European Journal of Statistics and Probability

Vol.5, No.1, pp.37-43, February 2017

Published by European Centre for Research Training and Development UK (www.eajournals.org)

STOCHASTIC MODELLING OF HUMAN UNDER-FIVE AGE MORTALITY WITH SPATIALLY STRUCTURED COUNTY FRAILTY EFFECTS IN KENYA.

Kiche J. , Orwa G. , Ngesa O. , Kihara P, Levi M.
Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture & Technology, P. O. Box 62000-00200, Nairobi, Kenya

ABSTRACT*:* *Survival analysis examines and models the time it takes for events to occur. The prototypical such event is death, from which the name 'survival analysis' and much of its terminology derives. This research study develops a predictive model and considers the factors associated with under-five child mortality rates in Kenya in order to provide the solutions and interventions to organizations concerned with demographic data. As the human lifespan increases, more and more people are becoming interested in mortality. The aim of this study is to estimate the robust and reliable estimates of level and trend in under-five mortality in Kenya . Survival analysis techniques and frailty modeling will be used as the statistical tools for analyzing the time-event data. Both the survival parametric and non –parametric models and the frailty models will be fit to help us draw the required conclusions based on under-five child mortality rates. The results of this study will be used to assist in formulating appropriate health programs and policies that will reduce under-five human mortality.*

KEYWORDS: Hazard function, Stochastic process, Frailty models, Parametric regression.

INTRODUCTION

Under-five mortality is a major public health challenge in developing countries. It is essential to identify determinants of under-five mortality because this will assist in formulating appropriate health programs and policies. Once we can model human under-five mortality , we can look for ways to extend our lifespan and counteract the negative aspects of mortality .That is if we find out that there are other factors which influence the death rate apart from the age factor , we can now give the recommendations on how to improve on them.

To accelerate the achievement of SDG (Sustainable Development Goals) number 3,there is need for an effort to accelerate child survival and provision for a framework to improve indicators for children. The strategy is guided by the national Health Sector strategic plan2,(NHSSP2) and the vision 2030 medium Term Plan that aim to reduce inequalities in the health care services and improve on the child health indicators. Existing data reveals that in the 1990's, infant and childhood mortality declined rapidly in Kenya as a result of various global initiatives to improve child health. After many years of declining health indicators, recent data is showing an improvement in the mortality indicators for Kenyan children. The Kenya Demographic Health Survey (KDHS) 2008/09 shows that compared to the 2003 KDHS, the Infant Mortality Rate (IMR) improved to 52 to 77 per 1000 live births and the Under Five mortality Rate (UFMR) improving to 74 from 115 per 1000 live births. However the neonatal mortality rate only reduced marginally from 33 to 31 per 1000 live births contributing to 42% of the under five mortality compared to 29% in 2003(KDHS). Child survival remains an urgent concern. It is unacceptable that about

16000 children still die every single day, this is equivalent to 11 deaths occurring every minute. (Estimates developed by the UN inter-agency group for child mortality estimation) , this has led to more research in this study area to assist in lowering the rate of infant and child mortality in Kenya.

Spatial analysis

Spatial analysis is the process of examining locations, attributes and relationships of features in spatial data through overlay and other analytical techniques in order to address a question or gain useful knowledge. Spatial analysis extracts or creates new information from spatial data. Spatial analysis is how we understand our world that is, mapping where things are, how they relate, what it all means and what actions to take. From computational analysis of geographic patterns to finding optimum routes, site selection and advanced predictive modeling, spatial analysis is at the very heart of geographic information system (GIS) technology. Spatial analysis or spatial statistics includes any of the formal techniques which study entities using their topological, geometric or geographic properties. In a more restricted sense, spatial analysis is the technique applied to structures at the human scale ,most notably in the analysis of geographic data. Spatial data also known as geospatial data, is information about a physical object that can be represented by values in a geographic coordinate system. Generally speaking, spatial data represents the location , size and shape of an object on planet earth such as a building, lake, mountain or township. Spatial data may also include attributes that provide more information about the entity that is being represented. Geographic Information System (GIS) can be used to access, visualize, manipulate and analyze geospatial data.

Frailty models for survival data.

A frailty model is a random effects model for time variables, where the random effect (the frailty) has a multiplicative effect on the hazard. It can be used for univariate (independent) failure times, that is to describe the influence of unobserved covariates in a proportional hazard model. Frailty models are the survival data analog to regression models, which account for heterogeneity and random effects. A frailty is a latent multiplicative effect on the hazard function and is assumed to have unit mean and variance , which is estimated along with other model parameters. As usually understood in survival analysis, frailty models are extensions of the proportional hazards model that incorporate unobserved random multiplicative components into the hazard function.

The concept of frailty offers a convenient way to introduce unobserved heterogeneity and associations into models for survival data. In its simplest form, frailty is an unobserved random proportionality factor that modifies the hazard function of an individual or a group of related individuals. Frailty models in survival analysis presents a comprehensive overview of the fundamental approaches in the area of frailty models. The notion of frailty provides a convenient way to introduce random effects, association and unobserved heterogeneity into models for survival data. In statistical terms a frailty model is a random effects model for time to event data where the frailty has a multiplicative effect on the baseline hazard function.

In essence, the frailty concept goes back to work of Greenwood and Yule(1920) on ‘accident proneness’.The term frailty itself was introduced by Vaupelet al.(1979) in univariate survival models and the model was substantially promoted by its application to multivariate survival data in a seminal paper by Clayton(1978)(without using the notion frailty)on chronic disease incidences in families. Frailty models are extensions of the proportional hazards model which is best known as the cox model(cox 1972),the most popular model in survival analysis.

Mortality rate-under 5 (per 1000) in Kenya.

Mortality is the state of being mortal, or susceptible to death. Mortality rate is a measure of the number of deaths (in general or due to a specific cause) in a population. Scaled to the size of that population per unit of time. Mortality rate is typically expressed in units of deaths per 1000 individuals per year; thus a mortality rate of 9.5 (out of 1000) in a population of 1000 would mean 9.5 deaths per year in that entire population ,or 0.95% out of the total. Under-five mortality rate is the probability per 1000 that a newborn baby will die before reaching age five , if subjected to current age-specific mortality rates. Mortality rate; under-5(per 1000) in Kenya was last measured at 70 in 2013 according to World Bank.

Broad categories that help to explain child mortality trends according to this study include:

- Fertility behavior like spacing births.
- Mother's age.
- Mother's education.
- Sex of the child.
- Birth order.
- Household wealth.
- Region(county/locality).
- Place of delivery.
- Place of residence (urban/ rural)
- Mother's occupation.
- Maternal (Marital status)
- Birth weight of the child.
- Household(family size, wealth index, sanitation)
- Religion .
- The use of health services by mothers and for her children.
- Environmental health conditions like outbreak of diseases, hygiene.
- Nutritional status ,breast feeding and infant feeding.
- Social economic status.

REVIEW OF THE MODELS

Modelling Binomial Data

Suppose $Y_i \sim Binomial(n_i, p_i)$ and we wish to model the proportions Y_i/n_i.

Then $E(Y_i/n_i) = p_i$ while $var(Y_i/n_i) = \frac{1}{n_i}p_i(1-p_i)$ So our variance function is

$$V(\mu_i) = \mu_i(1-\mu_i)$$

Our link function must map from $(0,1) \rightarrow (-\infty, \infty)$. A common choice is

$$g(\mu_i) = logit(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \quad (1)$$

Exponential Family

Most of the commonly used statistical distributions, e.g. Normal, Binomial and Poisson, are members of the exponential family of distributions whose densities can be written in the form

$$f(y;\theta,\emptyset) = exp\left\{\frac{y\theta - b(\theta)}{\emptyset + c(y,\emptyset)}\right\} \quad (2)$$

where $\emptyset$ is the dispersion parameter and θ is the canonical parameter.

It can be shown that $E(Y) = b'(\theta) = \mu$ *and* $var(Y) = \emptyset b''(\theta) = \emptyset V(\mu)$

For a generalized linear model (glm) where the response follows an exponential distribution we have $g(\mu_i) = g(b'(\theta_i)) = \beta_o + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$

The canonical link is defined as $g = (b')^{-1}$

$$\Rightarrow g(\mu_i) = \theta_i = \beta_o + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

Canonical links lead to desirable statistical properties of the glm hence tend to be used by default. However there is no a priori reason why the systematic effects in the model should be additive on the scale given by this link.

Estimation of the Model Parameters

A single algorithm can be used to estimate the parameters of an exponential family using maximum likelihood method .

The log-likelihood for the sample $y_1, \ldots, y_n$ is

$$l = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\emptyset_i} + \mathrm{c}(y_i, \emptyset_i)$$

The maximum likelihood estimates are obtained by solving the score equations

$$s(\beta j) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\emptyset_i V(\mu_i)} * \frac{x_{ij}}{g'(\mu_i)} = 0 \quad (3)$$

for parameters β_j. We assume that $\emptyset_i = \frac{\emptyset}{a_i}$ where $\emptyset$ is a single dispersion parameter and a_i are known prior weights; for example binomial proportions with known index n_i have $\emptyset = 1$ and $a_i = n_i$.

The estimating equations are then

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{V(\mu_i)} * \frac{x_{ij}}{g'(\mu_i)} = 0$$

A general method of solving score equations is the iterative algorithm Fisher's Method of Scoring (derived from a Taylor's expansion of $s(\beta)$)

In the $r-th$ iteration , the new estimate $\beta^{(r+1)}$ is obtained from the previous estimate $\beta^{(r)}$ by

$$\beta^{(r+1)} = \beta^{(r)} + s(\beta^{(r)})E\left(H(\beta^{(r)})\right)^{-1} \quad (4)$$

where H is the Hessian matrix: the matrix of second derivatives of the log-likelihood.

It turns out that the updates can be written as

$$\beta^{(r+1)} = \left(X^T W^{(r)} X\right)^{-1} \; X^T W^{(r)} z^{(r)}$$

i.e. the score equations for a weighted least squares regression of $z^{(r)}$ on X with weights $W^{(r)} = diag(w_i)$, where $z_i{}^{(r)} = \eta_i{}^{(r)} + \left(y_i - \mu_i{}^{(r)}\right) g'\left(\mu_i{}^{(r)}\right)$

$$and \quad w_i{}^{(r)} = \frac{a_i}{V(\mu_i{}^{(r)})\left(g'(\mu_i{}^{(t)})\right)^2}$$

The Cox model

Let Y_i denote the observed time (either censoring time or event time) for subject i. Let C_i be the indicator that the time corresponds to an event (i.e. if $C_i = 1$ the event occurred and if $C_i = 0$ the time is a censoring time). Let $X_i = \{ X_{i1}, \dots X_{ip} \}$ be the realized values of the covariates for subject i. The hazard function for the Cox proportional hazard model has the form

$$\lambda(t/X_i) = \lambda_0(t) exp\left(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) = \lambda_0(t)\exp(X_i.\beta) \quad (5)$$

This expression gives the hazard rate at time t for subject i with covariate vector (explanatory variables) X_i

Ignoring ties for the moment, conditioned upon the existence of a unique event at some particular time t the probability that the event occurs in the subject i for which $C_i = 1$ and $Y_i = t$ is

$$L_i(\beta) = \frac{\theta_i}{\sum_{j:Y_{j\geq Y_i}} \theta_j}$$

Where $\theta_j = exp(X_j.\beta)$. Observe that the factors of $\lambda_0(t)$ that would be present in both the numerator and denominator have canceled out.

Treating the subjects' events as if they were statistically independent, the joint probability of all realized events conditioned upon the existence of events at those times is the partial likelihood:

$$L(\beta) = \prod_{i:C_{i=1}} \frac{\theta_i}{\sum_{j:Y_{j\geq Y_i}} \theta_j}$$

The corresponding log partial likelihood is

$$l(\beta) = \sum_{i:C_i=1} \left(X_i.\beta - log \sum_{j:Y_{j\geq Y_i}} \theta_j \right) \quad (6)$$

This function can be maximized over β to produce maximum partial likelihood estimates of the model parameters.

The partial score function is

$$l'(\beta) = \sum_{i:C_i=1} \left(X_i - \frac{\sum_{j:Y_j \geq Y_i} \theta_j X_j}{\sum_{j:Y_j \geq Y_i} \theta_j} \right) \quad (7)$$

and the Hessian matrix of the partial log likelihood is

$$l''(\beta) = -\sum_{i:C_i=1} \left(\frac{\sum_{j:Y_j \geq Y_i} \theta_j X_j X'_j}{\sum_{j:Y_j \geq Y_i} \theta_j} - \frac{\left[\sum_{j:Y_j \geq Y_i} \theta_j X_j\right]\left[\sum_{j:Y_j \geq Y_i} \theta_j X'_j\right]}{\left[\sum_{j:Y_j \geq Y_i} \theta_j\right]^2} \right) \quad (8)$$

METHODOLOGY

Model developed

The survival time data is normally presented together with the other factors that aid in describing the survival time. These factors are such as; sex, health status , age, education , area of residence, religion, wealth index, mother's occupation, duration of breastfeeding, birth order ,birth weight of the child , among other factors. With this additional information one may decide to model the survival time including such information in the model , the resulting model after considering the available information (factors / covariates) is referred to as a regression model. A regression model can be parametric or semi-parametric.

Consider a vector of covariates $\boldsymbol{X}$, in this case the regression model linking the covariates to the hazard is given by:

$$h(t/\boldsymbol{X}) = h_0(t)\ \text{ɳ}(\boldsymbol{X}^T\beta)+u_i + v_j + e_{ij} \qquad (9)$$

where $h_0(t)$ is the baseline hazard, β is the vector of the regression parameters and $\text{ɳ}(\boldsymbol{X}^T\beta)$

is the link function, u_i represents county i and v_j represents randomly selected county that has more number of under five child mortality , e_{ij} is the error term .

The link function has to be a positive function since the hazard function is always positive.

The best choice for the link function is the exponential

$$\text{ɳ}(X) = \exp\ (\mathrm{X}) > 0, \quad \forall\, \mathrm{X}$$

from the above link exponential function , the model becomes

$$h(t/\boldsymbol{X}) = h_0(t)\ \exp\ (\boldsymbol{X}^T\beta)+ u_i + v_j + e_{ij} \qquad (10)$$

Cross-sectional data

in this research, we perform analysis of cross-sectional data which consists of comparing the differences among the subjects. Cross-sectional data, or a cross-section of a study population, in statistics, is a type of data collected by observing many subjects (such as individuals ,counties or regions) at the same point of time or without regard to difference in time.

Geographically Weighted Regression

Geographically weighted regression (GWR) is a local form of spatial analysis introduced in 1996 in the geographical literature drawing from statistical approaches for curve-fitting and smoothing applications. The method works based on the simple yet powerful idea of estimating local models using subsets of observations centered on a focal point. Since its introduction, GWR rapidly captured the attention of many in geography and other fields for its potential to investigate non stationary relations in regression analysis. The basic concepts have also been used to obtain local descriptive statistics and other models such as Poisson regression and probit. The method has been instrumental in highlighting the existence of potentially complex spatial relationships.

The evidence available suggests that GWR is a useful, if imperfect, tool for inferring spatial processes, and a relatively simple and effective tool for spatial interpolation. Related technical developments enhance GWR (e.g., autocorrelation tests and multiple comparison adjustments) and/or complement it (e.g., the expansion method). Other approaches provide alternatives to the use of GWR (e.g., kriging and Bayesian models).

Research Study Area

The study sites being considered in this research are the counties that are in Kenya.
The sites have resources which can increase the likelihood of the success of the study.