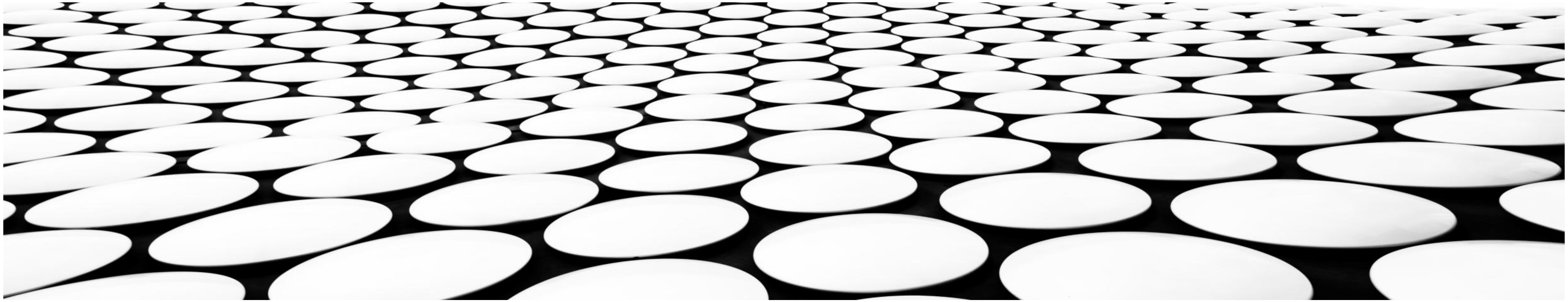
DATA MINING AND PREDICTIVE DATA ANALYTICS

CHAPTER-8

SIMPLE LINEAR REGRESSION



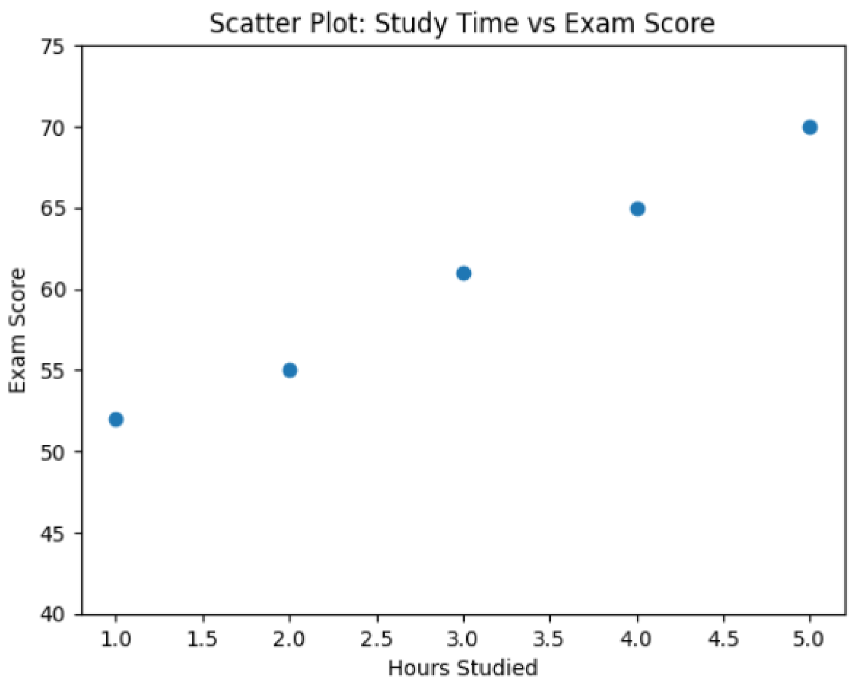
INTRODUCTION TO REGRESSION MODELING

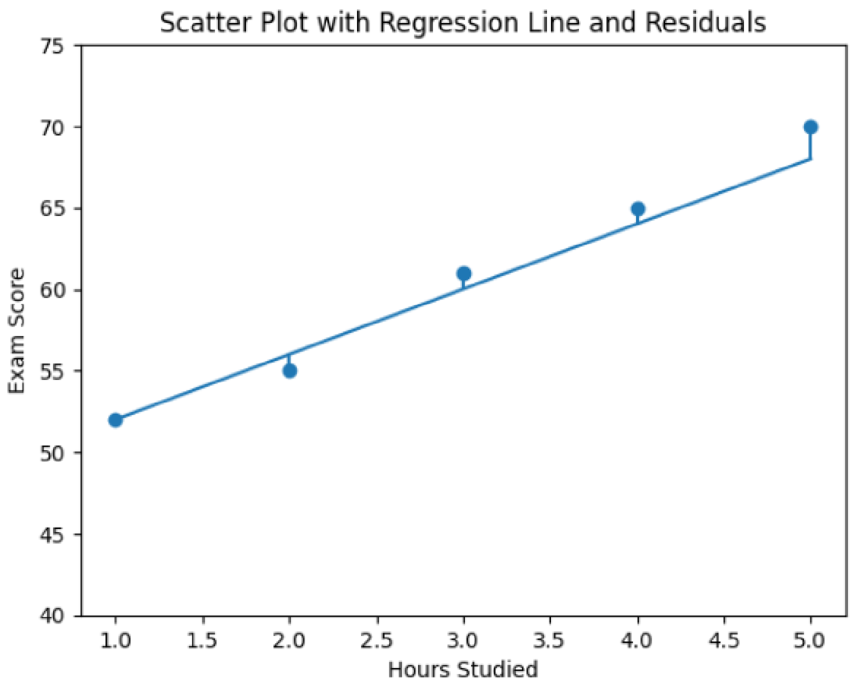
- **Introduction to Regression Modeling**
- Regression modeling is a powerful statistical technique used to estimate the value of a continuous target (response) variable based on one or more predictor variables.
- In **simple linear regression**, we study:
  - The relationship between one continuous predictor variable (x) and one continuous response variable (y)
  - Quantify how changes in one variable affect another
  - Predict the value of the response variable for the given predictor value
- A straight line, called Regression Line (Best-fitting straight line) is used to approximate this relationship.

GENERAL FORM OF THE LINEAR REGRESSION MODEL

- A Simple and Intuitive Example
- Suppose we want to estimate a student's exam score (y) based on the number of hours studied (x).

Student	Hours Studied (x)	Exam Score (y)
A	1	52
B	2	55
C	3	61
D	4	65
E	5	70





- Assume the estimated regression equation is: $\hat{y} = 48 + 4x$

Intercept ($b_0 = 48$) Slope ($b_1 = 4$)

GENERAL FORM OF THE LINEAR REGRESSION MODEL

- **Regression Equation**: The regression line is written in the form

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

where:

- $\hat{y}$ = estimated value of the response
- b_0 = estimated y-intercept
- b_1 = estimated slope
- b_0 and b_1 are called **regression coefficients**

- In our example, assume the estimated regression equation is: $\hat{y} = 48 + 4x$
- Prediction Using the Regression Equation
  - If a student studies for 3 hours. Then, $\hat{y} = 48 + 4(3) = 60$
  - Suppose a student studied 3 hours but actually scored 61 marks.
  - Hence**, <u>Residual (Prediction Error):</u>** $y - \hat{y} = 61 - 60 = 1$
  - Prediction Error is the vertical distance between the actual data point, and regression line

LEAST SQUARES ESTIMATES:

- **Least squares regression** is the most common method for regression that works by choosing the unique regression line that minimizes the sum of squared residuals (errors) over all the data points.
- **Least Squares Estimates:**
  - It determines model parameters by minimizing the sum of squared residuals, i.e., the squared differences between observed values and predicted values.
  - Key Property: Squaring residuals penalizes larger errors more heavily, leading to parameter estimates that best fit the data in an average sense.
  - Use Case: Widely used in linear regression to estimate coefficients that produce the best-fitting line through the data.

LEAST SQUARES ESTIMATES:

- The least-squares regression line for a data set consisting of n observations is given by: $\hat{y} = b_0 + b_1 x$
  - Formula to calculate Slope Estimate **(b_1):**

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

  - Formula to calculate Intercept Estimate (b_0):

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

Where

- n = number of observations
- x_i = i^{th} value of the predictor variable
- y_i = i^{th} value of the response variable
- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ = mean of the predictor
- $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ = mean of the response

HOW USEFUL IS THE REGRESSION?

- *Why Do We Need a Measure of Usefulness?*
  - A **least-squares regression** line can always be computed between two continuous variables.
  - However, the existence of a regression line does not guarantee that it is useful for prediction.
  - Hence, a natural question arises: How do we determine whether a regression equation provides good predictions?
  - To answer this, we develop a numerical measure of goodness of fit, known as **the coefficient of determination**, denoted by (r^2)

HOW USEFUL IS THE REGRESSION?

- Consider the data set which shows the distance in km traveled by a sample of 10 competitors, along with the elapsed time in hours.
- The estimated regression equation is: $\hat{y} = 6 + 2x$

			Predicted Score	Error in Prediction	(Error in Prediction)^2
Subject	X = Time	Y = Distance	$\hat{y} = 6 + 2x$	$(y - \hat{y})$	$(y - \hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	−1	1
5	4	14	14	0	0
6	5	15	16	−1	1
7	6	20	18	2	4
8	7	18	20	−2	4
9	8	22	22	0	0
10	9	25	24	1	1

SSE (Sum of Squared Errors)

$$\boxed{\mathrm{SSE} = \sum (y - \hat{y})^2 = 12}$$

HOW USEFUL IS THE REGRESSION?

- Interpretation of SSE
  - SSE measures the overall prediction error when using the regression equation
  - Smaller SSE indicates better predictive accuracy
  - However, SSE alone is not interpretable without a reference value
- A Baseline for Comparison: Ignoring the Predictor
  - Suppose we ignore the predictor variable (time) altogether.
  - In that case, the best estimate of distance for all competitors is simply the sample mean: $\bar{y} = 16 \text{ km}$
  - This corresponds to predicting the same distance regardless of time.

HOW USEFUL IS THE REGRESSION?

- <u>Total Variability in the Response: SST</u>
  - We now measure the total variability in the response variable without using x

Student	X = Time	Y = Distance	$\bar{y}$	$(y-\bar{y})$	$(y-\bar{y})^2$
1	2	10	16	−6	36
2	2	11	16	−5	25
3	3	12	16	−4	16
4	4	13	16	−3	9
5	4	14	16	−2	4
6	5	15	16	−1	1
7	6	20	16	4	16
8	7	18	16	2	4
9	8	22	16	6	36
10	9	25	16	9	81

Total Sum of Squares (SST)

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

$$\boxed{SST = 228}$$

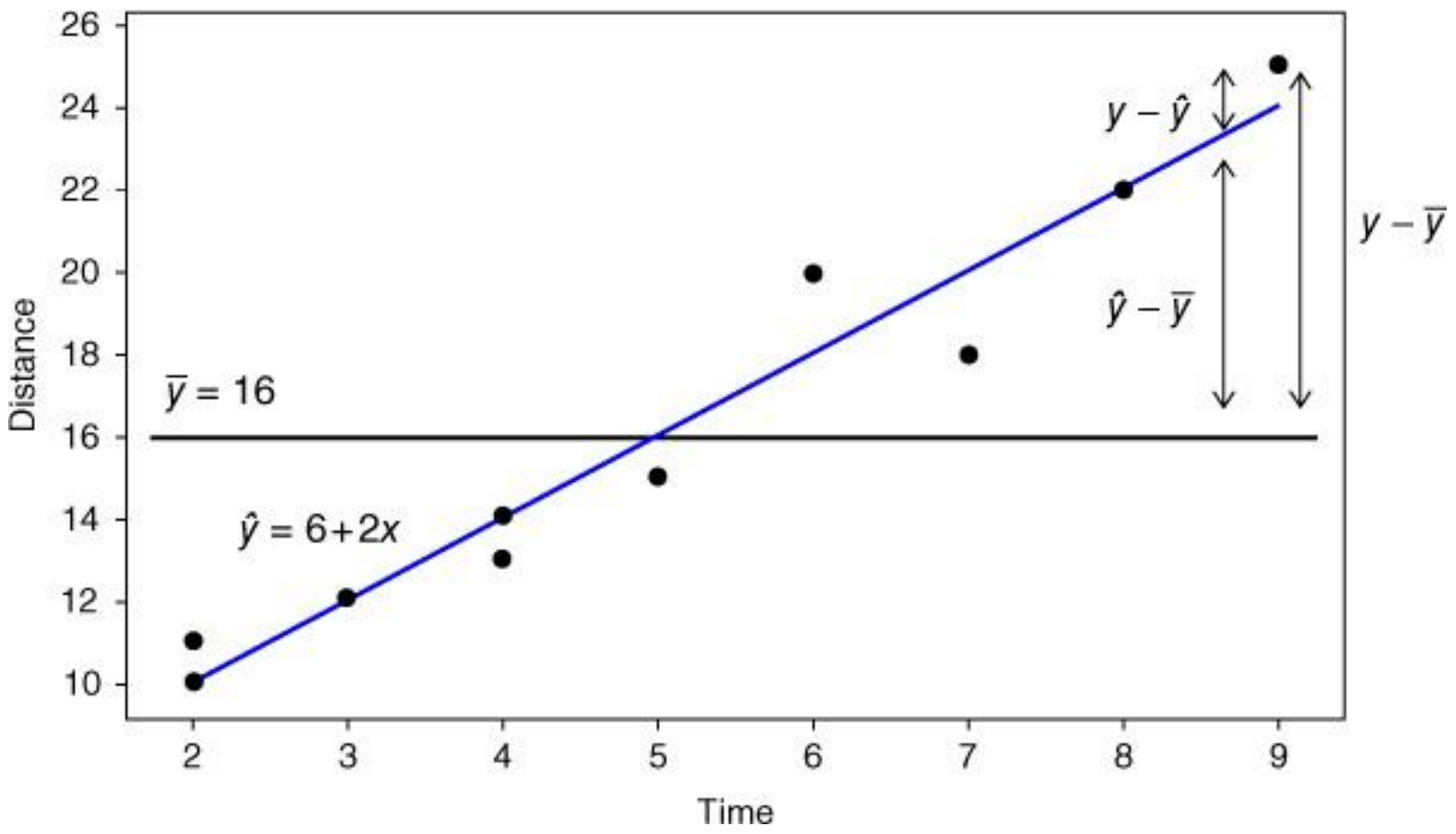
SST measures:

- Total variability in y
- Variability explained by **all sources combined**
- A **univariate** measure (ignores predictors)

HOW USEFUL IS THE REGRESSION?

- Comparing SSE and SST
  - SSE = 12 (using regression)
  - SST = 228 (ignoring predictor)
  - Since: SSE<<SST, we conclude: Using the predictor variable greatly improves prediction accuracy.
- Improvement Due to Regression: SSR
  - We now define the **Sum of Squares due to Regression** (SSR): $\boxed{\mathrm{SSR} = \sum (\hat{y} - \bar{y})^2}$
  - SSR measures:
    - The amount of variability explained by the regression
    - The improvement gained by using x

HOW USEFUL IS THE REGRESSION?



HOW USEFUL IS THE REGRESSION?

- <u>Fundamental Decomposition of Variability</u>
  - Using the identity:

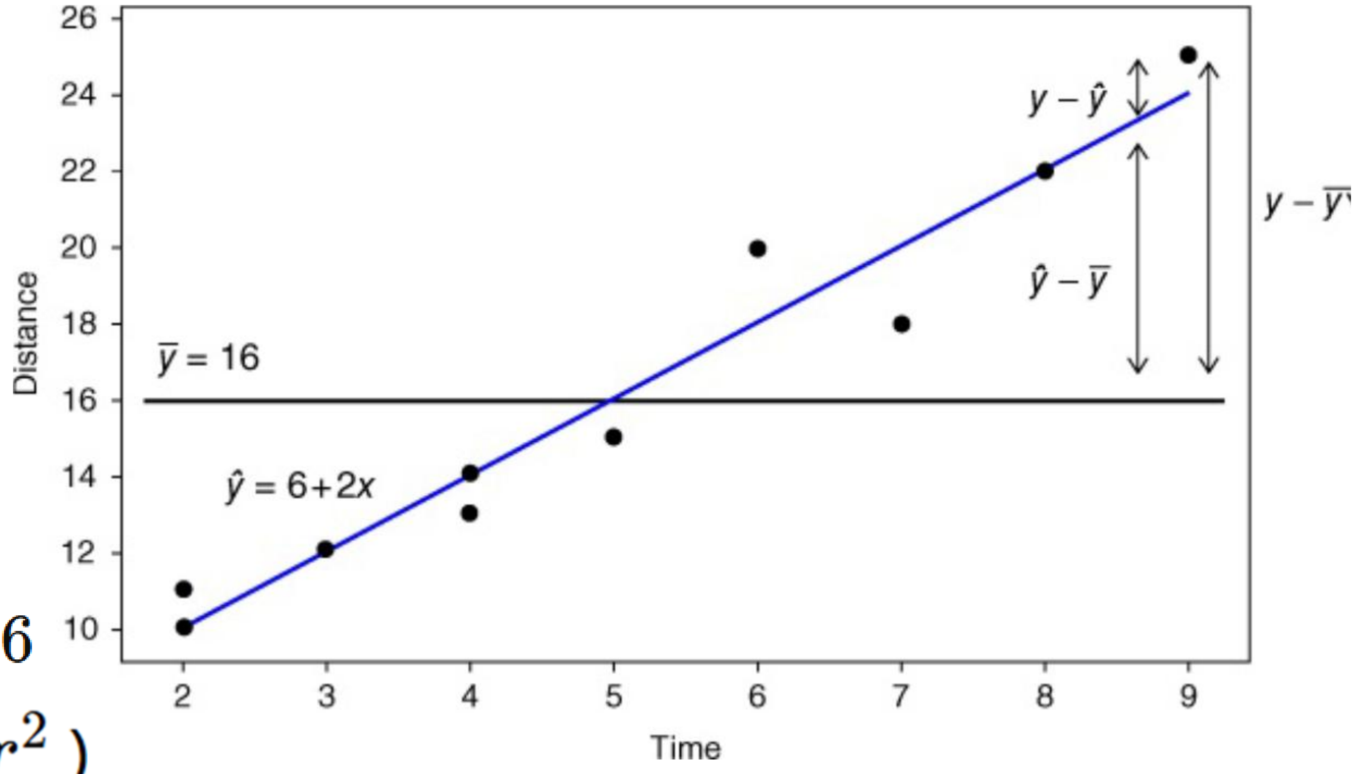
$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

  - We obtain

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

  - For this example:

$$\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE} = 228 - 12 = 216$$



  - **Coefficient of Determination** (r^2)
    - **Formula:** $\boxed{r^2 = \dfrac{\mathrm{SSR}}{\mathrm{SST}}}$

$$r^2 = \frac{216}{228} \approx 0.947$$

HOW USEFUL IS THE REGRESSION?

- **Coefficient of Determination** (r^2)
- Measures the goodness of fit of the regression as an approximation of the linear relationship between the predictor and response variables.

$$\boxed{r^2 = \frac{\text{SSR}}{\text{SST}}}$$

- represent the proportion of the variability in the y-variable that is explained by the regression

HOW USEFUL IS THE REGRESSION?

- Bounds and Meaning of (r^2)

Maximum Value

- $r^2 = 1$ when:
  - SSE = 0
  - All points lie exactly on the regression line
  - Perfect fit

Minimum Value

- $r^2 = 0$ when:
  - SSR = 0
  - Regression explains none of the variability
  - No improvement over the mean

STANDARD ERROR OF THE ESTIMATE

- In regression analysis:
  - The coefficient of determination (r^2) tells us how well the regression model fits the data in a relative sense—how much of the variability in the response variable is explained by the model.
  - However (r^2) does not indicate how accurate the individual predictions are in the original units of the response variable.
  - To assess the accuracy of predictions, we use another important statistic called the **standard error of the estimate**, denoted by s.

STANDARD ERROR OF THE ESTIMATE

- Mean Square Error (MSE)
  - To compute the **standard error of the estimate**, we first calculate the **Mean Square Error (MSE).** $$\text{MSE} = \frac{\text{SSE}}{(n - m - 1)}$$

    where:

    - SSE (Sum of Squared Errors) = $\sum (y_i - \hat{y}_i)^2$, the total squared residual error
    - n = number of observations
    - m = number of predictor variables
      - $m = 1$ for simple linear regression
      - $m > 1$ for multiple regression
  - MSE measures the average squared residual, adjusted for the number of predictors.
  - Because residuals are squared, MSE is expressed in squared units of the response variable (e.g., km^2, kg^2), which can make direct interpretation difficult.

STANDARD ERROR OF THE ESTIMATE

- **Standard Error of the Estimate (s)**
  - To return to the original units of the response variable, we take the square root of MSE, which is called the **Standard Error of the Estimate (s)**

$$s = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\mathrm{SSE}}{(n - m - 1)}}$$

  - s measures the typical prediction error
  - It estimates the average difference between observed values and predicted values
  - Smaller values of s indicate more precise predictions
  - A key advantage of s is that it is expressed in <u>the same units as the response variable</u>
  - Thus, the standard error of the estimate reflects the precision of the regression equation.

STANDARD ERROR OF THE ESTIMATE

- **Example: (sample of 10 competitors)**
  - $\mathrm{SSE} = 12$
  - $n = 10$
  - $m = 1$

$$s = \sqrt{\frac{12}{(10-1-1)}} = \sqrt{\frac{12}{8}} = \sqrt{1.5} \approx 1.2$$

- Practical Interpretation: On average, the predicted distance differs from the actual distance by about 1.2 km
- Key Takeaways
  - r^2 measures goodness of fit, not prediction accuracy
  - MSE measures average squared error but is hard to interpret directly
  - Standard error of the estimate (s) converts MSE into interpretable units
  - Smaller s values imply better predictive performance s is one of the most important diagnostic statistics in regression analysis

CORRELATION COEFFICIENT

- **The correlation coefficient *r*** (also known as the **Pearson product moment correlation coefficient**) is an indication of the strength of the linear relationship between two quantitative variables, and is defined as follows:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)\, s_x\, s_y}$$

- where Sx and Sy denote the sample standard deviations of the x- and y-values, respectively.
- Correlation Coefficient, indicates both the strength and direction of the linear relationship between two quantitative variables.

CORRELATION COEFFICIENT

- Interpreting Correlations
  - The correlation coefficient r always lies between -1 and $+1$.
  - Values of r close to $+1$ indicate a strong positive linear relationship, meaning that as x increases, y also tends to increase.
  - Values of r close to -1 indicate a strong negative linear relationship, where an increase in x is associated with a decrease in y.
  - When r is close to 0, there is little or no linear relationship, and changes in x do not systematically affect y.
  - In large datasets, even small absolute values of r may still be statistically significant, especially in data mining applications.