# Assignment of Regression

**Group 3**

* **Submitted to:**

**PROF. WASEEM HAKANI**

* **Submitted by:**

**AYESHA NAWAZ**

* **STUDENT ID: B**

**BSTF18EO17**

* **SUBJECT:**

**REGRESSION**

* **SEMESTER:**

**4TH(self support)**

**DEPARTMENT OF STATISTICS**

**UNIVERSITY OF SARGODHA**

QUESTION # 1

Calculate all multiple correlation coefficient for 10 observations each.

* **Answer:**Differentiate between correlation and rank correlation with 3 practical examples.
* **Answer:**
* **Correlation:**
* **Definition:**

Correlation means relationship between two variables such as “x” and “y”. This also relates to a mutual relationship or connection between two or more things. In other words two variables are said to be correlated if they tend to simultaneously vary in some direction.In statistics it is interdependence of variable quantities. Correlation is a statistical technique which shows whether and how strongly pairs of variables are related.

* **Types of correlation:**

1. **Positive correlation:**

As one variable increases so does the other.

1. **Negative correlation:**

As one variable increases, the other decreases.

1. **No correlation:**

There is no apparent relation between the variables.

1. **Examples:**
2. Height and weight are related ; taller people tend to be heavier than shorter people .The relationship isn’t perfect. It is used to measure the linear relationship between two variables height and weight.
3. Height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
4. The longer your hair grows , the more shampoo you will need.

**Rank correlation**

* **Definition:**

Rank correlation means relationship between two variables when theirs values have been ranked according to some characteristics. It is measured by cov (x, y).

In statistics ,a rank correlation is any of several statistics that measures the relationship between different variables or different ranking of same variable , where a ranking is assignment of ordering labels first, second, third etc. to different observations of particular variable. They are arranged according to some characteristics of interest .Such an ordered arrangement is called ranking or object is called its rank. The correlation between two such sets of ranking is known as rank correlation.

* **Examples:**

1. Suppose a coach trains long distance runners for one month using two methods .group a has 5 runners ,and group b has 4 runners. Find that runners from group a do indeed run faster, with following ranks : 1,2,3,4, and 6. The slower runners from group b thus have ranks of 5,7,8,9

The analysis is conducted on pairs ,defined as member of one group compared to member of other group .For example the fastest runner in study is a member of four pairs (1,5),(1,7),(1,8).

Group a is faster than runner of group b. There are total of 20 pairs, and 19 pairs support the hypothesis .the only pair which does not support the hypothesis is two runners with ranks 5 and 6. **By kerby** simple difference formula ,95% of data support hypothesis and 5% not support (1of 20) pairs than rank correlation will be r=95-05=90.

1. **Example:**

* By **spearsman rank** correlation :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **x** | **y** | **R1** | **R2** | **d=r1-r2** | **d2** |
| 50 | 60 | 4 | 3 | 1 | 1 |
| 75 | 80 | 3 | 1 | 2 | 4 |
| 80 | 50 | 2 | 4 | -2 | 4 |
| 93 | 70 | 1 | 2 | -1 | 1 |
| 40 | 45 | 5 | 5 | 0 | 0 |

**FORMULA:**

**Rs=**

=1-[6(10)/5(5-10)]

=0.5

1. **Example:**

An example is:

|  |  |  |  |
| --- | --- | --- | --- |
| **x** | **y** | **d** | **d2** |
| 3 | 4 | -1 | 1 |
| 2 | 6 | -4 | 16 |
| 5 | 3 | 2 | 4 |
| 6 | 1 | 5 | 25 |
| 7 | 2 | 5 | 25 |
| 4 | 7 | -3 | 9 |
| 1 | 5 | -4 | 16 |
|  |  |  | ∑d2=96 |

**Formula:**

**Rs=**

=1-576/336

= -0.71

QUESTION # 2

Calculate all multiple correlation coefficient for 10 observations each.

**Answer:**

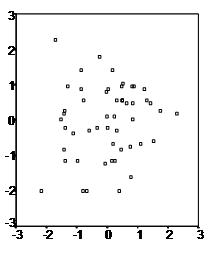
Explain the assumption of multiple regression.

* **Answer:**
* Following are assumptions of multiple regression

1. Linear relationship
2. Multivariate normality
3. No or little Multicollinearity
4. No auto correlation
5. Homoscedasticity

* **Linear relationship**:

A **linear relationship** (or **linear** association) is a statistical term used to describe a straight-line **relationship** between a variable and a constant. **Linear relationships** can be expressed either in a graphical format or as a mathematical equation of the form y = mx + b. Linear relationships can be expressed either in a graphical format where the variable and the constant are connected via a straight line or in a mathematical format where the independent variable is multiplied by the slope coefficient, added by a constant, which determines the dependent variable There must be a linear relationship between the outcome variable and the independent variables.  Scatterplots can show whether there is a linear or curvilinear relationship.



* **Multicollinearity:**

 there is no multicollinearity in the data.  Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be checked multiple ways:

1. **Correlation matrix :**

When computing a matrix of Pearson’s bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.

1. **Variance Inflation Factor (VIF)** :

The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.

If multicollinearity is found in the data, one possible solution is to center the data.  To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.

* **Multivariate Normality:**

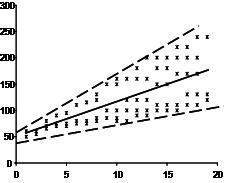
Multiple regression assumes that the residuals are normally distributed.

* **No Multicollinearity**:

Multiple regression assumes that the independent variables are not highly correlated with each other.  This assumption is tested using Variance Inflation Factor (VIF) values.

* **Homoscedasticity**:

This assumption states that the variance of error terms are similar across the values of the independent variables.  A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.



QUESTION # 3

Calculate all multiple correlation coefficient for 10 observations each.

**Answer:**Calculate rank correlation of 20 pairs and interpret the result.

**Answer:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **x** | **y** | **Rank of x** | **rank of y** | **d** | **d2** |
| 18 | 510 | 2 | 16 | -14 | 196 |
| 20 | 590 | 3 | 20 | -17 | 289 |
| 22 | 560 | 4 | 18.5 | -14.5 | 210.25 |
| 23 | 510 | 5.5 | 16 | -10.5 | 110.25 |
| 23 | 460 | 5.5 | 11.5 | -6 | 36 |
| 25 | 490 | 7 | 14 | -7 | 49 |
| 2 | 560 | 1 | 18.5 | -17.5 | 306.25 |
| 28 | 510 | 8 | 16 | -8 | 64 |
| 29 | 460 | 9 | 11.5 | -2.5 | 6.25 |
| 32 | 410 | 10 | 4.5 | 5.5 | 30.25 |
| 37 | 420 | 11 | 7 | 4 | 16 |
| 41 | 460 | 12 | 11.5 | 0.5 | 0.25 |
| 46 | 450 | 13 | 9 | 4 | 16 |
| 49 | 380 | 14 | 3 | 11 | 121 |
| 53 | 460 | 15 | 11.5 | 3.5 | 12.25 |
| 55 | 420 | 16 | 7 | 9 | 81 |
| 63 | 350 | 17 | 2 | 15 | 225 |
| 65 | 420 | 18 | 7 | 11 | 121 |
| 66 | 300 | 19 | 1 | 18 | 324 |
| 67 | 410 | 20 | 4.5 | 15.5 | 240.25 |
| **Formula:** |  |  |  |  |  |

**=1 -**

=1-1.84

**= -0.84**

**Reference:**

**TEXTBOOK- CORRELATION- AND- REGRESSION**

QUESTION # 4

Calculate all multiple correlation coefficient for 10 observations each.

**Answer:**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x1** | **x2** | **x3** | **(x1-Ⴟ1)** | **(x1-Ⴟ1)^2** | **(x2-Ⴟ2)** | **(x2-Ⴟ2)^2** | **(x3-Ⴟ3)** | **(x3-Ⴟ3)^2** | **(x1-Ⴟ1)(x2-Ⴟ2)** | **(x1-Ⴟ1)(x3-Ⴟ3)** | **(x2-Ⴟ2)(x3-Ⴟ3)** |
| |  | | --- | |  | | 28 | | 33 | | 21 | | 40 | | 38 | | 46 | | 48 | | 50 | |  | | |  | | --- | |  | | 74 | | 87 | | 69 | | 69 | | 81 | | 97 | | 96 | | 98 | |  | | |  | | --- | |  | | 5 | | 11 | | 4 | | 9 | | 7 | | 12 | | 14 | | 16 | |  | | |  | | --- | |  | | -10 | | -5 | | -17 | | 2 | | 0 | | 8 | | 10 | | 12 | |  | |  | | |  | | --- | |  | | 100 | | 25 | | 289 | | 4 | | 0 | | 64 | | 100 | | 144 | | |  | | --- | |  | | -9.87 | | 3.13 | | -14.87 | | -14.87 | | -2.87 | | 13.13 | | 12.13 | | 14.13 | | |  |  | | --- | --- | | 97.41 | | | 9.79 | | | 221.1 | | | 221.1 | | | 8.23 | | | 172.3 | | | 147.13 | | | 199.6 | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | |  | | | |  | | --- | | -2.75 | | 3.25 | | -3.75 | | 1.5 | | -0.75 | | 4.25 | | -1.75 | | 0.25 | |  | |  | | |  | | --- | | 7.56 | | 10.56 | | 14.06 | | 1.56 | | 0.56 | | 18.06 | | 3.06 | | 0.06 | |  | | |  | | --- | | 98.7 | | -15.65 | | 252.79 | | -29.74 | | 0 | | 105.04 | | 121.3 | | 169.56 | |  | | |  | | --- | | 27.5 | | -16.25 | | 63.75 | | 2.5 | | 0 | | 34 | | -17.5 | | 3 | |  | | |  | | --- | | 27.14 | | 10.17 | | 55.76 | | 18.58 | | 2.125 | | 55.8 | | -21.22 | | 3.53 | |  | |
| 304 | 671 | 78 | 0 | |  |  | | --- | --- | | 726 | sum=0.04 | | 0.04 | 1076.6 |  | =55.48 | 702 | 97 | 151.8 |

**☻**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | (x1-Ⴟ1)^2 | (x2-Ⴟ2) |
|  |  |  |  | 100 | -9.87 |
|  |  |  |  | 25 | 3.13 |
|  |  |  |  | 289 | -14.87 |
|  |  |  |  | 4 | -14.87 |
|  |  |  |  | 0 | -2.87 |
|  |  |  |  | 64 | 13.13 |
|  |  |  |  | 100 | 12.13 |
|  |  |  |  | 144 | 14.13 |
|  |  |  |  | sum=726 | sum=0.04 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |