Data Science – A Statistical Modelling in Health Science

Sarojini G. Deshmukh[1], U. B. Aithal[2]

[1,2]*Jagdishprasad Jhabarmal Tiberewala, University, Rajasthan*

Abstract: **In the developmental stage of fields of Science, Technology, Social Science, Business and management, Health Science etc., there is problem of dealing of Big Data which is one of the most promising trends and which has led to the development of a new interdisciplinary area called Data Science. Data science includes high performance computing, data processing, development and management of databases, data warehousing, mathematical representations, statistical modeling and analysis, and visualization with the goal of extracting information from the data collected for domain-specific applications. With a major Big Data objective of turning data into knowledge, statistics is an essential scientific discipline because of its sophisticated methods for statistical inference, prediction, quantification of uncertainty, and experimental design. The facts collected during enquiry are made flawless for the scientific decisions by using statistical modeling techniques and inferences. Statistical thinking helps to make scientific discoveries, quantifies the reliability, reproducibility and general uncertainty associated with the big data.**

In the health sciences most studies are concerned with the possible causes, remedies, determinants, related factors of a set of observations related to socio-economic-cultural, health, environmental status of a person facing any disease. Healthcare Scientists appeal to statistical modeling to confirm their hypotheses about possible causal relationships among the characteristics they consider, by taking the controlling relevant covariates and especially possible confounding factors to what extent can a statistical model performs about causal relationships among characteristics.

Keywords: **data science, big data, statistical modeling, health care analysis.**

1. INTRODUCTION

The new era of statistics is observing, exciting developments and new trends in statistical modeling and its applications in various fields of Science, Technology, Social Science, Business and Management, Health Science etc. In such developmental stage, there is problem of dealing of Big Data which is one of the most promising trends and which has led to the development of a new interdisciplinary area called Data Science. In the field of statistics, Cleveland [5] introduced this term as a new direction for the statistical area in 2001. It is used in the prominence of sagacity of statistical methods at large for collecting, analyzing and modeling of it. Data science includes high performance computing, data processing, development and management of databases, data warehousing, mathematical representations, statistical modelling and analysis, and visualization with the goal of extracting information from the data collected for domain-specific applications. One of the most active areas of data science research is related to very large data sets, or big data, which pose computational and statistical challenges to the investigator. [11] One of the examples of how big data have changed statistical inference is multiple testing, a tool of testing which include a wide variety of people, allowing interdisciplinary areas to be contained in multiple field having multiple hypothesis testing is a fundamental problem in statistical inference where the primary goal is to consider a set of hypothesis tests simultaneously. [8] For instance, health study incorporates concepts from both statistics and health science and it is possible for an individual working in this area to be both a statistician and a health scientist.

Health Science data tries to sense out of observations, the selection of what one observes depends upon underlying thrust of study and theoretical constructs. The facts thus collected are however often far from flawless for the scientific decisions. Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility and general uncertainty associated with these big data. Statistical thinking will be vital to Big Data challenges. With a major Big Data objective of turning data into knowledge, statistics is an essential scientific discipline because of its sophisticated methods for statistical inference, prediction, quantification of uncertainty, and experimental design. Such methods have helped and will continue to enable researchers to make discoveries in science, government, and industry. The Big Data Research and Development Initiative announced in USA in March, 2012 to help accelerate the pace of discovery in science and engineering, strengthen the national security, and transform teaching and learning. Since the launch of the initiative, the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Defense Advanced Research Projects Agency (DARPA) of USA have launched major Big Data programs. [2]

2. RESEARCH METHODOLOGY

A model is not personal mental image of some reality but is inter subjective constructions which can be applied with the aim of providing a simplified form of complex reality. A statistical model is based on stochastic representation of real world. It is made up of a set of assumptions under which the data are to be analyzed and interpreted.

Typical assumptions of statistical models are: whether the observed random variables follow identical distribution; the observations are independent; the basic sampling distributions are continuous and may pertain to a family characterized by a finite number of parameters.

In this research paper researcher is trying to highlight the key points of data science and the applications of statistical modelling related to health research. The information is searched and gathered by the researcher using secondary sources like published journals, online journals, published books, online books, etc. And the primary data of anaemic condition of the females of reproductive and non-reproductive age groups is collected from the Kolhapur city. This primary data is analysed using fitting of model to the data and testing its goodness using chi - square test.

3. OBJECTIVITY OF THE PROBLEM

In this paper an attempt has been made to find out different aspects of data of health and hence to fit a suitable model for the observed data of anaemic condition of women of different age groups and its inter-relations.

4. HEALTH SCIENCE SCENARIO

Data in Health science are massive, can vary from patient to patient in a very significant way. In the health sciences most studies are concerned with the possible causes, remedies, determinants, related factors etc. of a set of observations related to socio-economic-cultural, health, environmental status of a person facing any disease. In this field in particular, for policy or planning reasons, it is important to know what causes which effects and its inferences for future prophecy. The raw data of health are produced by complicated measurement technologies that introduce bias and variation that must be removed before the measurements represent biological quantities of interest. Similarly, to produce results that will replicate in future studies, it is critical to account for the biological variability between individuals and within an individual. Models are often characterized in terms of parameters—numerical characteristics of the model.

Health care Scientists appeal to statistical modeling to confirm or disconfirm their hypotheses about possible causal relationships among the variables they consider, taking controlling for relevant covariates and especially for possible confounding factors to what extent can a statistical model say something about causal relationships among variables. Finding a model that adequately describes the main features of the medical data is a process consisting of initial data analysis, graphical checks, parameter estimation and assessment of goodness of fit. [3]

In health research it is essential to use various multivariable modeling techniques. Multiple linear regression models, logistic regression analysis and proportional hazards regression models are mostly applicable in health research. Linear regression is used with interval outcomes (such as level of hemoglobin). With interval variables, equally sized differences on all parts of the scale are equal. Hemoglobin level is an interval variable. Logistic regression is used with dichotomous outcomes (yes or no; for example, severity of anaemia). Proportional hazards regression is used when the outcome is the length of time to reach a discrete event such as end point of disease or death.

There is a special class of statistical model called as structural model or causal model for describing multivariate data modelling. Structural equation modeling is a multivariate statistical analysis technique that is used to analyze structural relationships. Structural equation modeling technique is the combination of factor analysis and multiple regression analysis, and it is used to analyze the structural relationship between measured variables and latent constructs. Structural equation modeling is preferred by the researcher because it estimates the multiple and interrelated dependence in a single analysis. [8] In structural equation modeling, two types of variables are used endogenous variables and exogenous variables. In structural equation modeling, endogenous variables are equivalent to dependent variables. In structural equation modeling, exogenous variables are equal to the independent variable.

In structural equation modeling, theory can be thought of as a set of relationships providing consistency and comprehensive explanations of the actual phenomena. It is also termed as causal modeling.

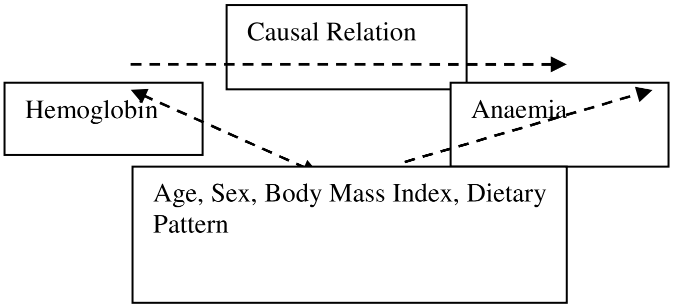
Causal models

If the analyst wants to draw causal inferences, then the approach has to move from a descriptive one to a modeling approach. Causal relationships can arise through a number of pathways. Models and estimators vary in sophistication with the degree of detail of the causal relationship the analyst is aiming to uncover. [10] For example, maternal education can affect child health either directly, through knowledge of healthy behavior, or indirectly, through preferences for child health. If the analyst is interested simply in whether educating women is an effective means of raising child health, irrespective of the mechanism through which it works, then the statistical model, and estimator, can be quite simple.

Constructs are defined at different levels, and the hypothesized relations between these constructs operate across different levels. These types of “multilevel” theoretical constructs require specialized analytic tools to properly evaluate. [8]

The medical model takes a reductionist view of health: Disease is seen simply as a defect in the person that is corrected by medical intervention. We can see how this plays out, for example, in how modern epidemiologists identify risk factors for disease. Epidemiology has constructed a number of authoritative design approaches (e.g. disease-control studies) and analytic tools (e.g. logistic regression that produces relative risk estimates) that are used to identify significant risk factors for development of, say, anaemia disease in women. Important risk factors for anaemia include genetic tendency; biology (low level of hemoglobin); dietary pattern (malnutrition); culture (background); and environment (access to health care). However, even though these factors clearly are operating at different levels, they are almost always measured at the individual-level.

Causal Relation between the variables Age, Sex, Body Mass Index, Dietary Pattern and hemoglobin.



5. USE OF COMPUTATIONAL STATISTICS WHILE DEALING WITH BIG DATA

Here the term “computational statistics” is used somewhat more broadly to include not only the methods of statistical computing, but also statistical methods like modelling that are computationally intensive. Thus, to some extent, “computational statistics” refers to a large class of modern statistical methods. [14]

While comparing between computational statistics and statistical computing the emergence of the field of computational statistics becomes coincidental with that of statistical computing, and would not have been possible without the developments in statistical computing. [7]

Predictive analytics, involves searching for meaningful relationships among variables and representing those relationships in models. There are response variables—things we are trying to predict. There are explanatory variables or predictors—things we observe, manipulate, or control that could relate to the response. Consider three general approaches to research and modeling as employed in predictive analytics: traditional, data-adaptive, and model-dependent. The traditional approach to research and modeling begins with the specification of a theory or model. Traditional methods, such as linear regression and logistic regression, estimate parameters for linear predictors. [4] Model building involves fitting models to data. After we have fitted a model, we can check it using model validity techniques for its future prophecy. In general, statistical analysis involves the prescribed expression of uncertainty in terms of probabilities. The reality that statistical analysis generates probabilities that there are relationships should not be seen in itself as an argument against the use of statistical evidence. [9]

Table showing age wise distribution of women having anaemia:

Age Groups	Number of women having anaemia	pi (proportion)
15-19	12	0.03428571
20-24	23	0.06571429
25-29	12	0.03428571
30-34	20	0.05714286
35-39	29	0.08285714
40-44	45	0.12857143
45-49	62	0.17714286
50-54	67	0.19142857
55-59	80	0.22857143
Total	350	1.0000000

A single predictor Poisson Regression model is fitted to the observed data to test the research hypothesis regarding the relationship between the age and anaemia.

$p_i = e^{(- \alpha + \beta x)}$

where p_i: probability of outcome of anaemia in different age groups,

β : regression coefficient

α: intercept

e = 2.7183 is the base of the system of natural logarithms.

X = can be categorical or continuous.

The analysis is carried out and the result showed that, Predicted p_i of Anemia = $e^{(- 4.35 + 0.0512 * Age)}$

The Chi-square goodness of fit test is used to check the effectiveness of the model that yielded a calculated value 11.62. At 5% level of significance the critical value is $\chi^2_{0.05}(8) = 15.507$ and is insignificant ($p > 0.05$). It shows that the model defined above is good for the data corresponding to anemia.

6. CONCLUSION

In this paper, we demonstrate that the Poisson regression modelling technique can be a powerful analytical technique for use when outcome variable is dichotomous. The effectiveness of the model was checked by predicted probabilities and inferential goodness of fit test.

REFERENCES

[1] Ahalt, S., et al. (2014), “Data to Discovery: Genomes to Health”, A White Paper from the National Consortium for Data Science. RENCI, University of North Carolina at Chapel Hill. http://dx.doi.org/10.7921/G03X84K4

[2] A Working Group of the American Statistical Association, (July 2, 2014), “Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society”, ASA.

[3] Bender R., Grouven. U. (1997), ‘‘Ordinal Logistic Regression in Medical Research’’, Journal of the Royal college of Physicians of London, Vol. 31 No.5.

[4] Chao-Ying J P, et al. (September/October 2002), ‘‘An Introduction to Logistic Regression Analysis and Reporting’’, The Journal of Educational Research, [Vol 96(No. 1)]

[5] Cleveland W. (2001). “Data science: An action plan for expanding the technical areas of the field of statistics”, Int. Statist. Rev, 69: 21-26.

[6] Douglas A. Luke, (2005), Luke Fina pdf, Book of Multilevel modelling, Saint Louis University School of Public Health.

[7] James E. Gentle, et al, (2012), “How Computational Statistics Became the Backbone of Modern Data Science”, http://sfb649.wiwi.hu-berlin.de ISSN 1860-5664

[8] J.D. Storey, (2002) “A direct approach to false discovery rates”, Journal of the Royal Statistical Society: Statistical Methodology 64 479–498.

[9] Naser A. Rashwan and Maie M. Kamel, (2011), “Using Generalized Poisson Log Linear Regression Models in Analyzing Two-Way Contingency Tables”, Applied Mathematical Sciences, Vol. 5, no. 5, 213 – 222

[10] Pearl J. (2000), Causality. Models, Reasoning, and Inference, Cambridge University Press, Cambridge.

[11] Tze Leung Lai (2013), “Data Science, Statistical Modeling, and Financial and Health Care Reforms”, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.