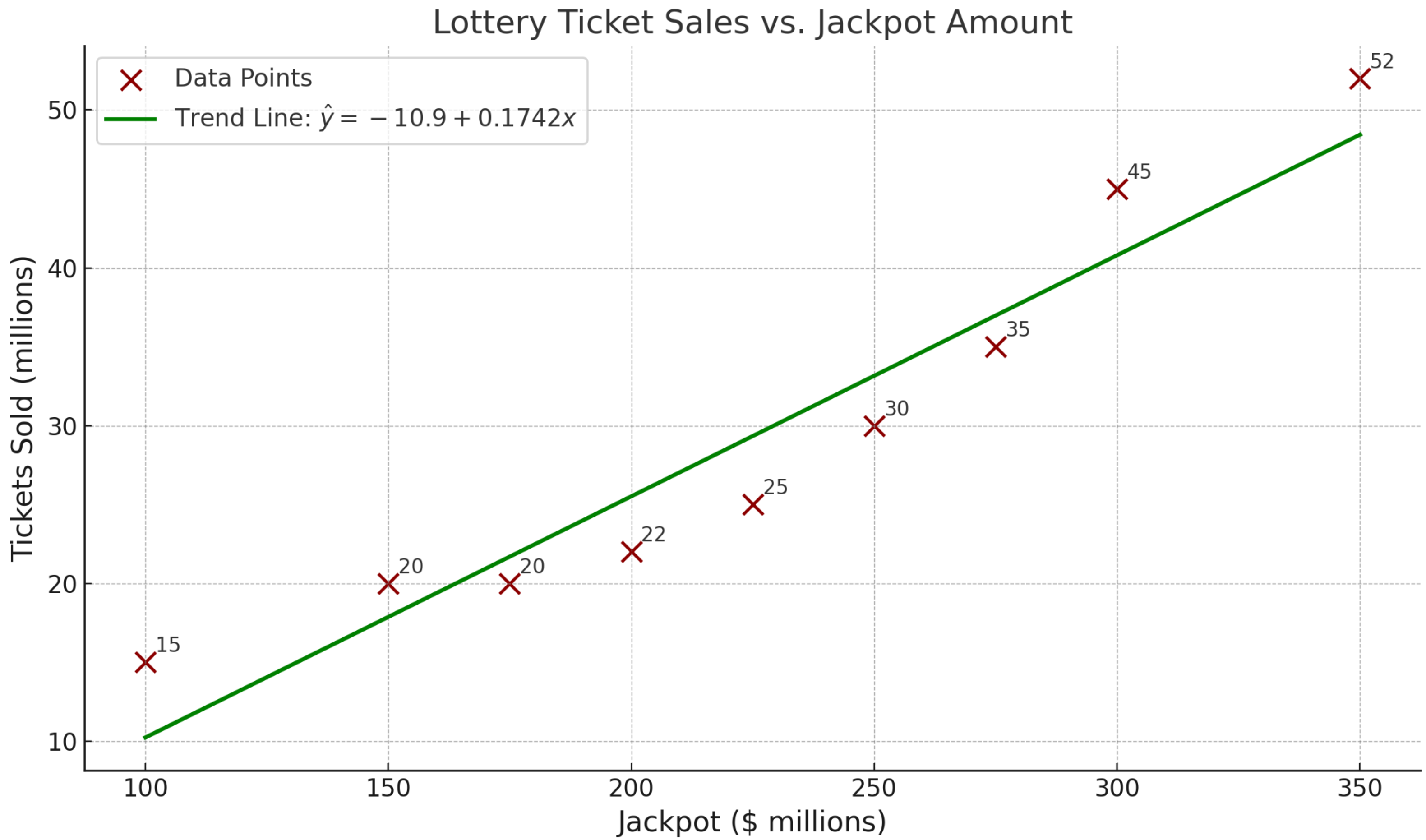
$$\hat{y} = -10.8716 + (0.1742)\mathbf{x}$$

**Or**

$$\hat{y} = -10.9 + (0.1742)\mathbf{x}$$

where $\hat{y}$ is the predicted number of tickets sold and ***x*** is the amount of the jackpot.

Graph the regression equation $\hat{y}$ =-10.9 +(0.1742)$x$ on the scatterplot of the jackpot/tickets data from Table 1 and examine the graph to subjectively determine how well the regression line fits the data.



**Problem:** A study was conducted at Virginia Tech to determine if certain **static arm-strength measures** have an influence on the **"dynamic lift"** characteristics of an individual. **Twenty-five** individuals were subjected to strength tests and then were asked to perform a **weightlifting test** in which weight was dynamically lifted overhead. The data are given in next two slides.

(a) Estimate $\beta_0$ and $\beta_1$ for the linear regression curve $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Find a point estimate of $\mu_{Y|30}$.

(c) Plot the residuals versus the *x*'s (arm strength). Comment.

| Individual | Arm Strength ($x$) | Dynamic Lift ($y$) |
|---|---|---|
| 1 | 17.3 | 71.7 |
| 2 | 19.3 | 48.3 |
| 3 | 19.5 | 88.3 |
| 4 | 19.7 | 75.0 |
| 5 | 22.9 | 91.7 |
| 6 | 23.1 | 100.0 |
| 7 | 26.4 | 73.3 |
| 8 | 26.8 | 65.0 |
| 9 | 27.6 | 75.0 |
| 10 | 28.1 | 88.3 |
| 11 | 28.2 | 68.3 |
| 12 | 28.7 | 96.7 |

| Individual | Arm Strength ($x$) | Dynamic Lift ($y$) |
|---|---|---|
| 13 | 29.0 | 76.7 |
| 14 | 29.6 | 78.3 |
| 15 | 29.9 | 60.0 |
| 16 | 29.9 | 71.7 |
| 17 | 30.3 | 85.0 |
| 18 | 31.3 | 85.0 |
| 19 | 36.0 | 88.3 |
| 20 | 39.5 | 100.0 |
| 21 | 40.4 | 100.0 |
| 22 | 44.3 | 100.0 |
| 23 | 44.6 | 91.7 |
| 24 | 50.4 | 100.0 |
| 25 | 55.9 | 71.7 |

| Individual | Arm Strength ($x$) | Dynamic Lift ($y$) | $xy$ | $x^2$ |
|---|---|---|---|---|
| 1 | 17.3 | 71.7 | 1240.41 | 299.29 |
| 2 | 19.3 | 48.3 | 932.19 | 372.49 |
| 3 | 19.5 | 88.3 | 1721.85 | 380.25 |
| 4 | 19.7 | 75.0 | 1477.5 | 388.09 |
| 5 | 22.9 | 91.7 | 2099.93 | 524.41 |
| 6 | 23.1 | 100.0 | 2310.0 | 533.61 |
| 7 | 26.4 | 73.3 | 1935.12 | 696.96 |
| 8 | 26.8 | 65.0 | 1742.0 | 718.24 |
| 9 | 27.6 | 75.0 | 2070.0 | 761.76 |
| 10 | 28.1 | 88.3 | 2481.23 | 789.61 |
| 11 | 28.2 | 68.3 | 1926.06 | 795.24 |
| 12 | 28.7 | 96.7 | 2775.29 | 823.69 |

| Individual | Arm Strength ($x$) | Dynamic Lift ($y$) | $xy$ | $x^2$ |
|---|---|---|---|---|
| 13 | 29.0 | 76.7 | 2224.3 | 841.0 |
| 14 | 29.6 | 78.3 | 2317.68 | 876.16 |
| 15 | 29.9 | 60.0 | 1794.0 | 894.01 |
| 16 | 29.9 | 71.7 | 2143.83 | 894.01 |
| 17 | 30.3 | 85.0 | 2575.5 | 918.09 |
| 18 | 31.3 | 85.0 | 2660.5 | 979.69 |
| 19 | 36.0 | 88.3 | 3178.8 | 1296.0 |
| 20 | 39.5 | 100.0 | 3950.0 | 1560.25 |
| 21 | 40.4 | 100.0 | 4040.0 | 1632.16 |
| 22 | 44.3 | 100.0 | 4430.0 | 1962.49 |
| 23 | 44.6 | 91.7 | 4089.82 | 1989.16 |
| 24 | 50.4 | 100.0 | 5040.0 | 2540.16 |
| 25 | 55.9 | 71.7 | 4008.03 | 3124.81 |
| | $\sum_{i=1}^{n} x_i$=**778.7** | $\sum_{i=1}^{n} y_i$ = **2050.0** | $\sum_{i=1}^{n} x_i y_i$= **65164.04** | $\sum_{i=1}^{n} x_i^2$= **26591.63** |

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

$$= \frac{(25)(65164.04)-(778.7)(2020)}{(25)(26591.63)-(778.7)^2}$$

$$\mathbf{b_1 = 0.5609}$$

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

or

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

$$= \frac{2020-(0.5609)(778.7)}{25}$$

$$b_0 = 64.53$$

$\mu_{Y|x} = \beta_0 + \beta_1 x$

Estimate of linear regression line is

$\boldsymbol{\mu_{Y|x}}$ = 64.53 +0.5609 $x$

Or

$\boldsymbol{\hat{y}}$ = 64.53 +0.5609 $x$

b) At $x$ = 30

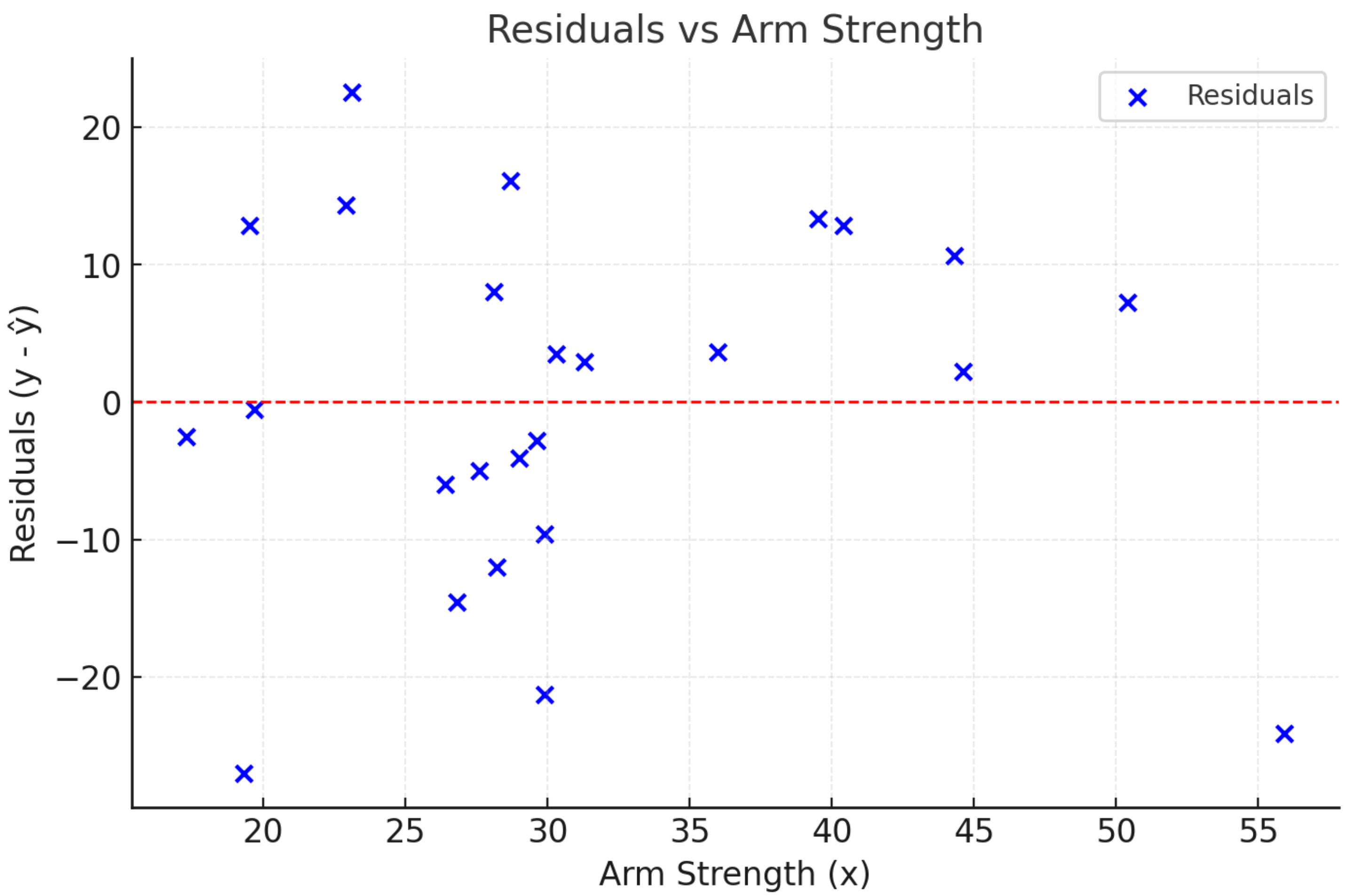
$\mu_{Y|30}$ = 64.53 +(0.5609) (30)

= 81.40

## c) No Clear Pattern:

The residuals appear randomly scattered around the horizontal line at zero.

.



**1. No Clear Pattern:**

The residuals appear randomly scattered around the horizontal line at zero.

This suggests that the regression model adequately captures the linear relationship between arm strength and dynamic lift.

## **2. Heteroscedasticity:**

Some variability in the spread of residuals is noticeable, particularly for smaller and larger values of arm strength.

This might indicate a **non-constant variance in the data**, suggesting that a **transformation or weighted** regression might improve the model.

### 3. Outliers:

A few points deviate significantly from zero.

These outliers could be investigated for potential measurement errors or unique conditions.

# What is Heteroscedasticity?

Heteroscedasticity occurs when the variance of residuals (errors) is **not constant across all levels** of the independent variable(s).

- **Homoscedasticity:** Constant residual variance.
- **Heteroscedasticity:** Residual variance changes with predictor values.

# Impact of Heteroscedasticity

## Effects of heteroscedasticity on regression analysis:

1. Does not bias the coefficients (e.g., $\beta_0$, $\beta_1$).

2. Increases standard errors, making hypothesis tests unreliable.

3. Can lead to incorrect conclusions about variable significance.

# Causes of Heteroscedasticity

- Presence of outliers.
- Skewed distribution of the dependent variable.
- Non-linear relationships not captured by the model.
- Measurement errors in the data.

# Detecting Heteroscedasticity

## 1. Residual Plot:

Plot residuals vs. fitted values or predictors.

Look for a funnel-shaped pattern (increasing or decreasing spread).

## 2. Statistical Tests:

Breusch-Pagan Test: Regress squared residuals on predictors.

White's Test: Tests for heteroscedasticity in a model.

## 3. Visual Inspection:

Look for systematic patterns in residuals.

# Solutions for Heteroscedasticity

**1. Transformations:**

Apply log, square root, or inverse transformations to stabilize variance.

**2. Weighted Least Squares (WLS):**

Assign weights inversely proportional to residual variance.

**3. Robust Standard Errors:**

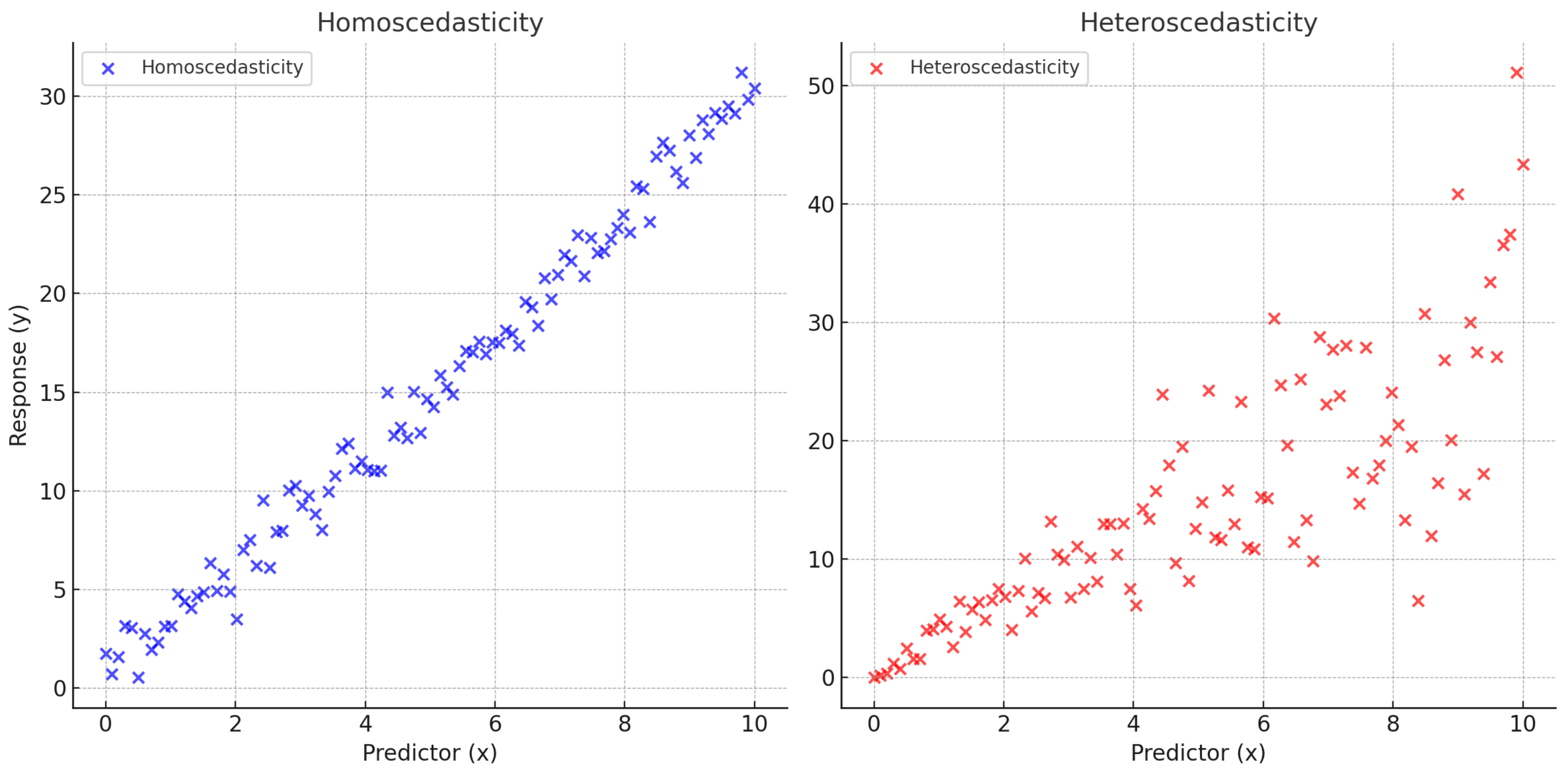
Use heteroscedasticity-robust standard errors to adjust hypothesis tests.

# Why Does It Matter?

- Heteroscedasticity undermines **reliability of statistical inference** (e.g., confidence intervals, p-values).

- Addressing it ensures accurate and meaningful interpretations of regression results.

# Homoscedasticity vs Heteroscedasticity



- **Homoscedasticity (Left Graph):**
  - Residual variance remains **constant** across all values of the predictor.
  - Points are evenly spread around the trend line.
- **Heteroscedasticity (Right Graph):**
  - Residual variance **increases** with the value of the predictor.
  - Points show a fan-like spread, indicating changing variance.