

WORKSHOP DOCKET

***Workshop on use of IBM-SPSS and other AI Tools for Research***

FACULTY: DR. ROHIT MUTKEKAR

* **Analysis of Variance (ANOVA)**
* **Assumptions**

1.The Dependent variable should be interval or ratio data type

2. The populations should be normally distributed

3. The population variances should be equal

* **One Way Analysis of Variance (One Way ANOVA) – Theory**

Here the null hypothesis is given by

H0: There is no significant difference between the population means

And the corresponding alternative hypothesis is given by

H1: There is a significant difference between atleast one pair of population means

Here the calculation is done using ANOVA table called One Way ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sources of Variation** | **Sum of Squares**  **(SS)** | **Degree of Freedom** | **Mean Sum of Squares**  **(MSS)** | **F-Ratio** |
| Between the Sample | SSB | k-1 | MSB=SSB/k-1 | MSB/MSE |
| Within the Sample | SSE | n-k | MSE=SSE/n-k |
| Total Variation | SST | n-1 |  |  |

Now the test statistics is given by-

Here,

n denotes total number of observations; k denotes number of entities under study

**Decision Making**

We are testing H0 Vs H1 at α level of significance, where we can reject H0 if

denotes F table value at α level of significance

**Examples on One Way ANOVA**

1. A clinical trial is run to compare weight-loss programs, and participants are randomly assigned to one of the comparison programs and counseled on the details of the assigned program. Participants follow the assigned program for 8 weeks. The outcome of interest is weight loss, defined as the difference in weight measured at the start of the study(baseline) and weight measured at the end of the study (after 8 weeks) in pounds. Three popular weight loss programs are considered viz. Low Calorie Diet, Low Fat Diet and Low Carbohydrate Diet. 15 participants volunteered to take up one of the program (5 participant assigned to each program). Data pertaining to weight lost after the program is given in the table below:

|  |  |  |
| --- | --- | --- |
| **Low Calorie Diet** | **Low Fat Diet** | **Low Carbohydrate Diet** |
| 8 | 2 | 3 |
| 9 | 4 | 5 |
| 6 | 3 | 4 |
| 7 | 5 | 2 |
| 3 | 1 | 3 |

Is there a significant difference in the mean weight loss among the three weight loss programs?

* **Two Way Analysis of Variance (Two Way ANOVA) without Replication – Theory**

Here the null hypotheses are given by

H0R: There is no significant difference between means with respect to row factor

H0C: There is no significant difference between means with respect to column factor

And the corresponding alternative hypothesis is given by

H1R: There is a significant difference between means with respect to row factor

H1C: There is a significant difference between means with respect to column factor

Here the calculation is done using ANOVA table called Two Way ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sources of Variation** | **Sum of Squares**  **(SS)** | **Degree of Freedom** | **Mean Sum of Squares**  **(MSS)** | **F-Ratio** |
| Between the Rows | SSR | r-1 | MSR=SSR/r-1 | MSR/MSE |
| Between the Columns | SSC | c-1 | MSC=SSC/c-1 | MSC/MSE |
| Residual | SSE | (r-1)(c-1) | MSE=SSE/(r-1)(c-1) |  |
| Total Variation | SST | n-1 |  |

Now the test statistics is given by-

n-number of observations (n=cr); r-number of rows; c - number of columns

**Decision Making**

1. We are testing H0R Vs H1R at α level of significance, where we can reject H0 if

denotes F table value at α level of significance for (r-1,(r-1)(c-1))degree of freedom

1. We are testing H0C Vs H1C at α level of significance, where we can reject H0 if

denotes F table value at α level of significance for (c-1,(r-1)(c-1))degree of freedom

* **Examples on Two Way Analysis of Variance (Two Way ANOVA) without Replication**

1. A hotel chain operating in different cities in India has collected data on revenue (in hundred thousand) of its different hotels for different types of rooms. The data is given as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Room Type** | **City** | | | |
| Mumbai | Delhi | Bangalore | Indore |
| Standard | 50 | 30 | 40 | 30 |
| Deluxe | 55 | 45 | 45 | 30 |
| Super Deluxe | 60 | 45 | 50 | 40 |

On the basis of the data given in the table, can it be concluded that-

1. There is difference in average revenue of the four cities?
2. There is difference in average revenue of the three-room types?

* **Example on Two Way Analysis of Variance (Two Way ANOVA) with Replication**

1. The following data refers to the yields of rice on two plots each with combination of the verity of rice and type of fertilizers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Fertilizer A** | **Fertilizer B** | **Fertilizer C** | **Fertilizer D** |
| **Verity 1** | 6 | 4 | 8 | 6 |
| 5 | 5 | 6 | 4 |
| **Verity 2** | 7 | 6 | 6 | 9 |
| 6 | 7 | 7 | 8 |
| **Verity 3** | 8 | 5 | 10 | 9 |
| 7 | 5 | 9 | 10 |

Test the above case at 1% l.o.s.

* **Multivariate Analysis of Variance (MANOVA) – Theory**

Multivariate Analysis of Variance (MANOVA) is an extension of ANOVA in a multivariate scenario. In ANOVA, we examine for statistical differences on one continuous dependent variable by one or more independent grouping variables.

In MANOVA, we extend this analysis by considering multiple continuous dependent variables, and bundles them together into a linear combination or composite variable. In MANOVA, we compare whether or not the newly created combination differs by the different groups or levels, of the independent grouping variable(s).

In other words, we essentially test whether or not the independent grouping variable(s) simultaneously explains a statistically significant amount of variance in the dependent variables as a whole.

**Assumptions:**

1. **Independent Random Sampling**: MANOVA assumes that the observations are independent of one another, there is not any pattern for the selection of the sample, and that the sample is completely random.

2. **Level and Measurement of the Variables**: MANOVA assumes that the independent variables are categorical and the dependent variables are continuous or scale variables.

3. **Absence of multicollinearity**: The dependent variables cannot be too correlated to each other. (***Tabachnick & Fidell (2012) suggest that no correlation should be above r = .90)***

4. **Normality**: Multivariate normality is present in the data.

5. **Homogeneity of Variance**: Variance between groups is equal

* **One Way – MANOVA**

Here the null hypothesis is given by

H0: There is no significant difference between the population means for the linear combination of the dependent variables

And the corresponding alternative hypothesis is given by

H1: There is a significant difference between atleast one pair of population means for the linear combination of the dependent variables

**Example on One Way MANOVA**

A study is conducted on set of random students from three universities on their anxiety and depression levels during their end term examination. Data pertaining to their anxiety and depression score (score out of 100) is tabulated in the table as below. Test whether their exits any significant difference between linear combination of the anxiety and depression score with respect to the three university students’ groups

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | University | Anxiety Score | Depression Score |
| 1 | 1 | 52 | 70 |
| 2 | 1 | 56 | 88 |
| 3 | 1 | 51 | 96 |
| 4 | 1 | 49 | 95 |
| 5 | 1 | 52 | 38 |
| 6 | 1 | 43 | 18 |
| 7 | 1 | 37 | 20 |
| 8 | 1 | 30 | 70 |
| 9 | 1 | 32 | 62 |
| 10 | 1 | 52 | 64 |
| 11 | 1 | 45 | 25 |
| 12 | 1 | 52 | 98 |
| 13 | 1 | 35 | 63 |
| 14 | 1 | 48 | 42 |
| 15 | 1 | 30 | 72 |
| 16 | 2 | 50 | 94 |
| 17 | 2 | 57 | 31 |
| 18 | 2 | 60 | 17 |
| 19 | 2 | 70 | 95 |
| 20 | 2 | 64 | 47 |
| 21 | 2 | 51 | 62 |
| 22 | 2 | 37 | 78 |
| 23 | 2 | 40 | 93 |
| 24 | 2 | 67 | 27 |
| 25 | 2 | 58 | 92 |
| 26 | 2 | 64 | 89 |
| 27 | 2 | 30 | 15 |
| 28 | 2 | 40 | 34 |
| 29 | 2 | 50 | 40 |
| 30 | 2 | 52 | 75 |
| 31 | 3 | 60 | 21 |
| 32 | 3 | 69 | 82 |
| 33 | 3 | 57 | 95 |
| 34 | 3 | 62 | 74 |
| 35 | 3 | 35 | 80 |
| 36 | 3 | 58 | 78 |
| 37 | 3 | 50 | 63 |
| 38 | 3 | 55 | 85 |
| 39 | 3 | 53 | 71 |
| 40 | 3 | 51 | 75 |
| 41 | 3 | 54 | 74 |
| 42 | 3 | 67 | 91 |
| 43 | 3 | 60 | 78 |
| 44 | 3 | 60 | 68 |
| 45 | 3 | 55 | 69 |

* **Two Way – MANOVA**

H01: There is no significant difference between the population means for the linear combination of the dependent variables w.r.t the first independent variable

H02: There is no significant difference between the population means for the linear combination of the dependent variables w.r.t the second independent variable

H03: There is no significant difference between the population means for the linear combination of the dependent variables w.r.t the joint effect of the independent variables

**Example on Two Way MANOVA**

A study is conducted on set of random students from three universities on their anxiety and depression levels during their end term examination. Inputs pertaining to the gender of the student was also collected. Data pertaining to their anxiety and depression score (score out of 100) is tabulated in the table as below. Test whether their exits any significant difference between linear combination of the anxiety and depression score with respect to the three university students’ groups and the gender of the respondents.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | University | Gender | Anxiety Score | Depression Score |
| 1 | 1 | 1 | 45 | 42 |
| 2 | 1 | 1 | 52 | 57 |
| 3 | 1 | 1 | 37 | 41 |
| 4 | 1 | 1 | 45 | 53 |
| 5 | 1 | 1 | 54 | 50 |
| 6 | 1 | 1 | 48 | 52 |
| 7 | 1 | 1 | 53 | 44 |
| 8 | 1 | 1 | 55 | 62 |
| 9 | 1 | 1 | 55 | 58 |
| 10 | 1 | 1 | 46 | 43 |
| 11 | 1 | 1 | 52 | 53 |
| 12 | 1 | 1 | 51 | 53 |
| 13 | 1 | 1 | 59 | 44 |
| 14 | 1 | 1 | 49 | 50 |
| 15 | 1 | 1 | 48 | 66 |
| 16 | 1 | 1 | 53 | 59 |
| 17 | 1 | 1 | 54 | 46 |
| 18 | 1 | 1 | 64 | 53 |
| 19 | 1 | 1 | 38 | 52 |
| 20 | 1 | 1 | 56 | 68 |
| 21 | 1 | 2 | 52 | 50 |
| 22 | 1 | 2 | 50 | 52 |
| 23 | 1 | 2 | 36 | 35 |
| 24 | 1 | 2 | 49 | 56 |
| 25 | 1 | 2 | 45 | 52 |
| 26 | 1 | 2 | 35 | 31 |
| 27 | 1 | 2 | 33 | 46 |
| 28 | 1 | 2 | 19 | 36 |
| 29 | 1 | 2 | 54 | 54 |
| 30 | 1 | 2 | 55 | 58 |
| 31 | 1 | 2 | 56 | 46 |
| 32 | 1 | 2 | 42 | 42 |
| 33 | 1 | 2 | 48 | 45 |
| 34 | 1 | 2 | 38 | 48 |
| 35 | 1 | 2 | 37 | 42 |
| 36 | 1 | 2 | 32 | 43 |
| 37 | 1 | 2 | 42 | 43 |
| 38 | 1 | 2 | 45 | 46 |
| 39 | 1 | 2 | 35 | 37 |
| 40 | 1 | 2 | 45 | 47 |
| 41 | 2 | 1 | 60 | 44 |
| 42 | 2 | 1 | 45 | 66 |
| 43 | 2 | 1 | 38 | 44 |
| 44 | 2 | 1 | 58 | 46 |
| 45 | 2 | 1 | 48 | 49 |
| 46 | 2 | 1 | 35 | 69 |
| 47 | 2 | 1 | 57 | 56 |
| 48 | 2 | 1 | 49 | 45 |
| 49 | 2 | 1 | 47 | 49 |
| 50 | 2 | 1 | 53 | 50 |
| 51 | 2 | 1 | 64 | 38 |
| 52 | 2 | 1 | 36 | 38 |
| 53 | 2 | 1 | 64 | 54 |
| 54 | 2 | 1 | 62 | 53 |
| 55 | 2 | 1 | 70 | 46 |
| 56 | 2 | 1 | 56 | 60 |
| 57 | 2 | 1 | 45 | 55 |
| 58 | 2 | 1 | 60 | 58 |
| 59 | 2 | 1 | 52 | 50 |
| 60 | 2 | 1 | 31 | 39 |
| 61 | 2 | 2 | 56 | 40 |
| 62 | 2 | 2 | 38 | 56 |
| 63 | 2 | 2 | 42 | 51 |
| 64 | 2 | 2 | 50 | 49 |
| 65 | 2 | 2 | 51 | 53 |
| 66 | 2 | 2 | 47 | 54 |
| 67 | 2 | 2 | 37 | 52 |
| 68 | 2 | 2 | 44 | 58 |
| 69 | 2 | 2 | 62 | 52 |
| 70 | 2 | 2 | 51 | 41 |
| 71 | 2 | 2 | 72 | 68 |
| 72 | 2 | 2 | 46 | 43 |
| 73 | 2 | 2 | 39 | 44 |
| 74 | 2 | 2 | 42 | 55 |
| 75 | 2 | 2 | 60 | 62 |
| 76 | 2 | 2 | 65 | 57 |
| 77 | 2 | 2 | 51 | 44 |
| 78 | 2 | 2 | 43 | 44 |
| 79 | 2 | 2 | 57 | 54 |
| 80 | 2 | 2 | 46 | 52 |
| 81 | 3 | 1 | 37 | 38 |
| 82 | 3 | 1 | 54 | 41 |
| 83 | 3 | 1 | 40 | 57 |
| 84 | 3 | 1 | 56 | 55 |
| 85 | 3 | 1 | 52 | 54 |
| 86 | 3 | 1 | 62 | 58 |
| 87 | 3 | 1 | 49 | 48 |
| 88 | 3 | 1 | 60 | 45 |
| 89 | 3 | 1 | 55 | 54 |
| 90 | 3 | 1 | 43 | 38 |
| 91 | 3 | 1 | 57 | 55 |
| 92 | 3 | 1 | 58 | 50 |
| 93 | 3 | 1 | 50 | 56 |
| 94 | 3 | 1 | 50 | 52 |
| 95 | 3 | 1 | 47 | 50 |
| 96 | 3 | 1 | 64 | 62 |
| 97 | 3 | 1 | 46 | 50 |
| 98 | 3 | 1 | 51 | 53 |
| 99 | 3 | 1 | 49 | 50 |
| 100 | 3 | 1 | 38 | 52 |
| 101 | 3 | 2 | 54 | 54 |
| 102 | 3 | 2 | 48 | 45 |
| 103 | 3 | 2 | 53 | 50 |
| 104 | 3 | 2 | 64 | 38 |
| 105 | 3 | 2 | 52 | 54 |
| 106 | 3 | 2 | 37 | 41 |
| 107 | 3 | 2 | 48 | 66 |
| 108 | 3 | 2 | 49 | 56 |
| 109 | 3 | 2 | 55 | 58 |
| 110 | 3 | 2 | 72 | 68 |
| 111 | 3 | 2 | 65 | 67 |
| 112 | 3 | 2 | 57 | 54 |
| 113 | 3 | 2 | 52 | 54 |
| 114 | 3 | 2 | 37 | 42 |
| 115 | 3 | 2 | 59 | 44 |
| 116 | 3 | 2 | 38 | 52 |
| 117 | 3 | 2 | 52 | 50 |
| 118 | 3 | 2 | 47 | 54 |
| 119 | 3 | 2 | 46 | 52 |
| 120 | 3 | 2 | 56 | 55 |

* **Simple Linear Regression Analysis**
* Introduction and Method of Least Square using Normal equations
* Std. Error, Co-efficient of Determination and Significance of the fitted model

**Examples on Simple Regression Analysis using Least Squares Method**

1. A national level organization wishes to prepare a manpower plan based on the ever-growing sales offices in the country. Data pertaining to manpower and the number of sales offices for previous is given below-

|  |  |  |
| --- | --- | --- |
| **Year** | **Manpower** | **Sales Offices** |
| 2009 | 370 | 22 |
| 2010 | 386 | 25 |
| 2011 | 443 | 28 |
| 2012 | 499 | 31 |
| 2013 | 528 | 33 |
| 2014 | 616 | 38 |

Fit a regression model for the given data and estimate the manpower required if the organization targets to have 43 sales offices at the end of 2015. Comment on the significance of the regression model.

1. The following data is relating to training and performance of salesmen employed in a company.

|  |  |
| --- | --- |
| **Training (hrs.)** | **Performance (Avg weekly sales)** |
| 20 | 44 |
| 5 | 22 |
| 10 | 25 |
| 13 | 32 |
| 12 | 27 |

Fit a regression model and determine the weekly sales that is likely to be attained by a salesman who is given 16 hours of training. Comment on the significance of the regression model.

* **Multiple Regression Analysis**
* Introduction
* General form of Multiple Regression Model
* Standard Error for Multiple Regression Model
* Adjusted Coefficient of Determination
* Multicollinearity in Multiple Regression
* Selection of Independent Variables in a Regression Model

**Examples on Multiple Linear Regression Analysis using Least Squares Method**

1. A sample of 30 students was randomly selected from a MBA class of a top business school. In the given case we plan to access their pay package based on their academic performance. Their marks were taken as a measure of their academic performance. There were given standard tests to assess them on three others parameters i.e., Communication ability, General awareness and IQ level. Fit a regression model for the same and interpret the same.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Student** | **Salary (Lac/Ann)** | **Marks (%)** | **Communication** | **General Awareness** | **IQ** |
| 1 | 5.5 | 65 | 88 | 78 | 105 |
| 2 | 16.5 | 90 | 93 | 95 | 120 |
| 3 | 4.5 | 65 | 63 | 56 | 100 |
| 4 | 7 | 68 | 79 | 66 | 100 |
| 5 | 3.5 | 67 | 58 | 77 | 100 |
| 6 | 6 | 62 | 85 | 87 | 110 |
| 7 | 4 | 84 | 77 | 44 | 105 |
| 8 | 7.5 | 82 | 78 | 66 | 105 |
| 9 | 2.75 | 75 | 53 | 68 | 100 |
| 10 | 10.5 | 81 | 95 | 54 | 115 |
| 11 | 5 | 74 | 68 | 50 | 110 |
| 12 | 9 | 88 | 78 | 78 | 115 |
| 13 | 4 | 65 | 63 | 45 | 100 |
| 14 | 4.5 | 84 | 68 | 56 | 100 |
| 15 | 2.5 | 68 | 58 | 43 | 95 |
| 16 | 8.5 | 71 | 89 | 68 | 110 |
| 17 | 12.5 | 92 | 91 | 76 | 120 |
| 18 | 3 | 53 | 55 | 71 | 100 |
| 19 | 8.75 | 77 | 77 | 77 | 100 |
| 20 | 6.5 | 68 | 79 | 45 | 105 |
| 21 | 13.5 | 89 | 89 | 56 | 115 |
| 22 | 5.5 | 68 | 64 | 43 | 100 |
| 23 | 4 | 63 | 62 | 40 | 100 |
| 24 | 5 | 65 | 69 | 55 | 100 |
| 25 | 7.6 | 74 | 71 | 68 | 105 |

1. A 10-year study conducted by the American Heart Association provided data on how age, blood pressure, and smoking relate to the risk od strokes. Assume that the following data are from a portion of this study. Relevant data is given below. Risk is interpreted as the probability (times 100) that the patient will have a stroke over the next 10-year period.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient ID** | **Risk** | **Age** | **Pressure** | **Smoker** |
| 1 | 12 | 57 | 152 | No |
| 2 | 24 | 67 | 163 | No |
| 3 | 13 | 58 | 155 | No |
| 4 | 56 | 86 | 177 | Yes |
| 5 | 28 | 59 | 196 | No |
| 6 | 51 | 76 | 189 | Yes |
| 7 | 18 | 56 | 155 | Yes |
| 8 | 31 | 78 | 120 | No |
| 9 | 37 | 80 | 135 | Yes |
| 10 | 15 | 78 | 98 | No |
| 11 | 22 | 71 | 152 | No |
| 12 | 36 | 70 | 173 | Yes |
| 13 | 15 | 67 | 135 | Yes |
| 14 | 48 | 77 | 209 | Yes |
| 15 | 15 | 60 | 199 | No |
| 16 | 36 | 82 | 119 | Yes |
| 17 | 8 | 66 | 166 | No |
| 18 | 34 | 80 | 125 | Yes |
| 19 | 3 | 62 | 117 | No |
| 20 | 37 | 59 | 207 | Yes |

1. Develop an estimated regression equation that relates risk of a stroke to the patient’s age, blood pressure and whether the patient is a smoker.
2. Is smoking a significant factor in the risk of a stroke? Use 5% l.o.s and explain.

* **Logistic Regression (LR)**

LR is used to predict the outcome of a dependent variable (which is discrete) based on some independent variables which may be continuous, discrete or mixed.

Thus, the dependent variable is categorical in nature having two or more discrete outcomes.

Examples where LR can be used-

1. Based on diagnostic test, patients may be classified as positive or negative indicating the presence of a disease or not.
2. Predicting chances of a heart attack based on the parameters related to one’s life style.
3. Predicting the outcome of a game viz. Win/Loss/Draw

**Binary Logistic Regression (BLR)**

In a BLR model the dependent variable has only two outcomes. If the number of outcomes is more than two then it’s a case of Multinominal Logistic Regression Model. Basically, in a LR model we compute the log odds for a particular outcome.

Here the odds of an outcome are given by the ratio of the probability of happening of an event to not happening of the event.

i.e. , where p is probability of the event under study

The above equation is with respect to one independent variable. This can be extended to ‘n’ independent variables.

**Examples on Binary Logistic Regression Analysis**

1. Suppose that a researcher wants to test the hypotheses regarding the relationship between size and age of a firm and its performance in a pharmaceutical industry. Size is measured by the number of employees (in 100s) working in the firm, age is the number of years for which the firm has been operating. The performance variable is however, measured with the help of a survey in which the CEO of the firm has to mention if his firm is making profit or loss. A response of one indicates profit and zero indicates loss. A sample of 50 firms is selected at random. Data on these variables is given in table. Here the scope of the problem is to predict the performance of the company based on its size and age.

|  |  |  |  |
| --- | --- | --- | --- |
| **Firm ID** | **Performance** | **Size** | **Age** |
| 1 | 1 | 20 | 40 |
| 2 | 0 | 2 | 41 |
| 3 | 1 | 20 | 43 |
| 4 | 1 | 19 | 25 |
| 5 | 0 | 3 | 29 |
| 6 | 0 | 33 | 7 |
| 7 | 0 | 4 | 30 |
| 8 | 1 | 29 | 57 |
| 9 | 1 | 87 | 15 |
| 10 | 1 | 29 | 48 |
| 11 | 0 | 11 | 44 |
| 12 | 1 | 20 | 34 |
| 13 | 0 | 17 | 16 |
| 14 | 0 | 8 | 18 |
| 15 | 0 | 13 | 16 |
| 16 | 1 | 66 | 17 |
| 17 | 1 | 8 | 40 |
| 18 | 1 | 85 | 16 |
| 19 | 0 | 1 | 44 |
| 20 | 0 | 5 | 34 |
| 21 | 0 | 7 | 22 |
| 22 | 1 | 61 | 31 |
| 23 | 1 | 21 | 16 |
| 24 | 1 | 93 | 12 |
| 25 | 0 | 2 | 17 |
| 26 | 1 | 19 | 43 |
| 27 | 1 | 4 | 50 |
| 28 | 0 | 8 | 44 |
| 29 | 0 | 6 | 35 |
| 30 | 0 | 9 | 24 |
| 31 | 1 | 65 | 29 |
| 32 | 0 | 4 | 43 |
| 33 | 0 | 8 | 41 |
| 34 | 0 | 7 | 48 |
| 35 | 1 | 28 | 18 |
| 36 | 1 | 15 | 32 |
| 37 | 0 | 9 | 32 |
| 38 | 1 | 78 | 14 |
| 39 | 1 | 99 | 12 |
| 40 | 0 | 8 | 15 |
| 41 | 0 | 6 | 19 |
| 42 | 0 | 1 | 21 |
| 43 | 1 | 4 | 86 |
| 44 | 1 | 1 | 56 |
| 45 | 1 | 51 | 21 |
| 46 | 0 | 1 | 22 |
| 47 | 1 | 7 | 61 |
| 48 | 1 | 5 | 88 |
| 49 | 0 | 1 | 11 |
| 50 | 1 | 66 | 25 |

1. A medical researcher has collected data from a set of random patients (who had visited the hospital to check whether they had any heart related diseases) from a hospital record. The data pertaining to the same with other details are as provided below. Fit an appropriate regression model to predict the chances of detection of heart related diseases.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Heart Diseases** | **Age** | **Total Cholesterol** | **Sys\_BP** | **Dia\_BP** | **BMI** | **Heart Beats** | **Sugar** |
| 0 | 39 | 195 | 106 | 70 | 27 | 80 | 77 |
| 0 | 46 | 250 | 121 | 81 | 28.7 | 95 | 76 |
| 0 | 48 | 245 | 127.5 | 80 | 25.3 | 75 | 70 |
| 1 | 61 | 225 | 150 | 95 | 28.6 | 65 | 103 |
| 0 | 46 | 285 | 130 | 84 | 23.1 | 85 | 85 |
| 0 | 43 | 228 | 180 | 110 | 30.3 | 77 | 99 |
| 1 | 63 | 205 | 138 | 71 | 33.1 | 60 | 85 |
| 0 | 45 | 313 | 100 | 71 | 21.7 | 79 | 78 |
| 0 | 52 | 260 | 141.5 | 89 | 26.4 | 76 | 79 |
| 0 | 43 | 225 | 162 | 107 | 23.6 | 93 | 88 |
| 0 | 50 | 254 | 133 | 76 | 22.9 | 75 | 76 |
| 0 | 43 | 247 | 131 | 88 | 27.6 | 72 | 61 |
| 0 | 46 | 294 | 142 | 94 | 26.3 | 98 | 64 |
| 0 | 41 | 332 | 124 | 88 | 31.3 | 65 | 84 |
| 0 | 39 | 226 | 114 | 64 | 22.4 | 85 | 87 |
| 1 | 38 | 221 | 140 | 90 | 21.4 | 95 | 70 |
| 0 | 48 | 232 | 138 | 90 | 22.4 | 64 | 72 |
| 1 | 46 | 291 | 112 | 78 | 23.4 | 80 | 89 |
| 0 | 38 | 195 | 122 | 84.5 | 23.2 | 75 | 78 |
| 0 | 41 | 195 | 139 | 88 | 26.9 | 85 | 65 |
| 0 | 42 | 190 | 108 | 70.5 | 21.6 | 72 | 85 |
| 0 | 43 | 185 | 123.5 | 77.5 | 29.9 | 70 | 74 |
| 0 | 52 | 234 | 148 | 78 | 34.2 | 70 | 113 |
| 0 | 52 | 215 | 132 | 82 | 25.1 | 71 | 75 |
| 0 | 44 | 270 | 137.5 | 90 | 22 | 75 | 83 |
| 1 | 47 | 294 | 102 | 68 | 24.2 | 62 | 66 |
| 0 | 60 | 260 | 110 | 72.5 | 26.6 | 65 | 65 |
| 0 | 35 | 225 | 132 | 91 | 26.1 | 73 | 83 |
| 1 | 61 | 272 | 182 | 121 | 32.8 | 85 | 65 |
| 0 | 60 | 247 | 130 | 88 | 30.4 | 72 | 74 |
| 0 | 36 | 295 | 102 | 68 | 28.2 | 60 | 63 |
| 0 | 43 | 226 | 115 | 85.5 | 27.6 | 75 | 75 |
| 1 | 59 | 209 | 150 | 85 | 20.8 | 90 | 88 |
| 1 | 61 | 175 | 134 | 82.5 | 18.6 | 72 | 75 |
| 0 | 54 | 214 | 147 | 74 | 24.7 | 96 | 87 |
| 0 | 37 | 225 | 124.5 | 92.5 | 38.5 | 95 | 83 |
| 0 | 56 | 257 | 153.5 | 102 | 28.1 | 72 | 75 |
| 0 | 52 | 178 | 160 | 98 | 40.1 | 75 | 225 |
| 0 | 42 | 233 | 153 | 101 | 28.9 | 60 | 90 |
| 0 | 36 | 180 | 111 | 73 | 27.8 | 71 | 80 |
| 0 | 43 | 243 | 116.5 | 80 | 26.9 | 68 | 78 |
| 0 | 41 | 237 | 122 | 78 | 23.3 | 75 | 74 |
| 1 | 52 | 175 | 148 | 92 | 25.1 | 70 | 254 |
| 0 | 54 | 195 | 132 | 83.5 | 26.2 | 75 | 100 |
| 1 | 53 | 311 | 206 | 92 | 21.5 | 76 | 215 |
| 0 | 49 | 208 | 96 | 63 | 20.7 | 65 | 98 |
| 0 | 65 | 252 | 179.5 | 114 | 30.5 | 90 | 87 |
| 0 | 46 | 261 | 119 | 77.5 | 23.6 | 75 | 74 |
| 1 | 63 | 179 | 116 | 69 | 22.2 | 95 | 75 |

* **Factor Analysis**

A researcher wants to determine the underlying benefits consumer seek from purchase of a toothpaste. A pretest sample of 30 respondents was taken. The respondents were asked to indicate their degree of agreement with the following statements using a 7-point scale (1 – strongly disagree 7- strongly agree)

i : It is important to buy toothpaste that prevents cavities.

ii : I like toothpaste that gives shiny teeth.

iii : A toothpaste should strengthen your gums.

iv : I prefer a toothpaste that freshens breath.

v : Prevention of tooth decay is not an important benefit offered

by a toothpaste.

vi : The most important consideration in buying toothpaste is

attractive teeth.

i ii iii iv v vi

7 3 6 4 2 3

1 3 2 4 5 4

6 2 7 4 2 4

4 5 4 6 1 3

1 2 2 3 6 5

6 3 6 4 1 2

5 3 6 3 4 4

6 4 7 4 1 3

3 4 2 3 5 4

2 6 2 6 7 3

6 4 7 3 2 6

2 3 1 4 5 3

7 2 6 4 1 4

4 6 4 5 3 3

1 3 2 2 6 6

6 4 6 3 3 4

5 3 6 3 3 4

7 3 7 4 1 4

2 4 3 3 6 4

3 5 3 6 4 3

1 3 2 3 5 6

5 4 5 4 2 3

2 2 1 5 4 4

4 6 4 6 4 4

6 5 4 2 1 7

3 5 4 6 4 4

4 4 7 2 2 7

3 7 2 6 4 5

4 6 3 7 2 3

2 3 2 4 7 7