An Open Access Journal available online

First Birth Interval: Cox Regression Model with Time Varying Covariates

Adeniyi O.I.[1*] & Akinrefon A.A.[2]

[1]University of Ilorin, Kwara State,
[2]Modibbo Adama University of Technology, Yola, Adamawa State
* E-mail: adeniyi.oi@unilorin.edu.ng

Abstract: The Cox regression model has been widely used for the analysis of time to event data with their associated risk factors, it assumes a constant hazard ratio over time and that the risk factors are independent of time. When the assumptions are violated, the estimates of the hazard ratio of the Cox regression estimates of the hazard ratios becomes misleading. In this study, we use a modified Cox regression model that incorporates time dependent covariate which measures the interaction of exposure with time.

Birth interval between marriage and first birth for the ever married women after marriage, taken from NDHS 2013 women data is fitted using the Cox regression model with time varying covariates due to the failure of existence of proportionality assumption. This model performs better compared to Cox regression model.

Keywords: Time to event, Hazard Ratio, Time-varying covariates, proportionality assumption

Introduction

The Cox Proportional hazard models requires that the hazard ratio is constant over time, which implies that the hazard for an individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. However, the Cox Proportional hazard model gives a misleading conclusions when the assumption is violated particularly in the presence of long follow-up period.

In order to avoid misleading estimates of the hazard ratio due to the presence of time-dependent variables, checking the proportionality of the hazards assumptions should be an integral part of a survival analysis by a Cox regression model. Even thoug

Cox regression model has been widely used recent publications [1, 2 &3] suggest that the test of the validity of the assumptions must be verified before its use.

To evaluate the proportional hazard assumption, we use the residuals measures like Schoenfeld residuals [4] to 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 to sometime t or if the size of its effect changes over time. For this time varying covariates, the Cox regression model with time varying covariate is used instead of the traditional one. We illustrate our discussion with a study on birth interval between marriage and first birth for ever-married women extracted from women data, NDHS 2013.

Methodology

Cox Proportional Hazard Model

The proportional hazards model is a regression model with time to event as dependent variable. It allows inclusion of information about known (observed) covariates in models of survival analysis and is the most applied model in this area. To investigate the relation between the survival time and some risk factors called covariates, the Cox proportional hazards model is used. In this model, the relative risk is described parametrically and the hazard function is described non-parametrically. The hazard function for individual *i* is written as:

$$h(t, X_i) = \mathrm{h}_0(t)\exp(\beta X_i) \qquad 1$$

$h_0(t)$ is a baseline hazard function, left unspecified; $\exp(\beta X_i)$ is the relative risk of individual *i* with X_i as the covariate vector. In this model, covariates act multiplicatively on the baseline hazard, adding additional risks on an individual basis. Coefficient vectors of the covariates are estimated by maximizing a partial likelihood function [5]. The model parameter β are interpreted by the hazard ratio assumed to be constant over time which is given as;

$$HR = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} \qquad 2$$

Where X^* is the set of predictors for one individual and X is the set of predictors for the other individual.

Regression models for time to event data have 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 is violated, therefore methods that do not assume proportionality must be used to investigate the effects of covariates on survival time. The significance of the estimated parameter of the Cox regression model does not implies that the model is well fitted and satisfies the proportional hazard assumption and vice versa, thus, Cox proportional hazards with time varying covariates is used.

Cox Proportional Hazard Model with Time Varying Covariate

In the Cox regression model, when time-dependent variables are used to assess the proportional hazard assumption for time- independent variables, the Cox regression model cannot be used because it can no longer satisfy the proportional hazards assumption. Therefore, Cox regression model that incorporate time-varying covariates should be used instead. A time-dependent variable is defined as any variable whose value for a given subject may differ over time (t) [6].

Given a survival analysis situation involving both time-independent and time-dependent predictor variables, the Cox proportional hazard model that incorporate both type of variables is given as

$$h\left(t, X(t)\right) = h_0(t)\exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \sigma_j X_j g_j(t)\right] \quad 3$$

Where $X(t) = \left(X_1, X_2, ..., X_{p1}\right)$ are the time-independent and $\left(X_1(t), X_2(t), ..., X_{p2}(t)\right)$ time-dependent variables. The term $\sigma X(t)$ is an interaction term between the covariate X and some function $\sigma(t)$ of time. The hazard ratio for Cox model with time varying covariates is given as

$$HR(t) = \frac{h\left(t, X^*(t)\right)}{h\left(t, X(t)\right)} = \exp\left[\sum_{i=1}^{p_1} \beta_i \left[X^*_i - X_i\right] + \sum_{j=1}^{p_2} \delta_j \left[X^*_j(t) - X_j(t)\right]\right] \quad 4$$

This model allows the hazard ratio to change over time giving greater flexibility than proportional hazards assumption in Eq. (2).

Likelihood estimation

Like the Cox regression model, parameters of the Cox regression model with time varying covariates can also be estimated by maximizing the partial likelihood of the model.

$$L(\beta) = \prod_{j=1}^{n} \frac{\exp\left(\beta X\left(t_j\right)\right)}{\sum_{l \in R(t_j)} \exp\left(\beta X_i\left(t_j\right)\right)} \quad 5$$

Application to data on birth interval

Dataset from the 2013 Nigeria Demographic and Health Survey (NDHS) were analysed. Data on interval of marriage to first birth were available for 26738 women aged 15-49. The survey was designed to provide these information at national, regional, and state or district levels, for both urban and rural areas. If a woman is married but has not given birth, the difference between her current age and age at marriage is used and is recorded as censored observation. We applied the methodology of Cox regression model to dataset on marriage to first birth interval (which is recorded in months).

The geopolitical zone, location of residence, religion, highest educational qualification, economic status, respondent age at marriage and working status were considered as explanatory variables. Three categories were created for Economic Status variable which comes from wealth index in NDHS data by combining 'poorest' and 'poorer' as 'poor', 'middle' are same as 'middle' and 'richer' and 'richest' are combined as rich. Also, the women's age at marriage was categorized into three arbitrary group as less than 18 years old women, 18-24 years old women and above 24 years old women. The two major religion being practiced were considered as Christianity and Islam while the highest educational qualification are categorised as No education, Primary, Secondary and Higher. The geopolitical zone in the country are North-central, North-east, North-west, South-east, South-south and South-west respectively while location of residence is classified as Urban and Rural. The working status of the women as categories into employed and nit employed.

Checking the Proportional Hazard Assumption

To test the hypothesis that the proportion hazard assumption is the valid, the following statement of hypothesis is given;

$H_0 : \delta_1 = \delta_2 = ... = \delta_{p2}$ (Assumption is valid)

$H_1 :$ at least one of the $\delta_i^{'} s$ is not equal to zero (Assumption not valid)

We use residual measures to investigate the departure from proportionality assumption. Schoenfeld residuals was used to test the assumption of proportional hazards. Schoenfeld residuals are usually calculated at every failure of time under the proportional hazard assumption, and usually not defined for censored observation [7, 8 & 9]. The overall significance test named as ‘global test’ of the model in Eq. (3) was performed from Schoenfeld residual shown in Table 1. The columns are the explanatory variables, categories of the explanatory variables, the Pearson correlation (rho) 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 as defined by Schoenfeld in 1982 and the corresponding p-value are shown for the null-hypothesis of proportionality.

Table 1: Test of Proportional Hazard Assumption

Explanatory Variable	Categories	rho	Chisq.	p-value
Zone	North-central			
	North-east	0.0128	3.95	0.0468
	North-west	0.0725	126.84	<0.0001
	South-east	-0.0202	9.88	0.0017
	South-south	0.0006	0.01	0.9198
	South-west	-0.0245	14.57	0.0001
Location of Residence	Urban			
	Rural	0.0025	0.15	0.6951
Highest Educational Qualification	No Education			
	Primary	-0.0023	0.12	0.728
	Secondary	-0.008	1.59	0.2017
	Higher	-0.0167	6.71	0.096
Religion	Islam			
	Christianity	0.0111	3.11	0.0778
Economic Status	Poor			
	Middle	-0.0104	2.63	0.1047
	Rich	0.0014	0.05	0.8219

Working Status	Not Employed			
	Employed	0.0142	4.88	0.0272
Age at Marriage	less than 18 years			
	18 to 24 years	-0.0322	24.73	<0.0001
	Above 24 years	-0.0491	56.57	<0.0001
Global Test			909.4	<0.0001

From the p-values reported in Table 1, it was revealed that covariates zone, highest educational qualification, working status and age ta marriage 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. Therefore, for the violation of proportional hazard assumption, a Cox regression with time varying covariate is used.

Cox Regression with time-varying covariates

We assume that $g_j(t) = t$, which implies that for each X_j in the model as main effect, there is a corresponding time dependent variable in the model of the form $\mathrm{X}_j * t$. The Cox proportional hazard model with time varying covariate is of the form

$$h(t, X(t)) = h_0(t) \exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j \left(X_j * t\right)\right] \qquad 6$$

Results

Table 2 presents the parameter estimates of Cox proportional Hazard model and Cox Model with time-varying covariates. The Akaike information criterion (AIC) [10] and 2LogL was used to select the preferred model between the Cox proportional hazard model and Cox Model with time-varying covariates. The values of the selection criteria shows that Cox model with time-varying covariates is preferred. Therefore, discussion of results is upheld for the parameter estimates from Cox model with time-varying covariates.

Table 2: Parameter Estimates for Cox PH Model and Cox with Time-varying Covariates

		Cox PH			Cox with time-varying covariates					
Explanatory Variable	**Categories**	**Hazard Ratio**	**β**	**p-value**	**Hazard Ratio**	**β**	**p-value**	**Hazard Ratio**	**δ**	**p-value**
Zone	North-central									
	North-east	0.8614	-0.1492	<0.001	0.8434	-0.1703	<0.001	1.0012	0.0012	0.228
	North-west	0.7567	-0.2788	0.001	0.6147	-0.4866	<0.001	1.0072	0.0072	<0.001
	South-east	1.0628	0.0609	0.038	1.1882	0.1724	<0.001	0.9945	-0.0055	<0.001 0.255
	South-south	1.0104	0.0104	0.699	1.037	0.0363	0.343	0.9985	-0.0015	

	South-west	1.2832	0.2494	<0.001	1.4023	0.3381	<0.001	0.9941	0.0059	<0.001
Location of Residence	Urban									
	Rural	0.9631	-0.0376	0.03	0.9623	-0.0384	0.027			
Highest Educational Qualification	No Education									
	Primary	1.1462	0.1365	<0.001	1.1993	0.1817	<0.001	0.9985	-0.0015	0.061
	Secondary	1.1373	0.1287	<0.001	1.2151	0.9148	<0.001	0.9968	-0.0032	0.001
	Higher	1.0706	0.0682	0.038	1.1858	0.1704	<0.001	0.9955	-0.0045	0.004
Religion	Islam									
	Christianity	0.9103	-0.094	<0.001	0.9074	-0.0972	<0.001			
Economic Status	Poor									
	Middle	1.1123	0.1064	<0.001	1.1215	0.1147	<0.001			
	Rich	1.0263	0.026	0.256	1.0392	0.03845	0.093			
Working Status	Not Employed									
	Employed	0.9941	-0.0059	0.704	0.9602	-0.0406	0.058	1.0016	0.0016	0.005
Age at Marriage	less than 18 years									
	18 to 24 years	1.1959	0.1789	<0.001	1.3711	0.3156	<0.001	0.9938	-0.0062	<0.001
	Above 24 years	0.9696	-0.0309	0.267	1.3497	0.2999	<0.001	0.9841	-0.016	<0.001
-2LogL		441976.52						441202.68		
AIC		442051.52						441232.68		

From Table 2, the results for the time varying covariates has it that the estimated hazard ratio for women for North-east is $HR = \exp\left(-0.1703 + 0.0012t\right)$, which implies that the estimated hazard ratio will increase exponentially by 0.0012 as the time increases compare to women form the North-central zone. Also, the hazard ratio for North-west women increases by 0.0072 as time increases while it decreases by 0.0015 and 0.0059 for women for South-south and South-west as time increases compare to women from the North-central. The hazard ratio decreases with time as the educational qualification improves by 0.0015, 0.0032 and 0.0045 for primary, secondary and higher educational qualification respectively compared to women with no formal education. The hazard ratio for employed women increases by 0.0016 as time increases compare to women who are unemployed while the hazard decreases with time by 0.0062 and 0.016 for women whose age at marriage are between 18 to 24 years and above 24 years respectively.

For the covariates that are not time varying, the hazard ratio decreases by 0.0377 for women living in the rural areas compare to women living in the urban areas. The hazard ratio decreased by 0.0926 for Christian women compare to Muslim women while the hazard increase by 0.2115 and 0.3922 for the middle and rich economic status compare to the poor status.

Conclusion

Cox regression model been the most popular approach in analysing survival data may give misleading estimates if the underlying assumptions are validated. The power of the tests is reduced for the covariates which are not satisfying the proportionality assumption. Once it is established that the assumptions are not valid, a Cox model that incorporate time-varying covariates will give a better estimate of the parameter. From the study carried out on birth interval between marriage using dataset from 2013 Nigeria Demographic and Health Survey (NDHS), it was revealed that factors like geopolitical zone, highest educational qualification, working status and age at marriage were time-varying among other factors that were considered to affect the interval of marriage time to first birth of women. The interest of the study is to found out the covariate that are time-dependent and fit an appropriate survival model to predict the hazard ratios

References

Ata, N. A. (2007). Cox regression models with non-proportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics* 36, 157-167.

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.

Rahman, A and Hosque, R (2015). Fitting Time to First Birth Using Extended Cox Regression Model in Presence of Non-proportional Hazard. Dhaka University. *Journal of Science* 63(1): 25-30

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

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

Kleinbaum, D.G. and Klein, M. (2005). Survival Analysis: A Self-Learning Text. Springer.

Stablein, D., Carter, W., and Novak, J. (1981). Analysis of survival data with non-proportional hazard functions. *Control Clinical Trials* 2: 149–159.

Grambsch, P. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 515-526.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* 67, 145-153.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716-723.