SECTION - A
PART - A

1a) **REGRESSION**: Regression is a statistical technique that models and analyzes the relationship between a dependent (target) variable and one or more independent (predictor) variables. The goal is to determine the strength and character of the relationship, predict futures values, and identify trends.

CORRELATION: Correlation quantifies the degree to which two variables move in relation to each other. It ranges from -1 to 1, where 1 indicates a perfect positive relationship, -1 indicates a ~~pref~~ perfect negative relationship, 0 indicates no relationship.

METHODS OF ESTIMATION OF REGRESSIONS

1. Ordinary least squares (OLS): minimizes the sum of squared residuals.
2. Maximum Likelihood Estimation: (MLE) Finds parameters values that maximize the likelihood of making the observations given the parameters.
3. Generalized Least squares (GLS): Generalizes OLS to allow for heteroscedasticity or autocorrelation.
4. Ridge Regression: Adds a penalty equal to the square of the magnitude of co-efficients to the OLS objective function to prevent overfitting.
5. Lasso Regression: Similar to ridge regression but uses an L1 penalty, which can shrink some co-efficients to zero, effectively performing variable selection.

papergrid

Date: / /

1.b) The assumptions of OLS are:

1. Linearity: The relationship between the dependent variable and the independent variables is linear.
2. Independence: Observations of the dependent variables are independent of each other.
3. Homoscedasticity: The variance of the error terms is constant across all the levels of the independent variables.
4. No Multicollinearity: Independent variables are not perfectly linearly related.
5. Normality of Errors: The error terms are normally distributed.

1.c) Detection:

1. Linearity: Use scatter plots of observed v.s predicted values or residuals vs predicted values.
2. Independence: Durbin-Watson test for auto-correlation of residuals.
3. Homoscedasticity: Plot residual vs fitted values or conduct the Breusch-Pagan Test.
4. Multicollinearity: check variance Inflation Factor (VIF); VIF values above 10 indicate significant multicollinearity.
5. Normality: Use Q-Q plots ~~or~~ conduct the Shapiro-Wilk test.

Correction:

1. Linearity: Apply ~~transitions~~ transformations (eg logarithmic, polynomial) to the values.
2. Independence: Use time series models or include lagged variables if data is time dependent.

papergrid

Date: / /

3. Homoscedasticity: Use robust standard errors, transform the dependent variables or apply weighted least squares.
4. Multicollinearity: Remove or combine correlated variables, or use regularization techniques like ridge regression.
5. Normality: Transform the dependent variable (eg log transformation) or use non-parametric method if normality cannot be achieved.

1.d) R^2 (Coefficient of Determination):
• Definition: R^2 is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model.
• Formula $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where SS_{res} is the sum of squares of the residuals
SS_{tot} is the total sum of squares.

INTERPRETATION:
• 0 to 1 Range: Ranges from 0 to 1
• An R^2 of 0 indicates that the model does not explain any of the variance in the dependent variable.
• An R^2 of 1 indicates that the model explains the variance.

USAGE:
• Goodness-of-fit: Helps assess how well the model fits the data
• Comparisons: Used to compare the explanatory power of different models.
• Limitations: Does not indicate whether a regression model is adequate; higher R^2 does not imply causation

1.e) Parametric Tests :

- Assumptions : assume that data follows a certain distribution (typically normal distribution)
- Examples :
  1. t-test : compares the means of two groups.
  2. ANOVA (analysis of variance) : compares the means of three or more groups.
  3. Pearson correlation : Measures the linear relationship between two continuous variables.

Non Parametric Tests :

- Assumptions : Do not assume a specific distribution for the data.
- Examples :
  1. Mann-Whitney U Tests : compares differences between two independent groups when the dependent variable is ordinal or continuous but not normally distributed.
  2. Kruskal - Wallis Test : compares differences between three or more independent groups on a non-normally distributed dependent variable.
  3. Spearman correlation : Measures the strength and direction of the association between two variables.

SECTION - B
PART - A

2) 1. **Probability Distributions**: A probablity distribution is a function that describes the likelihood of obtaining the possible values that a random variable can take.

TYPES OF PROBABILITY DISTRIBUTIONS:

1. Discrete Probability Distributions: where the random variable can take on a finite or countable number of values.
   - Examples.
     1. Binomial Distribution
     2. Poisson Distribution
     3. Geometric Distribution.

2. Continuous Probability Distributions: Distributions where the random variable can take on an infinite number of values within a given range.
   - Examples
     1. Normal Distribution
     2. Exponential Distribution
     3. Uniform Distribution.

2.2) Parameters of a Probability Distribution:

1. Mean(μ): The average or expected value of the random variable.
   - Impact: shifts the distribution left or right along the x-axis without altering its shape.

2. Variance (σ^2) and standard deviation (σ)

measures of the spread or dispersion of the distribution

• Impact: variance & standard deviation determine the ~~d~~ width of the distribution. A larger variance / standard deviation result in a wider, flater dist, while a smaller variane/standard deviation results in narrower, taller dist.

3. Skewness: measures the assymetry of the distribution.

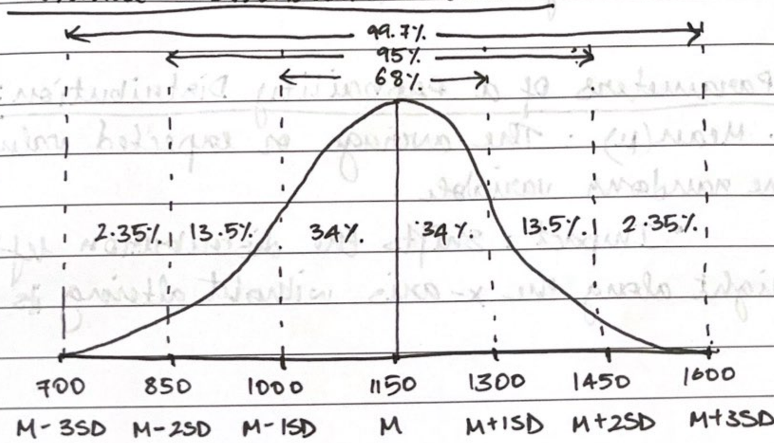
• Impact: positive skewness indicates a longer right tail, while negative skewness indicates a longer left tail.

4. Kurtosis: measures the "tailedness" of the distribution.

• Impact: High kurtosis indicates heavy tails and a sharper peak, while low kurtosis indicates light tails and a flatter peak.

2) 3. Total Area Under the Normal Distribution Curve: The total area under the normal distribution curve is 1

Normal Distribution curve.



2.4) • Mean ± 1 standard deviation: Aprox 68.27% of the data falls ~~in~~ within this range.

• Mean ± 2 standard deviation: Appox 95.45% of the data falls in this range.

• Mean ± 3 standard deviation: Appox 99.73% of the data falls within this range.