Vol.4.Issue.4.2016 (October -December)

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

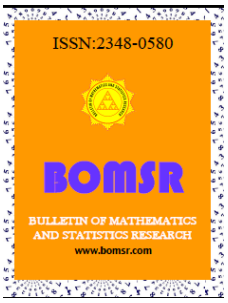
http://www.bomsr.com
Email:editorbomsr@gmail.com

ISSN:2348-0580

REVIEW ARTICLE

CONTEMPORARY STATISTICAL MODELS IN EPIDEMIOLOGICAL STUDIES: A REVIEW

VINAY KUMAR
Department of Statistics
M.D. University, Rohtak (Haryana) -124001
vinay.stat@gmail.com



ABSTRACT

Over the years, several statistical models are being used to determine the disease pattern in epidemiological studies. Estimation of the effect and prevalence of the disease has also been a major area in these studies. Such modelling and estimation is immensely useful for planning the health policies and a sound health status of the society. This paper is an attempt to present an overview of various statistical models and their uses in different epidemiological set-ups. An attempt is also made to present an annotated bibliography of the same.

Key Words: Logistic regression, Poisson regression, Gibbs sampler and Hazard model

1. INTRODUCTION

Numerous approaches have been used to clarify the epidemiologies and epidemiological determinants have been explored using different techniques of statistical methods. Here we aim to classify the contribution of statistical modelling and estimation under the different epidemic and share knowledge of the basic's theory regardless of the reader's statistical background. The present paper deals with an overview of the some of statistical modelling approach in epidemiological research, highlighting on some of the generally employed models for the analysis of the data like general modelling approach, frequentist approaches, Bayesian approaches etc. Readers requiring more detailed statistical presentation may refer to standard text books (Breslow and Day, 1980, 1987).

2. Statistical Modelling

2.1 General modelling approach: General analytical approach involved in the modelling is to express some function of disease occurrence (dependent variable). The model takes in the general form

$$f(z) = a_0+a_1x_1+----------+a_kx_k,$$

where z represents the risk, rate or odds of the disease in persons with characteristics x_1, x_2, x_3.....x_k and a_1 represents regression coefficient for x_1. Similarly a_2.....a_k represents the corresponding regression coefficients for the characteristics x_2,.....,x_k respectively. The vital role of the above model are that the specific function of the outcomes of interest depends on the exposures only through the quantity ($a_0+a_1x_1+$----------$+a_kx_k$) which is defined as linear predictor.

The family of models employing a logarithmic transformation is particularly suited to estimate ratio, rate and odds of disease or multiplicative measures of effect. Hence a logarithmic transformation is used so that the dependent variable in the model is log (z) i.e. logarithm to base "e". The transformation yields a function that has theoretical range of minus infinity to plus infinity. The coefficients in the model are obtained by the method of maximum likelihood and the method is based on likelihood function; which represents the probability of observing the data as a function of the unknown parameters ($a_0+a_1x_1+$----------$+a_kx_k$). The maximized value of the logarithm of the likelihood can also be used to obtained the log likelihood statistics which is also known as deviance (Breslow and Day, 1980)

The modelling facilitates the consideration of the simultaneous effects of several different exposure variables on risk factors, recognise the role of chance mechanism, help to control of confounding, and in estimating the effect modification between several factors. Estimates obtained by model fitting have greater numerical stability than those computed from standardised rates. There are some models which have better outfits while fitting which are following

2.1.1 Logistic regression model:

When dependent variable (outcomes variable) happens to be binary in nature i.e. an event occurring or not , taking the values unity or zero, the assumption necessary for fitting multiple linear regression model of the type Z= α + $\sum_{i=1}^{k} \beta iXi$, is violated as it is unreasonable to assume that distribution of errors as normal. For such data, instead of multiple linear regression analysis, multiple logistic regression (LR) analysis is used as a multivariate procedure. We prefer LR models while using dependent variable such as we develop a model based on the Logit transformation of the dependent variable to satisfy the needed assumptions. Thus, in LR model we predict the proportion of subjects with particular characteristics or equivalents, the probability individual with those characteristics for any combination of the explanatory variable.

$$\text{Logit } p(Z= 1/x) = \text{Logit } p= \log_e (p/ 1-p)$$
$$= \alpha+ \sum_{i=1}^{k} \beta iXi$$

The above model enables us to estimate the probability of happening of an event (Z= 1/x) :

$$P(Y=1/x) =\exp (\alpha+ \sum_{i=1}^{k} \beta iXi)/ 1+\exp (\alpha+ \sum_{i=1}^{k} \beta iXi)$$

Where P(Y=1/X) denotes the disease probability in stratum i for an individual with set of regression variables x_i, α is a constant term and represents the log odds of the disease risk for a person with standard set when all the regression variable are zero. The β's are regression coefficients and indicates the fraction by risk is increased (or decreased) for every unit change in x_i, exponential of βi represents the odds ratio.

2.1.2 **Modelling of data of survival studies**:

In many longitudinal studies the outcomes variable is the time elapsed between the entry of a subjects into study and the occurrence of an event thought to be related treatment. To attain a uniform terminology in such studies, the event of interests has been referred to as "death (failure)" and the outcomes variable as survival time. A serious complication in the analysis of survival times is the possibility of censoring, that is, of subjects not being observed for the full period until the

occurrence of the event. This leads to incomplete data due to censored observations. In survival studies censoring is more apart of the Exponential and Wiebull distributions were introduced to model survival experiences of homogenous populations incorporating the censoring schemes.

2.1.3 Exponential and Wiebull regression models

The Exponential distribution has been generalized to obtain regression model by allowing the hazard rate to be a function of the covariate "X". The exponential distribution model is :

$$h(t, x) = he^{\beta X}$$

Where h(t , x) is the hazard at time "t" for individual with a given set of covariates, β is a vector of unknown regression coefficients and h is a constant. The model assumes the multiplicative relationship between the hazard function and the effect of covariates and specifies that log failure rate is a linear function of the covariates. The hazard rate over time period is constant. Where as, the wiebull distribution is a generalization of the exponential distribution. However, it has a hazard rate which may have different shapes. For p=1, the distribution has a constant hazard rate (exponential distribution). The model is $h(t , x) = hp (ht)^{p-1} e^{\beta X}$

2.1.4 The Cox-proportional hazards model

The exponential and weibull models involve stronger distributional assumption than are suitable and inference procedures may not be sufficiently robust to departures from these assumptions. The distribution of survival times must be known to apply these models. In most studies, however, the distribution of survival times is unknown and can vary from one disease to another. Hence, in order to take into diversity of situations, which are encountered in practice, Cox in 1972 developed modelling procedures termed as Cox- proportional hazards model. This model is an important tool in the cohort and survival studies for modelling the effect of risk factors when the outcome of interest occurs with time. It measures the relative risk of outcomes under the assumption that the relative risk is constant over the follow- up period. In this model, the hazard for an individual is a product of a common baseline hazard and a function of a set of risk factors. The model utilizes both the risk factors and the rank order of time of occurrence.

The mathematical expression of the model is:

$$H (t , x) = h_0(t) \exp (b_1X_1+----------+b_kX_k).$$

Where $X =(X_1, ----------, X_k)$ is a k dimensional vector of covariates (risk factors), $h_0(t)$ is the base line hazard rate at time t when all the covariates are zero, b_j , j= 1,2,--------,k are regression estimates and h(t, x) is the hazard rate for an individual at time t with covariates X_i , i=1.2.-----,k, the estimates of b_i, depends on the rank ordering of occurrence of outcomes event and does not depend on the exact time of occurrence of events.

2.1.5 Poisson regression model

Poisson regression (PR) model is an important method of analysis for data set in the cohort study. In this model, the logarithm of the incidence rate is modelled as a linear combination of a set of risk factors. It measures the relative risk of outcome under the assumption that the number of outcome events is small in comparison with the total cohort.

The mathematical model is:

$\log_e\lambda_j = \alpha_j + X_{jk}b_k$

Where λ'^s are the unknown true rates of the outcome of interest, "α_j" are the nuisance parameters specifying the effects of the stratification variables.

2.2 FREQUENTIST APPROACH: GENERALIZED LINEAR MIXED MODELS

Estimation in classical GLMs is the likelihood based while the GLMMs estimation is based on quasi-likelihood approximation. These models incorporate a random component to account for the correlation within cluster or units, heterogeneity and overdispersion from the longitudinal,

hierarchical or clustered data structures. A GLMM is a member of a class of statistical models that combines GLMs and the ideas from the Linear Mixed Model with normal random effects. GLMMs assume data is from exponential family of distributions. Linear models characterize fixed effects only apart from model errors. In GLMM, as in the LMM, the linear predictor can contain random effects such that

$$\eta = X\beta + Zu$$

where β is a vector of fixed effects while u is the vector of random effects, X and Z are design matrices for fixed effects and random effects, respectively. The conditional mean $\mu|u$, relates to linear predictor through a link function

$$g(\mu|u) = \eta$$

The expected value of the random vector Y conditional on u is given by

$$E(Y|u) = \mu$$

Where $Y = (Y_{i1}, \ldots\ldots, Y_{ini})$ is from cluster i and the variance is given by

$$Var(Y|u) = \phi V(u),$$

where ø is dispersion parameter. The distribution of the random effects u, is assumed to be *N(0, G).* The conditional distribution of the data is a member of an exponential family. Estimation of parameter is not that different from that of the GLMs except that now the linear predictor includes an extra term representing the random effects.

2.3 Bayesian approach: The Gibbs Sampler and its implementation

The Gibbs sampler is a Markovian updating scheme for extracting samples from posterior, typically available up to proportionally, obtained as a product of the likelihood function and a prior. The scheme proceeds via iterated sampling from the various full conditional forms again specified up to proportionality from the joint posterior, treating, for each unknown in turn, every other quantity as fixed known constant.

The roots of the MCMC methods come from the Metropolis Algorithm attempted by physicists to compute complex integrals by expressing them as expectations for some distribution and then estimate this expectation by drawing samples from that distribution. The Gibbs sampler has its origins in image processing. The Gibbs sampler is a special case of the Metropolis-Hasting algorithm. Gibbs sampler has been found to be very useful in many multidimensional applications. In Gibbs sampler one needs only to consider the univariate conditional distributions. These conditional distributions have simple forms and are easier to simulate than complex joint distributions.

Consider a bivariate random variable (*x,y*) and suppose we wish to compute one or both marginal s*p(x)* and *p(y)* . The idea behind the sampler is that it is far easier to consider a sequence of conditional distributions p(x|y) and p(y|x) than it is to obtain the marginal's by integration of the joint density p(x,y). The sampler starts with some initial value y_o for y and x_o for x by generating a random variable from the conditional distribution $p(x|y)=y_o$). The sampler uses x_o to generate a new value of y_1, drawing from the conditional distribution based on the value x_o , $p(y|x= x_o)$. The sampler proceeds as follows

$$X_i \sim P(x|y = y_{i-1})$$

$$Y_i \sim P(y|x = x_i)$$

This process is repeated K times, generating a Gibbs sampler of length k, where the subset of points (x_i,y_j) for $1 \le j \le m < k$ are taken as the simulated draws from the full joint distribution. One iteration of all the univariate distributions is often called a 'scan' of the sampler. The Gibbs sampler sequence converges to a stationary distribution that is independent of the starting values and by construction this stationary distribution is the target distribution we are trying to simulate from. Powers of the Gibbs sampler to address a wide variety of statistical issues have been studied. The Gibbs sampler

can be thought of as a stochastic analog to the Expectation-Maximization approaches used to obtain likelihood functions when missing data are present. In the sampler, random sampling replaces the expectation and maximization steps. Any feature of interest for marginal can be computed from the m realizations of the Gibbs sequence.

3. Application

With the introduction of number of computer statistical packages, such as SPSS, SAS, STATA, Logistic Regression and Survival Analysis techniques have become accessible to wider audiences of investigators. However, the model should be attempted under an expert guidance in order to arrive at proper interpretations. The above models should been widely employed in the analysis of data in the field of epidemiological investigations for evaluating the risk factors and defining the high risk groups. Caution about modelling: first these techniques have underlying statistical assumptions that may not be valid for the data under consideration. Secondly, a particular model may provide an inadequate description of the true relationship under investigation. Thirdly, erroneous conclusions can be made from the results of statistical computer packages if the user is unfamiliar with the coding scheme employed for categorical data.

4. Caution about Modelling:

First, these techniques have underlying statistical assumptions that may not be valid for the data under consideration. Secondly, a particular model may provide an inadequate description of the true relationship under investigation. Thirdly, erroneous conclusions can be drawn from the results of statistical computer packages if the user is unfamiliar with the coding scheme employed for categorical data. The model building should be attempted under an expert guidance in order to arrive at proper interpretations.

References

[1]. Aalen, Heterogeneity in survival analysis.Stat.in Medicine. 7, 1121-1137, 1988

[2]. Altman and De Stavola B, Practical problems in fitting a proportional hazards model to data with updated measurements of covariates. Stat. Med. 13,301-341, 1994

[3]. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian adaptive methods for clinical trials*. Chapman & Hall: Boca Raton,2010.

[4]. Breslow NE and Day NE, Statistical methods in cancer research. Vol1, The analysis of case control studies, international Agency for Research on Cancer, Lyon, 1980

[5]. Breslow NE and Day NE, Statistical methods in cancer research. Vol 2, The analysis of cohort studies, International Agency for Research on Cancer, Lyon, 1988

[6]. Clayton D, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 65, 141–151, 1978.

[7]. Clayton D , A Monte Carlo method for Bayesian inference in frailty models. Biometrics. 47, 467–485, 1991.

[8]. Cox DR, Regression models and life tables; J R Stat Soc. B34;187-220,1972

[9]. Draper, D , Bayesian Hierarchical Modelling ,2000

[10]. Mccullagh,P and Nelder ,J.A ; Generalized Linear Models, Chapman and Hall, 1989

[11]. Gelman, A , Carlin, J.B , Stern, H.S , Rubin,D.B , Bayesian Data Analysis ; Chapman and Hall.1998

[12]. KleinbaumDavid,G , Survival Analysis: Statistics in Health Sciences. Springer-Verlag, New York,Inc,1996

[13]. Walsh, B , Markov Chain Carlo and Gibbs sampling, Lecture Notes, 2004