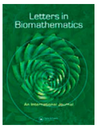
LETTERS IN BIOMATHEMATICS
https://doi.org/10.1080/23737867.2019.1581104



Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Statistical modelling for cancer mortality

Sarada Ghosh and G. P. Samanta

Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India

ABSTRACT

The usefulness of log-linear models for contingency table analysis is evident from its current general popularity among statisticians. We have extracted U.S. vital rates and survival data in cancer mortality from the Surveillance between the year of 1975–2015. This paper has two distinct fields: (i) Survival and (ii) Contingency table analysis in a single analytical framework based on log-linear model. In this paper, the effects of gender and different types of cancer on death rate have been demonstrated. Testing and estimation are also applicable here. The purpose of the underlying Cox model is to evaluate simultaneously the effect of several factors on survival, i.e. it can examine how specified factors influence the rate of happening of a particular event (e.g. death) during this time interval. The purpose of this work is not to develop new methodologies, but rather to present new uses and interpretations. Simulation is based on *R*-software.

ARTICLE HISTORY
Received 17 October 2018
Accepted 5 February 2019

KEYWORDS
Colon and Rectum cancer; Kidney and Renal Pelvis cancer; log-linear models; residual deviance; log-likelihood; Cox-proportional hazards model

1. Introduction

In the world, cancer is one of the deadliest diseases. Our body is made up of different types of cells. Generally, these cells are grown and divided under controlled conditions to create more cells as they are needed to keep the body healthy. When cells become old or damaged, they die and new cells can take place in there. Cancer is the uncontrolled growth of abnormal cells anywhere in a body. These abnormal cells are known as cancer cells, malignant cells or tumour cells. These cells can infiltrate normal body tissues. Cancer refers to cells that grow out of control and invade other tissues. Cells become cancerous due to the accumulation of defects, or mutations, in genetic material (DNA) of a cell. Tobacco use is one of the causes of cancer deaths. Apart from this, another reasons are due to obesity, poor diet, lack of physical activity or excessive drinking of alcohol. Other factors including certain infections, exposure to ionizing radiation and environmental pollutants are also responsible for cancer. Typically, many genetic changes are required for cancer development. Some cancers are due to inherited genetic defects from a people's parents. There are over 200 types of cancer and each is classified by the type of cell that is initially affected. Many cancers and the abnormal cells that compose the cancer tissue are further identified

CONTACT Sarada Ghosh ✉ saradaghosha111@gmail.com 🖃 Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, India

by the name of the tissue that the abnormal cells originated from (e.g. breast cancer, lung cancer and colon cancer). But some cancers do not form tumours (as e.g. Leukemia is a cancer of the bone marrow and blood). Some types of cancer begin in the skin or in tissues that line or cover internal organs (skin, lung, colon, pancreatic, ovarian cancers, etc.). Some of them begin in bone, cartilage, fat, muscle, blood vessels or other connective or supportive tissue (bone, soft tissue cancers). Some cancers begin in the cells of the immune system (i.e. T-cell lymphomas, B-cell lymphomas, etc.) and it can also begin in the tissues of the brain and spinal cord. In human body, the hollow area in the centre of each kidney is renal pelvis. It is a top part of the ureter. The ureter is a kind of long tube which makes a connection between the kidney and bladder. Actually, renal pelvis is very much related with kidney. Transitional cell cancer of the ureter and renal pelvis is a disease in which malignant (cancer) cells form in the ureter and renal pelvis.

Cancers may be prevented by not smoking, maintaining a healthy weight, not drinking too much alcohol, eating whole grains and fresh vegetables, fruits, vaccination against certain infectious diseases, not eating too much processed and red meat and avoiding too much sunlight exposure. Cancers can be treated with some combinations of radiation therapy, surgery, chemotherapy, immunotherapy, monoclonal antibody therapy, targeted therapy, etc. The choice of therapy is based on the location and grade of the tumour and the stage of the disease, as well as the medical state and age of the affected people. The aim of the treatment is to complete removal of the cancer cells without damaging the rest of the body. But it is unfortunate that most of the cancer treatments have a negative effect on normal body cells.

Chemotherapy is the treatment of cancer with one or more cytotoxic anti-neoplastic drugs through some protocol. Chemotherapy also affects cells that divide rapidly under normal circumstances: cells in the bone marrow, digestive tract and hair follicles. Targeted therapy is also a form of chemotherapy that targets specific molecular differences between cancer and normal cells. The efficacy of chemotherapy depends on the type of cancer and the stage. The combination of chemotherapy and surgery is useful in different types of cancer (including breast cancer, colorectal cancer, pancreatic cancer, testicular cancer, ovarian cancer and certain lung cancers). Besides it is curative for some cancers, such as some leukaemia (a cancer caused by an overproduction of damaged white blood cells). The effectiveness of chemotherapy is frequently limited by its toxicity to other tissues in the body. Even when chemotherapy does not provide a permanent cure, it may be helpful to reduce symptoms such as pain or to reduce the size of an incurable tumour in the hope that surgery will become possible in future.

Radiation therapy involves the use of ionizing radiation in an attempt to either cure or improve symptoms. As per the clinical evidence this therapy is useful for half of the cases. The radiation can be of two types: (i) from internal sources and (ii) from external sources. Generally, the radiation is most low energy X-rays for treating skin cancers, while higher energy X-rays are used for cancers within the body. It can be damaging or killing the DNA of cancerous tissue. To spare normal tissues (such as skin or organs, which radiation must pass through to treat the tumour), shaped radiation beams are focused from multiple exposure angles to intersect at the tumour. It will provide a much higher absorbed dose at the location of the tumour than in the surrounding (healthy tissue).

Surgery is the primary method of treatment to cancer therapy. It is an important part of definitive diagnosis and staging of tumours. In localized cancer, surgery typically attempts to remove the entire mass along with, in certain cases, the lymph nodes in the area. For some types of cancer it is sufficient to eliminate the cancer.

Immunotherapy is one of the most recent approaches for cancer. It depends on the generally-accepted hypothesis: 'the immune system is the best tool that humans have for fighting disease'. It can help for stimulating or helping the immune system to fight cancer. The approaches include antibodies, checkpoint therapy and adoptive cell transfer. In this work we have fit log-linear model to predict effects of gender and different types of cancer on death rate. Some of the more attractive features of this modelling system are the ease of model specification and reduction which provide the flexibility in treating both dependent and independent variables; and the fact that maximum likelihood estimates can be collectively characterized for an assortment of sampling distributions, including Poisson, multinomial and product multinomial (Agresti, 2002, 2007, 2010). Apart from this, the purpose of the underlying Cox model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g. infection, death) at a particular point in time. This rate is commonly referred as the hazard rate. Predictor variables (or, factors) are usually termed covariates in the survival-analysis literature. The simulation of this work is based on *R*-software. Finally, the last section consists of the general discussions and conclusions of the paper.

2. Model derivation and preliminaries

2.1. Log-linear models for Poisson count data

In most studies, there are several explanatory variables which may be continuous as well as categorical. The main purpose is usually to describe their effects on response variables (Ghosh & Samanta, 2019). When a model is good fitted, it evaluates the effects, includes relevant interactions, and provides smoothed estimates of response probabilities. The family of generalized linear models (GLM), which contains the most important models for categorical responses as well as standard models for continuous responses, is most important part in the statistical field for investigations. It is used increasingly in a wide variety of applications. The simplest distribution for count data is the Poisson distribution. It can take any non-negative integer value. Let Y denotes a count and let $\psi = E(Y)$. The Poisson pmf (probability mass function) for Y is

$$f(y;\psi) = \frac{e^{-\psi}\psi^{y}}{y!}, \quad \text{where} \quad y = 0, 1, 2, \ldots \tag{1}$$

The Poisson log-linear model is defined as

$$\log \psi_i = \sum_j \beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, \cdot \because, N. \tag{2}$$

The mean of Poisson distribution is non-negative. Although a GLM can model a positive mean using the identity link, it is more common to model the log of the mean. The log mean can take any real value like linear predictor $\theta + \beta X$. A Poisson log-linear GLM assumes a Poisson distribution for Y and uses the log link. The Poisson log-linear model when we consider X as explanatory variable and ϵ as the error term be

$$\log \psi = \theta + \beta x + \epsilon \tag{3}$$

The mean satisfies an exponential relationship for this model: $\psi = \exp(\theta + \beta x) = e^{\theta}(e^{\beta})^{x}$ A 1-unit increase in x has a multiplicative impact of e^{β} on ψ (i.e. the mean at x+1 equals the mean at x multiplied by e^{β}). A common use is for 'modelling cell counts' in contingency tables, since it is generally familiar with log-linear models for contingency table analysis (Agresti, 2010; Laird & Olivier, 1981). The models specify how the expected count depends on levels of the categorical variables for that cell as well as associations and interactions among those variables (Anderson, 1984; McCullagh, 1980; Powers & Xie, 2000). The purpose of log-linear modelling is the analysis of association and interaction patterns.

When a response count n_i has index equal to t_i, where t_i is the time or space (e.g. days in the community), the sample rate of occurrence is n_i/t_i. The expected value is ψ_i/t_i.

i.e.

$$\begin{aligned} E\left(\frac{n_i}{t_i}\right) &= \frac{1}{t_i}E(n_i) \\ &= \frac{\psi_i}{t_i}. \end{aligned}$$

When x is the explanatory variable, the Poisson log-linear regression model for expected rate of the occurence of events is

$$\begin{aligned} &\log\left(\frac{\psi_i}{t_i}\right) = \theta + \beta x_i \\ &\Leftrightarrow \log \psi_i - \log t_i = \theta + \beta x_i \\ &\Leftrightarrow \log \psi_i = \theta + \beta x_i + \log t_i \\ &\Leftrightarrow \psi_i = t_i \exp(\theta + \beta x_i), \end{aligned} \tag{4}$$

Sometimes, $\log t$ is called 'offset'. The mean is proportional to the index, with proportionality constant depending on the value of x. But when we consider the identity link, then the model is of the form:

$$\begin{aligned} &\frac{\psi_i}{t_i} = \theta + \beta x_i \\ &\Rightarrow \psi_i = \theta t_i + \beta x_i t_i. \end{aligned} \tag{5}$$

Here, $x_i t_i$ are explanatory variables and there is no intercept.

The likelihood function for n independent Poisson observations is obtained by taking a product of pmf given by Equation (1), then taking logs and ignoring a constant involving $\log(y_i!)$. Thus the log-likelihood function is

$$\log L(\beta) = \sum [y_i \log(\psi_i) - \psi_i], \tag{6}$$

where ψ_i depends on the covariates of x_i and a vector of p parameters β through the log link of Equation (2). Generally, the log is the canonical link for the Poisson distribution. If we take derivatives of the log-likelihood function with respect to the elements of β, and then taking the derivatives equal to zero, the maximum likelihood estimates ($\widehat{\psi}$) in log-linear Poisson models satisfy the estimating equations:

$$X^T y = X^T \widehat{\psi}, \tag{7}$$

where X is the model matrix, with one row for each observation and one column for each predictor, including the constant (if any), y is the response vector and $\widehat{\psi}$ is a vector of fitted values, calculated from the maximum likelihood estimator's $\widehat{\beta}$ by exponentiating the linear predictor $\xi = X^T\widehat{\beta}$. This estimating equation occurs not only in Poisson log-linear models, but also in any GLM (with canonical link), including logistic regression models for binomial counts and linear models for normal data.

Deviance is a measurement of discrepancy between observed and fitted values. For Poisson responses the deviance takes the form:

$$D = 2\sum_i \left[y_i \log \frac{y_i}{\widehat{\psi}_i} - (y_i - \widehat{\psi}_i) \right]. \tag{8}$$

The first term of the right side is identical to the binomial deviance, representing 'twice a sum of observed times log of observed over fitted' and the second term is 'a sum of differences between observed and fitted values' (usually zero as maximum likelihood estimators in Poisson models have the property of reproducing marginal totals). When the sample is large, the distribution of the deviance is approximately a chi-squared with $(n - p)$ degrees of freedom, where n is the number of observations and p is the number of parameters. Therefore, the deviance can be used directly to test the goodness of fit of the underlying model. An alternative measure of goodness of fit is Pearson's chi-squared statistic denoted and defined as:

$$X^2 = \sum_i \frac{(y_i - \widehat{\psi}_i)^2}{\widehat{\psi}_i}, \tag{9}$$

where the numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. It has the same form for binomial and Poisson data. For large samples, the distribution of Pearson's statistic is also approximately chi-squared with $(n - p)$ degrees of freedom. One of the advantages of deviance over Pearson's chi-squared is that it can be used to compare nested models. A good-fitting log-linear model provides a basis for describing and making inferences about associations among categorical responses. Standard methods apply for checking fit and making inference about model parameters. As usual, X^2 and G^2 (Likelihood Ratio chi-squared statistic) test whether a model holds by comparing cell fitted values to observed counts. Here 'degrees of freedom' equals 'the number of cell counts minus the number of model parameters'.

For log-linear models, the likelihood ratio tests can easily be constructed in terms of deviances. Apart from this, in this work we have constructed Wald tests based on the fact that the maximum likelihood estimator $\widehat{\beta}$ has a multivariate normal distribution with mean equal to the true parameter value β and variance-covariance matrix var $(\widehat{\beta}) = X^T WX$ (approximately in large samples). Here X is the model matrix and W is the diagonal matrix of estimation weights.

2.2. Modelling survival times

In this work, we have also considered a method for modelling survival times related to the Poisson log-linear model for rates that focuses on times until death rather than on numbers of deaths. In survival (and reliability) analysis, the hazard and survival functions are very useful. Let T denotes the time to some event, such as death or such as product failure in the context of reliability study. Let $f(t)$ denotes the pdf (probability density function) and $F(t)$ is the cdf (cumulative density function) of T. There exists a good connection between maximum likelihood estimation using a Poisson likelihood for numbers of events and a negative exponential likelihood for T (Aitkin & Clayton, 1980). The survival function is denoted and defined as:

$$S(t) = 1 - F(t) \tag{10}$$

and the hazard rate or hazard function is denoted and defined as:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \tag{11}$$

There are several models for the censoring mechanism that lead to non-informative censoring, and informative censoring models as well (Lagakos, 1979). Let, w_i be the indicator function, where $w_i = 1$ for death and $w_i = 0$ for censoring for subject i. Now, the survival-time likelihood for n independent observations is as follows:

$$\prod_{i=1}^{n} f(t_i)^{w_i}[1 - F(t_i)]^{1-w_i}. \tag{12}$$

Then the log likelihood equals to

$$\sum_i w_i \log[f(t_i)] + \sum_i (1 - w_i)\log[1 - F(t_i)]. \tag{13}$$

Further analysis is required for a parametric form for f and a model for the dependence of its parameters on independent variables. Most survival models focus on the rate at which death occurs rather than expectation of T. Now, consider the log likelihood (13) with $f(t)$ equals to the negative exponential density with parameter $\zeta \exp(\beta^T x)$, where $h(t;x) = \zeta \exp(\beta^T x)$ is the hazard function for a negative exponential survival distribution including explanatory variable x. For subject i, let $\psi_i = t_i \zeta \exp(\beta^T x_i)$ and by substitution, the log likelihood becomes

$$\begin{aligned} &\log\left[\left\{\left(\frac{\psi_1}{t_1}\right)^{w_1} e^{-\psi_1}\right\}\left\{\left(\frac{\psi_2}{t_2}\right)^{w_2} e^{-\psi_2}\right\}\cdots\right] \\ &= \sum_i \log\left\{\left(\frac{\psi_i}{t_i}\right)^{w_i} e^{-\psi_i}\right\} \\ &= \sum_i w_i \log\psi_i - \sum_i \psi_i - \sum_i w_i \log t_i. \end{aligned} \tag{14}$$

The first two terms of the last line involve β and is identical to the log likelihood for independent Poisson variates $\{w_i\}$ where expected values are $\{\psi_i\}$. In this application $\{w_i\}$ are binary rather than Poisson, but that is irrelevant to the process of maximizing with respect to β. This process is equivalent to maximizing the likelihood for the Poisson log-linear model:

$$\log \psi_i - \log t_i = \log \zeta + \beta^T x_i \tag{15}$$

with offset $\log(t_i)$, using observations $\{w_i\}$. The summation of the terms in the log-likelihood (for subjects having a common value of x) yields the observed data as the numbers of deaths $\left(\sum_i w_i\right)$ for each x, and the offset is the log $\left(\sum_i t_i\right)$ at each setting.

The Cox model is expressed by the hazard function denoted by $h(t)$. Briefly, the hazard function can be interpreted as the risk of dying at time t. The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an 'intercept' term that varies with time. The quantities $\exp(\beta_i)$ are called hazard ratios (HR). A value of β_i is greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases. Put another way: a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival. A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. For good prognosis means the negative sign that the hazard is lower. When hazard ratio is equal to 1, then there is no effect. Therefore,

(i) A covariate with hazard ratio > 1 (i.e. when $\beta_i > 0$) is called bad prognostic factor.
(ii) A covariate with hazard ratio < 1 (i.e. when $\beta_i < 0$) is called good prognostic factor.

Generally, a separate hazard rate is used for each piece of the time scale. The piecewise exponential approach is a natural one for life-table analysis where the period of follow-up is divided into intervals, since a common assumption is that the hazard function is approximately constant within interval (Holford, 1976). Perhaps the most appealing as well as popular feature (in survival analysis) of the hazard function is that it allows a needful way for specifying the effect of covariates on survival. The 'proportional hazards model' introduced by Cox (1972) is as follows:

$$h(t; x) = h_0(t)e^{x^T\beta}, \tag{16}$$

where $h_0(t)$ is the underlying hazard function which is chosen from any parametric family (such as exponential, Weibull, etc.) or it may be left unspecified and β is a column vector of unknown parameters specifying the effect of covariates (Cox, 1972). The ratio of hazard functions for any two individuals with covariate vectors X_1 and X_2 is $\exp\,(X_1 - X_2)^T\beta$ which is independent of t. It is noted that 'the ratio of hazard functions' does not depend on t provides a convenient way of summarizing the effect of a covariate on survival. 'Non-proportional hazards' models can be constructed by allowing $h_0(t)$ to depend on X, whereas time varying covariates allow X to depend on t.

3. Materials and methodology

The mortality data of cancer is provided by National Cancer Institute mentioned in the following link: https://seer.cancer.gov. In this work, we have considered the rates of population as per 1 lakh in the year between 1975 and 2015.

Here we have considered cancer patients who are classified by types of cancer [(i) Kidney and Renal Pelvis and (ii) Colon and Rectum] and by gender (male, female). To use gender and different types of cancer as predictors in a model for frequency of death, the proper baseline is not the number of subjects but rather the total time that subjects were at risk. Therefore, we model the rate of death. The time at risk for a subject is their follow-up time of observation. For a given gender and different types of cancer, the total time at risk is the sum of the times at risk for all subjects in the cell those who are died. We can get death rate by dividing deaths with the time at risk. We now model effects of gender and cancer type on the rate. Let g be a dummy variable for gender, with $g_1 = 0$ for male and $g_2 = 1$ for female. Let c be a dummy variable for cancer type, with $c_1 = 0$ for Kidney and Renal Pelvis and $c_2 = 1$ for Colon and Rectum. Let n_{ij} denotes the number of deaths for gender g_i and cancer type c_j, with expected value λ_{ij} for total time at risk t_{ij}. Given t_{ij}, the expected rate is λ_{ij}/t_{ij}. In this work, we have considered the model:

$$\log\left(\frac{\lambda_{ij}}{t_{ij}}\right) = \theta + \beta_1 g_i + \beta_2 c_j. \tag{17}$$

The corresponding model of Equation (17) with identity link be

$$\lambda_{ij} = \theta t_{ij} + \beta_1 g_i t_{ij} + \beta_2 c_j t_{ij}. \tag{18}$$

Now we fit the data for Poisson regression model considering log link and then the corresponding model with identity link can be fit which is shown in the following table:

It is assumed that there is a lack of interaction in the effects in the model. Here, model fitting uses standard iterative methods, treating n_{ij} as independent Poisson variates with means λ_{ij}. This is done conditional on t_{ij}.

In clinical investigations, there are many situations where several known quantities (known as covariates) potentially affect patient prognosis. In this work, two groups of patients are compared. For this purposed the established data are suitably constructed as well as computed for the underlying model. Let us consider two different types of gender (i) male ($= 0$) and (ii) female ($= 1$). For censoring status: 0 = Censored, 1 = Dead. Type-1 is the annual rate of affecting and dying people in Kidney and Renal Pelvis cancer, similarly Type-2 is the annual rate of affecting and dying people in Colon and Rectum cancer.

4. Concluding remarks

This work demonstrates how model fitting, estimation and testing methods have been developed for log-linear contingency table analysis. The methods for handling ordered categories, such as those described in Fienberg (1977), would be very useful. Table 1 presents the fitted death counts and estimated rates. The estimated effects are $\widehat{\beta_1} = -0.39$ (standarderror $= 0.283$) and $\widehat{\beta_2} = 1.56$ (standarderror $= 0.367$). It can be concluded that when the gender type is given, the estimated rate for Colon and Rectum

Table 1. Fit for Poisson regression models.

		Log link		Identity link	
Gender		Kidney & Renal Pelvis	Colon & Rectum	Kidney & Renal Pelvis	Colon & Rectum
Female	Number of deaths	3.63	17.36	2.78	19.5
	Death rate	0.007	0.036	0.005	0.041
Male	Number of deaths	5.36	25.63	6.5	23.21
	Death rate	0.011	0.053	0.013	0.048

cancer is $\exp(1.56) = 4.75$ times that for the Kidney and Renal Pelvis cancer. The 95% Wald confidence interval for β_2 of $1.56 \pm 1.96(0.3675)$ translates to $(2.4, 10.5)$ for the true multiplicative effect $\exp(\beta_1)$ and the likelihood-ratio confidence interval is $(2.4, 9.8)$. The present study contains much censored data. From the 133 patients, only 52 died during the study period. Both effect estimates are imprecise. The analysis uses all 133 patients through their contributions to the times at risk.

Goodness-of-fit statistics comparing n_{ij} to fitted values $\widehat{\lambda_{ij}}$ are $G^2 = 0.2$ and $X^2 = 0.3$. The residual degrees of freedom is equal to 1, since the four response counts have three parameters. The mild evidence of lack of fit corresponds to evidence of interaction between gender and cancer type. However, the model only considered different types of cancer and omitting the effect of gender (i.e. $\beta_1 = 0$), then it fits nearly as well, with $G^2 = 2.14$ and $X^2 = 2.16$ and degrees of freedom equals to 2. We also conclude that the corresponding model with identity link shows a good fit with $G^2 = 0.3$ and $X^2 = 0.3$ (degrees of freedom $= 1$), the table shows the fit. The estimate $\widehat{\beta_1} = -0.007$ (standard error $= 0.005$) then represents an estimated difference in death rates between the female and male gender for each type of cancers. Also, the estimate $\widehat{\beta_2} = 0.034$ (standarderror $= 0.007$) then represents an estimated difference in death rates between the different type of cancers for each type of gender.

The greatest usefulness of the Cox model in any application may be in describing the relationship between survival time of patients and various explanatory variables. Results of this study indicate that the Cox model is very useful for cancer mortality. Generally, the Cox proportional hazards regression analysis works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time. This survival rate of cancer patients is synonymous to the mortality rate of patients. The present work also expands to include an analysis of the significance of the variables used in influencing the survival rate of cancer patients. The 'survival rate' and the 'relationship between independent variables and survival rate' lead to an estimate of the reliability of the survival rate of cancer patients which is obtained in this study.

In the multivariate Cox analysis, the covariates 'Gender' and 'Type-1' (for Kidney and Renal Pelvis) remain significant (p-value < 0.01). The p-value for Gender < 0.01 with a hazard ratio HR $= \exp(coef) = 0.18$ indicates a relationship between patients' Gender and decreased risk of death. The hazard ratios of covariates are interpretable because of multiplicative effects on the hazard. If holding the other covariate as constant, being female (Gender $= 1$) reduces the hazard by a factor of 0.18, or approx 82%. Therefore, we come to the conclusion: being female is associated with good prognostic. Similarly, the p-value for Type-1 < 0.01 with a hazard ratio HR $= 0.89$ indicates a strong relationship between

Table 2. Summary of fitted models for survival.

Cancer type	Test	Value	Df	p-Value
Type-1: Kidney and Renal Pelvis	LR test	63.61	2	$< .01$
	Wald test	15.82	2	$< .01$
	Score test	43.49	2	$< .01$
Type-2: Colon and Rectum	LR test	23.43	2	$< .01$
	Wald test	22.34	2	$< .01$
	Score test	25.25	2	$< .01$

Type-1 value and decreased risk of death. Holding the other covariate as constant, a higher value of Type-1 is associated with a good survival. The summary outputs of the performed computation through R provides upper and lower 95% confidence intervals for the 'hazard ratio (exp($coef$)) of Gender': lower 95% bound $= 0.13$, upper 95% bound $= 0.58$. Similarly for the 'Type-1' case, we get lower 95% bound $= 0.86$, upper 95% bound $= 0.94$. The covariates 'Gender' and 'Type-2' (for Colon and Rectum cancer) remain significant (p-value < 0.01). The p-value for Gender < 0.01 with a hazard ratio HR $= \exp(coef) = 0.27$ indicates a relationship between the patients' Gender and decreased risk of death. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard. If holding the other covariate as constant, being female (Gender $= 1$) reduces the hazard by a factor of 0.27, or approx 73%. It can also be concluded that being female is associated with good prognostic. Similarly, the p-value for 'Type-2' < 0.01 with a hazard ratio HR $= 0.43$ indicates that there exists a relationship between 'Type-2' value and decreased risk of death. The summary outputs also gives upper and lower 95% confidence intervals for the hazard ratio (exp($coef$)) of Gender: lower 95% bound $= 0.02$, upper 95% bound $= 1.5$. Similarly, for 'Type-2' case, we get lower 95% bound $= 0.2$, upper 95% bound $= 0.8$.

The performed tests evaluate the null hypothesis 'all of the betas (β) are zero'. In this work, the test statistics are in good agreement with 'the null hypothesis is soundly rejected'. Finally, the output gives p-values for three alternative tests for overall significance of the model: (i) Likelihood-ratio test, (ii) Wald test, and (iii) Score log-rank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The 'Likelihood ratio test' has shown better behaviour for small sample sizes, so it is generally preferred for the Kidney and Renal Pelvis cancer (as evident from Table 2). But the value of score test is high for small sample sizes as evident from Table 2, so it can be concluded that the score test is useful for the Colon and Rectum cancer. For short time intervals, the piecewise exponential approach is essentially non-parametric, making no assumption about the dependence of the hazards on time. From the view point of data modelling, the piecewise exponential models are very flexible. If nothing is assumed about the underlying survival distribution, then an essentially non-parametric analysis can be implemented by making the time intervals sufficiently small. For future developments in this area, more formal methods for model simplifications should be included based on fitting a time curve to the estimated time effect. From the analytical point of view, it is both intuitively appealing and conceptually expedient to include survival analysis in the general counted data framework. The most obvious advantage is that the whole class of log-linear models used to characterize 'counted data structures' carries over directly to characterize 'survival data'. In this general framework, 'competing risk analyzes' and 'time varying covariates' can be easily and suitably handled.

The Poisson regression was used on flood occurrences (responsible variable) and the set of explanatory variables under consideration were tested in the work of Cupal, Deev, and Linnertova (2015). But in the present work, not only to determine the dependent variables of the underlying Poisson regression model with relatively input parameters but also Cox model is applied to evaluate simultaneously the effect of several factors on survival.

Our goal is to control infection and thereby prevent reproductive health problems and also aware the people about the diseases. So, the development of various cancer therapies and identification of the most effective therapy against the spread of tumour cells should be formed as a part of future research.

Acknowledgements

The authors are grateful to the anonymous referees and Editor in Chief Prof. Olcay Akman for their careful reading, valuable comments and helpful suggestions, which have helped them to improve the presentation of this work significantly. The first author (Sarada Ghosh) is thankful to the University Grants Commission, India for providing JRF.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Agresti, A. (2002). *Categorical data analysis*. 2nd ed. New York, NY: John Wiley and Sons.

Agresti, A. (2007). *An introduction to categorical data analysis*. 2nd ed. New York, NY: John Wiley and Sons.

Agresti, A. (2010). *The analysis of ordinal categorical data*. 2nd ed. New York, NY: John Wiley and Sons.

Aitkin, M., & Clayton, D. G. (1980). The fitting of exponential, weibull and extreme value distributions to complex censored survival data using GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(2), 156–163.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of Royal Statistical Society Series B*, *46*, 1–30.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Cupal, M., Deev, O., & Linnertova, D. (2015). The Poisson regression analysis for occurrence of floods. *Procedia Economics and Finance*, *23*, 1499–1502.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge, MA: MIT Press. 5, (2), pp. 263–264.

Ghosh, S., & Samanta, G. P. (2019). Fitting cumulative logit models for ordinal response variables in retail trends and predictions. *International Journal of Economics and Statistics*, *20*(1), 32–49.

Holford, T. R. (1976). Life tables with concomitant information. *Biometrics*, *32*(3), 587–597.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, *35*(1), 139–156.

Laird, N., & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis technique. *Journal of the American Statistical Association*, *76*(374), 231–240.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B*, *42*, 109–142.

Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. SanDiego, CA: Academic Press.

Appendix

```
# Poisson Regression for Rates

Cancer_data=read.csv("C:\\Users\\Acer-PC\\Desktop\\New folder (2)\\Cancer.csv")

fit.rate<-
glm(Deaths~Gender+Cancertype+offset(log(Exposure)),family=poisson,data=Cancer_data)

summary(fit.rate)

attach(Cancer_data)

 mhat<-fitted(fit.rate)

exphat<-fitted(fit.rate)/Exposure

temp<-rbind(mhat,exphat) array(temp,dim=c(2,2,2),dimnames=list(c("Deaths","Risk"),
Gender=c("Female","Male"), Cancertype=c("Type1","Type2"))

fit.idk<-glm(Deaths~I(Gender * Exposure) + I(Cancertype * Exposure) + Exposure -1,

family=poisson(link=identity), data=Cancer_data)

 summary(fit.idk)

# Modeling Survival Times

library(survival)

res.cox_gen=coxph(Surv(time,status)~Gender)

summary(res.cox_gen)

res.cox_type=coxph(Surv(time,status)~Cancertype)

summary(res.cox_type)
```