IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A STATISTICAL ANALYSIS OF MORTALITY USING GROUPED LOGIT MODEL

K. Sudha[1] and C. Geetha[2*]

[1]Research Scholar, [2*]Associate Professor and Head, Department of Statistics

Government Arts College (Auto), Salem - 7

ABSTRACT: This research was conducted to estimate the parameters of medically certified deaths for the years 2015 &2017 in India. Here, we used the Grouped Logit Model (G logit) and the weighted least squares method for estimation of the mortality variable due to several age groups. The data set was taken from the Census of India website. This study led to the identification of among the causes of death, which have the most and least amount of deaths occurring due to age.

Keywords: Mortality, Age group, Causes of death, G-Logit model.

1. INTRODUCTION

In the regression models, we assumed that the dependent variable Y is quantitative, while the explanatory variables are qualitative, quantitative (or dummy), or a mixture. In a model if Y is quantitative, our objective is to estimate the expected value given the values of the explanatory variables. When Y is qualitative, our objective is to estimate the probability that a particular outcome occurs. "Logistic regression is a powerful model used to binomial outcome (takes the value 0 or 1) as a function of one or more explanatory variables. Logistic regression analysis is often used to investigate the relationship between a binary response variable and a set of independent variables. Since dichotomous or binary response variable doesnot follow a normal distribution and it was analysed by using a link function named LOGIT. Eric A. Roth and K. Balan Kurup(1990) analysed data from a 1985 survey in two major population centres in Southern Sudan, reveals that significant positive associations between child survival and immunization, oral rehydration therapy and maternal education. Juba and Wau, assessed childhood mortality levels and evaluate the impact of UNICEF's health care programme. The logistic regression analysis showed that components are significantly associated with child survival and maternal education emerged as the most important determinant of child survival. Rohini P. Pande (2003) examines the role of the sex composition of surviving older siblings on gender differences in childhood nutrition and immunization, using data from National Family Health Survey, India (1992-1993). The author used Logit and ordered logit models and shows selective neglect of children with certain sex and birth-order combinations that operate differentially for girls and boys. Also suggested that parents want some balance in sex composition. Nur Ain Abd Aziz, Zalilaali, Norlida Mohd Nor, Adam Baharum and Maizurah Omar (2016), studied smokers characteristics after a smoke-free Melaka campaign in the area of Melaka based on demographic variables like race, marital status, occupation etc., by using multinomial logistic regression. Also a comparative study is examined between smoking behaviour of smoking and non-smoking family. Odds ratio reveals that Malay respondents smoking behaviour in the presence of non-smoking family as compared to other races respondents is high. Pandey A. P(2016), studied the analyse the socio-economic factors to motivate small holder farmers to motivate in contract farming mechanism. Using logit model, the hypothesis with gender, education level of farmers, loaning, electricity and cropping factors the study found that household size, gender, education, loaning, electricity, employability are significant whereas off farm income, components of contract farming are not significant. Opemo Damian Otieno, Juma Shem Godfrey (2019), used logistic regression analysis to examine the association between categorical variables and concluded that the probability of death was closely associated with HIV infections among fishing communities.

SezginOzcan(2014), a multivariate analysis of logistic regression is studied on servicemen of the U.S army in casuality to assess the risk in hostile incidents and concluded the males and married servicemen are more likely to be involved in hostile incidents. Sandri Sperandei(2013), explained logistic analysis model using an example with multiple explanatory variables. Sharareh R. NiakanKalhori, MahshidNasehi, Xiao-Jun Zeng (2010), A logistic model is used to predict the probability of failure in tuberculosis treatment course completion in Iran's health data and concluded that there is a detection of high-risk patient at DOTS therapy. Layla Aziz Ahmed (2017), studied in determining death variables from road traffic accidents and its effect through logistic regression. It concludes that explanatory variables with accident fatality and humanity victims had high significant effects. The main objective of this paper using grouped logit model for estimating the parameters of deaths due to various causes among various age groups in India at 2015 and 2017 respectively. The results represent the comparative analysis between years among age and causes.

2. MATERIALS AND METHODS

2.1 LOGIT MODEL

In logistic regression, binary variable does not follow normal distribution then it is called logit model for regression. since logit be a link function, then logit model of regression is as follows. Let the Linear probability Model (LPM) was,

$$P_i = \beta_1 + \beta_2 X_i \quad \ldots(2.1.1)$$

where X_i be the independent variable and P_i = E (Y_i = 1/X_i) be the conditional probability of dependent variable Y_i. Now consider the following representation

$$P_i = \frac{1}{1+e^{-(\beta_1 + \beta_2 X_i)}} \quad \ldots(2.1.2)$$

It can be written as, $P_i = \frac{1}{1+e(-Z_i)}$, where $Z_i = \beta_1 + \beta_2 X_i$

multiply and divide by e^{Z_i} , then $P_i = \frac{e^{Z_i}}{e^{Z_i}} \cdot \frac{1}{1+e^{-Z_i}} = \frac{e^{Z_i}}{e^{Z_i}+1} = \frac{e^Z}{1+e^Z}$ …(2.1.3)

and called as the (cumulative) logistic distribution function.

Here Z_i range from – ∞ to ∞ and P ranges from 0 to 1. Since, P_i is non – linearly related to Z_i (ie X_i), we cannot use the OLS to estimate the parameters. Hence,

$$\Longrightarrow \frac{P_i}{1-P_i} = \frac{e^{Z_i}/\ (1+e^{Z_i})}{1/\ (1+e^{Z_i})} = e^{Z_i}, \quad \ldots(2.1.4)$$

here, $\frac{P_i}{1-P_i}$ be the odds Ratio.

by taking log on both sides in (2.1.4) we get,

$$L_i = \log_e \left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_1 + \beta_2 X_i \quad \ldots(2.1.5)$$

is known as the log of odds ratio which is named as Logit.

2.2 GROUPED LOGIT MODEL

Grouped Logit model is used to estimate in favour of grouped variables or replicated variables. For estimation purposes, we write equation (2.1.5) as,

$$L_i = \log \left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i + U_i \quad \ldots(2.2.1)$$

where U_i = error term, Using Method of Least squares (MLE) P_i can be estimated under individual data level. For Grouped data P_i can be estimated by Weighted Least Squares (WLS). For empirical purposes, we will replace P_i by $\hat{P}_i$ and $\sigma^2 = \frac{1}{N_i \hat{P}_i (1-\hat{P}_i)}$ as an estimator of σ^2. The procedure for Grouped Logit (Glogit) is as follows. First we obtain for each X_i compute $\hat{P}_i = n_i / N_i$, secondly for each X_i obtain $\hat{L}_i = \log [\ \hat{P}_i/(1- \hat{P}_i)]$, thirdly the transformation of equation(2.2.1) as,

$$\sqrt{W_i}\,L_i = \beta_1\sqrt{W_i} + \beta_2\sqrt{W_i}\,X_i + \sqrt{W_i}\,U_i \quad \ldots(2.2.2)$$

ie., $$L_i^* = \beta_1\sqrt{W_i} + \beta_2 X_i^* + V_i \quad \ldots(2.2.3)$$

where $W_i = N_i\hat{P}_i(1-\hat{P}_i)$then estimate equation (2.2.2) by OLS ie., WLS is OLS on the transformed data since there is no intercept in (2.2.3) we have to use regression through origin method.

2.3 METHOD OF WEIGHTED LEAST SQUARES

Consider, $Y_i = \beta_1 + \beta_2 X_i + U_i$ be the linear model, and estimation of parameters are done through unweighted Least squares method. That is Errors should be minimized to obtain the estimates.

$$\Sigma\, u_i^2 = \Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad \ldots(2.3.1)$$

The method of least squares is used to minimize the weighted Residual sum of squares (RSS)

$$\Sigma w_i\, u_i^2 = \Sigma w_i\,(y_i - \widehat{\beta_1}* - \widehat{\beta_2}* X_i)^2 \quad \ldots(2.3.2)$$

where $\widehat{\beta_1}*$ & $\widehat{\beta_2}*$ are weighted least squares estimators. Since, $w_i = 1/\sigma_i^2$ ---(3) then V (u_i/x_i) is V$(y_i/x_i) = \sigma_i^2$. Differentiating equation (2.3.2) with respect to $\widehat{\beta_1}*$&$\widehat{\beta_2}*$.

$$\Sigma w_i y_i = \widehat{\beta_1}* \Sigma w_i + \widehat{\beta_2}* \Sigma w_i x_i \quad \ldots(2.3.3)$$

$$\Sigma w_i\, x_i y_i = \widehat{\beta_1}* - \Sigma w_i\, x_i + \widehat{\beta_2}* \Sigma w_i\, x_i^2 \quad \ldots(2.3.4)$$

Solving these equations simultaneously we obtain the weighted linear model as,

$\widehat{\beta_1}* = \bar{Y}^* - \widehat{\beta_2}*\,\overline{X}^*$ and the estimated value is, $\widehat{\beta_2}* = \frac{(\Sigma w_i)(\Sigma w_i x_i y_i) - (\Sigma w_i x_i)(\Sigma w_i y_i)}{(\Sigma w_i)(\Sigma w_i x_i^2) - (\Sigma w_i x_i)^2}$

Also, V $(\widehat{\beta_2}*) = \frac{(\Sigma w_i)}{(\Sigma w_i)(\Sigma w_i x_i^2) - (\Sigma w_i x_i)^2}$. Here $\overline{Y}^* = \frac{\Sigma w_i y_i}{\Sigma w_i}$, $\overline{X}^* = \frac{\Sigma w_i x_i}{\Sigma w_i}$ be the means of the regression.

2.4 SOURCES OF DATA

The secondary data is collected from MCCD report for the years 2015 and 2017, Census of India.

3. LOGIT MODEL FOR MORTALITY

Since the data are grouped, it is called a grouped logit mode(Glogit). In linear model P_i is linearly related to X_i whereas in Logit model L_i is linearly related to X_i. Let X_i ~ Binomial distribution be independent variable of age group. Then,

$$\hat{L}_{it} = \log\left(\frac{P_{it}}{1-P_{it}}\right) = \hat{\beta}_1 + \hat{\beta}_2 X_{it} + u_{it} \quad ;\ i=1,2,...10;\ t = 2015,2017 \quad \ldots(3.1)$$

is a grouped logit model for the death caused by a disease on various time points t (t=2015, 2017). Hence u_{it} ~ N $\left(0, \frac{1}{N_{it}P_{it}(1-P_{it})}\right)$ which is heteroscedastic. Instead of using ordinary least squares, the method of weighted least squares is used. Let $\hat{P}_{it} = \frac{n_{it}}{N_{it}}$ be relative frequency and $\sigma^2 = \frac{1}{N_{it}\hat{P}_{it}(1-\hat{P}_{it})}$ be the variance. Where N_i be the total number of deaths in the age groupX_i , i=1 to 10. and n_i be the number of deaths due to certain diseases in a specified age group. The transformed equation for mortality is given by,

$$\sqrt{W_{it}}\,L_{it} = \beta_1\sqrt{W_{it}} + \beta_2\sqrt{W_{it}}\,X_{it} + \sqrt{W_{it}}\,U_{it} \quad \ldots(3.2)$$

ie., $$L_{it}^* = \beta_1\sqrt{W_{it}} + \beta_2 X_{it}^* + V_{it} \quad \ldots(3.3)$$

where $W_{it} = N_{it}\hat{P}_{it}(1-\hat{P}_{it})$.

4. STATISTICAL ANALYSIS AND RESULT DISCUSSIONS

Table 1: Estimates using G-logit model for Medically Certified Deaths (2015) in India

S.No	Causes of Death	β_1	β_2	exp(β_2)
1	Certain Infectious and Parasitic	-1.43243	-0.01241	0.98766
2	Neoplasms	-3.12964	0.00622	1.00624
3	Blood-forming organs & certain disorders involving immune mechanism	-3.11479	-0.02319	0.97708
4	Endocrine, Nutritional and Metabolic diseases	-4.47386	0.02415	1.02444
5	Mental and Behavioural disorders	-5.72677	-0.00810	0.99193
6	Nervous system	-2.95245	-0.01973	0.98047
7	Circulatory system	-2.15070	0.02855	1.02897
8	Respiratory system	-2.62739	0.00636	1.00638
9	Digestive system	-2.42207	-0.00970	0.99035
10	Skin and subcutaneous tissue	-6.58185	0.01111	1.01117
11	Musculoskeletal system and connective tissue	-6.19133	-0.00943	0.99062
12	Genitourinary system	-3.98500	0.01135	1.01142
13	Congenital, Malformations, Deformations and Chromosomal Abnormalities	-3.21959	-0.04983	0.95139
14	Symptoms, Signs & Abnormal clinical & Laboratory findings, not elsewhere classifies	-2.29013	0.00703	1.00706
15	Injury, Poisoning and certain other consequences of external causes	-0.70674	-0.03901	0.96174

Table 1 shows the estimated values of the parameters using equation (3.3). The value of exp(-0.01241) = 0.98766 can be interpreted as the death cause by certain infectious and parasitic disease was decreased by 1.23% times due to age. The value of exp(0.00622) = 1.00623 can be interpreted as the death cause by neoplasms was increased by 100.62% times due to age. The value of exp(-0.02319) = 0.97708 can be interpreted as the death cause by blood-forming organs and certain disorders involving immune mechanism was decreased by 2.29% times due to age. The value of exp(0.02415) = 1.02444 can be interpreted as the death cause by Endocrine, Nutritional and Metabolic diseases was increased by 102.44% times due to age. The value of exp(-0.0081) = 0.99193 can be interpreted as the death cause by Mental and Behavioural disorders was decreased by 0.81% times due to age. The value of exp(-0.01973) = 0.98047 can be interpreted as the death cause by diseases of the Nervous system was decreased by 1.95% times due to age. The value of exp(0.02855) = 1.02897 can be interpreted as the death cause by diseases of the Circulatory system was increased by 102.90% times due to age. The value of exp(0.00636) = 1.00638 can be interpreted as the death cause by diseases of the Respiratory system was increased by 100.64% times due to age.The value of exp(-0.0097) = 0.99035 can be interpreted as the death cause by diseases of the Digestive system was decreased by 0.97% times due to age. The value of exp(0.01111) = 1.01117 can be interpreted as the death caused by Skin and subcutaneous tissue was increased by 101.12% times due to age. The value of exp(-0.00943) = 0.99062 can be interpreted as the death caused by musculoskeletal system and connective tissue was decreased by 0.94% times due to age.The value of exp(0.01135) = 1.01142 can be interpreted as the death cause by diseases of the Genitourinary system was increased by 101.14% times due to age.The value of exp(-0.04983) = 0.95139 can be interpreted as the death cause by Congenital,

Malformations, Deformations and Chromosomal Abnormalities was decreased by 4.86% times due to age.The value of exp(0.00703) = 1.00706 can be interpreted as the death cause by Symptoms, Signs & Abnormal clinical & Laboratory findings, not elsewhere classifies was increased by 100.71% times due to age. The value of exp(-0.03901) = 0.96174 can be interpreted as the death cause by Injury, Poisoning and certain other consequences of external causes was decreased by 3.83% times due to age. Let the null hypothesis be Ho : Age group as a risk factor for death due to disease in 2017 for India. For testing null hypothesis, Glogit model for mortality of equation (3.3) is used and the estimated parameter values are presented in table (2).

Table 2: Estimates using G-logit model for Medically Certified Deaths (2017) in India

S.No	Causes of Death	β_1	β_2	exp(β_2)
1	Certain Infectious and Parasitic	-1.35299	-0.01521	0.98490
2	Neoplasms	-3.33122	0.01308	1.01316
3	Blood-forming organs & certain disorders involving immune mechanism	-2.91525	-0.02249	0.97776
4	Endocrine, Nutritional and Metabolic diseases	-4.22267	0.02489	1.02521
5	Mental and Behavioural disorders	-5.74447	-0.01092	0.98914
6	Nervous system	-2.81906	-0.02012	0.98008
7	Circulatory system	-2.12009	0.02799	1.02839
8	Respiratory system	-2.71736	0.00837	1.00840
9	Digestive system	-2.24480	-0.01326	0.98683
10	Skin and subcutaneous tissue	-6.26393	0.00789	1.00792
11	Musculoskeletal system and connective tissue	-6.27652	-0.00100	0.99900
12	Genitourinary system	-3.76412	0.00780	1.00783
13	Congenital, Malformations, Deformations and Chromosomal Abnormalities	-2.89643	-0.06766	0.93458
14	Symptoms, Signs & Abnormal clinical & Laboratory findings, not elsewhere classifies	-2.22864	0.00132	1.00132
15	Injury, Poisoning and certain other consequences of external causes	-0.89054	-0.03667	0.96400

Table 2 shows the estimated values of the parameters using equation (3.3). The values of exp(-0.01521) = 0.98490 can be interpreted as the death caused by certain infectious and parasitic disease was decreased by 1.51% times due to age. The value of exp(0.01308) = 1.01316 can be interpreted as the death cause by neoplasms was increased by 101.32% times due to age. The value of exp(-0.02249) = 0.97776 can be interpreted as the death cause by blood-forming organs and certain disorders involving immune mechanism was decreased by 2.22% times due to age. The value of exp(0.02489) = 1.02521 can be interpreted as the death cause by Endocrine, Nutritional and Metabolic diseases was increased by 102.52% times due to age. The value of exp(-0.01092) = 0.98914 can be interpreted as the death cause by Mental and Behavioural disorders was decreased by 1.09% times due to age. The value of exp(-0.02012) = 0.98008 can be interpreted as the death cause by diseases of the Nervous system was decreased by 1.99% times due to age. The value of exp(0.02799) = 1.02839 can be interpreted as the death cause by diseases of the Circulatory system was increased by 102.84% times due to age. The value of exp(0.00837) = 1.00840 can be interpreted as the death cause by diseases of the Respiratory system was increased by 100.84% times due to age. The value of exp(-0.01326) = 0.98683 can be interpreted as the death cause by diseases of the

Digestive system was decreased by 1.32% times due to age. The value of exp(0.00789) = 1.00792 can be interpreted as the death caused by Skin and subcutaneous tissue was increased by 100.79% times due to age. The value of exp(-0.00100) = 0.99900 can be interpreted as the death caused by musculoskeletal system and connective tissue was decreased by 0.10% times due to age. The value of exp(0.00780) = 1.00783 can be interpreted as the death cause by diseases of the Genitourinary system was increased by 100.78% times due to age. The value of exp(-0.06766) = 0.93458 can be interpreted as the death cause by Congenital, Malformations, Deformations and Chromosomal Abnormalities was decreased by 6.54% times due to age. The value of exp(0.00132) = 1.00132 can be interpreted as the death cause by Symptoms, Signs & Abnormal clinical & Laboratory findings, not elsewhere classifies was increased by 100.13% times due to age. The value of exp(-0.03667) = 0.96400 can be interpreted as the death cause by Injury, Poisoning and certain other consequences of external causes was decreased by 3.60% times due to age.

5. CONCLUSIONS

For both years 2015 and 2017 some of the diseases having severe impact of death due to age group and some having lesser impact. For 2015, the diseases caused by mental and behavioural disorders were decreased by 0.81% times due to age. The government has to take remedial actions to prevent these risk factors and review the mental health problems to increases the understanding of our knowledge about risk factors.

For 2017, the diseases caused by musculoskeletal systems have a minimum amount of decrease due to age group. This is the common problem of the employees across the world. Regular physical activity and exercises will help to reduce these types of disorders. Also in 2015 & 2017, the disease caused by circulatory system has a large amount of increase due to age group. From all causes of deaths this cause of death is the major problem which has the highest mortality at all ages.

REFERENCES

1. Butler W.J, R.M Park(1987), Use of the Logistic Regression model for the analysis of proportionate mortality data, National Library of Medicine, 125(3):515-523.

2. Donald A. Pierce, William H. Stewart and Kenneth J. Kopecky(1979), Distribution free Regression Analysis of Grouped Survival Data, Vol.35(4),785-793

3. Eric A. Roth and K. Balan Kurup (1990), Child mortality levels and survival patterns from Southern Sudan, Journal of Bio social Science (1990),22,365-372

4. Kakoli Rani Bhowmik and Sabina Islam,(2016), Logistic Regression and Multiple Classification Analyses Risk factors of under-5 Mortality in Banglades, Proceedings of the Pakistan Academy of Science:B.Life and Environmental Sciences 53(1):21-34 ISSN:-0377-2969 (print), 2306-1448(online)

5.Layla Aziz Ahmed (2017), Using Logistic regression in determining the effective variables in traffic accidents, Applied Mathematical Sciences, Vol. 11(42), 2017.

6. Mohamed M. Shoukri, Sara N. Algatani, Abdelmoneim M. Eldali, Manan R. Almarzouqi,(2019), Analysis of Hospital Mortality Data: The Role of DRG's, Open Journal of Statistics, Vol.9(1).

7. Mozol A N(2025), Predictive model of mortality based on logistic regression of laboratory Indicators,Pacific Medical Journal, DOI:10.34215/1609-1175-2025-2-45-49

8. Nur Ain Abd Aziz, Zalilaali, NorlidaMohd Nor, Adam Baharum and Maizurah Omar (2016), Modeling Multinomial Logistic Regression on Characteristics of Smokers after the Smoke-free campaign in the area of Melaka, Advances in Industrial and Applied Mathematics, AIP conference proceedings, AIP publishing, 978-0-7354-1407-5, Jun 21, 2016. https//doi.org/10.1063/1.4954625

9. Opemo Damian Otieno, Juma Shem Godfrey (2019), Logistic Regression analysis of mortality among fishermen in Riperian counties of Lake Victoria, Kenya, Central African Journal of Public Health, Vol. 5(1) Pg. 46-51, 2019. ISSN: 2575-5773(print); ISSN 2575-5781(Online).

10. Pandey A. P(2016), Socio-Economic factors of Contrast farming: A Logistic Analysis, IRA-International Journal of Management and Social Sciences, ISSN 2455-2267, Vol. 3(3), 2016.

11. Rohini P. Pande (2003), Selective gender differences in childhood nutrition and immunization in rural India. The role of Sibblings, Demography. Vol.40(3), 395-418

12. Roins, JM,Blevins D(1987),Analysis of proportionate mortality data using logistic regression models, American Journal of Epidemiology, Vol.125(3), 524-535

13. Sandri Sperandei(2013), Understanding logistic regression analysis, BiochemiaMedica, Vol 24(1), 2014.

14. SezginOzcan(2014), A Study on Casualty profile using Logistics Regression, Journal of Military and Information Science, Vol 2(1),2014.

15.Sharareh R. NiakanKalhori, MahshidNasehi, Xiao-Jun Zeng (2010), A Logistic regression model to predict high risk patients to fail in Tuberculosis treatment course completion, International Journal of Applied Mathematics, Vol 40(2), 2010