Dhaka Univ. J. Sci. 63(1): 25-30, 2015 (January)

Fitting Time to First Birth Using Extended Cox Regression Model in Presence of Non-proportional Hazard

Md. Arif Rahman* **and Md. Rashedul Hoque**

Institute of Statistical Research and Training, Dhaka University, Dhaka-1000, Bangladesh

(Received : 24 March 2014; Accepted : 3 June 2014)

Abstract

The Cox regression model, which is widely used for the analysis of factor effects with censored survival data, makes the assumption of constant hazard ratio. Different methods should be used to deal with non-proportionality of hazards when this assumption is violated. In this study, we use the Extended Cox regression model where time dependent covariate terms with fixed functions of time are considered. Time to first birth for the ever married women after marriage, taken from BDHS 2011 women data is fitted using Extended Cox regression model due to the failure of existence of proportionality assumption. This model performs well as expected compared to Cox regression model.

Keywords: Non proportional hazard, Time dependent covariate, Heaviside function

I. Introduction

The Cox regression model relies on the proportional hazards assumption, implying that the factors investigated has a constant impact on the hazard over time. We emphasize the importance of this assumption and the misleading conclusions that can be inferred if it is violated and this is particularly essential in the presence of long follow-up periods.

The Cox regression model allows one to describe the survival time as a function of multiple factors related to the events (Cox, 1972). This model relies on a fundamental assumption, the proportionality of the hazard ratio, implying that the covariates which we need to investigate has a constant impact on the hazard ratio over the time. If time-dependent variables are included without appropriate modeling, the proportional hazard assumption is violated.

As a result, misleading effects of estimate can be derived. Checking the proportionality of the hazards should thus be an integral part of a survival analysis by a Cox regression model. Even though the Cox regression model has been widely used (more than 25000 citations since the publication of the original paper by Cox) recent publications (Ata, 2007; Bellera, 2010) suggest a growing interest in the quality of its applications. To test this non-proportional covariate we use the residuals measures like Schoenfeld residuals, whether the individual covariates pass the proportional hazard assumption and whether the model as a whole (global test) passes the assumption. Non-proportional hazards can arise if some covariate only affects survival up until sometime *t* or if the size of its effect changes over time. For this time varying covariates, Extended Cox regression model is used instead of the usual one. We illustrate our discussion with a study on time to first birth for ever-married women extracted from women data, BDHS 2011.

II. Mothodology

Sample Design

The survey of the 2011 BDHS is based on a two-stage stratified sample of households. In the first stage, 600 define only EAs were selected with probability proportional to the EA size listed by the Bangladesh Bureau of Statistics (BBS), with 207 clusters in urban areas and 393 in rural areas. In the second stage of sampling, a systematic sample of 30 households on average was selected per EA to provide statistically reliable estimates of key demographic and health variables for the country as a whole, for urban and rural areas separately, and for each of the seven divisions. With this design, the survey selected 18,000 residential households, which were expected to result in completed interviews with about 18,000 ever-married women.

Statistical Model

Cox regression model is used when the proportionality assumption regarding hazard holds. When non-proportional hazard occurs then we move our choice to Extended Cox regression model. Both of these are discussed below.

Cox Regression Model

The Cox regression model is the most common approach to model covariate effects on survival is the proposed by Cox, (1972) that

$$(t \mid x) = {}_0(t)\exp({}_1x_1 + {}_2x_2 + \circ + {}_px_p) \quad (1)$$

The baseline hazard function ${}_0(t)$ in the model can take any shape as a function of *t*. The only requirement is that ${}_0(t) > 0$, which is the non-parametric part of the model and the model is referred to as a semi-parametric model. Coefficient vectors of the covariates are estimated by maximizing a partial likelihood function.

The model parameter are interpreted by the hazard ratio (HR). The hazard ratio for two subjects with a fixed covariate vectors x_i and x_j,

$$HR = \frac{(t \mid x_i)}{(t \mid x_j)} = \frac{{}_0(t)\exp(x_i)}{{}_0(t)\exp(x_j)} = \exp(x_i \quad x_j) \quad (2)$$

which is constant over time, so the model is known as the *proportional hazards model*. Then the logarithm of hazard ratio can also be expressed as

$$\log \frac{(t \mid x_i)}{(t \mid x_j)} = (x_i \quad x_j) \quad (3)$$

*Author for Correspondence. *e-mail*: arahman6@isrt.ac.bd

From the Equation (2) we can said that $\exp(\ _k)$ is the hazard ratio associated with one unit increase in x_k. It can also be said that $\exp(\ _k)\ 1$ are the percentage change in hazard with one unit increase in x_k while adjusting for other covariates.

Regression models for survival data have traditionally been based on the Cox regression model, which assumes that the underlying hazard function for any two levels of some covariates is proportional over the time. If hazard ratios vary with time, then the assumption of proportional hazards may not be justified and we need to use methods that do not assume proportionality to investigate the effects of covariates on survival time. In this case significance of the estimated parameter of the Cox regression model does not mean that the model is well fitted and satisfies the proportional hazard assumption and vice versa. For non-proportional hazards Extended Cox regression model is used to handle the time dependent covariates.

Extended Cox Regression Model with Time Dependent Covariate

In the Cox regression model, there can be variables which involve time *t*. Such variables are called time-dependent variables. If there are time-dependent variables in the model, the Cox regression model can be used but can no longer satisfy the proportional hazards assumption. Therefore, Extended Cox regression model should be used instead.

Suppose we want to test whether the hazard ratio changes over time. Consider the following model:

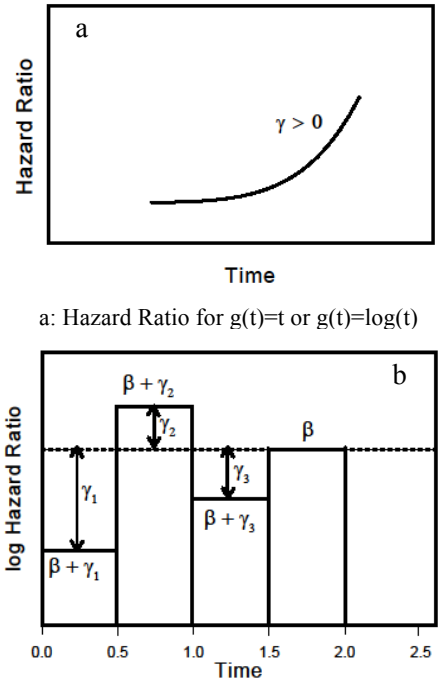
$$(t \mid x) = \ _0(t)\exp(\ X + \ Xg(t)), \tag{4}$$

where *g(t)* is some specified function of time chosen by the data analyst. The term *Xg(t)* is an interaction term between the covariate *X* and some function *g(t)* of time. For such a model the log hazard ratio is

$$\log\frac{(t \mid x_i)}{(t \mid x_j)} = (x_i \quad x_j)(\ + g(t)) \tag{5}$$

This model allows the log hazard ratio (5) to change over time giving us greater flexibility than proportional hazards assumption in Equation (3). In addition, testing whether or not is significantly from zero allows us the opportunity to evaluate the proportional hazards assumption.

Now the model in Equation (4) can be viewed as a proportional hazards model with two covariates, one is the time-independent covariate *X* and the other is the time-dependent covariate *Xg(t)*. We know that *g(t)* is defined as a function of time. It can be direct time *t*, log(*t*), Heaviside function or polynomial function etc. In Extended Cox regression model, the critical decision is the form that the functions *g(t)* should take. We use the forms of *g(t)* as log(*t*) and Heaviside (step) function. For Heaviside function at first we partitioning our time axis into K intervals by choosing (K-1) time points: $0 < \ _1 < \ _2 < \circ \ < \ _{K\ 1}$; and we obtain the following Heaviside or indicator functions as $g_1(t) = I[t \quad [0, \ _1)]$, $g_2(t) = I[t \quad [\ _1, \ _2)]$ and so on.



a: Hazard Ratio for g(t)=t or g(t)=log(t)

b: Hazard Ratio for Heaviside function

Fig. 1. Hazard Ratio of Extended Cox regression Model

Then the Extended Cox regression model can be defined as

$$(t \mid x) = \ _0(t)\exp\begin{matrix} X + \ _1Xg_1(t) + \ _2Xg_2(t) \\ +\circ \ + \ _{K\ 1}Xg_{K\ 1}(t) \end{matrix} \tag{6}$$

K-1 interaction terms between the covariate *X* and the indicator function of time intervals are included here and one indicator function must be excluded to avoid over-parameterization. The hazard ration of Extended Cox regression model are changeover the time shown by Fig. 1. For the continuous time function the hazard rate are according to Fig. 1 (a). When such a Heaviside function is used, the hazard ratio formula yields constant hazard ratios for different time intervals. For example, there are four time group $g_1(t), g_2(t), g_3(t), g_4(t)$ and fit the Extended Cox regression model according to Equation (6) and obtain the parameters $\ , \ _1, \ _2, \ _3$. Then the log hazard ratio of $g_1(t), g_2(t), g_3(t), g_4(t)$ time group is $(\ + \ _1), (\ + \ _2), (\ + \ _3), (\ )$ respectively which are graphically represented in Fig. 1 (b) (Kleinbaum, 2005).

Like as the Cox regression model, parameters of the Extended Cox regression model can also be estimated by maximizing the partial likelihood of the model.

III. Data and Analysis Results

Data for Women of Bangladesh demographic and health Survey (BDHS 2011) have been used for this study. Data are taken from DHS website. We apply the methodology of

Extended Cox regression for non-proportional hazard on the time to the first birth for the ever married women after marriage (Marriage to first birth interval) which is given in month. The data set we worked on excludes information of women who give their birth before marriage (Negative interval) or during the month of marriage (time is 0 month). So we get only 16025 ever married women who give their birth after marriage. Education, Religion, Economic Status, Age at Marriage, Age of Respondent, Respondent working status, Contraceptive use and Place of Resident are considered here as explanatory variables. Here Economic Status variable comes from wealth index in BDHS data by combining 'poorest' and 'poorer' as 'poor', 'middle' are same as 'middle' and 'richer' and 'richest' are combined as rich. This recoding is done to work with fewer categories and also for better understanding. Also we categorize the women of reproductive age (15-49) groups into three arbitrary group as 13-29 years old women, 29-39 years old women and 39-49 years old women to see the change pattern of the first birth result in different generations. These categories are termed as first generation, second generation and third generation, respectively.

Fitting Ordinary Cox Regression Model

Classical Cox regression model is fitted for this data to see the possible inaccuracy. The Cox regression model according to Equation (1) is:

$$(t \mid x) = {}_0(t)\exp({}_1x_1 + {}_2x_2 + \circ + {}_8x_8) \quad (7)$$

We get the maximum likelihood estimates with the corresponding standard errors, hazard ratio for different covariate and p-value. In Table 4 last column shows the hazard ratio of Cox regression model. The rate of giving first birth for Rich and Middle class women were the same compare Poor women. Women who worked outside had 8% less rate of giving first birth than women who do not work outside. From the results of Cox regression model we found Education, Age at Marriage and Contraceptive Use as significant variables. These covariates may have good effect on the time as meaning that it may not satisfy the model assumption. So we need to check the proportional hazard assumption.

Checking for Proportional Hazard Assumption

Each and every technique for testing the non-proportionality has its advantages and limitations. This can be checked by many numerical or graphical methods. Graphical approach requires categorical variables and also they do not provide any formal diagnostic checking whereas numerical approach involves for example testing for time dependent covariates or to find the existence of a trend in the residuals. We use residual measures to investigate the departure from proportionality assumption.

Table 1. Checking for Proportional Hazard Assumption

Explanatory Variables	Categories		Chi-sq	p-value
Education	No education			
	Primary	0.0278	12.41	0.0004
	Secondary	0.0794	103.49	< 0.0001
	Higher	0.0939	141.54	< 0.0001
Religion	Other religion			
	Islam	-0.0047	0.36	0.5490
Economic Status	Poor			
	Middle	0.0185	5.46	0.0195
	Rich	0.0176	4.96	0.0259
Age at Marriage	Age	-0.0662	59.92	< 0.0001
Age of Respondent	1st Generation			
	2nd Generation	-0.0093	1.39	0.2390
	3rd Generation	0.0219	7.67	0.0056
Respondent Working Status	No			
	Yes	0.0018	0.05	0.8220
Contraceptive use	No			
	Yes	0.0352	19.91	< 0.0001
Place of Resident	Urban			
	Rural	0.0169	4.57	0.0326
GLOBAL			294.70	< 0.0001

Schoenfeld residuals are specially used for testing the assumption of proportional hazards and also cumulative Schoenfeld residuals seems to be a more effective approach in detecting covariates with time-varying effects. Schoenfeld residuals are usually calculated at every failure of time under the proportional hazard assumption, and usually not defined for censored observation (Grambsch, 1994; Schoenfeld, 1980). Here, we perform the overall significance test named as 'global test' of the model in Equation (7) from Schoenfeld residual shown in Table 1. The column

is the Pearson correlation of scaled Schoenfeld residual and time. Scaled Schoenfeld residual means that it normalizes with mean from the fitted Cox regression model. The chisq is the Chi-square test of scaled Schoenfeld residual defined by Schoenfeld in 1982 and the corresponding p-value are shown for the null-hypothesis of proportionality.

From the p-values reported in Table 1 we see that most of the variables showed non-proportionality character and also the global test suggested strong evidence of non-proportionality (p-value <0.0001). These numerical findings suggest a non-constant hazard ratio for these variables. So, for the violation of proportional hazard assumption we will use the Extended Cox regression model.

Extended Cox Regression Model Results

Situation 1: $g(t)=\log(t)$

At first let $g(t)=\log(t)$ in the Extended Cox regression model. According to the Extended Cox regression model (8) we can express our model with the parameter $(\ _1,\ _2,\circ\ ,\ _8)$ as the time independent parameter and $(\ _1,\ _2,\circ\ ,\ _8)$ are time dependent parameter in the model.

$$(t\,|\,x)=\ _0(t)\exp\ \begin{matrix}\ _1x_1+\ _1x_1\log(t)\\ +\circ\ +\ _8x_8++\ _8x_8\log(t)\end{matrix}\ +\ _2x_2+\ _2x_2\log(t) \qquad (8)$$

Dummy variables are multiplied by the logarithm of time implying that the multiplicative result is equal to $\log(t)$ when the response of the dummy variable is 1. The estimated parameters of the model are shown in Table 2. The hazard ratio of the model for any covariate is $HR=\exp(\ +\ \log(t))$.

From the table it can said that the rate giving first birth of Primary educated women is $HR=\exp(2.32\ \ 0.68\log(t))$ when for the uneducated women rate giving first birth is 1. The secondary educated women are almost same as the primary educated women. The hazard ratio of Secondary women is $HR=\exp(2.10\ \ 0.60\log(t))$. But the hazard ratio of the higher educated women is increasing over the time. The hazard ratio of higher educated women is $HR=\exp(0.30+0.02\log(t))$.

Table 2. Extended Cox regression model for $g(t)=\log(t)$

Explanatory Variables	Categories		p-value		p-value
Education	No education				
	Primary	2.32	< 0.0001	-0.68	< 0.0001
	Secondary	2.10	< 0.0001	-0.60	< 0.0001
	Higher	0.30	0.0626	0.02	< 0.0001
Religion	Other religion				
	Islam	8.27	< 0.0001	-2.37	< 0.0001
Economic Status	Poor				
	Middle	1.22	< 0.0001	-0.41	< 0.0001
	Rich	0.44	< 0.0001	-0.17	< 0.0001
Age at Marriage	Age	0.75	< 0.0001	-0.26	<0.0001
Age of Respondent	1st Generation		< 0.0001		
	2nd Generation	1.11	< 0.0001	-0.37	<0.0001
	3rd Generation	2.35	< 0.0001	-0.75	<0.0001
Respondent Working Status	No				
	Yes	0.38	< 0.0001	-0.16	< 0.0001
Contraceptive use	No				
	Yes	2.22	< 0.0001	-0.68	< 0.0001
Place of Resident	Urban				
	Rural	2.03	< 0.0001	-0.66	< 0.0001

Situation 2: g(t) is Heaviside Function

For this method we need to define some Heaviside function for different time interval. The range of the data in this study is divided into four time interval according to quantile at 25%, 50%, 75% as $0 < 12 < 22 < 37 < °$; and defined as four time interval $g_1(t), g_2(t), g_3(t), g_4(t)$. Now we define regression model for the waiting time for first birth data which is our main intention. In the model (Eq. 9) we multiply each covariate with time group. Here, $_{11}$, $_{12}$, $_{13}$ are the parameter of Education at $g_1(t), g_2(t), g_3(t)$ time group and $_1$ are treated as the reference parameter of Education, also it can be said that it is the parameter of $g_4(t)$ time interval. According to this all parameters $= (,)$ are defined similar to this and $(x_1, x_2, \circ, x_8)$ are the covariates.

$$(t \mid x) = {}_0(t)\exp \begin{matrix} {}_1x_1 + {}_{11}x_1g_1(t) + {}_{12}x_1g_2(t) + {}_{13}x_1g_3(t) \\ +\circ + {}_8x_8 + {}_{81}x_8g_1(t) + {}_{82}x_8g_2(t) + {}_{83}x_8g_3(t) \end{matrix} \quad (9)$$

Table 3. Results of fitting the Extended Cox regression model for Heaviside function

Explanatory Variables	Categories	exp (β)	p-value	exp (γ_1)	p-value	exp (γ_2)	p-value	exp (γ_3)	p-value
Education	No education								
	Primary	0.875	0.0005	1.628	< 0.0001	1.312	< 0.0001	1.044	0.4526
	Secondary	1.041	0.3552	0.854	0.0218	0.978	0.7503	0.843	0.0082
	Higher	1.656	< 0.0001	0.144	< 0.0001	0.427	< 0.0001	0.638	< 0.0001
Religion	Other religion								
	Islam	0.379	< 0.0001	18.537	< 0.0001	3.885	< 0.0001	2.006	< 0.0001
Economic Status	Poor								
	Middle	0.864	0.0007	1.410	< 0.0001	1.313	< 0.0001	1.103	0.1134
	Rich	0.787	< 0.0001	1.324	< 0.0001	1.454	< 0.0001	1.231	0.0003
Age at Marriage	Age	0.804	< 0.0001	1.500	< 0.0001	1.364	< 0.0001	1.190	< 0.0001
Age of Respondent	1st Generation								
	2nd Generation	0.610	< 0.0001	2.132	< 0.0001	1.573	< 0.0001	1.413	0.0000
	3rd Generation	0.429	< 0.0001	3.315	< 0.0001	2.188	< 0.0001	1.922	< 0.0001
Respondent Working Status	No								
	Yes	0.745	< 0.0001	1.666	< 0.0001	1.376	< 0.0001	1.160	0.0207
Contraceptive use	No								
	Yes	0.845	< 0.0001	1.793	< 0.0001	1.346	< 0.0001	1.126	0.0071
Place of Resident	Urban								
	Rural	0.654	< 0.0001	2.562	< 0.0001	1.841	< 0.0001	1.360	< 0.0001

After estimating the parameters shown in Table 4 we found that most of the parameters associated with the covariates are significant. The significant parameters are not only our importance. The important thing is that the covariates are time-dependent and they are significantly different in different time interval compare to reference time interval. This is clearly seen in hazard ratio comparison (Table 4).

For the variable Education, the hazard is 42% higher for the primary educated women, 11% less for the secondary educated women and 76% less for the higher educated women, all are in comparison with the uneducated women in $g_1(t)$ time interval. The hazard is 15% higher for the primary educated women, 2% higher for the secondary educated women and 29% less for the higher educated women, compare to uneducated women in $g_2(t)$ time interval and so on. There is a trend seen in the hazard ratio in different time intervals. For primary educated people the hazard ratio is decreasing with the increasing waiting time. For higher educated women the hazard ratio is increasing with the increasing time. It can also be seen that in $g_1(t)$ time interval the hazard of primary educated women are higher and for higher educated women hazard is so much lower. But in $g_4(t)$ time interval totally reverse result is found.

IV. Comparison

From the results of Extended Cox regression model we see that the interaction parameters γ (Table 2 and Table 3) are significant that means the covariates are time dependent. For Heaviside function in Extended Cox regression model the hazard ratio are significantly different for different time interval but it is constant in Cox regression model (Table 4).

Table 4. Hazard Ratio for Extended Cox regression model (Heaviside function) and Cox regression model

Covariate	Value level	Extend Cox model				Cox PH exp (□)
		$g_1(t)$ exp (□ + γ_1)	$g_2(t)$ exp(□ + γ_2)	$g_3(t)$ exp(□ + γ_2)	$g_4(t)$ exp(□)	
Education	No education					
	Primary	1.42	1.15	0.91	0.875	1.06
	Secondary	0.89	1.02	0.88	1.042	0.94
	Higher	0.24	0.71	1.06	1.657	0.64
Religion	Other religion					
	Islam	7.04	1.48	0.76	0.380	1.01
Economic Status	Poor					
	Middle	1.22	1.13	0.95	0.864	1.06
	Rich	1.04	1.14	0.97	0.788	1.06
Age at Marriage	Age	1.21	1.10	0.96	0.805	1.07
Age of Respondent	1st Generation					
	2nd Generation	1.30	0.96	0.86	0.611	0.71
	3rd Generation	1.42	0.94	0.83	0.430	0.51
Respondent Working Status	No					
	Yes	1.24	1.03	0.86	0.745	0.92
Contraceptive use	No					
	Yes	1.52	1.14	0.95	0.845	1.15
Place of Resident	Urban					
	Rural	1.68	1.21	0.89	0.655	1.02

Table 5. -2log (L) and AIC Values of the Cox Regression Model and Extended Cox Regression Model

	Cox Regression Model	Extended Cox Regression Model	
		g(t) = log *t*	g(t) is a Heaviside function
AIC	274500.8	212004.2	239641.6
-2log *L*	274476.8	211956.2	239545.6

In survival analysis, comparisons between the models can be made on the Akaike's information criterion (AIC) or $2\log(L)$, which is merely a model selection criterion. Here L defines the likelihood. The values of AIC and $2\log(L)$ for the Cox regression model, Extended Cox regression model for two different function of time are given in Table 5. According to the AIC values, our study shows that Extended Cox regression model is found to be more appropriate model than the Cox regression model if the proportional hazards assumption does not hold.

V. Conclusion

Ignoring non-proportional hazards in Cox regression model can lead us completely wrong results. Using a Cox regression model without ensuring that the underlying assumptions are validated may result in negative implications on the estimates. The power of the tests is reduced for the covariates which are not satisfying the proportionality assumption, that is, we are less likely to conclude for a significant effect when there exist real one. If the assumption is violated Extended Cox regression model is appropriate because it is more flexible to handle time dependent variables. Extended Cox regression model is fitted well in our study as proportionality assumption fails to exist. In our study we only use $\log(t)$ and Heaviside function as the time function while applying for Extended Cox regression model but it can be some other function of time through a proper choice. From our study we find that various factors like education, economic status, age at marriage etc. that effect the time to first birth of women. So this is a big issue that time to first birth has some impact on the age at first birth or median age at first birth. More importantly our main goal was not to show the significant contributors on first birth for ever married women but also their different behavior over different time interval. We found that our covariates were time dependent and fitting Extended Cox regression model for non-proportional hazard, arises in such time dependent covariate situation performs better (fitted better) than traditional Cox regression model.

References

1. Ata, N. a. 2007. Cox regression models with nonpropor-tional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics, **36***, 157-167.
2. Bellera, C. A.-P. 2010. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology, 10*, 20.
3. Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187-220.
4. Grambsch, P. M. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, **81***, 515-526.
5. Kleinbaum, D. a. 2005. *Survival Analysis: A Self-Learning Text.* Springer.
6. Schoenfeld, D. 1980. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika, **67***, 145-153.

7. Schoenfeld, D. 1982. Partial residuals for the proportional hazards regression model. *Biometrika, **69***, 239-241.

Dhaka Univ. J. Sci. 63(1): 25-30, 2015 (January)

*Author for Correspondence. *e-mail*: arahman6@isrt.ac.bd