Bangladesh-Bharat Digital Service and Employment Training

Test-2

Total marks: 100 Time: 2 hours

Instructions:

- Answering all the questions are mandatory.
- **Part-I** contains 20 MCQ questions, each question has 2 marks.
- **Part-II** contains 6 problem solving questions, each question has 10 marks.

Part-I: Choose the correct option and justify your answer with one or two sentence(s) *(20 X 2 = 40)*

1. Which of the following statement(s) is(are) NOT true?
   a) Pearson's correlation analysis is applicable to only numeric data.
   b) Spearman's correlation analysis is applicable to only ordinal data.
   c) $\chi 2$ correlation analysis is applicable to only categorical data.
   d) Any non-parametric statistical learning approach is applicable when the entire population is known.

Correct Answer: b
Explanation:
Spearman's correlation analysis is applicable to both ordinal and numerical data because in both the cases, the rank of data can be calculated.

2. The value of correlation coefficient (r) lies between?
   a) 0 to 1 b) -1 to 1
   c) $-\alpha$ to $+\alpha$, where α is any value d) 1 to 5

Correct Answer: b
Explanation:
The value of correlation coefficient (r) lies between -1 to 1

3. Which of the following three cases depicts 'negative correlation' between the two variables X and Y?

   a) Case 1 (Plot in the left) b) Case 2 (Plot in the center)
   c) Case 3 (Plot in the right) d) None of the above plots

Correct Answer: b
Explanation:
The left plot = Positive correlation
The central plot = Negative correlation
The right plot = No correlation (As there is no effect on the value of variable Y, as value of variable X changes)

4. In an Auto-regression model for forecasting, the number of lags used as regressors is called the?
   a) order of auto-regression b) degree of auto-regression
   c) freedom of auto-regression d) None of the above

Correct Answer: a
Explanation:
In the case of Auto-Regression Model for Forecasting, the number of lags used as regressors is called the order of auto-regression.

5. In a regression analysis if the coefficient of determination $R^2 = 1$, then sum of squares of the errors (SSE) must be equal to?

a) 1
b) 0
c) Any positive value
d) Infinity

Correct Answer: b

Explanation:

The relationship between the coefficient of determination and SSE is given by

$R^2 = 1 - \frac{SSE}{SST}$

If $R^2 = 1$ then $SSE = 0$.

6. In order to find out the correlation between an independent variable X and a dependent variable Y, following information is available.

$$\sum(Y_i - \bar{Y})(X_i - \bar{X}) = 498, \sum(X_i - \bar{X})^2 = 338, \sum(Y_i - \bar{Y})^2 = 1212$$

What is the value of Karl Pearson's coefficient of Correlation between X and Y?

a) -0.78
b) 0.78
c) 0.55
d) -0.55

Correct Answer: b

Explanation:

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} = \frac{498}{\sqrt{338 \times 1212}} = \mathbf{0.778}$$

7. In If the difference between ranks of ith pair of the two variables is given by d_i, and total number of pairs of observations is n, then the Spearman's rank correlation coefficient is given by

a) $r_s = \frac{6\sum d_i^2}{n(n^2+1)}$
b) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$
c) $r_s = 1 + \frac{6\sum d_i^2}{n(n^2+1)}$
d) $r_s = \frac{6\sum d_i^2}{n(n^2+1)}$

Correct Answer: b

Explanation:

The rank correlation can be defined as

$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

where $d_i = Difference\ between\ ranks\ of\ ith\ pair\ of\ the\ two\ variables$

$n = Number\ of\ pairs\ of\ observations.$

8. The square of the correlation coefficient r, that is, r^2 will always be positive and is called?

a) Regression coefficient
b) Coefficient of determination
c) Covariance
d) Confidence level

Correct Answer: b

Explanation:

The square of the correlation coefficient is called the coefficient of determination.

9. A simple linear regression model of the form $Y = a + bX$ is used to compute the relationship between the variables X and Y. Suppose there are n sample points, $(x_i, y_i), i = 1,2, \ldots, n$, and $\bar{x}$ and $\bar{y}$ are their corresponding means. The value of linear regression model coefficient b is given by?

a) $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
b) $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
c) $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$
d) $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Correct Answer: a
Explanation:

The value of linear regression model coefficient b is given by $b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

10. A study is conducted to find the relationship between the number of hours spent in physical exercise and passing the fitness examination. Data is collected for a total of 10 Indian army aspirants, and shown in the following table.

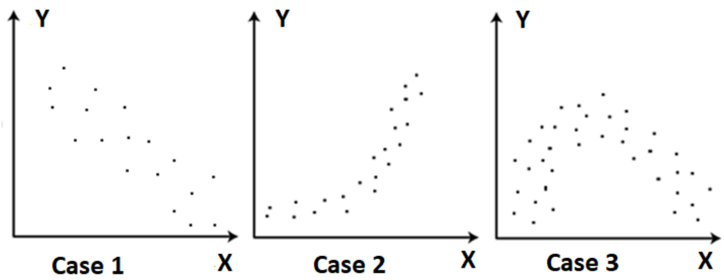
Hrs. exercise	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
Pass/ Fail	F	F	F	P	F	F	P	P	P	P

To study how does the number of hours spent in physical exercise affects the probability of the aspirant passing the fitness test, among below, which kind of analysis is most suitable?

a) Multiple non-linear regression
b) Multiple linear regression
c) Binary logistic regression
d) Multinomial logistic regression

Correct Answer: c
Explanation:
'P' and 'F' can be represented as '1' and '0'. The reason for using logistic regression for this problem is that the values of the dependent variable, P or F, while represented as '1' and '0', are not cardinal numbers. If the problem was changed so that P/F was replaced with the marks obtained in the test, (like from 0 to 100), then simple linear regression can be used.

11. Which of the following three cases depicts a 'non-monotonic relationship' between the two variables X and Y?



a) Case 1 (Plot in the left)
b) Case 2 (Plot in the center)
c) Case 3 (Plot in the right)
d) None of the above plots

Correct Answer: c
Explanation:
Case 1: As variable X increases, Y decreases monotonically: - Monotonic relationship
Case 2: As variable X increases, Y increases monotonically: - Monotonic relationship
Case 3: As variable X increases, variable Y initially increase. But after a certain value of X, Y decreases with the increase of X: - Non-monotonic relationship

12. In regression analysis, the variable that is being predicted is called as?

a) Response
b) Regressor
c) Independent variable
d) Dependent variable

Correct Answer: a, d
Explanation:
In regression analysis, the variable that is being predicted is called as Response or Dependent variable.

13. What is(are) the required assumption(s) for the auto-regression analysis?
   a) The time series under consideration is nonstationary
   b) The time series under consideration is non-uniform
   c) The time series under consideration is stationary, but not uniform
   d) The time series under consideration is both stationary and uniform

Correct Answer: d
Explanation:
Auto Regression analysis assumes that the time series under consideration is both uniform and stationary.

14. If the sample data in a $\chi 2$ test contains m rows and n columns, then the degree of freedom will be
   a) $m \times n$
   b) m
   c) $(m-1) \times (n-1)$
   d) $(m \times n - 2)$

Correct Answer: c
Explanation:
The $\chi 2$ statistics tests the hypothesis that A and B are independent. The test is based on a significance level, with (n-1) ×(m-1) degrees of freedom., with a contingency table of size n*m.

15. SST represents
   a) The error in the fitted model.
   b) The proportion of variability of the fitted model.
   c) The variation in the response values.
   d) The coefficient of determination.

Correct Answer: c

Explanation:

SST represents variation in response values. $SST = \sum_{i=1}^{n}(y - \bar{y})^2$

16. If a regression model has more than one independent variable with linear equation, then it is called
   a) Auto regression model.
   b) Linear regression model.
   c) Multiple linear regression model.
   d) Multiple non-linear regression model.

Correct Answer: c
Explanation:
Multiple regression models consist more than one independent variable and is linear in coefficient.

17. Which regression model can be used for time series data?
   a) Multiple non-linear regression.
   b) Simple linear regression.
   c) Auto-regression.
   d) Simple non-linear regression.

Correct Answer: c
Explanation:
Regression analysis for time-ordered data is known as Auto-Regression Analysis

18. The second lag of y_t in auto-regression is denoted as
   a) $y_{(t-1)}$
   b) $y_{(t-3)}$
   c) $y_{(t-n)}$
   d) $y_{(t-2)}$

Correct Answer: d
Explanation:
Lags are where results from one time period affect following periods.

19. How many model parameters are to be learned when a simple non-linear regression model is to be built with a training set of size n?

a) n
b) 2
c) 3
d) Cannot be told

Correct Answer: c
Explanation:
It depends on non-linearity, that is, degree of the regression model. For a simple linear model, the parameters are b0 and b1, that is, two parameters. For a simple non-linear model, it should be at least 3, for example, $y = a + bx + cx^2$.

20. For a given dataset, to compute the relationship between the variables x and y, following two regression models are obtained.

Model 1: $Y = \beta_2 X^2 + \beta_1 X + \beta_0$, with $R^2\ score = 0.68$

Model 2: $Y = \alpha_3 X^3 + \alpha_2 X^2 + \alpha_1 X + \alpha_0$ with $R^2\ score = 0.87$

Which model is more acceptable?

a) Model 1
b) Model 2
c) Both models are equally acceptable
d) None of the model is acceptable

Correct Answer: b
Explanation:
The model that has higher correlation with training data, is more acceptable.

Part-II: Solve the following problems *(10 X 6 = 60)*

1. Consider the table given below:

X	-4	-3	-2	-1	0	1	2	3	4
Y	-10	-8	-4	-2.5	0	2.5	4	8	10

Find Karl Pearson correlation in this data. (10)

Answer:

We know, $r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}$

Here, $\bar{X} = 0, \bar{Y} = 0$

$\therefore r = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}} = \frac{40+24+8+2.5+0+2.5+8+24+40}{\sqrt{16+9+4+1+0+1+4+9+16} \times \sqrt{100+64+16+6.25+0+6.25+16+64+100}} = \frac{149}{\sqrt{60}\sqrt{372.5}} = \frac{149}{149.5} = \mathbf{0.997}$

2. The marks for 8 students on mid-term and end-term examinations in Data Analytics course are given in table below

Mid-term	82	73	95	66	84	89	51	82
End-term	76	83	89	76	79	73	62	89

a) Obtain the simple linear regression analysis to predict the score on the end-term examination from the mid-term examination score. (1+2+3)
b) It is suggested that if the regression is significant, then there is no need to have final examination. How you test the significance level of your regression analysis? (2+2)

Answer:

a) Simple linear regression model to predict the marks of end-term scores takes the following form:

Assume, End-term= Y and Mid-term = X

So, the simple LR model to predict the marks of end-term score looks like

$Y = \beta X + \alpha$

Expression for the model parameters are:

$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

$\alpha = \overline{y} - \beta\overline{x}$

Calculated value of the model parameters:

$\overline{x} = 77.75,\ \overline{y} = 78.375$

$\beta = 0.44$

$\alpha = 78.375 - (0.44 \times 77.75) = 44.135$

$\therefore Y = 44.135 + 0.44X$

b) The validity of the model can be done as follows:

SSE= Residual sum of the squared error

$= \sum_{i=1}^{n}(actual\ output - predicted\ output)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 289.12$

SST= Total corrected sum of squares

$= \sum_{i=1}^{n}(actual\ output - average\ of\ the\ output)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 555.88$

$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{289.12}{555.88} = 1 - 0.52 = 0.48$

The regression is not significant.

3. Happiness Index (HI) is measured as low (L), medium (M), high (H) and very high (VH). A survey is conducted among a population of varied age groups and data observed are recorded in table given below.

Age-group	80-90	90-100	10-20	20-30	30-40	40-50	50-60	60-70	70-80
HI	H	VH	VH	VH	M	L	L	M	H

a) Which correlation analysis is applicable to check if there is any correlation exists between age-group and happiness index. (2)

b) Calculate the coefficient of determination and interpret your result. (3+2+2+1)

Answer:

a) For the given data, Spearman correlation analysis is applicable. The sample data are of ordinal type. And for ordinal data, the Spearman Correlation analysis is applicable.

b) **Calculation of coefficient of deamination:**
The contingency table form the given data

Sample#	$Rank_x$	$Rank_y$	Diff=d	d^2
1	2	4.5	-2.5	6.25
2	1	2	-1	1
3	9	2	7	49
4	8	2	6	36
5	7	6.5	0.5	0.25
6	6	8.5	-2.5	6.25
7	5	8.5	-3.5	12.25
8	4	6.5	-2.5	16.25
9	3	4.5	-1.5	2.25

Calculation of coefficient of correlation:

$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{6*119.5}{9*80}$ = **0.00416**

Calculation of coefficient of determination:

The coefficient of determination is $(r^s)^2 = 0.000017$

$$t = r\sqrt{\frac{n-1}{1-r^2}} = \mathbf{0.0117}$$

Interpretation of the result obtained: Almost 0% pair is correlated.

4. A survey was conducted among 500 students who are studying either in "government funded collages" (GVT) or "privately funded colleges" (PVT). The objective of the survey to see the choice of "classroom-based learning" (C) over the "Internet based learning" (I). The survey results are summarized in the table given below.

Learning

Colleges	C	I	
GVT	75	125	200
PVT	60	240	300
	135	365	500

Calculate the χ^2 –value from the sample data shown in the above table. (10)

Answer:

Calculation of χ^2value from the given data.

The contingency table showing observed and expected frequencies are shown in the form of a contingency table.

Learning

Colleges	C	I	
GVT	75 (54)	125 (146)	200
PVT	60 (81)	240 (219)	300
	135	365	500

The formula for the χ^2-value is:

$\chi^2 = \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$, o_{ij} = Observed frequency and e_{ij} = Expected frequency

The calculated value of χ^2-value in this case is:

$\chi^2 = \frac{(75-54)^2}{54} + \frac{(125-146)^2}{146} + \frac{(60-81)^2}{81} + \frac{(240-219)^2}{219}$ = 8.16 +3.02 +5.44 +2.01 =18.63

5. A set of data in 2D-space is given below. Here, *Y* is the regressor.

Y	50	11	30	40	25
X	10	3	5	8	4

Two regression models are given below.

$Y1 = 0.6 + 0.35x_1$

$Y2 = 1 + 0.2x_1 + 0.5{x_1}^2$

Which will be the best regression model? Justify your answer. (10)

Answer:

$R_1{}^2$ = 1- $\frac{SSE1}{SST}$ = 1- $\frac{4827.31}{878.8}$ = - 4.5

$R_2{}^2$ = 1- $\frac{SSE2}{SST}$ = 1-$\frac{533.461}{878.8}$ = 0.393

Second regression model is better.

6. A study is conducted to examine if the presence or absence of a shopping mall in a particular locality is affected by the average family income per year (in lakhs) of that society, or not. A locality is surveyed to find out the average family income of the locality, and then the presence/absence of shopping malls. The independent variable is the average family income level. A total of 50 of such localities are surveyed. The number of localities at each average family income (N) and the number of the localities having a shopping mall is shown in the table.

Suppose the average family income level for a locality is 13 lakhs per year. Find out the probability that a shopping mall is present in that locality or not, using logistic regression analysis method. (10)

Average family income per year in lakhs (x)	Total localities that falls under the given income level (N)	Number of localities not having a shopping mall (NSM)	Number of localities that does have a shopping mall (SM)
9.5	14	13	1
10.5	7	3	4
11.5	12	6	6
12.5	14	4	10
13.5	3	0	3

Answer:

x	**N**	**NSM**	**SM**	$\boldsymbol{Odds}$ **(SM/NSM)**	$\boldsymbol{ln(Odds)}$ **(y)**	$\boldsymbol{x-\bar{x}}$	$\boldsymbol{y-\bar{y}}$	$\boldsymbol{(x-\bar{x})(y-\bar{y})}$	$\boldsymbol{(x-\bar{x})^2}$
9.5	14	13	1	0.077	-2.56395	-2	-2.56916	5.1383268	4
10.5	7	3	4	1.333	0.287432	-1	0.282219	-0.2822186	1
11.5	12	6	6	1	0	0	-0.00521	0	0
12.5	14	4	10	2.5	0.916291	1	0.911078	0.9110776	1
13.5	3	0	3	(3/0=)4/1=4*	1.386294	2	1.381081	2.7621612	4

Note: * However, one of the classes has zero occurrences of 0, creating an undefined odds ratio (3 ÷ 0). Since the ln(odds) is undefined, we followed a common practice of adding 1 to both the numerator and denominator counts in the calculation of all the ln(odds).

From this table:

$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = 8.529347, \quad \sum_{i=1}^{n}(x_i-\bar{x})^2 = 10, \quad \bar{x} = 11.5, \quad \bar{y} = 0.0052$

We know, $Y = \beta_0 + \beta_1 X$

So, using the formula for the linear regression,

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{8.529347}{10} = \mathbf{0.8529}$$

Now,

$\beta_0 = \bar{y} - \beta_1\bar{x} = 0.0052 - 0.8529 \times 11.5 = -9.803$

$\therefore Y = -9.803 + 0.8529 \times (Given\ average\ family\ income\ per\ year\ in\ lakhs)$

$= -0.9803 + 0.8529 \times 13 = 1.2877$

So, the probability $= \frac{e^{1.2877}}{1+e^{1.2877}} = \frac{3.62}{4.62} = \mathbf{0.78}$