**Assignment on Regression**

[submission deadline: **11:59 pm,** **28th August 2021**]

1. Describe a problem related to your organisation where the multiple regression technique can be applied.
2. Identify Independent and Dependent variables and justify the reasons for selecting them through methods discussed in class.
3. Collect relevant data and establish the regression equation. The sample python codes are provided to you for your easy reference.
4. Show how you can use this equation for prediction as well.

**Solution to Assignment**

Name: Chinmay Harkawat  
Organization: Sardar Vallabhbhai National Institute of Technology (SVNIT) Surat  
Role: Student (2nd Year, Mechanical)  
Contact: [u20me157@med.svnit.ac.in](mailto:u20me157@med.svnit.ac.in) [[LinkedIn](linkedin.com/in/chinmay-harkawat-45290a1b4)] [Phone No: +91 9619750652]

**Introduction**

During these troubled times the news of the vaccine trial completion was a relief to many. However, vaccine producers struggled to keep up with the rising tide of cases.

Keeping that requirement in mind I tried to define the problem statement as follows.

**Problem Statement**: Forecasting vaccine requirements & classifying at risk individuals. Here, a person is considered at risk if he/she has taken 1st dose of vaccine.

**Proposed Solution**: We use Multiple Regression Technique to indirectly forecast vaccine requirement by predicting number of positive cases possible using parameters given in Table 1.1. We use stage of vaccination and number of doses administered / planned as data points to cluster risk of getting infected into threat levels Table 1.2.

|  |  |
| --- | --- |
| **Possible** Parameters | Type |
| No. of Positive Cases | Dependent |
| 1st Dose Covaxin OR Covidshield – **Data1** | Independent |
| 2nd Dose Covidshield – **Data2** | Independent |
| Positivity Rate - **Rate** | Independent |
| Reason for Selection: The data parameters were selected after running the data through   Excel and observing their Correlation Matrix and then P-Values.   After one parameter was eliminated using the Correlation Matrix the   remaining parameters were run repeatedly through the regression   module and eliminated until 3 remained. | |

**Table 1.1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Hierarchical Clustering** | | | |
| Parameter Considered | No of Clusters | Scatter Plot | Remarks |
| Positivity Rate | 5 |  |  |
| 1st Dose Completion 18+ |
| Positivity Rate | 5 |  | Not a lot of change |
| 1st Dose Completion |
| Positivity Rate | 3 |  | Drastic Change in observation. |
| 1st Dose Completion |

**Table 1.2**

|  |  |
| --- | --- |
| **Dataset**  The dataset used for analysis is a customized combination of data taken from https://data.gov.in/ | |
| Tamil Nadu COVID Infection Data [District-wise] | <https://tn.data.gov.in/node/6887329#web_catalog_tabs_block_10> |
| Tamil Nadu Vaccination Data [District-wise] | <https://tn.data.gov.in/node/6887218#web_catalog_tabs_block_10> |
| Final Dataset Used | <https://github.com/janeka1122/AI_ML_SCM/blob/main/Assignment%202/Assignment2_Dataset.csv> |

**Regression Analysis**

|  |  |  |
| --- | --- | --- |
| No. of variables | Data1 | 2 |
| Data2 |
| Equation | **Y** = β0 + **X1**β1 + **X2**β2 + ε | |
| β0 ( Y-Intercept) | - 15.4263 | |
| β1 | 0.00018272 | |
| β2 | -0.00287228 | |
| R2 | - 0.1199 | |
| Prediction | 2.80112172 | |
| Error | >80% | |
| Adding Another parameter ‘Rate’ | | |
| No. of variables | Data1 | 3 |
| Data2 |
| Rate |
| Equation | **Y** = β0 + **X1**β1 + **X2**β2 + **X3**β3 + ε | |
| β0 ( Y-Intercept) | -41.9156 | |
| β1 | 2.56675514e+01 | |
| β2 | 1.63965989e-04 | |
| β3 | -2.27187871e-03 | |
| R2 | 0.6064 | |
| Prediction | 17.3938 | |
| Error | <10% | |

**Final Repo:** [**https://github.com/janeka1122/AI\_ML\_SCM/tree/main/Assignment%202**](https://github.com/janeka1122/AI_ML_SCM/tree/main/Assignment%202)