Anekant Education Society's

Tuljaram Chaturchand College of

Arts, Commerce and Science, Baramati

(Autonomous)

QUESTION BANK

FOR

M.Sc SEM-II

STATISTICS

PAPER: STAT- 4204

Regression Analysis - 4 Credit

(With effect from June 2019)

Q.1 A) Choose the correct alternative of the following: **(1 each)**

i) Consider the following statements

I) The eigen values of hat matrix are -1 and 1

II) Cov $\left(\hat{\beta}_0, \hat{\beta}_1\right) = 0$

a) I and II are true   b) I true and II false

c) I false and II true   d) I and II are false

ii) The least squares estimator of the regression coefficient in no intercept model is

a) $\dfrac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i^2}$ b) $\dfrac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$ c) $\dfrac{\sum y_i x_i}{\sum y_i}$ d) $\dfrac{\sum y_i x_i^2}{\sum x_i}$

iii) The hat matrix $H = X\left(X'X\right)^{-1}X'$ is

a) Symmetric and idempotent matrix

b) Symmetric and orthogonal matrix

c) Skew symmetric matrix

d) None of these

iv) Any model that is not linear in the unknown parameters is a _____ regression model.

a) linear   b) non-linear   c) multiple linear   d) none of these

v) The sum of the residual in any regression model that contains an intercept β_0 is always

a) greater than zero   b) zero   c) one   d) less than zero

vi) Logistic regression model is an appropriate when the response variable is distributed as

a) Normal   b) Poisson   c) Gamma   d) Binomial

vii) The multicollinearity problem in a linear regression concerns with

a) error terms   b) predictor variables   c) response variable   d) none of these

viii) The range of partial correlation coefficients is

a) 0 to 1   b) 0 to ∞ c) -1 to 1   d) - ∞ to ∞

ix) The hat matrix $H = X(X^{'}X)^{-1}X^{'}$ is

a) Orthogonal b) Symmetric c) Skew Symmetric d) Upper triangular

x) The problem of multicollinearity is related to

a) errors b) regressors c) response variable d) both errors and regressors

xi) Consider the following statements:

III) The coefficient of determination R^2 may be interpreted as the proportion of total variability in the response variable Y that is accounted for by the predictor variable X.

IV) R^2 is defined as $R^2 = 1 - \frac{SST}{SSE}$

where SST = Total sum of square

SSE = Sum of squares due to error

a) I is true and II is false b) I is false and II is true

c) I and II are false d) I and II are true

xii) The range of multiple correlation coefficient is

a) -1 to 1 b) - ∞ to ∞ c) 0 to ∞ d) 0 to 1

xiii) The least square estimator of the slope in no intercept regression model is

a) $\frac{\sum_{i=1}^{n} y_i x_i}{\sum x_i^2}$ b) $\frac{\sum_{i=1}^{n} x_i y_i}{\sum y_i^2}$ c) $\frac{\sum_{i=1}^{n} y_i x_i^2}{\sum y_i^2}$ d) $\frac{\sum_{i=1}^{n} y_i^2 x_i}{\sum x_i^2}$

xiv) Autocorrelation is the problem related to

a) regressors b) responses c) errors d) both regressors and responses.

xv) In logistic regression model with single covariate the odds ratio ψ is related to the regression coefficient β_1 is

a) $\psi = e^{\beta_1}$ b) $\psi = \beta_1$ c) $\psi = \ln \beta_1$ d) $\psi = e^{\beta_0}$

xvi) The sum of the residuals in any model with intercept β_0 is always

a) one b) zero c) greater than zero d) none of these

xvii) The least square estimator for the multiple linear regression model $Y = X\beta + \in$ can be expressed as

a) $\beta + (X^{'}X)^{-1}X^{'} \in$ b) $\beta + (X^{'}X) \in$ c) $\beta + X \in$ d) $\beta + (X^{'}X)^{-1} \in$

xviii) The hat matrix $X\left(X'X\right)^{-1}X'$ is

a) idempotent b) skew symmetric c) both a and b d) orthogonal

xix) The range of multiple correlation coefficient is

a) 0 to 1 b) -1 to 1 c) 0 to ∞ d) -∞ to ∞

xx) Autocorrelation is the problem related to

a) regression b) errors c) response variable d) both a and c

B) State whether the following statements are True or False. [1 each]

i) Logistic regression is used when the response variable is dichotomous

ii) $|x'x| = 0 \Rightarrow$ multicollinearity is present among repressors.

iii) Autocorrelation is the problem related to response variable.

iv) F test is used to test the significance of a individual regression coefficients in the multiple linear regression model.

v) The range of partial correlation coefficient is [0, 1].

vi) The log transformation is suitable for linearizing the function $y = \beta_0 e^{\beta_1 x}$.

vii) Normality assumption is not required for obtaining prediction interval.

viii) Residuals are useful in detecting outliers in response.

ix) The model $y = \beta_0 X^{\beta_1} \in$ can be linearized by using square root transformation.

x) In case of simple linear regression model the least squares estimator $\hat{\beta}_1$ is an unbiased estimator of β_1

xi) The model $y = \theta_1 e^{\theta_2 x} + \varepsilon$ is not intrinsically linear.

xii) A horizontal regression line has no slope.

xiii) Mean sum of square due to residual is not biased estimator of σ^2 in simple linear regression.

xiv) The hat matrix is idempotent but not symmetric.

xv) The model $y = \beta_0\, e^{\beta_1 x}\varepsilon$ is intrinsically linear model.

xvi) The R^2 indicates the proportion of variability around $\bar{y}$ explained by regression.

xvii) The ridge estimator $\hat{\beta}_R$ is unbiased.

xviii) The weighted least squares estimator is $\left(X'X\right)^{-1} X'WY$

xix) Variance inflation factors are useful in detecting autocorrelation.

xx) The hat matrix H is skew symmetric matrix.

xxi) A generalized linear model with log-link function is the classical linear model.

Q.2 Define the following terms with illustration: **(2 each)**

i) Estimation space
ii) Polynomial regression
iii) Logit transformation
iv) Residual
v) Conditional indices and conditional number of $X^{'}X$ matrix.
vi) Link function.
vii) Logit transformation.
viii) Hat matrix.
ix) Residuals
x) Leverage point
xi) Polynomial regression
xii) Multiple correlation coefficient.
xiii) Hat matrix
xiv) Link function
xv) Studentized residual
xvi) Odds ratio
xvii) Conditional indices and conditional number.
xviii) Link function.
xix) Partial correlation coefficient.
xx) Model Deviance.

Q.4 Questions for 4 marks.

Unit 1 :

1) State steps used in regression analysis.
2) Explain why R^2_{adj} is better than R^2.
   a) Interpret the value of $R^2 = 0.95$.
   b) Interpret the regression coefficients in the following fitted model
      Height = 87.88 + 2.464 width.
3) Prove or disprove: "The sum of residuals in any regression model that contains intercept β_0 is always zero."
4) Consider simple linear regression model with first order autocorrelated errors. How will you estimate parameters β_0 and β_1 in this model.
5) Distinguish between R^2 and adj R^2
6) Explain Weighted lest squares method in simple linear regression model.
7) Define simple regression model stating all assumptions. Also derive the least squares estimator of intercept and slope.

Unit 2 :

1) Describe the test procedure to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ v/s
   H_1: at least one $\beta_j \neq 0$. Where β_j' s are regression coefficients.
2) Derive the expression for Mallow's C_p Statistic.
3) Define estimation and error space in the contest of multiple linear regression model. Also show that estimation space and error space are orthogonal to each other.
4) With usual notation show that $e = (I – H) \varepsilon$
5) Define variable selection problem in regression analysis. Describe stepwise procedure used in subset selection.
6) Describe the backward elimination method for the variable selection in regression.
7) With usual notations outline the procedure of testing a general linear hypothesis.
8) Obtain 100 $(1 - \alpha)$ % confidence interval of the mean response in the multiple linear regression.
9) How to interpret regression coefficients in multiple linear regression model with k regressor.

Unit 3 :

1) What is the need of transformation of variables in regression analysis. Also state transformation used in regression analysis.

2) What is data transformation? Describe Box-Cox method of transforming the response variable.
3) What are consequences of multicollinearity on least squares estimates?
4) Describe the detection of multicollinearity using variance inflation factors.

Unit 4 :

1) Define canonical link in GLM. Also state common canonical links for generalized linear model.
2) Explain Wald test to test individual model coefficients.
3) Explain link functions and its role.
4) Explain how the odds-ratio is related to the parameter β in a logistic regression with single covariates.
5) Write a note on Generalized Linear Model.
6) What is polynomial regression?
7) Explain the concept of inverse regression.
8) Define logistic regression model. Give real life situation where this regression model can be used.
9) Write a note on interpretation of the parameters in a logistic regression model.
10) Discuss generalized linear model.
11) Give interpretation based on odds ratio.

Q.5 Questions For <u>Seven</u> Marks

Unit 1 :

1) Define regression through origin. Obtain least squares estimator of regression coefficient in regression through origin model. Also state 100 (1- α) % confidence interval on regression coefficient.
2) Explain lack of fit test in detail.
3) Explain plot of residuals against fitted.
4) Explain
   i) Homoscadasticity
   ii) Weighted least square
5) Consider the simple linear regression model

$$y = \beta_0 + \beta_1 x + \in, \quad E(\in) = 0 \quad V(\in) = \sigma^2 < \infty$$

Find: i) $E\left(\hat{\beta}_1\right)$ ii) $E\left(\hat{\beta}_0\right)$ iii) $V\left(\hat{\beta}_1\right)$ iv) $V\left(\hat{\beta}_0\right)$.

6) Discuss normal probability plot.
7) Obtain confidence interval on β_0, β_1 and σ^2 in simple linear regression model.
8) Write a note on measuring the quality of fit of a simple linear regression model.
9) Explain simple linear regression model and no intercept regression model with suitable examples.
10) Consider the model $y = \beta_0 + \beta_1 x + \in$ with $E(\in) = 0 \; and \; \mathrm{var}(\in) = \sigma^2 < \infty$ then show that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators of β_1 and β_0 respectively. Also find var ($\hat{\beta}_0$) and var ($\hat{\beta}_1$) where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least square estimators of β_0 and β_1 respectively.
11) Explain the following terms in the context of regression.
    i) Heterosecdasticity
    ii) Autocorrelation
    iii) Function belonging to error
12) Define null distribution. Derive the null distribution of simple correlation coefficient.
13) Explain a formal test for the lack of fit in regression analysis.
14) Explain a formal test for the lack of fit in regression analysis.
15) Define simple regression model. Derive the least squares estimators of intercept and slope.
16) Describe the procedure of testing of hypothesis about parallelism (slope) and equality of intercepts with reference to simple linear regression.

17) What are the uses of residual plots? Write a note on normal probability plot and plot of residuals against the fitted values.
18) Explain test for lack of fit in detail.
19) Show that the criteria of minimum MS_{Res} and maximum adjusted R^2 equivalent.

Unit 2 :

1) State assumptions used in multiple linear regression model. The following graphs are used to verify some assumptions of the ordinary least squares regression of Y on $x_1, x_2, \dots x_p$:
   i) The scatter plot of Y versus each predictor x_j, $j= 1, 2, \dots p$
   ii) Scatter plot matrix of the variables $x_1, x_2, \dots x_p$
   iii) The residuals versus fitted values for each of the above graphs what assumption can be verified by the graph.
2) Define model deviance. Describe the test procedure to test model adequacy based on model deviance.
3) Describe a linear regression model with k variables. Write the model in matrix form. State all assumptions. Obtain the least square estimates of regression coefficients.
4) For a multiple linear regression model, obtain with stating required assumptions.
   i) A confidence interval for the mean response at particular point $x_0 = [x_{01}, x_{02}, \dots x_{0k}]$
   ii) A confidence interval for the regression coefficients β_j, $j = 1, 2, \dots, k$
5) State and prove Gaurs-Markov theorem.
6) Define Mallow C_p. Derive the same.
7) Describe forward selection method for variable selection in linear regression.
8) Define multiple linear regression model. Explain the least squares method to estimate parameters in multiple linear regression model.
9) Write a note on tests on individual regression coefficients.
10) State and prove any two properties of least squares estimators in case of multiple linear regression model.
11) Derive expression for Mallow's Cp Statistic.
12) Define variable selection problem in regression analysis. Explain backward elimination method to select appropriate regressors in the model.
13) Define model deviance. Write a note on testing of hypotheses on subset of parameters using deviance.Define the multiple linear regression model. State all assumptions involved in it. With usual notations show that $\hat{\beta} = (x^1x)^{-1} x^1y$.

14) Define model deviance. Describe the test procedure based on model deviance to test H_0 : fitted model is adequate v/s H_1 : fitted model is not adequate in logistic regression.
15) Derive the least squares estimators of regression coefficients in multiple regression model. Also describe a procedure to test the hypothesis $H_0 : \underline{\beta} = \underline{0}$ V/s $H_1 : \underline{\beta} \neq \underline{0}$.
16) Explain variable selection problem in regression. Derive the Mallow's C_p-Statistic.
17) What is the need of transformation in regression analysis. Also state transformations used in regression analysis.
18) Write a note on 'Analysis of Variance for Significance of Regression in Multiple Linear Regression'. Derive the expression for Mallow's C_p Statistic.

Unit 3 :

1) How multicollinearity will be detected on the basis of following methods:
   i) Variance Inflation Facto ii) Eigen system Analysis of x'x matrix
2) Derive the null distribution of sample multiple correlation coefficient.
3) Explain the problem of multicollinearity in the connection with linear regression model. Discuss its consequences on least square estimates.
4) What is autocorrelation? Derive the Durbin Watson test. What are its limitations?
5) Derive the null distribution of sample correlation coefficient.
6) Discuss various methods of detecting multicollinearity.
7) Define multiple correlation coefficient and partial correlation coefficient. State and prove the relationship between them in case of n variables.
8) Derive the null distribution of simple correlation coefficients.
9) Write short note on the following:
   i) Box-Cox power transformation ii) Weighted least squares method.
10) Explain the concept of autocorrelation in regression analysis. Describe the Durbin-Watson test to determine there is positive autocorrelation in the errors.
11) Define the term multicollinearity. Explain the effects of multicollinearity.
12) Derive the null distribution of sample multiple correlation coefficient.
13) Define problem of multicollinearlity. Explain any two methods of detecting multicollinearity in the data.
14) Define partial correlation coefficient and multiple correlation coefficient. Derive the relationship between them.
15) What is multicollinearity. Discuss various sources of multicollinearity.
16) Describe the test of significance related to
    i) Simple correlation coefficient
    ii) Multiple correlation coefficient
    iii) Partial correlation coefficient

Unit 4 :

1) Define non linear and intrinsically linear models. Examine whether the following models are intrinsically linear or not.

i) $y=\theta_1 e^{\theta_2 x}+\varepsilon$ ii) $y=\theta_1 e^{\theta_2 x}\varepsilon$

2) Explain
   i) Generalized linear model
   ii) Non- linear regression model
3) Discuss least squares method for parameter estimation in non- linear regression model with suitable example.
4) Define the logistic regression model. Derive the Likelihood Ratio test for it.
5) Define non linear and an intrinsically linear model. Give one example of each.
6) Describe linearization technique for the estimation of parameters in non linear regression model.
7) Define logistic regression model. Derive Maximum Likelihood Estimator for logistic regression model with single covariates.
8) Derive the likelihood ratio test for testing of the coefficients of logistic regression model with single covariate.
9) Define generalized linear model (GLM). Derive the score equations.
10) Explain the following:
    i) Studentized residuals
    ii) Standardized residuals
    iii) Press residuals
11) Distinguish between linear and non linear regression models. Write a note on linearization of the non linear function.
12) Define non-linear model and intrinsically linear model. Examine the following models are intrinsically linear or not.

    i) $y=\theta_1 e^{\theta_2 x}\in$

    ii) $y=\theta_1 e^{\theta_2 x}+\in$
13) What is a non-linear regression model. Describe the non-linear least square method for estimation of parameters in non-linear regression model.
14) Define logistic regression model. Derive the likelihood ratio test with reference to logistic regression.
15) Define generalized linear model (GLM) and obtain maximum likelihood estimate of parameters of GLM.
16) Define exponential family of distribution. Show that following families are member of exponential family (i) Binomia (ii) Poisson (iii) Normal.

17) Write short note on

i) Forward selection method for variable selection.

ii) Box-Cox power transformation.

iii) Polynomial and inverse regression

18) Explain the following terms:

i) logit transformation

ii) odds ratio

iii) probit transformation