Joint Modeling of Mortality Incidence and Survival.

A.A.Sunethra, M.R. Sooriyarachchi
Department of Statistics, University of Colombo

Abstract - Analysis of medical data mostly consider survival and mortality count as response variables in identifying factors that are associated with the survival times and with the death count of patients. However, it is quite possible and the literature has shown evidence for these two responses to be correlated and share common factors. Therefore, joint modeling of these two responses simultaneously in one model can provide improved results than fitting two univariate models since the correlation between the two responses can be captured in a joint model. The literature did not consist of any such situation where joint modelling of survival time and the death count was considered. This manifested the objective of developing a method for jointly modeling survival and count responses for which a bivariate Poisson model was proposed. This method was facilitated by the equivalence of the log-likelihoods of survival and Poisson models. The suggested method was fitted for a data set of Dengue patients where factors associated with survival times of dengue patients and death count of patients were identified by the joint model. For comparing the performance of the proposed joint model with two univariate models that can be fitted separately for the two responses, the Akaike Information Criterion (AIC) was used. It was confirmed that the performance of the joint model surpasses the fit of two univariate models since the AIC of the joint model was lower than the total of the AICs of the two univariate models.

Keywords- bivariate poisson, joint modeling, mortality incidence, survival

I. Introduction

'Incidence of Mortality' and 'Survival' are two of the most common measurements that are associated with the analysis of medical data particularly in Epidemiology. Here the incidence of mortality refers to the number of deaths of patients while survival refers to the lifetime of patients. Therefore, these two measurements have played the role of dependent variables in most of the statistical models fitted to medical data. For example, analysis focussed on the incidence of mortality would use number of deaths as the dependent variable and it is of interest to identify the factors that effect the death count of patients having a particular disease, whereas the analysis focussed on the life time of patients would use survival time as the dependent variable in the models and in both instances it is of interest to identify the factors that have an impact on the two responses. However, it is often the case that these two responses (survival and mortality incidence) are related to each other (i.e. Correlated) and can have common factors associated with the number of deaths and time to death. When two variables are dependent and share common variables, it is more efficient to fit a joint bivariate model to such data with count as one response and survival time as the other response. This is the motivation behind this research. The work is novel since no such analysis was found in the literature where joint modeling of mortality (death count) and time to death was considered. This leads to the formulation of the objective of developing a methodology for the joint modelling of morality incidence and survival. It is noteworthy that the method suggested here is not restricted only for medical/health data since it can be applied to data from any other scenario which has a count response variable and a duration/time to event response variable that are correlated.

II. Literature review

A comprehensive literature review was undertaken which manifested the pathway to formulate the methodology for jointly modelling of a survival and count response. In the past, the analysis of survival and event counts was done in such a way that the survival time was taken as the dependent variable of the model while the count variable was taken as a covariate of the model [1]. This has the inherent weakness of treating the count variable as a fixed effect, whereas actually the count (i. e. the number of deaths) is not a fixed effect parameter. As an improvement, [2] suggested to take the count as a covariate measured with some error. But, [3] indicated through a simulation study that joint models provide more precise estimates than standard survival models which are constrained on taking count as a covariate. The joint modeling of a survival and a count variable was found in [4] where frailty proportional hazard model was used for jointly modeling a survival variable and count variable where the count resembled number of hospitalizations and survival variable was the time to death. Reference [5] demonstrated the use of Poisson process approach for jointly modeling a survival and a count variable where the survival referred to time to death and count resembled the number of Epileptic seizures experienced by some patients with Epilepsy. The difference in this study when compared to [4] is that it required the timings of the events which made the count variable, i.e. timing of hospitalizations while our study is based on the cumulative counts of such events (count of death). With the availability of the timings of the events, the joint modelling of survival and count can be achieved by joint modeling of two survival/time-to-event processes where one time-to-event process relates to the death and the other relates to the timing of hospitalizations or seizures.

The approach undertaken here considered the count variable as a Poisson random variable while an indirect approach was undertaken to model the survival process. There is a nice coincidence between Poisson random variables and survival random variables where the log-likelihood of a Poisson random variable is equivalent to the log-likelihood of a survival random variable under the assumption of proportional hazards of survival data. This equivalence is explained and used for modeling survival data by [6] and [7] where they used the Poisson model for estimating survival models. Based on this approach, the joint modelling of survival and count can be achieved by joint modelling of two Poisson variables. Therefore, the literature for joint modeling of two Poissons is considered.

Among the several distributions for joint distribution of two Poisson, the most widely used is the bivariate Poisson for which the definition is not unique [8]. The trivariate reduction method [9] is used in this study, which has been used by [10], [11] and [8] for analyzing correlated count data.

III. Methods

The key considerations on joint/multivariate modeling are the features of the variables that make up the multivariate/joint response and on obtaining the joint distribution of those responses. When the responses are from different families of distributions leads to difficulties in deriving a joint distribution of the joint responses. This is the case in the bivariate response of survival and count as well. The presence of censored observations in survival responses is a major problem when combining a survival response with any other type of responses. Thus, formulating a joint model for a survival response and a count response was a challenging task mainly due to the difficulty in obtaining a joint distribution of the two responses. Existing software for multivariate survival modelling could not be utilized for combining with a count variable though it was possible to jointly model two or more survival responses

This difficulty was overcome by using a two-stage approach of modelling a survival distribution through a Poisson distribution. This approach was facilitated mainly due to the equivalence of the log-likelihoods of a survival model and a Poisson model under the assumption of proportional hazard in survival data. That is, when the proportional hazard assumption holds for survival data, estimates for a survival model (parametric or semi-parametric) can be achieved by the fit of a Poisson model which is described in the following section [6]; [7].

Equivalence of Survival and Poisson Log-likelihoods

Let T_i denote the survival time of the ith individual

Let δ_i be an indicator variable taking the value 1 for actual survival times(uncensored) and the value 0 for censored survival times.

he Log-likelihood of survival data can be derived as [4] :

$$l(\beta \mid b) = \sum_i \delta_i \log(\mu_i) - \mu_i + \delta_i \log\left[\frac{h_0(t_i)}{H_0(t_i)}\right] \qquad (1)$$

$$\mathrm{l}(\beta|\mathrm{b}) = \sum_{\mathrm{i}} \delta_{\mathrm{i}} \log(\mu_{\mathrm{i}}) - \mu_{\mathrm{i}} + \delta_{\mathrm{i}} \log\left(\frac{\mathrm{h}_0(\mathrm{t}_{\mathrm{i}})}{\mathrm{H}_0(\mathrm{t}_{\mathrm{i}})}\right)$$

Where μ_i denotes the mean survival time, $h_0(t_i)$ denotes the baseline hazard function and $H_0(t_i)$ is the cumulative baseline hazard.

In the case of semi-parametric survival models (Cox-model) and with Exponentially distributed survival data, the last term on the right hand side of (1) does not involve any unknown parameters and hence does not influence the maximum likelihood estimation [7].

Now consider the log likelihood of ‘n’ independent Poisson random variables $w_i \sim Poisson(\mu_i^*)$ which reduces to [8]:

$$\text{l}=\sum_i w_i \log(\mu_i^*) - \mu_i^* - k \qquad (2)$$

It is noteworthy that these two likelihoods are identical with respect to maximization when δ_i (Censoring indicator) is regarded as $\delta_i \sim \text{Poisson}(\mu_i{}^*) \delta_i \sim Poisson(\mu_i^*)$. Such a Poisson model ($\delta_i \sim \text{Poisson}(\mu_i{}^*) \delta_i \sim Poisson(\mu_i^*)$) can be used to estimate a survival model.

Though this equivalence holds for parametric models [6] as well as semi-parametric models [7], the estimation procedure explained below is considered only for semi-parametric models [7] as the Cox proportional hazard model [12] is by far the most popular model for survival data.

For Cox proportional hazard model, $h_0(t) \, h_0 \ (t)$ is an arbitrary function.

$$h_i \ (t) = h_0 \ (t) \exp(\eta_i) \qquad (3)$$

As per [12], the parameters of this model can be estimated via the partial likelihood estimation which yields the following form of the partial likelihood:

$$L = \prod_{i \in F} \frac{\exp(\eta_i)}{\sum_{R(t_i)} \exp(\eta_{i'})} \qquad (4)$$

Where R(t) is the set of individuals at risk (the risk set) and F is the set of individuals whose survival times were observed (non-censored).

Now, consider the specification of likelihood for the following Poisson random variable.

Let t_k k=1,2,….,L be the times where actual survival times have occurred. For each such actual survival time, create the following variables for the observations belonging to R(t_k).

$$y_{i(k)} = \begin{cases} 1, & \textit{for actual survival times} \\ 0, & \textit{for censored survuval times} \end{cases}$$

Now consider :

$$y_{i(k)} \sim Poisson\left(\mu_{i(k)}\right) = \exp\left(\alpha_k + \eta_i\right))$$

And suppose that these ($y_{i(k)}$) are independent. Therefore, likelihood for survival time t_k can be written as [13]:

$$l_k^* = \prod_{i(k)} \frac{e^{-\mu_{i(k)}} {\mu_{i(k)}}^{y_{i(k)}}}{y_{i(k)}!} = \frac{e^{\alpha_{k+}\eta_i}}{exp\left[e^{\alpha_k} \sum_{i\grave{o}R(t_k)} \exp\left(\eta_i\right)\right]} \qquad (5)$$

According to the properties of the Poisson distribution, $\sum_{i\varepsilon R} y_{i(k)}$ also follows a Poisson distribution with mean. $\mu_k = \sum_{i\grave{o}R} \mu_{i(k)}$ Assuming that there are no tied observations, the following holds:

$$\sum_{i\grave{o}R(t_k)} \mu_{i(k)} = e^{\hat{\alpha}_k} \qquad \sum \exp\left(\eta_i\right) = 1$$

Therefore, $e^{\hat{\alpha}_k} = {}^1/_{\sum_{i \in R(t_k)} \exp(\eta_i)}$

Thus, the likelihood function at the survival reduces to the following form:

$$l_k^* = \frac{{}^1\!/\!_{\sum_{i\grave{o}R(t_k)} \exp\left(\eta_i\right)} {}^* \exp\left(\eta_i\right)}{\exp[1]} = \frac{\exp\left(\eta_i\right)}{\sum_{i\grave{o}R(t_k)} \exp\left(\eta_i\right)}$$

(6)

Then, the likelihood of such ‘L’ Poisson random variables can be written as:

$$L = \prod_{k=1}^{L} \frac{\exp(\eta_i)}{\sum_{i\grave{o}R(t_k)} \exp\left(\eta_i\right)} \qquad (7)$$

which is equivalent to the partial likelihood function of the Cox- model. Therefore, parameter estimation for the Cox model can be achieved by fitting the Poisson model [4]; [12].

As per the details given above, the estimation of the survival model parameters can be done through the fit of a Poisson model. Therefore, through this approach the joint modelling of survival and count can be achieved by the joint modelling of two Poisson random variables. This now reduces the problem to obtaining the joint bivariate distribution of two responses from the same family.

The following section describes the joint modelling of two Poisson random variables.

Joint Modelling of Two Poissons

As per [13], “Joint Modelling of two or more counts data has received a great deal of attention in recent years”. Bivariate count models are used when two count variables are correlated.

Among the several models for joint distribution of two Poissons, the Bivariate Poisson has achieved more popularity [14]. While the definition of bivariate Poisson model is not unique, the trivarite reduction method of constructing the joint distribution is used in this study[8].

Let $X_1, X_2, \ldots, X_n$ be random variables which follow independent Poisson distributions.

Consider the following random variables:

$$X = X_1 + X_3$$
$$Y = X_2 + X_3$$

Then, X and Y jointly follow a bivariate Poisson distribution $BP(\lambda_1, \lambda_2, \lambda_3)$, of the form (8).

Where $\lambda_1, \lambda_2, \lambda_3$ denotes the mean of the three independent Poisson X_1, X_2, X_3 respectively.

$$f_B(x, y) = P(X = x, Y = y)$$

$$= e^{(-\ddot{e}_1-\ddot{e}_2-\ddot{e}_0)} \frac{\ddot{e}_1^{x}}{\ddot{e}x!} \frac{\ddot{e}_2^{y}}{\ddot{e}y!} \sum_{i=0}^{\min(x,y)} \binom{x}{i}\binom{y}{i} i! \left(\frac{\ddot{e}_0}{{}_1\quad{}_2}\right)^{i} \qquad (8)$$

The marginal distributions of the two count variables (X and Y) are Poisson and this model can only accommodate positive correlation between the two count variables.

The bivariate Poisson regression model is used for incorporating the effect of covariates on the bivariate responses.

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}); log\lambda_{ki} = w_i^T \alpha_k \quad \text{k=1,2,3}$$

where $i = 1, 2, \ldots, n$ $i = 1, 2, \ldots, n$ denotes the observations, w_i denotes the vector of explanatory variables of length ‘p’ and α_k is the corresponding vector of regression parameters. It is important to note that λ_3 denotes the covariance between the two variables X and Y. Thus, fit of the Poisson model for λ_3 gives the explanatory variables that constitutes the covariance between the two Poison responses of X and Y [14]. Simplifying the above detailed explanation on the methodological development, it can be noted that the method suggested here is mainly based on two aspects i.e modelling the survival variable as a count variable and jointly modeling the two count variables.

The suggested method is illustrated for some data on Dengue patients reported in Sri Lanka in the years 2006-2008.

IV. Example

The final data set was obtained by combining data on Dengue patients obtained from the Epidemiological Unit, which had records of Dengue patients reported in 2006-2008 and climate data that was obtained from the Meteorological Department of Sri Lanka. Since the analysis here is mainly focussed on illustrating the suggested methodology, the variables that had previously been identified to be significant in the literature were used in this study where the table 1 gives a list of the variables used in the study, respective categories and coding used in the analysis.

Table 1: Data Description

Variable	Notation	Categories	Code
Survival Time	SURVIVAL	< 7 days	1
		1-9 Days	2
		> 9 days	3
Outcome	OUTCOME	Died	1
		Discharged	0
Place Treated Initially	PATTREAT	Government Hospital	0
		Private Hospital	1
Fever	FEVER	Yes	1
		No	0
White blood Cell Count	WBCL	< 4700	0-(low/ Moderate)
		> 4700	1 (high)
Platelet Count	PLTL	< 72000	0 (low/ moderate)
		> 7200	1 (high)
Packed Cell Volume	PCVH	< 45	0 (low/ moderate)
		> 45	1 (high)

Classification	CLASSIFI	Dengue Fever	1
		Dengue Homoerotic Fever	2
Rainfall	Rain	< 270.9	0 (low/ moderate)
		> 270.9	1 (high)
Temperature	Temp	< 260.5	0 (low/ moderate)
		> 260.5	1 (high)
Humidity	Humid	< 83.5	0 (low/ moderate)
		> 83.5	1 (high)
Death count	COUNT		

It is noteworthy that data was categorized as above mainly since the data were readily available in categorized form, but the method suggested is not restricted only for categorical responses or covariates.

The bivariate Poisson model was regressed to have count as one Poisson response variable and outcome variable as the other Poisson variable with the distinction of the Poisson model for outcome should be fitted with an offset of log(risk set) which is the logarithm of the number of patients at risk at each survival time. The inclusion of the offset could be achieved by fitting the Bivariate Poisson model using the SAS Proc NLMIXED procedure [8].

Initially all the explanatory variables were introduced to the model for both the responses, and the most insignificant variables were removed step by step which resulted following set of variables in the final model selected (Table 2). It can be seen that only few variables were significantly associated with the survival time though all the variables were significantly associated with the count of death. The covariance between the two variables (survival and death count) was associated with the place treated initially.

Table 2: Results of the Final Model

Response	Variable	Coefficient	P-value
Survival	intercept	-4.24	<. .0001
	Pattreat	-1.186	<. .001
	Pcvh	0.4431	.0055
	classifi	1.27	< .0001
	Temp	1.12	< .0001
	humid	0.53	.0138
Count	intercept	- 0.04	.1319
	Pattreat	.0.31	< .0001
	Fever	0.224	< .0001
	Wbcl	- 0.128	< .0001
	Platl	0.27	< .0001
	Pcvh	0.46	< .0001
	Classify	0.79	< .0001
	Rainfall	-0.05	< .0001
	Temp	1.05	< .0001
	Humid	0.39	< .0001
Covariance (Survival, ount)	Intercept	-3.46	< .0001
	Pattreat	-0.61	.0037

As per the parameter estimates obtained above, the marginal models for survival time (proportional hazard model) and death count (Poisson model) can be written as below

$$h(t) = h_0(t)\exp(-4.24 - 1.186 * \text{pattreat} + 0.44 * \text{pcvh} + 1.27 * \text{classifi} + 1.12 * \text{temp} + 0.53 * \text{humid}) \quad (9)$$

$$\log(\text{death count}) = -0.04 + .31 * \text{pattreat} + .224 * \text{fever} - .13 * \text{wbcl} + .27 * \text{platl} + .46 * \text{pcvh} + .79 * \text{classifi} - .05 * \text{rainfall} + 1.05 * \text{temp} + .39 * \text{humid} \quad (10)$$

$$\text{Cov}(\text{survival, count}) = \exp(-3.46 - 0.61 * \text{pattreat}) \quad (11)$$

A single covariate, namely, 'place treated' will be interpreted here. The rest of the covariates in the model can be similarly interpreted. The risk of death among the patients who were initially treated at private hospitals is 3.27 times (exp(1.186)) higher than the patients who were treated at government hospitals, while the expected death count is higher among the patients who were treated initially at private hospitals by an amount of 1.36 than those treated at government hospitals. The covariance between survival and the count is 0.017 for those treated in private hospitals and is 0.031 for those treated in government hospitals.

This example was mainly drawn for the purpose of showcasing the methodology suggested for joint modeling of survival and count variables which was expected to provide improved performance than two univariate models fitted for such data. Therefore, the fit of the joint model was compared with the fit of two univariate models for which the Akaike Information Criteria (AIC) was used. It is important to note that the AIC of the joint model was 168353 while the two univariate models had an AIC of 165800 for death count model while it was 2718 for the survival model which resulted a total of 168518 (165800+2718). Since the AIC of joint model is less than the total of the AICs of the two univariate models, it can be suggested that the joint model is more efficient than the two univariate models.

V. Discussion

The main objective of the study was to formulate a method for jointly modelling a survival and a count response variable which was motivated by the dependence between the survival times and the number of deaths of patients with particular diseases. A novel method which was developed using the equivalence of the log-likelihoods of survival and count data under the assumption of proportional hazards in survival data were suggested and applied. This required fitting a specialized form of a bivariate Poisson regression model. It was observed that the joint model is better than the two univariate models that can be fitted separately for the two responses.

Extensions to the study can be suggested mainly in two aspects of assuming parametric proportional hazard models for survival data and using another joint distributional form for the joint distribution of two Poisson variables, whereas this research only considered the Cox proportional hazard model for survival data and Bivariate Poisson model for the joint distribution of two Poisson variables.

References

[1]. Verity, C. M., Hosking, G., & Easter, D. J. "A multicentrecomparative trial of sodium valproate and carbamazepine in paediatric epilepsy", Developmental Medicine & Child Neurology, 37(2), 97-108, 1995.

[2]. Carroll, R . J., Ruppert, D . and Stefanski, L . A. Measurement Error in Nonlinear Models. London:C hapman and Hall, 1995.

[3]. Cowling, B. J. Survival models for censored point processes (Doctoral dissertation, University of Warwick), 2003.

[4]. Liu, L., Wolfe, R. A., & Huang, X. Shared frailty models for recurrent events and a terminal event.Biometrics, 60(3), 747-756, 2004.

[5]. Cowling, B. J., Hutton, J. L., & Shaw, J. E. H. "Joint modelling of event counts and survival times", Journal of the Royal Statistical Society: Series C (Applied Statistics), 55(1), 31-39, 2006.

[6]. Aitkin, M., Clayton, D. "The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM", Applied Statistic 29(2) :156-63, January 1980.

[7]. Whitehead, J. "Fitting Cox's regression model to survival data using GLIM", Applied Statistics, 29(3) : 268–275, 1980.

[8]. AlMuhayfith, F.E, Alzaid, A.A, Omair, M.A. "On Bivariate Poisson Regression Models", Journal of King Saud University- Science, 2015, doi: http://dx.doi.org/10.1016/j.jksus.2015.09.003.

[9]. Johnson, N. L., Kotz, S., &Balakrishnan, N. Discrete multivariate distributions, New York: Wiley, Vol. 165, 1997.

[10]. Karlis, D., &Ntzoufras, I. "Analysis of sports data by using bivariate Poisson models", Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381-393, 2003.

[11]. Karlis, D., &Ntzoufras, I. "Bivariate Poisson and diagonal inflated bivariate Poisson regression models" in: R. Journal of Statistical Software, 14(10), 1-36, 2005.

[12]. Cox, D. R. "Regression models and life-tables", Journal of the Royal Statistical Society. Series B (Methodological), 187-220, 1972.

[13]. Da Silva, G. T., & de Lima, A. C. P. 'Using SAS software for Multilevel Models in Survival Analysis', PharmaSUG SAS Users Group Conference, Miami, Florida, May 4-7, 2003.

[14]. Chou, N. T. Bivariate Count Data Regression Models–A SAS® Macro Program.