Regression & Correlation

1. Purpose of Regression

- Regression is used primarily for prediction.
- When no additional information is available, the mean is the best predictor.
- When an explanatory variable is available, regression improves prediction accuracy.
- It models relationships found in scatterplots.

2. Scatterplots & Correlation

- Scatterplots visualize relationships between two quantitative variables.
- Correlation coefficient (r) measures:
    - Direction of association
    - Strength of linear association
- r is based on standardized x and y values.

3. Properties of the Correlation Coefficient

- $-1 \le r \le 1$
- $r > 0$ → positive association
- $r < 0$ → negative association
- $|r|$ near 1 → strong linear relationship
- $|r|$ near 0 → weak linear relationship
- r is unit-free and unaffected by scaling.
- r measures only linear relationships.
- Correlation does not imply causation.

4. Regression Line & Least Squares

- Regression line equation: $\hat{y} = a + bx$
- Least squares minimizes the sum of squared residuals.
- Slope = $r \times (SD_Y / SD_X)$
- The regression line predicts the average value of y for a given x.

5. Predicting y from x & x from y

- Prediction is made by plugging x into the regression equation.
- Predictions depend on distance from the mean and the value of r.
- There are two different regression lines:
    - Predicting y from x
    - Predicting x from y
- These lines are not the same and cannot be reversed.

6. Normal Approximation in Regression

- When the scatter is football-shaped, y-values near a given x follow a normal distribution.
- Center of distribution = predicted value ($\hat{y}$).
- Spread = $\sqrt{(1 - r^2)} \times SD_Y$.
- Used for probability-based predictions.

7. Residuals & Diagnostic Plots

- Residual = observed y − predicted y.
- Residual plots check the validity of the regression model.
- Random scatter → good model.
- Curved pattern → nonlinear relationship.
- Fan-shaped pattern → heteroscedasticity.
- Transformations (log, square root) may fix problems.

8. Outliers, Leverage & Influential Points

- Outliers: unusual y-values.
- High leverage points: extreme x-values.
- Influential points: strongly affect the regression line.
- Influential points may not show large residuals.
- Avoid extrapolation beyond observed x-range.
- R^2 represents the proportion of variation explained by the model.

9. Data Science in Medicine – Industry Perspective

- Traditional drug discovery is slow and biology-driven.

- Data science accelerates hypothesis generation using large datasets.
- Main challenge is cultural resistance, not technology.
- Education is required for adoption in medical fields.