COX PROPORTIONAL HAZARD MODEL WITH TIME-DEPENDENT COVARIATES: A REVIEW

Rinku Saikia[1] and Manash Pratim Barman[2]
[1]Research Scholar, Department of Statistics, Dibrugarh university.
[2]Assistant Professor, Department of Statistics, Dibrugarh University.

Abstract

Cox Proportional Hazard Model is a popular semi- parametric model for determining the relationship between survival time with a set of covariates. The main feature of the model is its proportional hazard assumption i.e., the covariates are time-independent. In this paper an attempt has been made to study an extension of the Cox Proportional Hazard model which is called Extended Cox Model. A detailed description of the model is incorporated in this paper together with its estimation procedure. To study the practical utility of the model a set of data is generated by using simulation technique considering a particular situation.

Keywords: Survival Analysis, Semi-Parametric, Covariates, Time-Dependent, etc.

1. Introduction:

In statistical literature, it is observed that a good number of models have been developed for analyzing survival data or life time data. The most commonly used model in this context is the Cox Proportional Hazard model (PH) [Cox, 1972]. It is basically

used in medical and bio- statistical fields to study the effect of different possible factors (covariates) on the time to happening of a particular event. The Cox Regression model says that the hazard at a time (t) can be expressed as the product of two quantities. First is the baseline hazard function and the second quantity is the exponent of the linear sum of multiplication of regression coefficients with explanatory variables $X_i (i = 1, 2, \ldots, p)$. Proportional hazard assumption, the main feature of the model is that the baseline hazard is a function of time (t), but the exponent portion which contains the explanatory variable that does not involve time. Thus the explanatory variables use in the Cox PH model are time-independent. However, modification of the Cox regression model has been done so that it can accepts explanatory variable that involve time. This type of explanatory variables is called time-dependent as they may change their values over time. But the Cox Regression model involving time-dependent explanatory variables no longer satisfy the proportional hazard assumption and it is called extended Cox model. The extended Cox model can incorporate both time-independent and time-dependent explanatory variables. Extended Cox Model is basically used to study the proportional hazard assumptions for the time independent explanatory variables and also to assess the effect of variables not satisfying proportional hazard assumption on time to happing of an event. When time-dependent variables are used to assess the proportional hazard assumption for a time-independent variable, the Extended Cox model contain product (i.e., interaction terms) involving time-independent variable being and some function of time. The parameters of the extended Cox model is estimated by using partial maximum likelihood estimation procedure. In this paper, an attempt has been made to provide a extensive review about Extended Cox Regression model. Its practical utility is also studied by using generated data by suing simulation in package R.

Review of the Earlier Work

With time, there have been occurring a lot of developments on Cox PH model. In this section, it is tried to have a bird's eye view on a lion's share of such developments.

Kaplan and Meier (1958) in his study described about the incomplete observations. In this paper it was assumed that the life time (age at death) was

independent of the potential loss time, in practice this assumption described careful scrutiny. Despite the incompleteness of the data, it was described to estimate the proportion $P(t)$ of items in the population whose lifetimes would exceeds t, without making any assumption about the form of the function $P(t)$. Cox (1972) considered the analysis of censored failure times. In this work he showed that the hazard function was taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. It also considered the conditional likelihood inferences about the unknown regression coefficients.

Cox (1975) gave the definition of partial likelihood with generalizing the ideas of conditional and marginal likelihood. He showed the application of life tables and inference in stochastic processes. The paper had also been discussed the usual large sample properties of maximum likelihood estimates and the apply when partial likelihood was used. Gill (1984) described an informal discussion about the martingale techniques which can be used to extend Cox's regression model and to derive its large sample properties. Pettitt and Daud (1990) considered the investigation of time-dependence in Cox's proportional hazards model using the residuals to consider time-dependent effects. For data analysis and model development the authors considered smoothing these residuals to consider time-dependent effects. In this work the authors included a model which considered time-dependent modulation of the linear predictor was introduced and suggested for use. Lin (1991) showed a class of estimation functions for the vector of regression parameters in the Cox proportional hazard model with possibly time- dependent covariates by incorporating the weight function which was commonly used in weighted log rank tests into the partial likelihood score function. In this paper he also discussed that when the Cox model was inappropriate however the estimators with different weight function generally converge to non identical constant vectors. Lin et al. (1993) discussed a new class of graphical and numerical methods for checking the adequacy of the Cox regression model. The procedures were derived from cumulative sums of martingale based residuals over follow up time and covariate values. Fisher and Lin (1999) discussed the use of time dependent covariates, which offer additional opportunities but must be used with caution. The author represented the functional form

of the Extended Cox model which was much more complex then the Cox model with fixed (time-dependent) covariates. Clarkson and Jennrich (2000) discussed about the computation of extended maximum likelihood in the Cox proportional hazards model. In their works authors discussed about computation of extended maximum likelihood in the situation of infinite parameter estimates were occurred when linear combination of the covariates monotonically increases or decreases with the failure times.

Ata and Sozer (2007) in his work discussed the stratified Cox regression model and extended Cox regression model, which used time-dependent covariate terms with fixed function of time. The results of their work were illustrated by an analysis of lung cancer data in order to compare these methods with respect to Cox regression model in the presence of non proportional hazards.

Xin (2011) in his work investigated ties and time - varying covariates in survival analysis. According to the author there were two types of ties: ties between event times (type 1 ties) and ties between event times and the time that discrete time varying covariates change or "jump"(Type 2 ties). In the study author discussed the effect of the Type 2 ties on Cox's partial likelihood, the current default method to treat Type 2 ties in statistical packages like SAS and R and proposes alternative methods (Random and Equally Weighted) for Type 2 ties. The author also discussed the effect of the percentage of Type 1 and Type 2 ties on the Random and Equally Weighted methods for handling both types of ties. Austin (2012) described the data generating processes for the Cox proportional hazards model with time- varying covariates when event times followed an exponential, Weibull, or Gompertz distribution. In his work he considered three types of time varying covariates: first, a dichotomous time varying covariate that can change at most once from untreated to treated; second, a continuous time varying covariate such as cumulative exposure at a constant dose to radiation or to pharmaceutical agent used for a chronic condition; third, a dichotomous time varying covariate with a subject being able to move repeatedly between treatment states. He illustrated the utility of closed form expression for simulating event times by using Monte Carlo simulations to estimate the statistical power to detect as statistically significant the effect of different types of binary time- varying covariates.

Models and Methodology :

The Cox PH model has achieved wide spread use in the analysis of time to event data with censoring and covariate.

Statistically, the Cox proportional hazard regression model for time- dependent covariate is

$$\lambda\left(t/\bar{Z}\right)=\lambda_0\left(t\right)\exp\left(\beta' Z(t)\right)$$

Where β' is a set of unknown regression parameters and $\lambda_0\left(t\right)$ is an unspecified baseline hazard function and $\bar{Z}$ is the history of the vector of the time-dependent covariates up to time *t*.

As an example of a time- dependent covariate, suppose that researcher wants to consider the effect of pesticides among tea garden workers over time on mortality. A sample of workers in tea garden where pesticides were very much used monitored for a period of time and data were collected on survival and tea garden workers exposure. For the ith individual in the sample, the data could be summarized as

$$\left(X_t,\Delta_i,\bar{Z}_t\left(X_i\right)\right)$$

Where

X_i = Min (T_i, C_i) is the observed survival time or censoring time.

$\Delta_i = I\left(T_i \le C_i\right)$ is the failure indicator.

$\bar{Z}_l(X_i)$ is the history of tea garden workers exposure up to time X_i. This may, for example, be daily exposure collected on that individual every six months. This data may be collected up to the point that patient dies or until he/ she is censored, or until he/ she stops working at the tea garden.

Let us consider the following cox proportional hazards model with time-dependent covariates for the above situation

$$\lambda\left(t/\bar{Z}_t(t)=\lambda_0(t)\exp\left(\beta' y\left(\bar{Z}_l(t)\right)\right)\right.$$

Now the question is what should be the form of the function $y\left(\bar{Z}_l(t)\right)$. There are number of alternatives

* One may use cumulative exposure that is

$$y\left(\bar{Z}_l(t)\right)=\sum_j Z_i(u_{ij})(u_{ij}-u_{i(j-1)})$$

Where u_{ij} are days at which measurements were made prior to days t.

* One may use average exposure up to time t,

$$y\left(\bar{Z}_l(t)\right)=\frac{\sum_{u_{ij<1}} Z_t(U_{ij})}{measurments\ up\ to\ t}$$

* One may use maximum exposure up to time t,

$$y\left(\bar{Z}_l(t)\right)=\max\left\{Z_i\left(U_{ij}\right):U_{ij}<1\right\}$$

One may also want to consider models such as

$$y\left(t/\bar{Z}_t(t)\right)=\lambda_0(t)\exp\left(\beta_1 y_1\left(\overline{Z_t(t)}+\beta_2 y_2\left(\bar{Z}_l(t)\right)\right)\right)$$

Where $y_1\left(\bar{Z}_l(t)\right)$ = *cumulative exposure up to time t and*

$y_2\left(\bar{Z}_l(t)\right)$ = *maximum exposure up to time t*

This model may be used if one think that both of those components of the tea garden workers history may have an effect on survival. It also allow to test whether these different components of history are important on survival by testing whether the parameters β_1 or β_2 are significantly different from zero. For estimation of the model, the two methods are most frequently use. They are Partial likelihood Score function and Weighted Score function. In this paper an explanation of the partial likelihood score function is presented.

Partial Likelihood Score Function:

To obtain the estimates of the covariate parameters Cox (1972, 1975) developed a non-parametric method be called, Partial likelihood. Estimation of the parameter values is then obtained by use of maximum partial likelihood estimation. Partial likelihood is used because the likelihood formula considers probability only for those subjects who fail and does not explicitly consider probability for success.

Data: (X_i, Δ_i, z_i), $i = 1, 2, \ldots, n$ where for the ith individual

$$X_i = \min(T_i, C_i)$$

$$\Delta_i = I(T_i \le C_i)$$

$z_i = \left(z_{i1,} \ldots\ldots, z_{ip}\right)'$ is a vector of covariates.

Model: The proportional hazards model is

$$\lambda\left(t/z_i\right) = \lambda_0(t)\exp\left(z_i\beta'\right)$$

Where

$$\lambda\left(t/z_i\right) = \frac{\lim_{h\to 0} p\left[t \le T \le t+h/T_i \ge t, z_i\right]}{h}$$

Assume that C_i and T_i are conditionally independent given z_i then the cause specific hazard can be used to represent the hazard of interest. That is (in terms of conditional probabilities)

$$P\left[x \le X_i \le x+\Delta x, \Delta_i = 1/X_i \ge x, z_i\right] = P\left[x \le T_i \le x+\Delta x/T_i \ge x, z_i\right] \approx \lambda_{T_i}\left(x/z_i\right)\Delta x$$

To define some notation let us break the time axis (patient time) into a grid of points. Assume the survival time is continuous. The grid points dense enough so that at most one death can occur within any interval.

Let $dN_i(u)$ denote the indicator for the ith individual being observed to die in $[u, u + \Delta u)$.

Namely,

$$dN_i(u) = I\left(X_i \in [u, u+\Delta u), \Delta_i = 1\right)$$

Let $Y_i(u)$ denote the indicator for whether or not the ith individual is at risk at time u.

Namely,

$$Y_i(u) = I\left(X_i \geq u\right)$$

Let $dN(u) = \sum_{i=1}^{n} dN_i(u)$ denote the number of deaths for the whole sample occurring in [u, $u + \Delta u$). Since we are assuming Δu is sufficiently small, so $dN(u)$ is either 1 or 0 at any time u.

Let $Y(u) = \sum_{i=1}^{n} Y_i(u)$ be the total number from the entire sample who are at risk at time u.

Let $F(x)$ denote the information up to time x (one of the grid points)

$F(x) = \{(dN_i(u), Y_i(u), z_i), i = 1, \ldots . n;$ *for grid points* $u < x$ and $dN(x)\}$

The individual who was observed to die among those at risk at time x *if* $dN(x) = 1$.

Let $I(x)$ denote individual in the sample who died at time x if someone died. If no one dies at time x, *then* $I(x) = 0$. For example, if $I(x) = j$, then this means that j^{th} individual in the sample with covariate vector z_j *died in* $[x, x + \Delta x)$.

Let $F(\infty)$ denote all the data in the sample. Namely

$$F(\infty) = \left\{\left(X_i, \Delta_i, z_i\right), i = 1, 2, \ldots, n\right\}$$

If we let $u_1 < u_2 < \ldots$, denote the value of grid points along the time axis, then the data (with redundancy) can be expressed as

$$\left(F(u_1), I(u_1), F(u_2), I(u_2), \ldots, F(\infty)\right)$$

Denote the observed value of the above random variables by lower cases. Then the likelihood of the parameter $\lambda_0(t)$ and β can be written as

$$\begin{aligned} & P[F(u_1) = f(u_1);\ \lambda_0(.),\ \beta] \times P[I(u_1) = i(u_1)/F(u_1) = f(u_1);\ \lambda_0(.),\ \beta] \\ & \quad \times P[F(u_2) = f(u_2)/F(u_1) = f(u_1),\ I(u_1) \\ & \quad = i(u_1)\ ;\ \lambda_0(.),\ \beta] \\ & \quad \times P[I(u_2) = i(u_2)/F(u_1) = f(u_1),\ I(u_1) = i(u_1),\ F(u_2) = f(u_2);\ \lambda_0(.),\ \beta] \\ & \quad \times \ldots \end{aligned}$$

And the last term can be simplified as

$$\begin{aligned} & P[I(u_2) = i(u_2)/F(u_1) = f(u_1),\ I(u_1) = i(u_1),\ F(u_2) = f(u_2);\ \lambda_0(.),\ \beta] \\ & \quad = P[I(u_2) = i(u_2)/F(u_2) = f(u_2);\ \lambda_0(.),\ \beta] \end{aligned}$$

This is, the full likelihood can be written as the product of a series of conditional likelihoods.

The partial likelihood (as defined by D.R. Cox) consists of the product of every other conditional probabilities in the above presentation . That is

$$PL = \prod_{\{all\ grid\ point\ u\}} P\left[I(u) = i(u)/F(u) = f(u); \lambda_0(.), \beta\right]$$

There are two cases in calculating the partial likelihood

Case I. Suppose conditional on $F(u)$*if* $dN(u) = 0$. That is, no death is observed at time u. in such case, $I(u) = 0$ *with probability* 1.

Hence for any grid point u where, $dN(u) = 0$, we have

$$P[I(u) = 0/F(u) = f(u)] = 1$$

Therefore, the partial likelihood is not affected at any point u such that $dN(u) = 0$

Case II $dN(u) = 1$. Conditional on $F(u)$, if one individual dies at time u, then it must be one of the individuals still at risk (alive and not censored) at time u; i.e.; among the following individuals

$$\{i: Y_i(u) = 1\}$$

Also conditional on $F(u)$, we know the covariate vector z_i associated to each individual *i such that* $Y_i(u) = 1$.

Let A_i = the event that subject I is going to die in $[u, u + \Delta u)$ given that he/ she is still alive at u. if a patient is not at risk at u (i.e.; $Y_i(u) = 0$, then $A_i = \phi$. since let us choose Δu to be small that there is at most one death in $[u, u + \Delta u)$ so one may know $A_1, A_2, \ldots A_n$ mutually exclusive.

Because of the independence of survival times and censoring times, those $Y(u)$ patients who are at risk at u (not censored and still alive at u) make up a random sample of the subpopulation consisting of the patients who will survive up to u (and with the same covariate value). Under independent censoring assumption, the cause specific hazard is same as the hazard of interest; i.e.;

$$\lambda\left(u, \delta_i = 1/z_i\right) = \lambda\left(u/z_i\right)$$

Since Δu is chosen to be very small, so

$$P\left[A_i\right] \approx Y_i(u)\lambda\left(u, \delta_i = 1/z_i\right)\Delta u$$

$$= Y_i\left(u\right)\lambda_0\left(u\right)\exp\left(z_i'\beta\right)\Delta u\,,$$

Where the last equation is due to the assumption of the Cox model. Therefore

$$P\left[I(u)=i(u)/F(u)=f(u);\lambda_0(.),\beta\right]$$

$$\approx \frac{\lambda_0(u)\exp\left(z'_{l(u)}\beta\right)\Delta u,}{\sum_{i=1}^{n} Y_i(u)\lambda_0(u)\exp(z'_l\beta)\Delta u}$$

$$= \frac{\exp\left(z'_{l(u)}\beta\right)}{\sum_{i=1}^{n}\exp\left(z_l\beta\right)' Y_{l(u)}}$$

Hence $Y_{i(u)}(u) = 1$ since this patient had to be at risk at u (since this patient died in $[u, u + \Delta u)$).

Combining these cases, the partial likelihood can be written as

$$PL(\beta) = \prod_{\{all\ grid\ point\ u\}}\left[\frac{\exp\left(z'_{l(u)}\beta\right)}{\sum_{i=1}^{n}\exp\left(z_l\beta\right)' Y_l(u)}\right]^{dN(u)}$$

Other equivalent ways of writing the partial likelihood include. Let $t_1, \ldots, t_d$ define the distinct death, then

$$PL(\beta) = \prod_{j=1}^{d}\left[\left[\frac{\exp\left(z'_{l(u)}\beta\right)}{\sum_{i=1}^{n}\exp\left(z'_l\beta\right)Y_l(t_j)}\right]\right];$$

$$PL(\beta) = \prod_{i=1}^{n}\prod_{\{all\ grid\ pt\ u\}}\left[\frac{\exp(z'_l\beta)}{\sum_{i=1}^{n}\exp(z_l\beta)'Y_l(u)}\right]^{dN_i(u)}$$

$$PL(\beta) = \prod_{i=1}^{n}\left[\frac{\exp\left(z'_i\beta\right)}{\sum_{l=1}^{n}\exp\left(z_l\beta\right)' Y_l(x_l)}\right]^{\delta_i}$$

The importance of using the partial likelihood is that this function depends only on β, the parameter of interest, and is free of the baseline hazard $\lambda_0(t)$, which is infinite dimensional nuisance function. Cox suggested treating partial likelihood function and inference on β accordingly. For example, maximize the partial likelihood to get the estimate of β, often called maximum likelihood estimate and use the minus of the second derivative of the log partial likelihood as the information matrix etc.

A Practical Situation :

Here an attempt has been made to generate survival time of happening of an adverse event of individuals expose to two sets of drugs prescribed by two doctors over a period of one year. To create such a situation the researcher generates survival time of 1000 individuals with their daily exposure history for a period of one year together with a binary time-independent explanatory variable sex. The simulation process for generating the survival times is conducted in *R* by using permutational algorithm introduced by Abrahamowicz et. al. (1996). The programme for generating the survival data are presented in Appendix I. Cox proportional hazard model for time-dependent covariates is fitted for the data generated by using *R*. For fitting the model the *coxph* function and the survival library of *R* is used. The hazard ratios estimated by fitting the model for the time-dependent as well as time-independent covariates are presented in the table.

Table I : Estimation of Cox proportional Hazard Model with time-dependent covariates

Covariates	Coefficient	Hazard Ratio (HR)	SE (HR)	Z	P-value
Sex	0.6340	1.8852	0.0747	8.49	<0.001
Prescription 1	0.0224	1.026	0.0182	1.23	0.22
Prescription 2	0.0126	1.0127	0.0181	0.70	0.48

From table I, it has been observed that the hazard ratio (HR) for sex is 1.8852 which is statistically significant (p-value<0.001). It can be concluded that sex has significant influence on the time of adverse effect. The hazard ratio for time dependent covariate prescription 1 is 1.1.026 which is not significant (p-value: 0.22), thus the exposure status through drugs of prescription1 has no significant influence on the time of happening of an adverse effect. A similar type of conclusion can be drawn in case of the time dependent covariate prescription 2 as of prescription1.

To test the goodness of fit of the overall model, the likelihood ratio test is conducted. The results is,

Likelihood ratio test = 69.4 on 3 df, p-value < 0.001

From the results of likelihood ratio test it can be concluded that the data generated to study the effect of sex and exposure to two types of drugs on the time of happening of an adverse effect fits well with Cox Proportional Hazard model having time-dependent covariates.

Conclusion:

Cox Proportional Hazard model is the most popular and widely used regression model for survival data. It modeled the hazard of happening of an event rather than the actual survival time. The prime property of this regression model is that the covariates of this model follow a proportional hazard assumption i.e., their value remains same with course of time. In this is paper an extensive review on extension of Cox model is conducted which can accommodate time-varying covariates together with its estimation procedure. To check it practical utility, the model is fitted with generated data considering a particular situation and the model is found to be properly fit.

References:

1. Abrahamowicz M., MacKenzie T., Esdaile J.M. (1996) : "Time-dependent hazard ratio: modelling and hypothesis testing with application in lupus nephritis". JASA 91:1432-9
2. Ata Nihal and Sozer , M.Tekin (2007) : "Cox Regression Models With Non Proportional Hazards Applied To Lung Cancer Survival Data", Hacettepe Journal of Mathematics and Statistics, Volume 36(2) , pp. 157-167.
3. Austin Peter C. (2012) : "Generating Survival Times To Simulate Cox Proportional Hazards Models With Time- Varying covariates" , Statistics in Medicine.
4. Clarkson, Douglas B. and Jennrich Robert I (2000) : "Computing Extended Maximum Likelihood Estimates for Cox Proportional Hazards Model" Institute of Mathematical Statistics pp. 205-217.
5. Cox, D.R.(1972) : "Regression Models and life tables", Journal of the Royal Statistical Society, Series B,(Methodological), Vol 34,No.2, pp.187-220.
6. Cox, D.R. (1975): "Partial Likelihood", Biometrika, Vol,62, No.2, pp. 269-176.
7. Fisher, Lloyd D. and Lin ,D. Y. (1999) : "Time Dependent Covariates in the Cox Proportional Hazards Regression Model", Annual Review Public Health 1999, 20:145-57 .
8. Gill,D. Richard (1984) : " Understanding Cox's Regression Model: A Martingale Approach", Journal of American Statistical Association, Volume 79, Issue 386, pp 441-447.
9. Kaplan, E.L.and Meier Paul (1958) : "Non Parametric Estimation from Incomplete Observations", Journal of the American Statistical Association, Vol 53, No. 282, pp. 457- 481.
10. Lin, D. Y. (1991) : "Goodeness of Fit Analysis for the Cox Regression Model Based on a Class of Parameter Estimators" , Journals of the American Statistical Association, Vol. 86, No. 415. Pp. 725-728.

11. Lin, D. Y., Wei L.J. and Ying Z.(1993): "Checking the Cox Model with Cumulative Sums of Martingale - Based Residuals", Biometrika , 80, 3, pp. 557-72.

12. Pettitt ,A.N. and Daud I. Bin (1990): " Investigating Time Dependence in Cox's Proportional Hazards Model", Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 39, No.3, pp. 313-329.

13. Xin Xin (2011) : "A Study Of Ties And Time-Varying Covariates In Cox Proportional Hazards Model", An M. Phil Dissertation. A Thesis presented to the Faculty of Graduate studies of the University of Guelph in the Partial fulfillment of requirements for the degree of Master of Science, August 2011.