PAGE NO.
DATE: / /

NAME: Rohan Arpit Dungdung
VID: V01106712
SCMA 632 : Statistical Analysis and Modelling
Date: 11/07/2024

SECTION - A

1
ans a) Regression can be defined as a statistical technique used to model the relationship between a dependent variable and one or independent variables. The goal is ~~to predict~~ find out the prediction. Whereas Correlation ~~coefficient~~ measures the strength and direction of a linear relationship between two variables. The coefficients ranges from ~~-1~~ -1 to +1
The different methods of estimation of regressions are: ① Ordinary Least Squares
② Maximum Likelihood estimation
③ Lasso Regression
④ Robust regression
⑤ Non linear regression
⑥ Bayesian regression
⑦ Partial least squares.

ans. 1(b) The assumptions of fitting a regression by OLS are:-

(1) Linearity - relationship between dependent variable and the independent variables should be linear.

(2) Independence - observations should be independent of each other.

(3) Homoscedasticity - residuals should have constant variance at every level of independent variables.

(4) No autocorrelation: residuals should not be correlated to each other.

(5) Exogeneity - independent variables should be uncorrelated with the error term.

ans. 1(c) For linearity, it is detected through scatter plots and residual plots. and in order to correct them, we need to apply transformations such as square root to the variables.

For independence, it is detected through Durbin-Watson test, they detect autocorrelation in residuals especially in time series data. and for correcting we add lagged terms of the independent and dependent variables.

for Homoscedasticity, it is detected through Breusch-Pagan Test and it is corrected through Weighted least squares.

for non autocorrelation, it is detected through Ljung-Box test and it is corrected through autoregressive integrated moving average models.

for exogeneity it is detected through Durbin-Wu-Hausman Test and it is corrected through instrumental variable

1(d) R^2 of a regression is a statistical measure which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is used for model evaluation, comparing models. In terms of value range $R^2 = 1$ means perfect fit whereas $R^2 = 0$ means no fit. High R^2 indicates a good fit, Low R^2 indicates a poor fit. Example of high R^2 is predicting the house prices based on no. of bedrooms example of low R^2 such as predicting students' grades based on hours he might have studied.

PAGE NO.
DATE : / /

1(g) ans Parametric tests are statistical tests that make certain assumptions about the parameters of the population distribution from which the sample is drawn whereas non-parametric tests are also known as tests that do not rely on assumptions about the parameters of the population distribution. These are used when assumptions of parametric tests cannot be met. They are more flexible and can be used with ordinal data. For examples t-test, where we compare means of 2 groups, ANOVA, used to compare three or more groups. Examples of non parametric are Kruskal Wallis Test, Mann-Whitney U test.

Signature .....

SECTION-B

ans 1(a) Probability distribution is a mathematical function that describes the likelihood of different outcomes in a random experiment. It specifies the probability of each possible outcome. Types of probability distribution are:

(i) Discrete probability distribution
Example: Binomial, Poisson, geometric

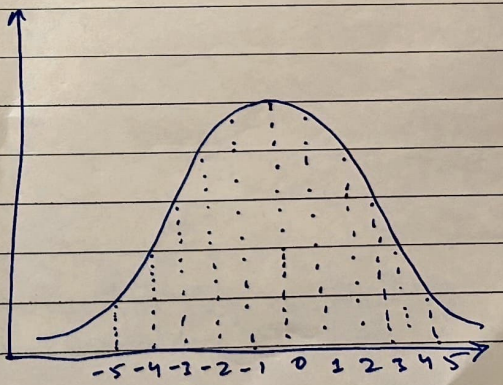
(ii) Continuous probability distribution
Example: Gaussian distribution,
Exponential distribution
Uniform distribution

ans 1(b) Parameters of probability distribution are mean, variance and standard deviation, skewness, Kurtosis. Impact of mean is that in discrete distributions the mean represents the average outcome of random variable. Impact of variance indicates that the data points are spread out over a wider range of values.

Signature .....

Impact of skewness is that tail on the right side of the distribution is longer than the left side.
Impact of Kurtosis is that it indicates heavy tails and a sharp peak.

ans 1(c) The total area under the normal distribution curve is 1. This represents that probability of all possible outcomes for a normally distributed random variable sums to 1.



Signature .....

PAGE NO.
DATE : / /

ans 1(d) Mean ± 1 falls under approx 68 % of the data falls within one standard deviation of the mean.

Mean ± 2 falls approx 95 % of the data within two standard deviations of the mean.

Mean ± 3 falls under 99.7 % of the data within 3 standard deviations of the mean.