JP Journal of Biostatistics
Volume 6, Number 1, 2011, Pages 19-30
Published Online: November 2011
Available online at http://pphmj.com/journals/jpb.htm
Published by Pushpa Publishing House, Allahabad, INDIA

ANALYSIS OF CANCER INCIDENCE DATA OF POPULATION BASED CANCER REGISTRY IN THE NORTH OF IRAN: APPLICATION OF POISSON REGRESSION MODEL

K. O. Hajian-Tilaki

Department of Social Medicine and Health
Faculty of Medicine
Babol Medical Science University
Babol, Iran
e-mail: drhajian@yahoo.com

Abstract

The conventional approach to modeling cancer incidence rates uses the direct and indirect methods of standardization to estimate the standardized risk ratio of exposure for cancer registry data. However, this method cannot deal with several confounding factors. An alternative approach is to apply a Poisson regression model to estimate the incidence rate ratio for counts of events from a cancer registry data. The exponential of coefficient of the Poisson regression model represents the risk ratio. The author applied the Poisson regression model to the cancer registry data from the north of Iran to estimate the age-sex incidence rate ratio using GLIM software. It was shown how one could use the model for counts of events that occurred in a population or a person-year of time. For this analysis with GLIM software, the numerator and denominator of rate are needed from each cell of a contingency table. This model is quite flexible for count data and can control for several confounding factors while detecting the factor which modifies the effect of exposure.

2010 Mathematics Subject Classification: 62P10.

Keywords and phrases: Poisson regression model, risk ratio, cancer registry data.

Received May 24, 2011

Introduction

Cancer of digestive tracts is one the most common cancer in Iran and as elsewhere in the developing countries its incidence is high. In particular, published studies from 1968 to 1978 have documented the high incidence rate of esophageal cancer in the eastern part of south of Caspian see, in the north of Iran (1-3). The International Agency for Research on cancer in Lyon, in collaboration with the Institute of Public Health and the Cancer Institute, Tehran University has conducted numerous scientific and valuable studies on cancer in northern of Iran (Babol Cancer Registry). In particular, for esophageal cancer, there is a large variation its incidence found over short distance (1-2). In area where poverty is a consequence of environment, the disease would seem to have common (1). The epidemiologic evidence would suggest that the nutritional imbalances are of importance in the etiology of esophageal cancer in the north of Iran (1-5). A recent published monograph from Babol Cancer Registry (6) showed that gastric cancer (15.17%), breast cancer (10.7%), esophageal cancer (9.9%) and Melanoma (8.5%) are the first forth common cancer in the province of Mazandaran while in the province of Golestan (Gorgan and Gonbad), the high risk area of esophageal cancer, the first forth common cancer was esophageal (22.1%), gastric (12.5%) melanoma (8.5%) and breast (6.4%). Based on recent cancer registry data in Babol Hajian-Tilaki and et al (2003) reported the mean (± SD) of esophageal cancer was increased to 65 ± 13 years in the north of Iran (7). This increasing age, perhaps is due to changing in nutrition and health behaviours during the three last decades in the north of Iran.

In analysis of cancer registry of incidence data, standardization is the conventional approach for dealing with confounding as an adjustment procedure for comparative studies in epidemiology (8). This approach is applicable when confounding factors are categorical. The goal of any adjustment procedure is to correct for differences in the distribution of confounding factors between two groups under comparison. For example the rate of death due to breast cancer can be compared for two groups of women. The groups may differ with respect to age. Standardization procedure estimates the death rate based on a common age distribution. This common distribution is taken from another group, known as a standard population.

If the confounding factor is numerical, standardization can be applied first by categorizing the confounding factors. The effect of this categorization may a loss of information resulting in small cell frequencies at each cell in a contingency table,

hence low precision (8, 9). In dealing with confounding either categorical or numerical without loss of information, an alternative approach is to use a family of generalized linear models with a log link function, such as the Poisson regression model (9, 10). The Poisson regression model includes log-linear models that are intrinsically nonlinear models (10, 11, 12). In this model the response variable is a count that follows Poisson distribution. The approach considered enables one to estimate the risk ratio and to describe the relation between the dependent variable and the predictor variables (8-13).

This article describes the basic method of Poisson regression analysis and its application in the analysis of cancer registry data in a population based study. First we illustrated how Poisson regression model can be applied to the cancer registry count data either incidence or death data from a population during a period of time or person time of follow up. In particular, we estimated the age-sex adjusted risk ratio of cancer incidence from population based cancer registry in a part of northern Iran.

Methods

Poisson regression model

Suppose, we have a count random variable Y with mean μ $(EY = \mu)$. The Poisson probability distribution as follows:

$$\Pr(Y, \mu) = \mu^{Y} e^{-\mu} / Y!, \qquad Y = 0, 1, 2, ..., n.$$

The Poisson distribution is often used to model the occurrence events for example incidence (or death) of cancer in a population or person times of follow-up. Suppose, we have counts of events (incidences or deaths) from a cancer registry in a contingency table, where Y_{ij} denotes the frequency of a cell related to the ith row category (confounding level) and jth column (exposure level) where $Ey_{ij} = \mu_{ij}$ and the log link of μ_{ij} is associated with a set of covariates and exposure with model $\ln(\mu_{ij}) = \beta X$, e.g., $Ln(\mu_{ij}) = \sum \alpha_i X_i + \beta E$, where X_i's is an indicator with $X_i = 1$ if age group i and 0 otherwise. Also, the indicator variable E is defined as 1 if exposed group $(j = 1)$ and 0 if nonexposed $(j = 0)$. For exposed group $(j = 1)$:

$$\ln \mu_{i1} = \alpha_i + \beta \quad \text{for} \quad i = 1, 2, ..., k$$

and for nonexposed group $(j = 0)$:

$$\ln \mu_{i0} = \alpha_i$$

and then, $\ln \mu_{i1} - \ln \mu_{i0} = \beta$.

Thus, $RR = \mu_{i1}/\mu_{i0} = \exp(\alpha_i + \beta)/\ln \alpha_i = \exp(\beta)$.

Thus, using a Poisson regression model, the risk ratio of exposure and age group (vs. baseline age) can be estimated by the exponential of the coefficients of regression. We can interpret $\exp(\beta)$ as an estimated overall risk ratio adjusted for age. The general fit for a Poisson regression model is obtained using the Poisson probability function to derive the likelihood function that can be maximized in order to estimate the parameters β's and its standard error (SE) as well as the information matrix using GLIM software. The user needs only specify the link function and the input counts of events and the population size (or person year) as denominator of the rate) for each cell of the contingency table respectively. Then, the adjusted estimates of the parameters $\ln(RR) = \beta$ and hence the adjusted risk ratio is estimated.

Modelling Rate using Poisson Regression

Suppose the data of number of event and the population (or person year of follow up) were presented in a contingency table. A general failure rate in each subgroup of interest can be estimated by $\lambda = Y/l$, where Y denotes the observed count data and l denotes the person time of follow up (or population size that events occurred) and the ratio of the two rates is commonly used as a rate ratio or incidence density ratio (hazard ratio).

Suppose Y_i denotes the number of events, and l_i the person time of follow up. Let $X_i = (X_{1i}, X_{2i}, ..., X_{ki})$ denote the components of covariates and exposure and $\beta = (\beta_0, \beta_1, ..., \beta_k)$ is the set of unknown parameters. Then

$$\ln \lambda_{ij}(X, \beta) = \beta_0 + \sum \alpha_i X_{ij}.$$

The expected number of events is:

$$EY_{ij} = \mu_{ij} - l_{ij}\lambda_{ij}(X, \beta).$$

Then, $\ln \mu_{ij} = \ln l_{ij} + \ln \lambda_{ij}(X, \beta)$

$$Ln\, \mu_{ij} = \ln l_{ij} + X\beta.$$

Now, suppose Y_{ij} denotes a dependent count of events in the contingency table in *i*th row (confounding variable level) and *j*th column (exposure level) and P_{ij} denotes the number of person year of follow up that have experienced Y_{ij} events. The dependent count variable Y_{ij} has a Poisson distribution with mean μ_{ij} within cell of *i*th row and *j*th column in contingency table with probability model:

$$P(Y_{ij}) = e^{-\mu ij} \mu_{ij}^{Yij} / Y_{ij}!.$$

The expected number of events in each cell is a function of the effect of *i*th row (confounding variables) and *j*th column (exposure):

$$\mu_{ij} = P_{ij} \exp(a + b_i + c_j).$$

Then, $\log \mu_{ij} / P_{ij} = a + b_i + c_j$, where a is a constant and $R_{ij} = \mu_{ij} / P_{ij}$ is the rate. Thus, $\ln Rate = a + b_i + c_j$.

Suppose, we want to compare the incidence rate of a type of cancer between males and Females $(j = 1$ male, $j = 2$ female$)$ and the incidence count are presented in 8 categories of age $(i = 1, ..., 8)$; b_i represents the coefficient for *i*th row of the age group and c_j represent the effect of the *j*th column for gender. The regression model for rate among males and females are respectively as follows:

$\ln$ rate(male) $= a + b_i + c_1$

$\ln$ rate(female) $= a + b_i + c_2$ and the log rate ratio (*RR*) as follows:

$$\ln RR = c_2 - c_1.$$

Thus, the difference of $\ln RR$ is constant across the age group. This is an inherent assumption of a multiplicative model.

Since the rate comprises of a numerator (counts of events) such as deaths or incidences, and the denominators (person years of follow up or population

abbreviated as *PY*) then:

$$\ln \text{count} - \ln PY = a + b_i + c_j$$

$$\ln \text{count} = \ln PY + a + b_i + c_j.$$

The coefficient of the term ln(*PY*) is equal to 1; this term is called the *offset* in the GLIM program.

Data and Analysis

The data were collected from registered cancer incidences in different diagnostic centers of Babol, in the north of Iran in 2002. We tabulated the frequency of cancer with respect to age and sex for overall cancer and for specific types of cancer such as esophageal cancer, gastric cancer and breast cancer. The age and sex distribution of the population of the catchment's area of cancer registry were derived from census data. The GLIM software was used with a program that is shown in the Appendix. The input data was the number of events, the size of the population within each stratum of the age-sex contingency table, the age group and gender status. The Poisson link function was used. An indicator variable for gender and $k-1$ indicator variables for *k* categories of age strata were defined and the first age group was defined as the base-line group.

Results

Table 1 shows that the age-sex distribution of the population size and overall number of cancer patients, specifically, the number of esophageal cancers and gastric cancers. The population size of the catchment area of the registry was 421,068 and the overall number of registered cancers was 632 subjects. The overall cancer rate was 175 per 100,000 for males and 125 per 100,000 for females. Both esophageal and gastric cancers were more common among males than females. Using the data of Table 1 as a numerator and denominator of the rate, and defining an indicator variable for gender (male = 1 and female = 0) and other indicator variables for age groups with a baseline age < 10 for overall cancers (and with baseline age < 40 years for gastric and esophageal cancer and age < 30 years for breast cancer), we estimated the age and sex adjusted effect using a log-linear model with a Poisson link function. The regression coefficients, their standard errors (SE's), the risk ratio $(RR = \exp(B))$ and its *P*-value are shown in Table 2. The age

adjusted risk ratio for males vs. females is 1.35 $(P < 0.01)$. The Figure 1 shows that the sex adjusted risk ratio significantly increases with increasing age for overall cancer. Table 2 shows that the results of the application of a Poisson regression specifically for gastric and esophageal cancer with base line age < 40 years. The Figure 2 again shows that the risk ratio increases exponentially with age. However, using the breast cancer data, Figure 3 shows that the risk ratio is maximized at 55 years and then the trend tends to decrease.

Table 1. Population size and cancer incidence during a year with respect to the age group and gender

Age group	Male				Female			
	Pop. size	No. cancer	No. Eso.	No. Gast.	Pop. size	No. cancer	No. Eso.	No. Gast.
0-9	44992	8	-	-	44030	9	-	-
10-19	55679	12	-	-	55421	14	-	-
20-29	33626	16	-	1	37533	16	-	2
30-39	28461	19	-	2	29460	27	2	-
40-49	17396	31	4	3	18589	44	2	1
50-59	11683	49	3	15	11635	50	4	9
60-69	9977	88	11	24	9121	49	12	10
> = 70	6848	143	17	46	6573	57	12	11
Total*	208680	366	37	91	212388	266	32	33

Eso.: esophageal; Gast.: gastric

*The total was counted with 18 subjects of male and 26 subjects of female with missing data of age group.

Table 2. Estimates of parameters of Poisson regression model and age-sex adjusted risk ratio for overall cancer

Parameter	Estimate	SE	RR*	P-value
1 (scale)	-8.72	0.25	-	-
Sex (M vs F)	0.30	0.08	1.35	P < 0.01
Age (10-19)	0.21	0.31	1.33	NS
Age (20-29)	0.87	0.30	2.38	P < 0.001
Age (30-39)	1.43	0.28	4.18	P < 0.001

Age (40-49)	2.39	0.27	10.91	P < 0.001
Age (50-59)	3.10	0.26	22.19	P < 0.001
Age (60-69)	3.62	0.26	37.34	P < 0.001
Age (> = 70)	4.35	0.25	77.48	P < 0.001

*Comparison group: 0-9 years

Table 3. Estimates of parameters of Poisson regression model and age-sex incidence risk ratio for gastric and esophageal cancer

Parameter	Estimate	SE	RR*	P-value
1 (scale)	-11.41	0.23	-	-
Sex (M vs F)	0.81	0.16	2.24	P < 0.001
Age (40-49)	2.74	0.47	15.48	P < 0.001
Age (50-59)	4.15	0.39	63.43	P < 0.001
Age (60-69)	5.08	0.37	156.02	P < 0.001
Age (> = 70)	5.86	0.36	350	P < 0.001

*Comparison group: 20-39 years

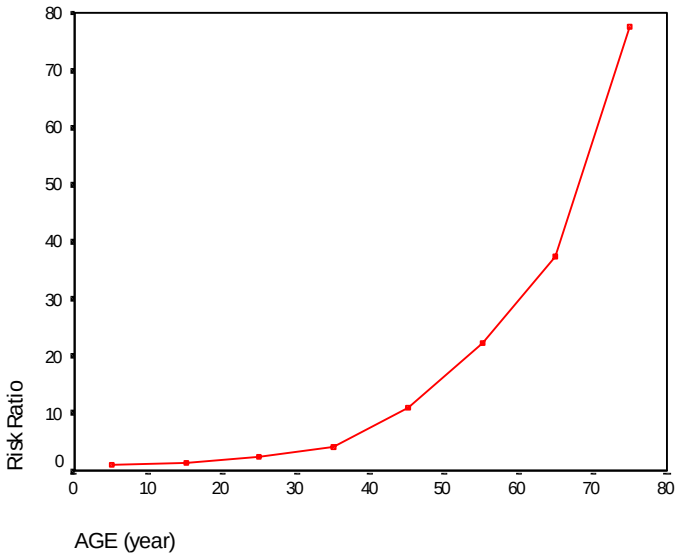

Figure 1. Sex adjusted risk ratio with age for overall cancer.

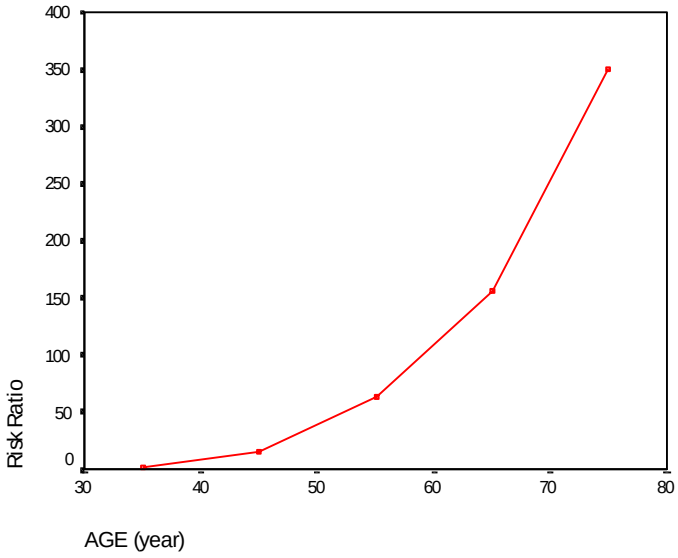

Figure 2. Sex adjusted risk ratio with age for gastric and esophageal cancer.

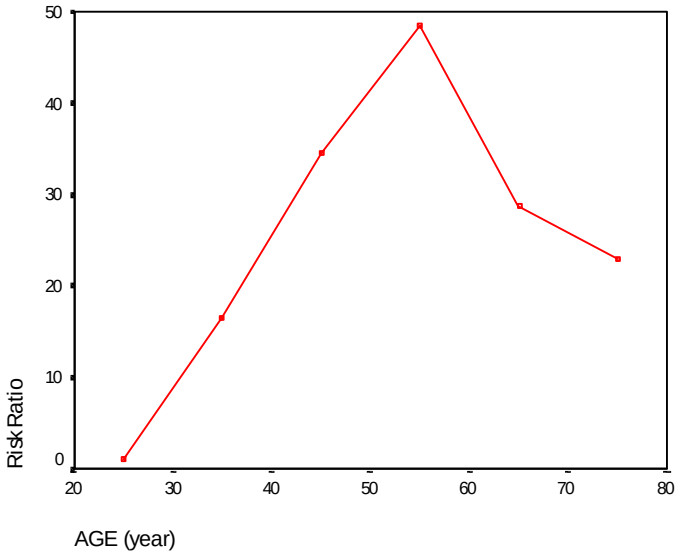

Figure 3. Risk ratio of breast cancer with age.

Discussion

In this study, the age-sex risk ratio was estimated in a population based cumulative incidence study. The adjusted risk ratio was estimated using a Poisson regression model. In contrast to the conventional approach of standardization of the rate, a Poisson regression model has several advantages. First, it can be applied to the counts of events with experience in person time of follow-up (an incidence density type study) or in a population at specific period of time – e.g., a year (cumulative incidence type study). Second, it is applicable when estimating the adjusted rate ratio for either categorical or continuous confounding factors. Third, it easily can be applied to estimate the interaction effect of exposure and covariates for example the interaction effect of age and gender and thus the users are able to determine whether a covariate is a modifier. However, the model inherently is a multiplicative model with a specific assumption. An important assumption is the log rate ratio is constant across strata. In some sense, this is the same assumption needed for the Cox proportional hazard regression model which is a semiparametric model without any assumption of underlying hazard. In contrast to the modelling of data by the Cox regression model, the Poisson regression model involves less computation in the maximization of the likelihood function for estimation of regression coefficients with categorization of count data that is presented in a contingency table which is a summarized data. While the dependent variable in a Cox regression model is a continuous time to event (censored data) and the conditional likelihood is constructed based on the risk sets at each failure time. Thus, the estimation procedure involves a huge computation. In spite of less computation for the Poisson regression model, these two models yield the same results of risk ratio (15). Even if the data were not presented in a contingency table, in particular with numerical covariates and continuous exposure variable, the Poisson regression model is applicable for estimating risk ratio with some other software for example Egret software. In particular, for cancer registry data of population based cumulative incidence type study, we show how the users can apply Poisson regression model to estimate adjusted risk ratio without data from a standard population. For this situation, the Poisson model needs the rare disease assumption that is almost satisfied for cancer count data. In another situation, this model is used for age-period-cohort analysis of breast cancer mortality rates (16) and the cohort effect was detected by fitting a Poisson regression model. Overall, this approach of analysis is applicable and can be fitted in any software package that estimates GLMs with user

defined link functions (including GLIM, SAS, Stat, Splus, R and Egret) and utilizes the theory of generalized linear model for assessing goodness of fit and regression diagnostics.

References

[1] N. E. Day, Some aspect of epidemiology of esophageal cancer, Cancer Research 33 (1975), 3304-3307.

[2] E. O. Mahboubi and B. Aramesh, Epidemiology of esophageal cancer in Iran, with special reference to nutritional and cultural aspects, Preventive Medicine 9 (1980), 613-621.

[3] A. Z. Keller, The epidemiology of esophageal cancer in the west, Preventive Medicine 9 (1980), 607-612.

[4] P. J. Cook-Mozafari, F. Azordegan, N. E. Day, A. Ressicaud, C. Sabai and B. Aramesh, Oesophageal cancer studies in the Caspian littoral of Iran: results of a case-control study, Br. J. Cancer 39 (3) (1979), 293-309.

[5] P. Ghadirian, G. F. Stein, C. Gorodetzky, M. B. Robetrfroid, G. A. Mahon, H. Bartsch and N. E. Day, Esophageal cancer studies in the Caspian littoral of Iran: some residual results, including opium use as risk factors, In. J. Cancer 35(5) (1985), 593-597.

[6] U. Yahyapour, J. Alijan-Tabar and H. Jharejo, Annual monograph of Babol health research, The Institute of Public Health and the Cancer Institute, Tehran University 16 (2001), 11-18.

[7] K. O. Hajian-Tilaki, A. R. Firouzjahi and M. T. Kia, Pattern of age distribution of different cancers in Babol, J. Shaheed Beheshti University of Medical Sciences/Iran 27(3) (2003), 239-245.

[8] S. Anderson, A. Auquier, W. Hauk et al., Statistical Methods for Comparative Studies, John Wiley and Sons, New York, pp. 113-134.

[9] D. G. Kleinbaum, L. L. Kupper and K. E. Muller, Applied Regression Analysis and other Multivariable Methods, PWS-KENT Publishing Company, Boston, 1988, pp. 512-519.

[10] E. L. Frome, The analysis of rates using Poisson regression models, Biometrics 39 (1983), 665-674.

[11] F. Kianifard and P. P. Gallo, Poisson regression analysis in clinical research, J. Biopaharm Stat. 5(1) (1995), 115-129.

[12] E. L. From and H. Checkoway, Use of Poisson regression models in estimating incidence rate and ratios, Am. J. Epidemiol. 121(2) (1985), 309-323.

[13] D. K. Blough, C. W. Madden and M. C. Hornbrook, Modelling risk using generalized linear models, J. Health Econ. 18 (2) (1999), 153-171.

[14] W. Garner, E. P. Mulvey and E. C. Shaw, Regression analysis of counts and rates: Poisson, overdispersed Poisson, and negative binomial models, Psychol. Bull. 118 (3) (1995), 392-404.

[15] P. W. Dickman, A. Sloggett, M. Hills and T. Hakulinen, Regression models for relative risk, Stat. Med. 23(1) (2004), 51-64.

[16] A. Cayuela, S. Dominguez Rodriguez, M. Ruiz-Borrego and M. Gili, Age-period-cohort analysis of breast cancer rates in Andalucia (Spain), Ann. Oncol. 15(4) (2004), 686-688.

Appendix: Program used in GLIM

```
?$ Input 12$
?$ data sex age count incidence$
?$ input 12$
File name ? A:cancer.dat
?$ Factor age 8$
?$ Factor Sex 2$
?$ Yv incidence$
?$ error Poisson$
?$ cal count = %log(count)$
?$ offset count $
?$ fit sex+age $
?$ display e r$
?$ stop $
```