Application of Scan Statistics as an Inferential and Multivariate tool to analyse the Cardiovascular Cases in India: A Data Science Approach

SET ID- 223214
Saumya Sinha
(23MDT0005)
Sriya S Pillai
(23MDT0008)

Under the guidance of:
Dr. Jitendra Kumar
(EMP ID: 15975)

INTRODUCTION

- In India, Cardiovascular Diseases are emerging as a major public health concern. Medical professionals and experts note that heart-related issues among Indians have nearly doubled over the past decade, affecting even younger individuals.
- The age-standardized death rate in India is higher than the global levels.
- The aim of this research paper is to detect cardiovascular disease hotspots across India using the Spatial Scan Statistics (SaTScan) software.
- In the realm of spatial epidemiology, employing Scan Statistics as an inferential and multivariate analytical tool assists in determining which clusters of alarms warrant further investigation and which are likely random occurrences.
- To analyse relation between different factors affecting CVD, the data has been modelled using multiple linear regression.

OBJECTIVES

- To do exploratory data analysis on cardiovascular cases.
- To do descriptive analysis.
- To design a model.
- To optimise our research by checking the major factors causing cvd.
- To find out the hotspots for cardiovascular diseases in India

RELEVANCE

PUBLIC HEALTH INTERVENTIONS

- The findings of this research can be used to inform public health interventions aimed at reducing the burden of CVD in India.
- These findings could be used to identify areas where there is a need for more resources, such as hospitals, clinics, and healthcare professionals.

RESOURCE ALLOCATION

- Efficient resource allocation.
- The government could allocate more funding to states and union territories with high clusters of CVD cases.

AWARENESS CAMPAIGNS

- The findings could also be used to develop focused awareness campaigns about CVD and its risk factors.
- The government could launch a campaign to educate people about the signs and symptoms of CVD and the importance of early detection and treatment.

DATASET

- The data has been collected from Longitudinal Ageing Study in India (LASI) Wave 1 for every State & UT. This data is for the age group 45 and above and was collected from the year 2017-2021.
- The dataset contains cases of different factors relating to CVD such as, blood pressure, diabetes, alcohol and tobacco consumption, along with other necessary details.
- Additionally, the geographical data, the estimated population and the no. of CVD cases in each state and union territory of India has been collected for scan statistics.

S.No.	STATE/UT	Estimated Population	Estimated Adult Population (>45)	LATITUDE	LONGITUDE	Cardiovascular diseases (CVDs)	Hypertension or high blood pressure
1	Andaman & Nicobar	415547	97446	10.2188344	92.5771329	26993	25141
2	Andhra Pradesh	53139259	15171258	15.9240905	80.1863809	6022989	5840934
3	Arunachal Pradesh	1538603	340031	28.0937702	94.5921326	120031	112550
4	Assam	34837717	7594622	26.4073841	93.2551303	1564492	1496141
5	Bihar	120976512	22985537	25.6440845	85.906508	7079545	6757748
6	Chandigarh	1134647	266075	30.72811	76.77065	64656	58537

METHODOLODY

❖ MULTIPLE LINEAR REGRESSION MODEL

For a given regressed variable, multiple linear regression model involves two or more regressor variables. It is created by calculating the regression coefficients, as indicated in eq1. The inference of the regressor variable on the regressed variable is suggested by these regression coefficients. Afterwards, prediction of regressed variable can be performed for a specific collection of regressor variables.

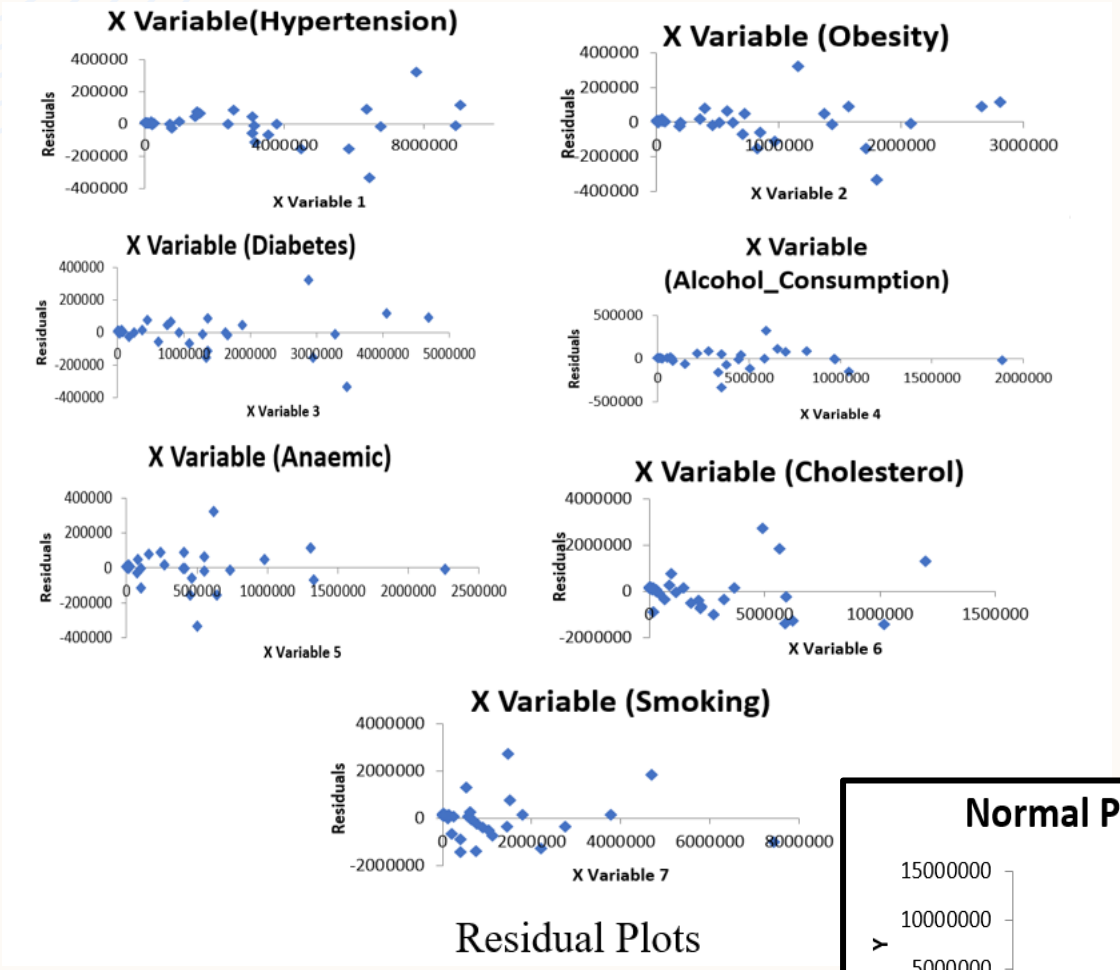
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

❖ MODEL ASSUMPTIONS

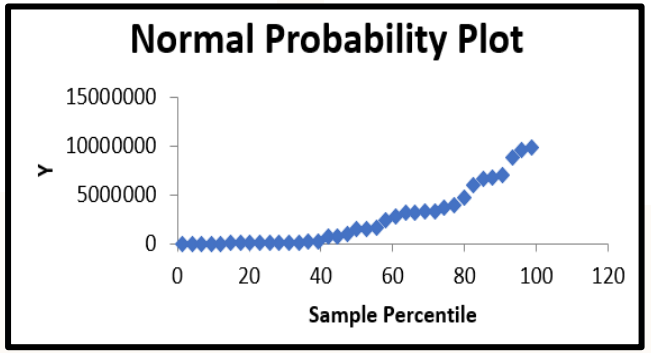
1. There is no Autocorrelation. Random error terms or disturbances are iid distributed.
2. $E[\varepsilon_i] = 0$
3. $V[\varepsilon_i] = \sigma^2$ There's assumption of Homoskedasticity.

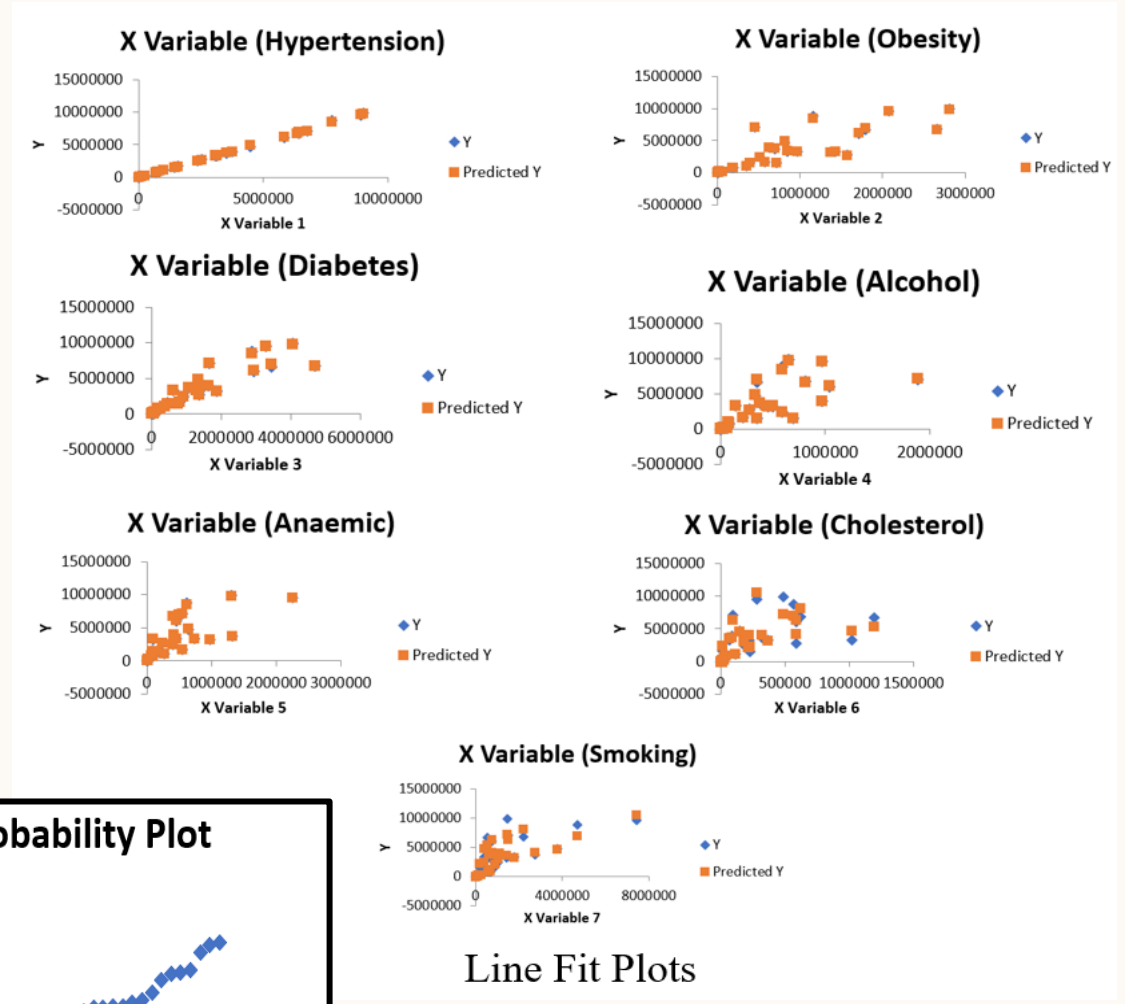
DATA VISUALIZATION

- Regression Plots of important variables are as follows:



Residual Plots





Line Fit Plots

❖ DESCRIPTIVE ANALYSIS

The mean and standard deviation is utilised to calculate Coefficient of Variance of each independent variable. Ranking is done on the basis of CV.

	Mean	Standard Deviation	Coefficient of Variance	Ranking
Hypertension or high blood pressure	2390211.5	2775212.926	116.1074202	1
Obesity by Anthropometric Indicators	656620.62	784902.9944	119.5367566	2
Diabetes or high blood sugar	1020418.2	1287862.081	126.2092439	3
Prevalence of heavy episodic drinking	327204.27	413584.8982	126.3996029	4
Anaemia	345212.86	484537.42	140.3590275	5
High Cholesterol	206192.73	294001.0353	142.5855488	6
Currently smoking	986662.86	1529223.861	154.9895021	7
Currently consuming tobacco	1779047.9	2959404.932	166.3476847	8

The top 5 variables are considered for regression analysis. The summary output of regression analysis includes regression statistics, ANOVA, residual output and probability output.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	3.21E+14	6.41E+13	6077.573	2.18289E-45
Residual	31	3.27E+11	1.06E+10		
Total	36	3.21E+14			

Similarly, Proportional Ranking has been calculated for each independent variable.

			Ranking
High Cholesterol	7629131	0.026738	1
Prevalence of heavy episodic drinking	12106558	0.04243	2
Anaemia	12772876	0.044766	3
Obesity by Anthropometric Indicators	24294963	0.085147	4
Currently smoking	36506526	0.127946	5
Diabetes or high blood sugar	37755473	0.132323	6
Currently consuming tobacco	65824771	0.230699	7
Hypertension or high blood pressure	88437826	0.309951	8
	285328124		

The top 5 variables are considered for regression analysis.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressior	5	2.99E+14	5.97E+13	82.35583	5.86347E-17
Residual	31	2.25E+13	7.25E+11		
Total	36	3.21E+14			

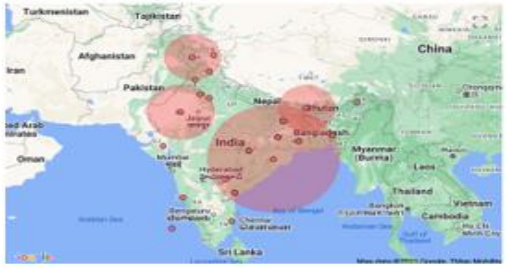
Comparative Study of models based on Consistency ranking and that based on Proportional ranking.

Statistical Criteria	Model based on CV	Model based on PR
R Square	0.998980896	0.929987668
Adjusted R Square	0.998816524	0.918695356
Standard Error	102722.416	851418.6308
p-value	2.18289E-45	5.86347E-17

❖ SCAN STATISTICS

The excel data has been used as input into SatScan application. Case file includes the number of cases of independent variables, taken separately, coordinate file contains the latitude and longitude, and lastly, population file contains the number of CVD cases of every State & UT. Discrete Poisson based model is used to identify the hotspots.

Clusters on the basis of:



CVD

Hypertension/BP

Alcohol Consumption

Obesity

Diabetes

Anaemia

RESULT AND DISCUSSION

- The multiple linear regression equation considering top 5 variables based on CV:
  $y = -6990.78 + 1.130775x_1 + (-0.02882)\,x_2 + (-0.037)\,x_3 + (-0.23315)\,x_4 + (-0.0423)\,x_5$
- The multiple linear regression equation considering top 5 variables based on PR:
  $y = -155527 + 1.438701x_1 + 2.478916\,x_2 + (-0.05643)x_3 + 1.444067x_4 + 0.676285x_5$

Comparing the 2 models, model based on **coefficient of variance gives better results as R square value** i.e. 0.998 is higher than that of the counterpart. Even the standard error is lesser in comparison.

The results suggest that the following 5 variables are important risk factors for cardiovascular disease in India**: High BP , Alcohol consumption, Anaemia, Obesity, Diabetes.** These 5 variables can explain a large proportion of the variation in cardiovascular disease cases in India.

The main objective of this research is to evaluate the statistical significance of disease cluster alarms and find the hotspots. After the completion of the calculations, a standard text-based results file is automatically generated, containing information about the clusters detected. Prominent clusters considering the factors have been detected.

❖ MODEL VALIDATION

Hypothesis testing

1. H_0: $\beta_0 = \beta_1 = \cdots = \beta_5 = 0$

2. H_1: At least one variable is significant

Let $\alpha = 0.05$

Test statistic:

Suppose we apply p-value criteria. In both the models

p-value is less than α (0.05). Hence, we reject H_0.

Similarly, suppose we apply F-statistic criteria.

$F_{table(5,31,0.05)} = 2.5336$, $F_{stat} = 6077.57$ & 82.35 for both the models respectively (Table5,6). Since $F_{stat} > F_{table}$, we reject our null hypothesis H_0.

This implies that both our models are valid.

CONCLUSION

This project involves examining cardiovascular data from 37 states and union territories. Regression analysis indicates a strong fit with 99% accuracy when considering major factors such as Hypertension, Diabetes, Alcohol Consumption, Obesity and Anaemia.

Hotspots, identified through SatScan, majorly include regions like **Gujarat, Andhra Pradesh, Telangana, Bihar, Karnataka, Maharashtra, Kerala, Chhattisgarh, Punjab and Tamil Nadu**. These analyses can be used to design a model for proper resource allocation in these critical regions and prevent increase in cardiovascular diseases.

REFERENCES

- [1] Kumar, Vinod, Lalotra, Gotam Singh. "*Predictive Model Based on Supervised Machine Learning for Heart Disease Diagnosis"*. 2021 IEEE International Conference on. :1-6 Dec, 2021
- [2] Fiaidhi, J., Mohammed S. *"Prognosis analysis of thick data: Clustering heart diseases risk groups case study"*. Department of Computer Science, Lakehead University, Ontario, Canada: June 2021
- [3] Lohaj, Oliver, Pella, Zuzana, Paralic, Jan. *"Data analytics methods for analyzing the impact of factors on early detection of cardiovascular risk"*. 2022 IEEE 20th Jubilee World Symposium (SAMI): Mar, 2022
- [4] Rustamov, Zahiriddin, Rustamov, Jaloliddin, Sultana, Most Sarmin, Ywei, Jeanne, Balakrishnan, Vimala, Zaki, Nazar. *"Cardiovascular Disease Prediction using Ensemble Learning Techniques: A Stacking Approach"*. 2023 19th IEEE International Colloquium (CSPA) : Mar, 2023
- [5] LASI, Wave 1: https://www.iipsindia.ac.in/lasi

THANK YOU!