📌 Key Takeaways from All Regression & Correlation Topics

📌 1. Simple Linear Regression

Used when **one independent variable** predicts **one dependent variable**.
Equation:

Y=a+bXY = a + bX

- aa → Intercept (constant term).
- bb → Slope (shows change in YY for one unit change in XX).
- **Example:** Predict salary (YY) based on experience (XX).
- **Evaluation Metric: R-squared (R2R^2)**, **Mean Squared Error (MSE)**.
- **Assumption:** Linear relationship, normally distributed residuals.

📌 2. Multiple Linear Regression

Used when **two or more independent variables** predict **one dependent variable**.
Equation:

Y=a+b1X1+b2X2+...+bnXnY = a + b_1X_1 + b_2X_2 + ... + b_nX_n

- X1,X2,...XnX_1, X_2, ... X_n → Multiple independent variables.
- **Example:** Predict house price based on **bedrooms, square footage, and distance to city**.
- **Assumptions:** Linearity, No multicollinearity, Homoscedasticity, Normality of errors.

📌 3. Least Squares Regression Line (LSRL)

- **Used to find the best-fit line** that minimizes errors in prediction.
- Equation: Y=a+bXY = a + bX
- **"Least squares" minimizes the sum of squared residuals** (difference between actual & predicted values).
- **Used in both Simple & Multiple Regression.**

📌 4. Pearson Correlation Coefficient (rr)

- Measures **linear relationship strength** between two variables.

r=∑(X−X‾)(Y−Y‾)∑(X−X‾)2∑(Y−Y‾)2r = \frac{\sum (X - \bar{X}) (Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}

- **Value Range:**
  - r=1r = 1 → **Perfect positive correlation**
  - r=0r = 0 → **No correlation**
  - r=−1r = -1 → **Perfect negative correlation**
- **Example:** Relationship between study time & exam score.

📌 5. Spearman’s Rank Correlation Coefficient

- Measures **monotonic relationships** (not necessarily linear).

rs=1−6∑di2n(n2−1)r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}

- **Based on rankings instead of actual values.**
- **Best for ordinal data (ranked data, not continuous).**
- **Example:** Relationship between customer satisfaction rank & product sales rank.

📌 6. Poisson Distribution

- **Used for counting events in a fixed interval (time/space).**
- **Formula:**

P(X=k)=λke−λk!P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}

- λ\lambda = Average rate of occurrence.
- kk = Number of occurrences.
- **Example:** Number of calls received at a call center per hour.

📌 7. Normal Distribution & CDF

- **Bell-shaped symmetric distribution** where most values cluster around the mean.

- **Properties:**
  - **Mean = Median = Mode**
  - **68-95-99.7 Rule (Empirical Rule)**
    - **68%** of data within **1 standard deviation**
    - **95%** within **2 standard deviations**
    - **99.7%** within **3 standard deviations**
- **Cumulative Distribution Function (CDF):** $P(X \leq x) = \frac{1}{2} \left[1 + \text{erf} \left(\frac{x - \mu}{\sigma \sqrt{2}} \right) \right]$
  - **Example:** Heights of people, IQ scores.

📌 8. Central Limit Theorem (CLT)

- **If we take multiple random samples from any distribution, the sample means will follow a Normal Distribution, regardless of the original distribution.**
- **Formula for sample mean distribution:**

$$\mu_{\bar{X}} = \mu, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- **$\mu_{\bar{X}}$** = Mean of sample means.
- **$\sigma_{\bar{X}}$** = Standard deviation of sample means.
- **Example:** The average height of students in different classrooms will be normally distributed.

📌 9. Regression Line & Pearson Correlation Connection

- If two regression equations are: $Y = b_{yx} X + a, \quad X = b_{xy} Y + c$ Then, Pearson Correlation Coefficient is: $r = \sqrt{b_{yx} \times b_{xy}}$
- Example:
  - Regression of **Y on X**: $3X + 4Y + 8 = 0$ ⟹ $b_{yx} = -3/4$.
  - Regression of **X on Y**: $4X + 3Y + 7 = 0$ ⟹ $b_{xy} = -3/4$.
  - **Pearson's r**: $r = \sqrt{\left(-\frac{3}{4}\right) \times \left(-\frac{3}{4}\right)} = \frac{3}{4} = 0.75$
  - **Indicates strong positive correlation.**

📌 Final Summary Table

Concept	Key Takeaways
Simple Linear Regression	One independent variable predicts a dependent variable.
Multiple Linear Regression	More than one independent variable predicts a dependent variable.
Least Squares Regression Line	Best-fit line minimizes squared residuals.
Pearson Correlation (rr)	Measures linear relationship strength (−1≤r≤1-1 \leq r \leq 1).
Spearman Correlation (rsr_s)	Measures ranked (monotonic) relationships.
Poisson Distribution	Probability of discrete events in fixed intervals.
Normal Distribution	Bell-shaped curve, mean = median = mode.
CDF of Normal Distribution	Calculates probability up to a given value.
Central Limit Theorem (CLT)	Sample means follow a normal distribution, regardless of population.
Regression-Pearson Connection	r=byx×bxyr = \sqrt{b_{yx} \times b_{xy}}.

📌 Why These Topics Are Important?

✅ Used in **Data Science, Machine Learning, Finance, Marketing, Engineering**.

✅ Helps in **prediction, hypothesis testing, and decision-making**.

✅ Pearson & Spearman correlations help **understand relationships between variables**.

✅ CLT is **foundation for statistical inference**.