International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 3 (March 2014), PP.64-71

Statistical Model For Risk Estimation

[1]Dr. Vahida Attar, [2]Dr. D. Datta

[1]Department of computer and IT, College of Engineering Pune ,Shivaji Nagar, Pune India.
[2]Department of atomic energy, BRNS Mumbai, India.

Abstract:- The study of the distribution and determinants of disease prevalence in man is done primarily in Radiation Epidemiology. Epidemiologists seek to relate risk of disease to different levels and patterns of radiation exposure. In this paper we examine statistical model of Poisson regression previously employed for the estimation of radiation risk. We examine different regression techniques, which overcome the underlying assumptions of Poisson Regression for risk estimation and propose Hurdle's Model for the same. The models need application of logarithmic transform to yield the additive model instead of multiplicative model, which is usually used in Risk Assessment and thus obtain the Linear Dose-Response Model.

I. INTRODUCTION

Epidemiology is concerned with study of distribution of disease and determinants of health-related states or events in specified human populations and application of this study for control of human health problems. It is observed that, people exposed to radiation usually suffer from cancer and other fatal diseases. For instance, there are two ways in which nuclear workers are exposed to radiation according to the Canadian Nuclear Safety Commission, either while working with sources of man-made radiation (nuclear industry, health care, research institutions or manufacturing) or they are exposed to elevated levels of natural radiation (mining, air crews construction).

Radiation is categorized as ionizig and non-ionizing.[8] Ionizing radiation is radiation with enough energy so that during an interaction with an atom it can remove bound electrons, i.e., it can ionize atoms. Examples are X-Rays and electrons.Non-ionizing radiation is radiation without enough energy to remove bound electrons from their orbits around atoms. Examples are microwaves and visible light.The ionizing radiation interacts with the cells and damages them, which in turn results in malignant growth in the body. Thus studying the levels of radiation and its corresponding effects can prove beneficial in setting the safety levels of exposure.

In our study we intend to develop a generalized model for risk evaluation or odds ratio for death due to cardiovascular disease and other cancer due to ionizing radiation. Radiation Effects Research Foundation has played an important role of carrying out cohort study on the Japanese Atomic Bomb survivors with a follow-up of 50 years. This report makes use of data obtained from the Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan. RERF is a private, non-profit foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the U.S. Department of Energy, the latter through the National Academy of Sciences. The objective is to apply suitable methods to gain further insight in the model developed by RERF on Cardiovascular diseases. Main outcome is to measure Mortality from stroke or heart disease as the underlying cause of death and dose response relations with atomic bomb radiation. This finding would significantly benefit the humanity as with growing technology, there are associated hazards. This is an interdisciplinary problem as it encompasses Epidemiology, Statistics and Data Analysis Methods. The paper starts discussion of prevalent risk models for low-ionized radiation. Aanalyse the strengths and limitations of the models used. This is followed by provision of theoretical background of Poisson Model, used for count data and estimation of relative risk. It comprises of multiplicative and additive models of relative risk. Finally it proposes the use of variations of Poisson Regression called Hurdle model.

II. RISK OF RADIATION EXPOSURE

Comparing the acute exposure experienced by atomic bomb survivors with the low dose rate exposures experienced gradually over time due to occupational, environmental or natural background circumstances is now frequently done. Wide ranges of risk estimates have been reported with some significantly lower than the estimate of cancer incidence among atomic bomb survivors and some higher but not significantly so. The differences are in large part related to chance, different dose ranges, different lengths of follow-up, differences in ethnicity and background cancer rates, and biases and confounding that play a much more important role when studying populations exposed to low doses.

Manmade verses Natural Resources radiation exposure: Whether it is taken from external exposure or from intakes of radioactive material . Depending on dose: The lower the dose, the lower the risk but the lower the dose, the greater the difficulty in detecting any increase in the number of cancers possibly attributable to radiation. In higher doses the risk is also higher but the chases of detecting the cancer are also more. At doses below 100 millisieverts (mSv), it is not possible to distinguish cancer due to radiation from that of the natural variation of the disease among the general population.

Radiation epidemiology has revealed that ' a single exposure can increase the lifetime risk of cancer; the young are more susceptible than the old (although not markedly so); females are more susceptible than males; the foetus is not more susceptible than the child; genetic (heritable) effects have not been found in humans to date; risks differ by organ or tissue and some cancers have not been convincingly increased after exposure. Many small exposures over years can significantly decrease lifetime.'

Epidemiologists use the term "risk" in two different ways to describe the associations that are noted in data. Relative risk is the ratio of the rate of disease among groups having some risk factor, such as radiation, divided by the rate among a group not having that factor. Absolute risk is the simple rate of disease among a population and Excess absolute risk (EAR) is the difference between two absolute risks.[2]

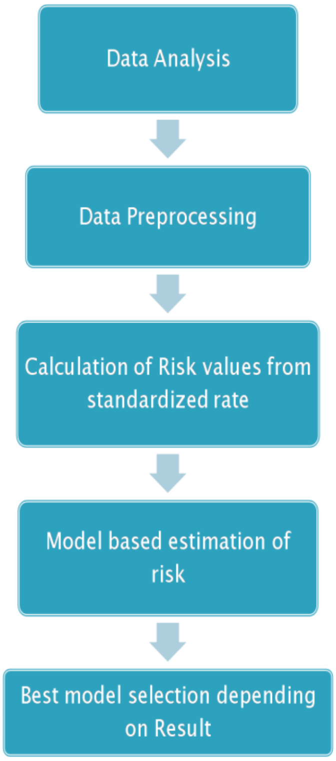

Data Analysis
Data Preprocessing
Calculation of Risk values from standardized rate
Model based estimation of risk
Best model selection depending on Result

Figure 1. Modelling of data

III. RISK MODELS

Poisson regression methods for grouped survival data were used to describe the dependence of risk on radiation dose and to evaluate the variation of the dose response with respect to city, sex, age at exposure, and attained age. Significance tests and confidence intervals (CI) were based on likelihood ratio statistics. The results were considered statistically significant when the two-sided $P < 0.05$. The models used here are as follows.

Excess Relative Risk (ERR) model:

$$\lambda_0\ (c, s, a, b)\ [1{+}ERR\ (d, e, s, a)] \qquad (1)$$

Excess Absolute Risk (EAR) model:

$$\lambda_0\ (c, s, a, b)\ [1{+}EAR\ (d, e, s, a)] \qquad (2)$$

Where λ_0 is the baseline or background mortality rate at zero dose, depending on city (c), sex (s), birth year (b), and attained age (a). λ_0 was modelled by stratification for the ERR model and by parametric function involving relevant factors for the EAR model. ERR or EAR depends on radiation dose (d) and, if necessary, effect modification by sex, age at exposure (e), and attained age.

Effect modification was described using multiplicative-function models as follows:

$$\varepsilon(e,s,a) = exp(\tau.e+\upsilon.ln(a))(1+\sigma.s) \qquad (3)$$

where τ, υ and σ were the coefficients for effect modification by age at exposure, attained age, and sex, respectively. The term that includes sex (s = 1 for men and s = -1 for women) as a modifier allows the β_1 parameter to represent sex-averaged risk estimates.[1] Therefore, ERR and EAR models were, respectively,

$$\lambda_0(c,s,b,a)[1+\beta_1 d.exp(\tau e+\upsilon.ln(a))\ (1+\sigma s)] \qquad (4)$$

$$\lambda_0(c,s,b,a)[\beta_1 d.exp(\tau e+\upsilon ln(a))\ (1+\sigma s)] \qquad (5)$$

3.1 Strengths

This study has several strengths, including a large population not pre-selected for existing disease or occupational fitness, a wide but relatively low dose range (0->3 Gy) and well characterized doses, a 53 year follow-up with virtually complete mortality ascertainment, and corroborative evidence from more detailed clinical and biomarker studies of risk of circulatory disease on a random subsample of the cohort. The analyses of radiation dose with stroke and heart disease mortality showed that the association is reasonably robust with respect to confounding by lifestyle, sociodemographic, or other health factors or misdiagnosis.

3.2 Limitations

The model also has several limitations and uncertainties. Ascertainment of circulatory disease from death certificates is of limited diagnostic accuracy and represents only a fraction of cases of incident disease. Some selection effects due to dose related early mortality from the bombs may have occurred, although the impact of these is likely to be small. Other limitations include unclear dose-response effects below about 0.5 Gy, inadequate information about possible biological mechanisms, and uncertainty about the generalisability of these results to Western populations because of differences in genetic factors, dietary and lifestyle risk factors, and baseline levels of risk for stroke and heart disease.

Another problem is excess zeros. In this situation, the distribution has more zeros than would be expected from a Poisson distribution. Often this is caused by two processes creating the data set, one of the processes having an expected count of zero. Thus model is always overly restrictive when it comes to estimating features of the population other than the mean, such as the variance or the probability of single outcomes.[1]

3.3 Model for Group Data

The data layout consists of a table with J rows (j = 1. . . J) And K columns (k = 1. . . K). Within the cell formed by the intersection of the j^{th} row and k^{th} column, one records the number of incident cases or deaths d_{jk} and the person-years denominators n_{jk} where j is used for indexing J age intervals and k for representing one of K exposure categories.

Observed rate $\lambda_{jk}=d_{jk}/n_{jk}$ (6)

d_{jk} =No of deaths, n_{jk} = person years

This is considered as an estimate of a true rate λ_{jk} that could be known exactly only if an infinite amount of observation time were available. The goal of the statistical analysis is to uncover the basic structure in the underlying rates λ_{jk}, and in particular, to disentangle the separate effects of age and exposure.[3]

Various possible structures for the rates satisfy the requirement of consistency. In particular, it holds if the effect of exposure at level k is to add a constant amount β_k to the age-specific rates λ_{j1} , for individuals in the baseline or non-exposed category (k = 1).The model equation is :

$$\lambda_{jk} = \alpha_j + \beta_k \tag{7}$$

Where $\alpha_j = \lambda_{j1}$ and β_k ($\beta_1 = 0$) are parameters to be estimated from the data. If additivity does not hold on the original scale of measurement, it may hold for some transformation of the rates. The log transform

$$log\, \lambda_{jk} = \alpha_j + \beta_k \tag{8}$$

yields the multiplicative model where

$$\lambda_{jk} = \Theta_j * \Psi_k \tag{9}$$

here, $\alpha_j = \log(\Theta_j)$, $\beta_k = \log(\Psi_k)$, Ψ_k = Relative risk of decease at exposure level k.

The excess (additive) and relative (multiplicative) risk models are ubiquitous models to describe the relationship between the effects of exposure and the effects of age and other factors that may account for background or spontaneous cases. These have been used to describe different aspects of radiation carcinogenesis in human populations (Committee on the Biological Effects of Ionizing Radiation, 1980).[2]. Due to the sharp rise in background incidence with age, relative risk estimates derived from current data generally predict a greater lifetime radiation risk than do the estimates of additive effect.[3]

3.4 Multiplicative Model for Rate

The basic data consist of the counts of deaths d_{jk} and the person-years denominators nik in each cell, together with p-dimensional row vectors $x_{jk} = (xjk^1,\ldots,xjk^p))$ of regression variables . These latter may represent either qualitative or quantitative effects of the exposures on the stratum-specific rates, interactions among the exposures and interactions between exposure variables and stratification (nuisance) variables.

A general form of the multiplicative model is:

$$log\, \lambda_{jk} = \alpha_j + x_{jk} * \beta \tag{10}$$

where the λ_{jk} are the unknown true disease rates, the $\alpha_{j\text{ are}}$ nuisance parameters specifying the effects of age and other stratification variables, and $\beta= (\beta_1,\beta_2,...,\beta_p)^T$ is a p-dimensional column vector of regression coefficients that describe the effects of primary interest. An important feature of this model is that the disease rates depend on the exposures only through the quantity $\alpha j + x_{jk} * \beta$, which is known as the linear predictor. If the regression variables x_{jk} depend only on the exposure category k and not on j, above equation specifies a purely multiplicative relationship such that the ratio of disease rates $\lambda_{jk}/\lambda_{jk}'$, for two exposure levels k and k', namely exp {(xk - xk')β), is constant over the strata.

3.5 Additive Model for Risk

The limitation of multiplicative model is that when applied with quantitative exposure variables, it leads to relative risk functions that increase exponentially with increasing exposure: RR(x) = exp(xβ). This need not be applicable to different diseases and thus one needs to apply suitable transform. Many of the quantitative dose-response relations actually observed in cancer epidemiology approximate a power relationship of the form

$$RR(x) = (x + x0)\hat{}\beta \qquad (11)$$

This relative risk function may be approximated by first transforming the dose to z = log (x + x0) and then fitting the multiplicative model in the form

$$log(\lambda_{jk}) = \alpha_j + z_k * \beta = \alpha_j + log(x_k + x_0) * \beta \qquad (12)$$

The choice xo= 1 is not uncommon as a 'starter' dose since it yields the usual RR(x) = 1 at the baseline level x = 0. xo may also be treated as an unknown parameter and the best fitting value, found by trial and error or some other more systematic technique. Certain formulations of multistage theory and other more general considerations lead to relative risk functions that are linear or quadratic in measured exposures, for example

$$RR(x) = 1+\beta x \text{ or } RR(x) = 1 + \beta_k + \gamma x^2 \qquad (13)$$

These are special cases of a general class of models of the form

$$\lambda_{jk} = exp\ (\alpha_j)\{l + x_{jk}\}\beta \qquad (14)$$

One drawback of these is that the range of the β parameters is necessarily restricted by the requirement that $x_{jk}*\beta > -1$ for all values of x_{jk}, since negative relative risks would otherwise result. This suggests that, wherever possible, the regression variables x_{jk} be coded so that they have positive coefficients.

As usually happens for models in which there is a range restriction on the parameters, the log-likelihood function is skewed and not at all like the quadratic, symmetric log-likelihood of the approximating normal distribution. Estimates of the parameters may be unstable, and standard errors that are determined from the usual likelihood calculations may be unhelpful in assessing the degree of uncertainty. With suitable transformation, we fit the additive relative risk model to Poisson rates and thus obtain a family of general relative risk modes which is given by

$$\lambda_{jk} = exp\ (\alpha_j)\ r(\ x_{jk}\ ,\ \beta) \qquad (15)$$

where the relative risk function is specified by the power relation

$$\begin{aligned} &log\ r\ (x_{jk}*\beta) = (1+ x_{jk}*\beta)\hat{}\rho/\rho \quad \rho != 0 \\ &log\ r\ (x_{jk}*\beta) = log(1+ x_{jk}*\beta) \quad \rho = 0 \end{aligned} \qquad (16)$$

This yields the additive relative risk model at ρ = 0 and the standard multiplicative model at ρ= 1.[3]

IV. PROPOSED MODEL

4. 1 Poisson Regression Model

The typical Poisson regression model expresses the natural logarithm of the event or outcome of interest as a linear function of a set of predictors. The dependent variable is a count of the occurrences of interest e.g. the number of cases of a disease that occur over a period of follow-up. Typically, one can estimate a rate ratio associated with a given predictor or exposure. A measure of the goodness of fit of the Poisson regression model is obtained by using the deviance statistic of a base-line model against a fuller model.

When the response variable is in the form of a count and for rare events the Poisson distribution (rather than the Normal) is more appropriate since the Poisson mean > 0. So the logarithm of the response variable is linked to a linear function of explanatory variables such that

$$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \qquad (17)$$

The basic Poisson regression model relates the probability function of a dependent variable y_i (also referred to as regress and, endogenous, or dependent variable) to a vector of independent variables x_i (also referred to as repressors, exogenous, or independent variable).[9] The standard uni-variates Poisson regression model makes the following assumptions:

- Logarithm of the disease rate changes linearly with equal increment increases in the exposure variable.
- Changes in the rate from combined effects of different exposures or risk factors are multiplicative.
- At each level of the covariates the number of cases has variance equal to the mean.
- Observations are independent.

This paper uses the Poisson Model with variation to obtain the risk coefficients. This section describes the statistical models used in estimation of risk. These risk estimates are of prime interest for the epidemiologist. Moreover, the coefficients obtained for risk, depend on the statistical model used for fitting the data. Therefore the type of model used for fitting the data determines the risk estimates and their accuracy.

Poisson regression is appropriate when the conditional distributions of Y are expected to be Poisson distributions. This often happens when you are trying to regress on count data. In fact, its applicability extends well beyond the traditional domain of count data. The Poisson regression model can be used for any constant elasticity mean function, whether the dependent variable is a count or continuous, and there are good reasons why it should be preferred over the more common log transformation of the dependent variable.

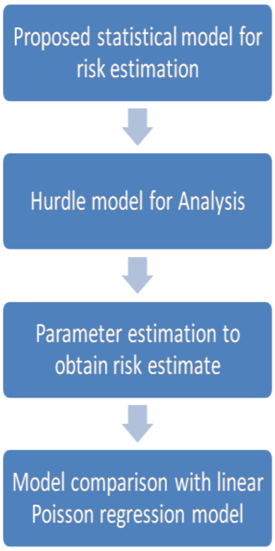
The simplicity of the Poisson regression model, an asset when modelling the mean, turns them into a liability, and more elaborate models are needed. There are two common difficulties in Poisson regression and they are both caused by heterogeneity in the data.

1. Over dispersion- This occurs when the variance of the fitted model is larger than what is expected by the assumptions (the mean and the variance are equal) of the Poisson model. Over dispersion is typically caused by a Poisson regression that is missing an important independent variable or by data being collected in clusters (like collecting data inside family units).

2. The second problem, also caused by heterogeneity, is excess zeros. In this situation, the distribution has more zeros than would be expected from a Poisson distribution. Often this is caused by two processes creating the data set, one of the processes having an expected count of zero. Poisson model is always overly restrictive when it comes to estimating features of the population other than the mean, such as the variance or the probability of single outcomes.

4.2 Hurdle Model

To overcome the limitations of Poisson Hurdles Model is used



Proposed statistical model for risk estimation
Hurdle model for Analysis
Parameter estimation to obtain risk estimate
Model comparison with linear Poisson regression model

Figure 2. Risk estimation Model

Hurdles model (each assuming either the Poisson or negative binomial distribution of the outcome) has been developed to cope with zero-inflated outcome data with over-dispersion (negative binomial) or without (Poisson distribution). Hurdle model deals with the high occurrence of zeros in the observed data,

A hurdle model is a modified count model in which there are two processes, one generating the zeros and one generating the positive values. The two models are not constrained to be the same. The concept underlying the hurdle model is that a binomial probability model governs the binary outcome of whether a count variable has a zero or a positive value. If the value is positive, the "hurdle is crossed," and the conditional distribution of the positive values is governed by a zero-truncated count model.
More formally, the hurdle model combines a count data model f_{count} (y; x, β) (that is left-truncated at y = 1) and a zero hurdle model f_{zero} (y; x,γ) (right censored at y = 1).

$$f_{\text{hurdle}}(y;\ x,z,\beta,\gamma) = \begin{cases} f_{\text{zero}}(0;\ z,\gamma),\ \text{if } y = 0 \\ (1\ -\ f_{\text{zero}}(0;\ z,\gamma)).\dfrac{f_{\text{count}}(y;x,\beta)}{1-f_{\text{count}}(0;x,\beta)}\ ,\text{if } y > 0 \end{cases} \tag{18}$$

The model parameters β, γ and potentially one or two dispersion parameters θ (if f_{count} or f_{zero} or both are negative binomial densities) are estimated by Maximum Likelihood, where specification of the likelihood has the advantage that the count and the hurdle component can be maximized separately . Since the risk estimates are exponential for doses groups, we would apply log transform on dose variables for obtaining the linear relationship.[9]

4.3 Comparison of hurdle with Poisson regression

The models are compared by applying Vuong test. The Vuong test shows that the hurdles model provides a better fit than Poisson regression model. From the comparison of AIC values it is observed that hurdles model provides better fit.

Table 1. Odds ratios of hurdle model

Age Categories	**Odds Ratio**
Agexcat8	4.2799616
Agexcat9	5.5625905
Agexcat10	7.4448636
Agexcat11	10.1254675
Agexcat12	10.2122343
Agexcat13	10.6221722
Agexcat14	10.1927220
Agexcat15	6.8081283
s2	0.6887824

Table 2. Comparison of Models

Model	AIC
Multiplicative Poisson	20687.0
Additive Poisson	20686.6
Additive hurdle	20686.2

V. CONCLUSION AND FUTURE WORK

Poisson regression has various limitations. There are several alternative models like Negative Binomial Regression and Zero-inflated Poisson Regression to fit the data to the model. We developed Hurdles model as an alternative to Poisson model for risk estimation and observed that it provides a better fit as compared to Poisson model. There are other models like Negative Binomial and ZIP to cope up with the problem of excess zeros and over dispersion. Future work includes fitting data to these models and comparing the model parameters for obtaining the best fit of data.

REFERENCES

[1]. Health risks from exposure to low levels of ionizing radiation-Background for Epidemiologic Methods,

[2]. Beir vii phase 2, national research council of the National Academics,THE NATIONAL ACADEMIC PRESS Washington D.C.

[3]. N.E. Breslow and N.E. Day, Methods in Cancer Research Volume II - The Analysis of Cohort Studies W.N. Venables, D.M.Smith and the R CORE Team, An Introduction to R www.rerf.or.jp Radiation Effects Research Foundation

[4]. Analysis of epidemiological data using R and Epicalc -Virasakdi Chongsuvivatwong Probability and Statistics- Walpole Steven L. Simon, Ph.D. Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute,

[5]. Introduction to Radiation Physics and Dosimetry Regression Models for count data in R.