Ann. Data. Sci. (2014) 1(2):173–189
DOI 10.1007/s40745-014-0013-9

Extended Exponential Geometric Proportional Hazard Model

Sadegh Rezaei · Sina Hashami · Lotfollah Najjar

Received: 3 May 2014 / Revised: 8 July 2014 / Accepted: 1 September 2014 /
Published online: 8 October 2014

Abstract Researchers often use ordinary least square and generalized linear models even for censored data. Cox (1972) presented a new method that is useful for cases which include censored data. Researchers began to use this model without considering baseline hazard models. These methods are described and compared with a new proportional hazard model. Conclusions on the performance are presented here. In this paper , we propose the Extended Exponential Geometric (EEG) proportional hazard model. We compared this model with: (1) the semi-parametric proportional hazard model (2) the Linear regression model with and without log translation, (3) Generalized linear models (GLM), for example, the use for the case with strictly positive values. Survival analysis examines and models the time required for events to occur. Survival analysis focuses on the distribution of survival times. There are well known methods for estimating unconditional survival distributions. Most interesting survival modeling examines the relationship between survival and one or more predictors, usually termed covariances in survival analysis literature.

Keywords Proportional hazard models · Generalized linear model · Ordinary least square · Extended exponential geometric distribution

S. Rezaei (✉) · S. Hashami
Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
e-mail: srezaei@aut.ac.ir

L. Najjar
Information System Quantitative Analysis, College of Information Science and Technology (IS&T), University of Nebraska, Omaha, NE, USA

1 Introduction

Survival analysis examines and models the time required for events to occur. Survival analysis focuses on the distribution of survival times. There are well known methods for estimating unconditional survival distributions. Most interesting survival modeling examines the relationship between survival and one or more predictors, usually termed covariances in survival analysis literature.

The Cox proportional hazards model [5] is now the most widely used for the analysis of survival data in the presence of covariances or prognostic factors. This is also the most popular model for survival analysis because of its simplicity, and because it is not based on any assumptions about the survival distribution. Part of the interest in survival methods for event times, costs, expenditures, and related outcomes is the presence of censoring of some sort in the data collection. As Kleinbaum etal. describe this in their survival books [13]. Up to now, many researchers have compared the performance of the Cox model with other regression models. Using data on 155 CABG cases, Dudley et al. [8] compared the Cox regression model with OLS models (with and without log transformation), and logit alternative. They found that the Cox model provides more accurate predictions of mean, median, and high cost cases.

In a study of Medicare data of stroke patients, Lispscomb et al. [15] compared the performance of several alternative estimators: OLS with and without log transformation, two-part models with and without log transformations of the positive expenses, and Cox proportional hazards models. Using several criteria for comparisons of alternative estimation approaches under cross validation, they found that the Cox and two-part model with log transformation outperformed the other alternatives. Some researchers have been concerned how it is possible apply survival methods if there is non-random censoring or other issues [9,10]. Others have used survival methods in cases without censoring, because they wanted to employ less parametric methods than have been typically used before (Anirban Basu et al. 2004).

The purpose of this paper is to propose a new form for baseline hazard function $h_0(y)$ that is related to Extension exponential geometric distribution(EEG distribution) [1] for proportional hazard models and accelerated failure time models.

EEG distribution can be used in modeling situations where the population's survival capacity decreases over time. Its ability to closely fit real data renders it a promising alternative to the popular weibull and gamma distributions. This distribution is useful for series systems with identical components and parallel systems with identical components. Some baseline hazards have been proposed in literature, for instance:

$$h_0(y) = \lambda\, \rho y^{\rho-1}$$

For parametric proportional hazards model this choice leads us to Weibull distribution, [7] and for parametric accelerated failure time models this choice is a particular assumption for $h_0(y)$. This choice also leads us to Weibull distributions [7]. The Weibull model is a popular parametric model which allows for the inclusion of covariates of survival times. Exponential distribution is also a special case of Weibull distribution; statisticians chose Exponential distribution to model life data because the statistical methods for it were fairly simple [14]. As earlier, we propose a new

parametric proportional hazard model and an accelerated failure time model based on EEG distribution. Finally we will use simulation techniques to see which of these models perform best. The models that we considered are: (1) EEG proportional hazard model (2) semi-parametric proportional hazard model (3) linear regression on $\ln(y)$ (4) Generalized linear models (GLM), specifically, the use of Gamma distribution with a log link $E(y \mid x) = \exp(x\beta)$ It is clear that all of these methods are used for cases with strictly positive values.

We use the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to evaluate the performance of these models and to determine which of these models is appropriate under different situations.

The plan for the paper is as follows. Sect. 2 describes the general modeling approaches that we consider. Then we present our new models in Sects. 3 and 4. Our simulation framework is then presented. The results of the simulations that focus on the survival time and two covariates that related to the survival time are discovered. Finally the conclusion follows.

2 Modeling Framework

We adopt the perspective that the purpose of the analysis is to show how the expected outcome, $E(y \mid x)$, responds to shifts in a set of covariates.

While many aspects of the following discussion apply for the more general case of nonnegative y , the discussion here is confined to the case with strictly positive values of y to streamline the analysis.

Note, right censoring is very common in survival time data, but left censoring is fairly rare. The term "censoring" will be used in this paper to mean in all instances " right censoring".

Our modeling framework includes: a gamma regression model with a log link (one of the generalized linear model, GLM); OLS on $\ln(y)$; and the Cox proportional hazard model with unspecified baseline hazard.

The gamma model estimates the $E[\ln(y)]$ directly, but the Cox model estimates the hazard rate that a case which has survived to time y will fail in the next time period.

2.1 Gamma Models

In generalized linear models (GLM) [17], we assume that $E[\ln(y)]$ exhibits an exponential conditional mean or log link relationship:

$$\ln(E(y \mid x)) = x'\beta$$

or

$$E(y \mid x) = \exp(x'\beta) = \mu(x; \beta)$$

Because of the work by Blough et al. [2], and by Manning and Mullahy [16], we will focus on the gamma regression model. The gamma distribution has a raw scale variance function that is proportional to the square of raw scale mean function.

Trough out this paper, we are assuming a log link for the expectation of y given x.

2.2 Linear Models

One of the oldest models that researchers use, is OLS on y . Where y is the dependent variable. The regression model is given by:

$$y = x'\beta + \epsilon$$

where x is a matrix of observations on covariates, β is a column vector of coefficients to be estimated, and ϵ is the column vector of error terms. We assume that $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$ and ϵ_i are independent. Then we can estimate β by:

$$\hat{\beta} = (x'x)^{-1}x'y$$

2.3 Log Linear Models

One of the popular models that researcher use is ordinary least suers or least suers variant with $\ln(y)$ as the dependent variable. The reason for log transform is that the resulting error term is approximately normal. The regression model is given by:

$$\ln(y) = x'\beta + \epsilon$$

where x is a matrix of observations on covariates, β is a column vector of coefficients to be estimated, and ϵ is the column vector of error terms.

We assume that $E(\epsilon) = 0$ and $E(x'\epsilon) = 0$. If the error term is normally distributed, then

$$E(y|x) = \exp\left(x'\beta + 0.5\sigma_\epsilon^2\right)$$

If ϵ is not normally distributed, but is i.i.d., or if $\exp(x'\beta)$ has constant mean and variance, then

$$E(y|x) = s\exp(x'\beta),$$

where $s = \exp(\epsilon)$. If $\exp(\epsilon)$ is some function $f(x)$ then

$$E(y|x) = f(x)\exp(x'\beta),$$

or equivalently,

$$E(y|x) = x'\beta + \ln(f(x)),$$

and in the log normal case,

$$E(y|x) = x'\beta + 0.5\sigma_\epsilon^2,$$

2.4 Semi-parametric Proportional Hazards Models

One very popular model in survival data is the Cox proportional hazards model, which is proposed by Cox. The Cox proportional hazard model is given by:

$$h_i(y) = h_0(y)\exp(x_i'\beta),$$

where $h_0(y)$ is the unspecified baseline hazard function when $\exp(x_i'\beta) = 0$ Note that the expected value of can be written as:

$$\begin{aligned} E(y) &= \int_0^\infty S(y)dy \\ &= \int_0^\infty \exp\left\{-\int_0^\infty h(u)du\right\}dy \end{aligned}$$

The estimates of the β parameters are obtained by maximizing the partial log likelihood, because the baseline hazard is separable from the part containing β. The partial likelihood was originally given by Cox (1972). The partial likelihood can be derived as a profile likelihood, i.e., first β is fixed and the survival likelihood is maximized as a function of $h_0(y)$ only to find estimators for the baseline hazard in terms of β. We write the partial likelihood as follows:

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(x_{(i)}'\beta)}{\sum_{j\in R(y_{(i)})}\exp(x_j'\beta)} \tag{1}$$

where $y_{(1)} < y_{(2)} < \cdots < y_{(r)}$ denote the ordered event times with corresponding covariates $x_{(1)}, \ldots, x_{(r)}$ and $R(y_{(i)})$ is the risk set at time $y_{(i)}$ that containing all the subject that are still at risk to experience the event at that time.The partial likelihood can be interpreted in terms of conditional probabilities (Klein and Moeschberger 1997). The properties of the partial likelihood estimator for β are well established (Gill 1984; Fleming and Harrington 1991).

Since we are interested in the expectation of y, not the hazard or survival function, per se, we need to estimate the baseline hazard and survival functions to predict survival time at various levels of the covariates X. For this purpose we have to use Breslow's estimators [4]. Although the Cox models uses an exponential form for the hazard function, the interpretation of the estimated coefficient is different that is in OLS on y or gamma models.

Note that the log normal and gamma distribution models do not satisfy the proportional hazard assumption.

3 Proportional Hazards Models

In first subsection we are going to obtain a proportional hazard model with a new baseline hazard function, and in second subsection, we wish to estimate parameters. These

parameters break down into two parts: The parameters that are related to distribution and the parameters that are related to efficiency of explanatory variables.

3.1 Parametric Proportional Hazard Model

The parametric proportional hazards model is the parametric version of the Cox proportional hazards model. In general in the presence of covariates the proportional hazard model can be written as:

$$h_i(y) = h_0(y)\exp\left(\sum_{i=1}^{p}\beta_i X_i\right) \tag{2}$$

$$= h_0(y)\exp\left(X_i'\beta\right) \tag{3}$$

In this model if $h_0(y)$ is assumed to follow a specific distribution, we will use the parametric proportional hazard model. We wish to estimate the parameters in this model by maximizing likelihood.

Now we are going to discuss about the parametric proportional hazard model with a new baseline hazard. suppose that:

$$h_0(y) = \frac{\lambda}{1-(1-\gamma)e^{-\lambda y}} \tag{4}$$

where $\lambda > 0$ and $\gamma > 0$.

The function $h_0(y)$ is monotonically increasing for $\gamma > 1$, decreasing for $0 < \gamma < 1$ and constant for $\gamma = 1$. As we said in introduction this choice refers to EEG distribution which proposed by Adamidis(2005). Then proportional hazard is given by:

$$h_i(y) = \frac{\lambda}{1-(1-\gamma)e^{-\lambda y}}\exp(X_i'\beta) \tag{5}$$

It follows that:

$$\begin{aligned} S_i(y) &= \exp\left(-\int_0^y \frac{\lambda}{1-(1-\gamma)e^{-\lambda s}}\exp(X_i'\beta)d_s\right) \\ &= \exp\left(\exp(X_i'\beta)\int_0^y \frac{-\lambda}{1-(1-\gamma)e^{-\lambda s}}d_s\right) \\ &= \left(\frac{\gamma e^{-\lambda y}}{1-(1-\gamma)e^{-\lambda y}}\right)^{\exp(X_i'\beta)} \end{aligned} \tag{6}$$

For last equality see appendix A.
We know that $f_i(y) = h_i(y)\cdot S_i(y)$, by (5),(6) we can write:

$$f_i(y) = \frac{\lambda}{1-(1-\gamma)e^{-\lambda y}}\exp\left(X_i^\tau\cdot\beta\right)\cdot\left(\frac{\gamma e^{-\lambda y}}{1-(1-\gamma)e^{-\lambda y}}\right)^{\exp\left(X_i^\tau\cdot\beta\right)} \tag{7}$$

3.2 Estimation by Maximum Likelihood

Now we are going to estimate the parameters by maximum likelihood. The form of likelihood expression is determined by the type of data that is available. We introduce g and G as notation for the density function and the cumulative distribution function of the censoring time, and also f and F for the density function and cumulative function of the event time respectively.

Under random censoring, survival data consist of a combination of event times and censored observations. The likelihood for a sample of size is therefore given by:

$$L = \prod_{i=1}^{n} \Big[\big(1 - G(y_i)\big) f(y_i)\Big]^{\delta_i} \Big[\big(1 - F(y_i)\big) g(y_i)\Big]^{(1-\delta_i)} \tag{8}$$

If we further assume that the distribution of the censoring times does not depend on the parameters of interest related to the survival function, or we have uninformative censoring, (Liang et al. 1995; Fleming and Harrington 1991), the factors $(1-G(y_i))^{\delta_i}$ and $(g(y_i))^{1-\delta_i}$ are not informative for inference on the survival function and, therefore, they can be deleted from the likelihood resulting in:

$$\begin{aligned} L &= \prod_{i=1}^{n} (f(y_i))^{\delta_i} \, (S(y_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\left(\frac{\lambda \exp\left(x_i'\beta\right)}{1-(1-\gamma)e^{-\lambda y_i}} \right) \left(\frac{\gamma e^{-\lambda y_i}}{1-(1-\gamma)e^{-\lambda y_i}} \right)^{\exp(x_i'\beta)} \right]^{\delta_i} \\ &\quad \left[\left(\frac{\gamma e^{-\lambda y_i}}{1-(1-\gamma)e^{-\lambda y_i}} \right)^{\exp(x_i'\beta)} \right]^{1-\delta_i} \end{aligned} \tag{9}$$

Then:

$$\begin{aligned} l = &\sum_{i=1}^{n} \delta_i \ln\left[\lambda \exp\left(x_i'\beta\right)\right] - \sum_{i=1}^{n} \left(\delta_i + \exp(x_i'\beta)\right) \Big(\ln\left[1-(1-\gamma)e^{-\lambda y_i}\right] \Big) \\ &+ \sum_{i=1}^{n} \left(\exp(x_i'\beta)\right) \ln\left[\gamma e^{-\lambda y_i}\right] \end{aligned}$$

The first derivatives of the log likelihood function with respect to the parameters are:

$$\begin{aligned} \frac{\partial l}{\partial \lambda} = &\frac{1}{\lambda} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \left(\delta_i + \exp\left(x_i'\beta\right)\right) \frac{(1-\gamma) y_i e^{-\lambda y_i}}{1-(1-\gamma)e^{-\lambda y_i}} \\ &- \sum_{i=1}^{n} \left(\exp\left(x_i'\beta\right)\right) y_i, \end{aligned} \tag{10}$$

$$\frac{\partial l}{\partial \gamma} = -\sum_{i=1}^{n} \left(\delta_i + \exp(x_i'\beta)\right) \frac{e^{-\lambda y_i}}{1-(1-\gamma)e^{-\lambda y_i}} + \frac{1}{\gamma}\sum_{i=1}^{n} \left(\exp(x_i'\beta)\right), \quad (11)$$

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i x_{ij} - \sum_{i=1}^{n} x_{ij} \exp(x_i'\beta) \ln\left(1-(1-\gamma)e^{-\lambda y_i}\right) + \sum_{i=1}^{n} x_{ij} \exp(x_i'\beta) \ln \gamma e^{-\lambda y_i} \quad (12)$$

for $j = 1, \ldots, p$

The maximum likelihood estimates can be obtained as the simultaneous solutions of the equations $\frac{\partial l}{\partial \lambda} = 0$, $\frac{\partial l}{\partial \gamma} = 0$ and $\frac{\partial l}{\partial \beta_j} = 0$. The solution of these three nonlinear equations must be obtained using a numerical method. The Newton-Raphson algorithm is one of the standard methods used to solve equations. The Newton Raphson iteration is a useful technique for finding zeros of function. It was first introduced by Newton around 1669, and later generalized by Raphson.

4 Accelerated Failure Time models

The accelerated failure time model is an alternative if the proportional hazards assumption does not hold. Different diagnostic tests have been developed to evaluate the proportional hazards assumption [12]. In contrast to the proportional hazard model, the accelerated failure time model is best characterized in terms of survival function. In the first subsection we will introduce accelerated failure time models, and obtain a new accelerated failure time model based on a new baseline hazard function. As in the previous subsection we will estimate parameters in the second subsection.

4.1 Parametric Accelerated Failure Time Model

Although parametric proportional hazard models are very applicable to analysis survival data, in some cases we cannot use this models. The accelerated failure time model is an alternative to the proportional hazard model in these cases.

Under accelerated failure time models we measure the direct effect of explanatory variables of survival time instead of hazard, as we do in proportional hazard models. Similar to the proportional hazard model, the accelerated failure time model describes the relationship between survival probabilities and a set of covariates. Accelerated failure time models are discussed in details in textbook [4],

Accelerated failure time models are fitted using the maximum likelihood method. The unknown parameters are found by maximizing likelihood function with the Newton-Raphson method in some software package, e.g., R and SAS.

4.2 Estimation by Maximum Likelihood

Now we are going to estimate parameters in this model, as in Sect. 2.2. In the previous subsection we showed that event time has EEG distribution, now we can estimate

parameters with maximizing likelihood. If we have uninformative censoring then likelihood is the form:

$$h_i(y) = \exp(X_i'\beta)h_0(\exp(X_i'\beta)y) \tag{13}$$

Then we have:

$$h_i(y) = \exp(X_i'\beta)\frac{\lambda}{1-(1-\gamma)e^{-\lambda\exp(X_i'\beta)y}} \tag{14}$$

It follows that:

$$\begin{aligned} S_i(y) &= \exp\left(-\int_0^t \frac{\lambda\exp(x_i'\beta)}{1-(1-\gamma)e^{-\lambda\exp(x_i'\beta)s}}d_s\right) \\ &= \frac{\gamma e^{-\lambda\exp(x_i'\beta)y}}{1-(1-\gamma)e^{-\lambda\exp(x_i'\beta)y}} \end{aligned}$$

For last equality see appendix B. We can obtain $f_i(y)$ by $h_i(y)$ and $S_i(y)$:

$$\begin{aligned} f_i(y) &= \frac{\lambda\exp(x_i'\beta)}{1-(1-\gamma)e^{-\lambda\exp(x_i'\beta)y}}\frac{\gamma e^{-\lambda\exp(x_i'\beta)y}}{1-(1-\gamma)e^{-\lambda\exp(x_i\prime\beta)y}} \\ &= \lambda\exp(x_i'\beta)\gamma e^{-\lambda\exp(x_i'\beta)}y\{1-(1-\gamma)e^{-\lambda\exp(x_i'\beta)}y\}^{-2} \end{aligned}$$

It means that Y has EEG distribution with $\lambda\exp(X_i'\beta)$ and γ. i.e.

$$Y \sim EEG\,(\lambda\exp x_i\prime\beta, \gamma) \tag{15}$$

4.3 Estimation by Maximum Likelihood

Now we are going to estimate parameters in this model like Sect. 2.2. In previous subsection we obtained that event time has EEG distribution, now we can estimate parameters with maximizing likelihood.

If we have uninformative censoring then likelihood is the form:

$$L = \prod_{i=1}^{n}\left[\frac{\lambda\gamma\exp(x_i'\beta)e^{-\lambda y_i\exp(x_i\prime\beta)}}{\left(1-(1-\gamma)e^{-\lambda y_i\exp(x_i'\beta)}\right)^2}\right]^{\delta_i}\left[\frac{\gamma e^{-\lambda y_i\exp(x_i'\beta)}}{1-(1-\gamma)e^{-\lambda y_i\exp(x_i'\beta)}}\right]^{1-\delta_i} \tag{16}$$

Then:

$$\begin{aligned} l = &\sum_{i=1}^{n}\delta_i\ln\left(\lambda\exp(x_i'\beta)\right) + \sum_{i=1}^{n}\ln\left(\gamma e^{-\lambda y_i\exp(x_i'\beta)}\right) \\ &-\sum_{i=1}^{n}(1+\delta_i)\ln\left(1-(1-\gamma)e^{-\lambda y_i\exp(x_i'\beta)}\right) \end{aligned} \tag{17}$$

The first derivatives of the log_likelihood function with respect to the parameters are:

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{n} \frac{\delta_i}{\lambda} - \sum_{i=1}^{n} y_i \exp\left(x_i'\beta\right) - \sum_{i=1}^{n}(1+\delta_i)\frac{(1-\gamma)y_i \exp(x_i'\beta)}{e^{-\lambda y_i \exp(x_i^{\tau}\cdot\beta)}[1-(1-\gamma)e^{-\lambda y_i \exp(x_i'\beta)}]}, \tag{18}$$

$$\frac{\partial l}{\partial \gamma} = \frac{n}{\gamma} - \sum_{i=1}^{n}(1+\delta_i)\frac{e^{-\lambda y_i \exp(x_i'\beta)}}{1-(1-\gamma)e^{-\lambda y_i \exp(x_i'\beta)}} \tag{19}$$

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i x_{ij} - \sum_{i=1}^{n} \lambda\, y_i x_{ij} \exp(x_i'\beta) - \sum_{i=1}^{n}(1+\delta_i)\frac{(1-\gamma)\,\lambda\, y_i x_{ij} \exp(x_i'\beta)e^{-\lambda y_i \exp(x_i'\beta)}}{1-(1-\gamma)e^{-\lambda y_i \exp(x_i'\beta)}}, \tag{20}$$

for $j = 1, \ldots, p$.

The maximum likelihood estimates can be obtained as the simultaneous solutions of the equations $\frac{\partial l}{\partial \lambda} = 0$, $\frac{\partial l}{\partial \gamma} = 0$ and $\frac{\partial l}{\partial \beta_j} = 0$. The solution of these three nonlinear equations must be attained using a numerical method. Newton-Raphson algorithm is one of the standard methods used to solve equations.

5 Simulation

We know that in linear regression models, the response variable is directly connected with the considered covariates, the regression coefficients and the simulated random errors. Thus, the response variable can be generated from the regression function, once the regression coefficients and the error distribution are specified. However, in the Cox model, which is formulated via the hazard function, the effect of the covariates have to be translated from the hazards to the survival times, because the usual software packages for estimation of Cox models require individual survival time data. A general formula describing the relation between the hazard and the corresponding survival time of the Cox model is derived. In EEG proportional hazard models, semi-parametric proportional hazard model, and generalized linear model (GLM), specifically, the use of gamma distribution with log link function and OLS on were employed in a series of simulation experiments to assess the performance of the proposed method. For each of experiments, 500 simulated data sets were generated; each experiment is different from the others. We repeated experiments for different sample sizes. At first we began by 25 observations, and then we did it for 50, 100, 150, 200 observations. Compared these models by means of AIC and BIC for each time. We also obtained MSE of two covariates in our models, presented in Table.2

All data were generated and all models were estimated using R version 2.15.3 software. The packages that we used are: maxLik package and survival package. Maxlik package was used to estimate the parameters in new models and in computing the value

of log likelihood. Survival package was used to estimate parameters and compute log likelihood for the semi-parametric proportional hazard models. By using these values of log likelihood, we can obtain AIC and BIC and then choose the appropriate model between other models.

Sample R code to implement the procedure used in this paper is provided in the appendix C.

The intention is to demonstrate the proposed procedure using two covariates, one of the covariate is discrete and the other one is continuous. The discrete covariate could capture, for instance, ages of a patients at the time of the first heart transplant, and the continuous covariate could capture, mismatch scores which measure the degree of tissue incompatibility between the initial donor and recipient hearts with respect to antigens.

The Stanford Heart Transplantation program was begun in October 1967, that 184 patients had received heart transplants, in that study researchers surveyed the relation between survival time, age , and mismatch scores between the initial donor, recipient, and age of each patient. As we said, simulation in the Cox model is different from linear regression models. We know that the survival function of the Cox proportional hazards model is given by:

$$S_i(y) = \exp\big[-H_0(y)\exp(x_i'\beta)\big] \tag{21}$$

where

$$H_0(y) = \int_0^y h_0(u)d_u \tag{22}$$

is the cumulative baseline hazard function. Thus, the distribution function of the Cox model is

$$F_i(y) = 1 - \exp\big[-H_0(y)\exp x_i'\beta\big] \tag{23}$$

Let Z be a random variable with distribution function F, then $U = F(Z)$ follows a uniform distribution on the interval 0 to 1 [18]. Let Y be the survival time of the Cox model, then

$$U = 1 - \exp\big[-H_0(Y)\big]\exp(x_i'\beta) \sim uniform[0, 1] \tag{24}$$

then

$$Y_i = H_0^{-1}\big[-log(1-U)\exp(-x_i'\beta)\big] \tag{25}$$

where U is a random variable with $U \sim uniform[0, 1]$. In previous sections we obtained $S_i(y)$, for EEG proportional hazard model, then we can write:

$$Y_i = -\frac{1}{\lambda}\ln\frac{\exp\frac{\ln(1-U)}{\exp\exp(x_i'\beta)}}{\gamma + \frac{\ln(1-U)}{\exp(x_i'\beta)}} \tag{26}$$

Then variables in data set generated by different distributions, X_1 generated by a Poisson distribution that we assumed with an arbitrary and fix parameter, and X_2 generated by a weibull distribution with fix parameters and censored time generated by an exponential distribution with constant rate.

Table 1 Simulation results for EEG PH, OLS on $ln(Y)$, semi-parametric, GLM with log link, GLM with log link linear and their AIC and BIC criteria

# Sims	n	AIC					BIC				
		EEG PH model	OLS ln(y)	Semipara PH	GLM log link	Linear model	EEG PH model	OLS ln(y)	Sempara PH	GLM-log link	Linear model
500	25	87.56	86.61	97.43	92.21	114.43	92.44	90.27	99.87	95.87	118.089
500	50	168.50	173.13	248.56	180.47	228.86	176.15	178.86	252.38	186.21	234.59
500	100	332.32	342.67	606.93	358.22	458.73	342.74	350.48	612.14	366.04	466.55
500	150	493.99	509.84	1012.59	532.39	685.82	506.04	518.88	1018.62	541.43	694.85
500	200	658.72	677.95	1443.63	710.64	912.89	671.91	687.84	1450.23	720.53	922.79

Table 2 Simulation results for EEG accelarated failure, Weibul, Exponential PH, Log normal, Log logistic Models and their AIC and BIC criteria

# Sims	n	AIC					BIC				
		EEG AFT	Wei PH	EXP PH	LNOR PPH	LLOG PPH	EEG AFT	Wei PH	EXP PH	LNORL PH	LLOG PH
500	25	88.13	83.16	85.66	87.42	86.24	93.01	89.18	87.92	91.54	92.13
500	50	169.18	165.30	166.57	171.30	168.94	176.83	174.70	170.83	177.74	177.72
500	100	332.99	327.16	329.20	338.51	332.70	343.41	337.64	334.32	350.51	347.67
500	150	494.76	487.24	492.25	505.90	499.59	514.60	502.63	498.41	519.30	512.79
500	200	659.50	649.77	652.98	674.17	665.08	672.70	662.57	660.81	685.88	679.47

We know that AIC and BIC are measures of the relative quality of a statistical model for a given set of data. They provide a means for model selection. AIC and BIC offer a relative estimate of the information lost when a given model is used to represent the process that actually generates the data.

We wish to select, from amount R candidate model, the model that minimizes the information loss; that is, we select models that have minimum AIC and BIC.

The results of the simulation are presented in Table.1. For small observations, OLS on is better than others, but this model has small differences with EEG proportional hazard models in AIC and BIC. For more than 50 observations, AIC and BIC for EEG proportional hazard model are less than others. It means that for data sets with large numbers of observations, the EEG proportional hazard model is better than OLS, with and without log transformation and also better than the semi-parametric proportional hazard model. Another important thing that we should note is that , if increases the different between AIC and BIC for EEG proportional hazard model and other models increase. We can use the EEG proportional hazard model in large data sets, and we can ensure that we will lose less information. Then we will have better prediction of survival time.

6 Conclusions

The high capacity of performing calculations by uses of computer allows the evaluation of statistical methods via simulation studies. One of the most important statistical models in medical research is the Cox proportional hazard model, which we can compare these models to other statistical models by simulation studies.

In previous section we have developed the general relation between the hazard and the survival time of the EEG proportional hazard model, and then we have generated appropriate survival times for this distribution and then we fitted the Cox proportional hazard model to these data with respect to the covariates.

In most of time the researchers do not pay attention to baseline hazard function, and they use partial likelihood to estimate the effect of covariates. However, there are a lot of practical situations, Where the use of more flexible distributions than the exponential distribution is required. Another point that we should note that is, when we use additional parameters we can calculate hazard rate whit more accuracy.

Base on simulations that we presented in previous section we found that EEG proportional hazard model is better than alternative models on large number of survival times with covariates and lost less information than other models. Therefore EEG proportional hazard model provided more accurate predictions of mean, median, and high cost cases and this model can replace to exponential hazard models and OLS with and without log translation and semi-parametric proportional hazard.

7 Appendix A

Let $h_i(y) = \frac{\lambda}{1-(1-\gamma)e^{-\lambda y}} \cdot exp(X_i^\tau \cdot \beta)$ and we know that $S_i(y) = \exp\left(-\int_0^y h_i(s)d_s\right)$, we can write:

$$\begin{aligned} S_i(y) &= \exp\left(-\int_0^y \frac{\lambda}{1-(1-\gamma)e^{-\lambda s}\exp(X_i'\beta)} d_s\right) \\ &= \exp\left(\exp(X_i'\beta)\int_0^y \frac{-\lambda}{1-(1-\gamma)e^{-\lambda s}} d_s\right) \end{aligned}$$

let:

$$\begin{aligned} I &= \int_0^y \frac{-\lambda}{1-(1-\gamma)e^{-\lambda s}} d_s \\ &= \int_0^{-\lambda y} \frac{1}{1-(1-\gamma)e^{u}} d_u \\ &= \int_{(1-\gamma)}^{(1-\gamma)e^{-\lambda y}} \frac{d_z}{z(1-z)} \\ &= \int_{(1-\gamma)}^{(1-\gamma)e^{-\lambda y}} \frac{d_z}{z} + \int_{(1-\gamma)}^{(1-\gamma)e^{-\lambda y}} \frac{d_z}{1-z} \\ &= \ln(1-\gamma)e^{-\lambda y} - \ln(1-\gamma) - \ln(1-(1-\gamma)e^{-\lambda y}) + \ln(\gamma) \end{aligned}$$

$$= \ln \frac{(1-\gamma)e^{-\lambda y}}{1-(1-\gamma)e^{-\lambda y}} + \ln \frac{\gamma}{1-\gamma}$$
$$= \ln \frac{\gamma e^{-\lambda y}}{1-(1-\gamma)e^{-\lambda y}}$$

then:

$$S_i(y) = \left(\frac{\gamma e^{-\lambda y}}{1-(1-\gamma)e^{-\lambda y}}\right)^{\exp(X_i^\tau \cdot \beta)} \tag{27}$$

8 Appendix B

Let $h_i(y) = \frac{\lambda \exp(x_i^\tau \beta)}{1-(1-\gamma)e^{-\lambda \exp(x_i^\tau \beta) y}}$ and we know that $S_i(y) = \exp\left(-\int_0^y h_i(s)d_s\right)$, we can write:

$$S_i(y) = \exp\left(-\int_0^y \frac{\lambda \exp(x_i^\tau \beta)}{1-(1-\gamma)e^{-\lambda \exp(x_i^\tau \beta) s}} d_s\right) \tag{28}$$
$$= \exp\left(\exp(x_i^\tau \beta) \int_0^y \frac{-\lambda}{1-(1-\gamma)e^{-\lambda \exp(x_i^\tau \beta) s}} d_s\right) \tag{29}$$

let $c = \exp(x_i^\tau \beta)$ and let:

$$I = \int_0^y \frac{-\lambda c}{1-(1-\gamma)e^{-\lambda cs}} d_s$$
$$= \int_0^{-\lambda cy} \frac{1}{1-(1-\gamma)e^{u}} d_u$$
$$= \int_{(1-\gamma)}^{(1-\gamma)e^{-\lambda cy}} \frac{1}{z(1-z)} d_z$$
$$= \ln \frac{\gamma e^{-\lambda cy}}{1-(1-\gamma)e^{-\lambda cy}}$$

Then:

$$S_i(y) = \frac{\gamma e^{-\lambda \exp(x_i^\tau \beta) y}}{1-(1-\gamma)e^{-\lambda \exp(x_i^\tau \beta) y}} \tag{30}$$

9 Appendix C

R commands presented in this section:

```
#load packages
library(maxLik)
library(survival)
# loglikelihood for EEG proportional hazard model
loglik<-function(eta,Y,v,x1,x2){
```

```
if((eta[3]>0)&(eta[4]>0)){
  k=c()
  for(i in 1:length(Y)){
k[i]=v[i]*log(eta[3]*exp(eta[1]*x1+eta[2]*x2))
-((v[i]+exp(eta[1]*x1+eta[2]*x2))*(log(1-(1-eta[4])
*exp(-eta[3]*Y[i]))))


+((exp(eta[1]*x1+eta[2]*x2))*(log(eta[4]
*exp(-eta[3]*Y[i]))))


}
 l=sum(k)}
else{l=-Inf}
l}


simureg<-function(eta){
     x1<-c(rpois(200,lambda=42))
     x2<-c(rweibull(200,shape=2.075,scale=1.27))
     u<-c(runif(200))
     y=(1/-0.05)*log((exp((log(1-u))
/(exp(0.03*x1+0.167*x2))))/(0.5+0.5*exp((log(1-
     u))/(exp(0.03*x1+0.167*x2)))))
      ycen<- c(rexp(200,rate=0.0742))
     T<-c(pmin(y,ycen))
     v<-c(as.numeric(y<=ycen))
  dataset<-data.frame(Y,v,x1,x2)
  tem1<-maxLik(loglik,start=m,x2=x2,v=v,x1=x1,Y=Y)
     a<-tem1$maximum
     tem10<- 8 - 2*a
  tem2<-glm( Y ~x1 + x2 , data = dataset)
     tem20<-tem2$aic
  tem3<-glm( Y ~x1 + x2 , data = dataset,
family = Gamma(link = ''log''))

     tem30<-tem3$aic
   tem4<-glm( log(Y) ~x1 + x2 , data = dataset)
     tem40<-tem4$aic
   tem5 <- coxph(Surv(Y, v)~ x1+ x2, data=dataset)
     b<-tem5$loglik[2]
     tem50<- 2-2*b
    c(tem1=tem10,tem2=20,tem3=tem30,tem4=tem40,
tem5=tem50)
  }
```

```
  x=matrix(nrow=500,ncol=5)
    for(i in 1:500){
    x[i,]=simureg(m)
print(i)
}
```

References

1. Adamidis K, Dimitrakopoulou T (2005) On an extension of exponential geometric distribution. Statistics and probability letters 73:259–269
2. Blough DK, Madden CW, Hornbrook MC (1999) Modeling risk using generalized linear models. J Health Econ 18:153–171
3. Buckley IV, James I (1979) Linear regression with censored data. Biometrika 66(3):429–436
4. Collett D, Collett D (1994) Modelling survival data in medical research. Chapman and Hall, London
5. Cox DR (1972) Regression models and life tables. J R Stat Soc 187–220 Series B 34
6. Cox DR, Oakes D (1984) Analysis of survival data. Chapman and Hall, London
7. Dachateau L, Janssen P (2008) The frailty model. Springer, New York
8. Dudley RA, Smith LR, Harrell FE Jr (1993) Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. J Clin Epidemiol 46(3):261–271
9. Etzioni RD, Feuer EJ, Sullivan SD (1999) On the use of survival analysis techniques to estimate medical care costs. J Health Econ 18:365–380
10. Hallstrom A, Sullivan SD (1998) On estimating costs for economic evaluation in failure time studies. Med care 36(3):433–436
11. Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiely, New York
12. Klein JP, Moeschberger ML (1997) Survival analysis techniques for censored and truncated Data. Springer, New York
13. Kleinbaum David G, Klein Mithchel (2005) Survival analysis a self learning text, 2nd edn. Springer, New York
14. Lawless JF (1982) Statistical models and methods for lifetime data analysis. Wiley, New York
15. Lipscomb J, Anuckiewicz M, Parmigiani G (1998) Predicting the cost of illness: a comparison of alternative models applied to storke. Med care 18(2):S39–S56
16. Manning WG, Mullahy J (2001) Estimating log models: to transform or not transform? J Health Econ 20(4):461–494
17. McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
18. Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics. Mc Graw Hill, Tokyo

Sadegh Rezaei received his B.Sc. degree in statistics and computer science from Shahid Chamran University, Ahvaz, Iran, in 1981, his M.Sc. degree in mathematical statistics from Tarbiyat Modaress University, Tehran, Iran, in 1986, and his Ph.D. degree in statistics from Adelaide University, Adelaide, Australia, in 1996. From 1996 to 2004 he was with the department of statistics in Shahid Chamran University of Ahvaz. He is currently an associate professor in the department of statistics in the faculty of mathematics & computer science of Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. His main research interests are statistical modeling, lifetime distributions, and statistical speech processing. He has supervised different students in the master's and Ph.D. programs, conducted research in statistical speech processes and statistical modeling, and has applied these models to real data sets such as earthquakes and voice activity detection (VAD) data.

Sina Hashami Allameh Tabatabaii University, 2005-2011. B.S., Statistics. Amirkabir University of thechnology (Tehran Polytechnic), 2011-2013. M.S., Statistics and Math. Thesis: Some regression models for life time random variable. The field of interests are: Mathematical Statistics, Reliability and Survival Analysis, Engineering and Public Health.

Lotfollah Najjar is an Associate Professor in the Department of Information Systems and Quantitative Analysis in the College of Information Science and Technology at the University of Nebraska at Omaha. He holds a Ph.D. in Industrial and Management Systems Engineering with supporting areas in MIS, and Operations Management from university of Nebraska-Lincoln. His research interests are in the areas of Quality Information Systems (Data Quality), Data Mining, Data Analytics, Big Data, Business Process Reengineering & IT, Software Quality and Reliability, System Quality, and Total Quality Management (TQM) & IT .Najjar's teaching interests are in Quality Information Systems, Data Analytics Business Process Reengineering & IT, Introduction to Management Information System, Quality Control, Production and Operations Management, Statistics, and Mathematics. He has been with UNO since 1989