Research Journal of Mathematical and Statistical Sciences ________________________________ISSN 2320-6047
Vol. **11(1),** 4-8, May **(2023)** *Res. J. Mathematical and Statistical Sci.*

Cox-proportional hazard model with flexible penalized spline: application to colon cancer

K.T. Amzat[1*] and A.U. Chukwu[2]
[1]Department of Mathematical Sciences, Crescent University, Abeokuta, Nigeria
[2]Department of Statistics, University of Ibadan, Ibadan, Nigeria
amzatkafayat@gmail.com

Available online at: www.isca.in, www.isca.me
Received 10th August 2022, revised 23rd January 2023, accepted 6th March 2023

Abstract

This study examines the inherent trends in non-proportional hazard model and how non-linear covariates effect and interaction with time could be estimated by penalized likelihood with B-spline basis function. Bootstrap simulation was used to assess the assumptions and the results shown that the continuous variable "age" exhibited a non-linear trend, so, assessment of Cox PH model is essential to avoid wrong statistical inference.

Keywords: Cox model, penalized-likelihood, Time-varying covariate, Bootstrapping.

Introduction

Cox model is the most useful model in analyzing survival data because it is simple and does not rely on the assumptions of survival distribution. The model imposes assumptions which have not been justifiable by most data set. Generalized Additive Model could be applicable to any likelihood-based regression model, having potential to uncovering non-linear covariates effects. Fisher and Lin[1] observed that modeling non-proportional models could be a powerful tool for determining predictive relationships using quantities that vary over time. Proportional hazard model may be inappropriate in some instances and alternatives such as stratified and time-dependent variables could be use in analyzing survival data. Therneau and Grambsch[2] estimating the parameters of proportional hazard model when assumptions are relaxed was considered by Abrahamowicz et al[3] in modeling time dependent and non-linearity effect simultaneously using regression spline product model concluded that effect of some continuous covariates accounting for time dependent shows strong evidence that non-linearity is violated.

The Proportional hazard assumptions are likely to be violated in long follow up, where the effect of some covariates may vary over time (Cox[4]). Ata and Sozer[5] observed that handling non-proportionality of hazards, varieties of methods could be employed. His findings revealed that using stratified and extended Cox model is more suitable than proportional hazard model. Most times, it is assumed that continuous covariate have a linear form, but this assumption is given less consideration. Gray[6] suggests that spline functions could be used as a way of modeling continuous covariate without meeting stringent assumptions. Additional references on additive modelling and others could be seen in Friedman & Silverman[7] and Amzat & Adeosun[8]. This study examines the non-linearity of a continuous covariates.

Methodology

The Cox proportional hazard model depicts the survival time as a function of several prognostics factors [1], expressed as:

$$\boldsymbol{h(t,x)} = \boldsymbol{h_o(t)exp\{\beta_1 x_1 + \cdots + \beta_x x_k\}} \tag{1}$$

where $\boldsymbol{h(t,x)}$ is the hazard function at time t for a subject with covariate values $\boldsymbol{x_1, \ldots x_k}$. $\boldsymbol{h_o(t)}$ is the baseline hazard function. The Cox PH model[1] imposes two assumptions. First, it assumes that the effect of each covariate on the hazard does not vary over time. Analytically, ratio of $\boldsymbol{h(t,x)}$ for two different covariates is given as:

$$\frac{h(t,x_i)}{h(t,x_j)} = \frac{h_o(t)exp\{\beta_1 x_{i1}+\cdots+\beta_k x_{ik}\}}{h_o(t)exp\{\beta_1 x_{j1}+\cdots+\beta_k x_{jk}\}} = exp\{\beta_1(x_{i1}-x_{j1}) + \cdots + \beta_k(x_{ik}-x_{jk})\} \tag{2}$$

Implies the ratio of two hazards is a constant which does not depend on time. Secondly, the effect of each continuous covariate is linear on logarithm of the hazard. That is, log hazard ratio for the *ith* individual to the baseline described as

$$log\left(\frac{\lambda_{i(t)}}{\lambda_{o(t)}}\right) = \beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_p Z_{pi} \tag{3}$$

This implies that the proportional hazard model is a linear model for the log of the hazard ratio.

Non Proportional Hazard: Non proportional hazards may arise as a result of model failures. In appropriate functional form of a covariate is a model failure that could affect diagnosis of non-proportionality[2,9,10]. To diagnose for non-proportionality of hazards, measures such as Schoenfeld residual plots, graphical and numerical approaches need be investigated.

The relationship between non proportionality and the appropriate functional form of covariates suggests a sequence of how analysts ought to perform diagnosis of Cox model.

Methods of Handling Non-Proportional Hazard: Stratified Model: Stratified Cox model involves splitting of samples into different subgroups base on categorical variable which is known as stratification of variable and the underlying hazard function which causes variation between these subgroups is re-estimated. Hence, stratified model for stratum *s* is defined as follows:

$$\lambda_s(t,x) = \lambda_{so}(t)exp(\beta x) \qquad (4)$$

where *s* = 1, 2, …, S and S is the number of subgroups base on stratification

Interaction with Time: Covariate interaction with time occurs when the effect of an explanatory variable on survival varies with time. This technique is an approach to identify covariates in a model[11]. Interacting covariates with time in a model enhances interpretation of parameters by taking into consideration covariates influence on hazard. Covariate interaction with time for a variable x is given as:

$$\mathrm{h}(t,x) = \mathrm{h}_o(t)exp(\beta_1 x + \beta_2 x f(t)) \qquad (5)$$

Penalized Spline: P-spline are regression splines such that a penalty is imposed on the coefficients of the piecewise polynomial[12-15]. The smoothing parameter λ influences the smoothness of the fit. To explore the nature of log hazard ratio, flexible penalized B-spline is explored to model the continuous covariate.

$$HR = \frac{\mathrm{h}(t/x_i)}{\mathrm{h}(t/x_0)} = \exp(\beta^{'}\mathrm{x}) \qquad (6)$$

Equation (5) as a smooth function compactly becomes

$$\mathrm{HR} = \exp f(\mathrm{x}) \qquad (7)$$

where f(x) is a smooth function given as :

$$f_j(x) = \sum_{i=1}^{k} \beta_i B_{ij} \qquad (8)$$

β_i are coefficients to be estimated, and B_{ij} are basis function.

The coefficient in (8) is estimated by penalized partial likelihood expressed as

$$l_{pen}(\beta,\lambda) = l_\beta - \frac{1}{2}\lambda \int (f^{''}(x))^2 \ dx \qquad (9)$$

Where $l_{(\beta)}$ is log partial likelihood.

Interaction of time with a continuous variable

Consider extended CPHM

$$\mathrm{h}\big(t,x,z(t)\big) = \mathrm{h}_0(t)exp(\beta^{'} + \alpha^{'}z(t)) \qquad (10)$$

Where z(t) is a smooth time varying covariate while α is unknown parameter. Approximating z(t) by a linear combination of B-spline, we have

$$HR(t) = exp\sum \alpha^{'}z(t) \qquad (11)$$

The unknown parameter $\alpha^{'}$ (10) is estimated by penalized partial likelihood given as

$$l_{pen}(\propto,\lambda) = l_\propto - \frac{1}{2}\lambda \int (f^{''}(t))^2 \ dt \qquad (12)$$

$l_\propto$ is log partial likelihood.

The Newton Raphson procedure is employed to obtain the estimates of parameters.

Data description on colon cancer

The colon cancer data set was employed in the study as an example, consisting of 13,011 patients along with covariates sex (female, male), age in years, status (alive or dead), Sub-site (Coecum and Ascending, Descending and Sigmond, others and Nos), year diagnose (75-84, 85-94), age group (0-44, 45-49, 60-74, 75+, survival time in years.

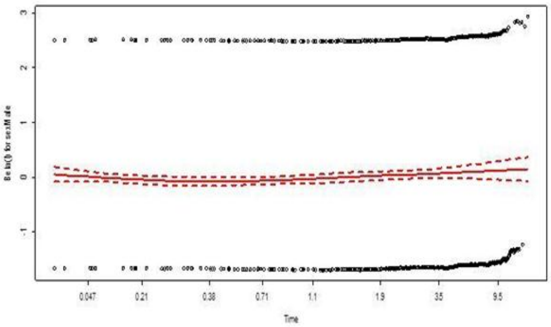

Figure-1: Plot of estimated effects of "Sex" covariate against survival times.

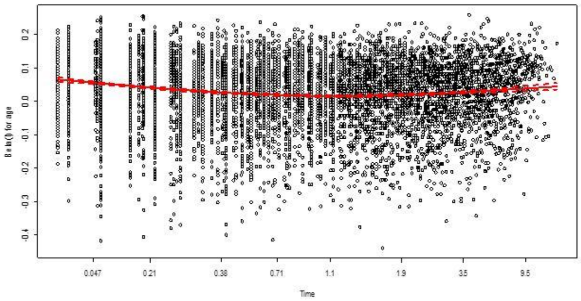

Figure-2: Plot of estimated effect of "Age" covariate against survival time.

Table-1: Estimates of Cox proportional hazard model.

Variables	β (p-value)	Schoenfeld Test ρ (p-value)	S.e (Coeff)	HR (95%CI)
Male	0.1359(2.15e-019)	-0.0053 (6.31e-01)	0.0227	1.146 (1.096-1.198)
Age	0.4369 (2e-16)	-0.0580 (2.38e-08)	0.0026	1.045 (1.040-1.050)
Sub site D&S	-0.0300 (0.2384)	0.0482 (9.94e-06)	0.0254	0.971 (0.923-1.020)
Sub site O&N	0.1508 (0.0014)	-0.741 (1.28e-11)	0.0473	1.163 (1.060-1.276)
Transverse	0.1698 (5.07e08)	-0.0310 (4.59e-03)	0.0312	1.185 (1.115-1.260)
Year 85-94	-0.4966 (2e-16)	-0.0212 (4.98e-02)	0.0223	0.609 (0.583-0.636)
45-59	-0.5000(7.16e-12)	-0.0259 (1.83e-02)	0.0729	0.607 (0.526-0.700)
60-74	-0.7571 (7.77e-16)	0.0447 (3.95e-05)	0.094	0.469 (0.390-0.564)
75+	-0.7253 (2.32e-09)	0.0374 (5.62e-04)	0.1214	0.484 (0.382-0.614)
Global		(0.00e+00)		

Table-2: Result of Cox Non-PH Model with P-Spline.

Variable	β (p-value)	S.e (coeff)	HR (95% CI)
Male	0.1403(6.06e-10)	0.0227	1.151 (1.101-1.203)
Pspline (age), L	0.0436(0.0e+00)	0.0025	0.586 (0.196-1.751)
Pspline (age), N L	0.00E+00		
Subsite D & S	-0.026 (3.0e-01)	0.0254	0.974 (0.927-1.024)
Subsite O & N	0.1403 (3.0e-03)	0.0473	1.151 (1.049-1.262)
Transverse	0.1748 (2.1e-08)	0.0312	0.840 (1.120-1.266)
Year Diagnose			
Age 85-94	-0.492(0.0e+00)	0.0223	0.611 (0.585-0.639)
45-59	0.0719 (5.4e-01)	0.1164	1.075 (0.855-1.350)
60-74	0.0502 (7.3e-01)	0.1463	1.051 (0.790-1.401)
75+	0.0203 (9.0e-01)	0.1563	1.021 (0.751-1.386)

Table-3: Analysis of Deviance**.**

	Loglik	p value
Model PH	-74034	
Model Non PH	-73990	-2.2e-16

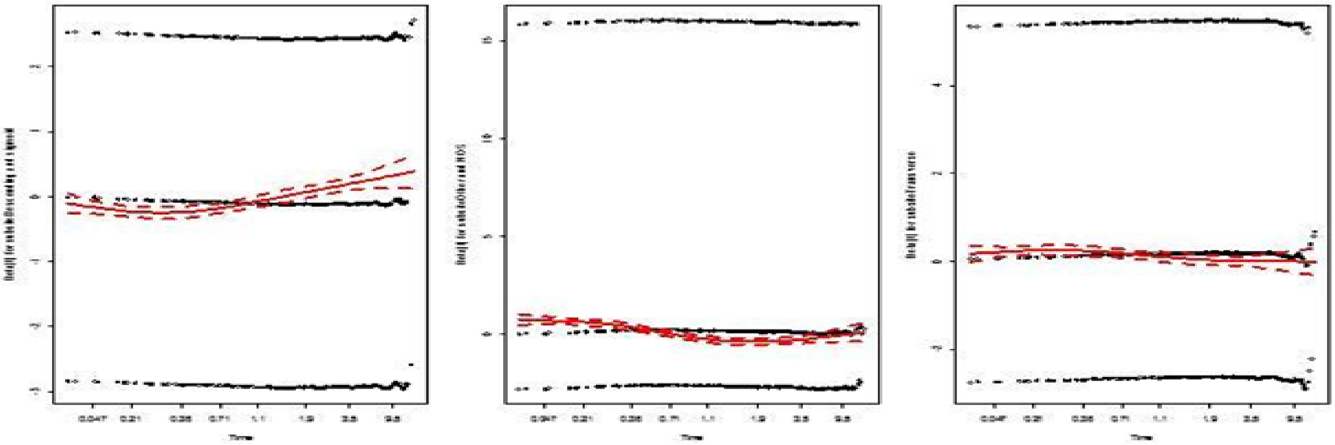

Figure-3: Plot of estimated effects of “Sub-site” covariate against transformed survival times.

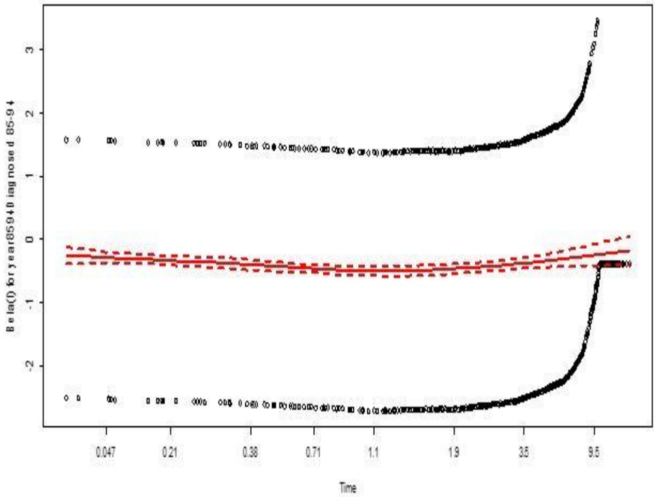

Figure-4: Plot of estimated effects of “Year 8594” covariate against survival time.

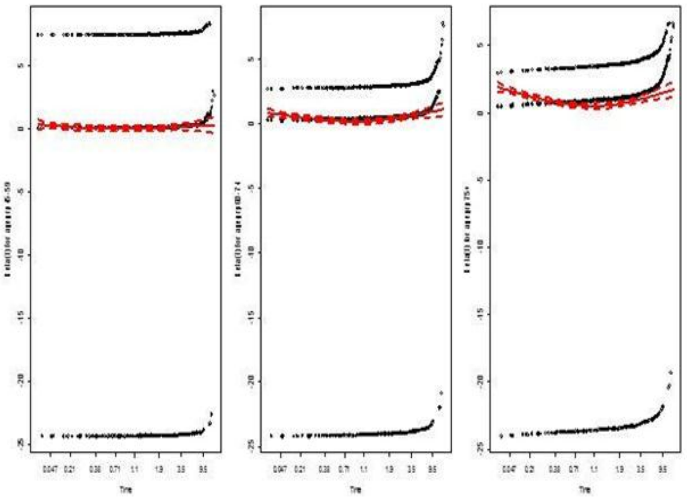

Figure-5: Plot of estimated effects of “Year” covariate against survival times.

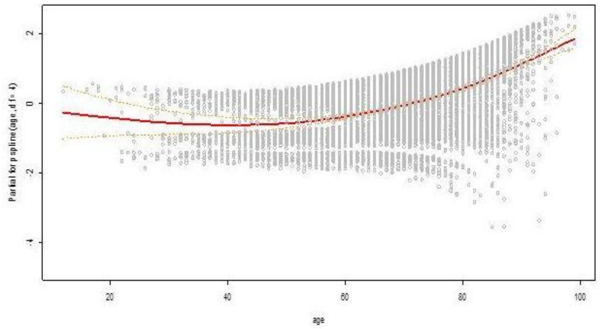

Figure-6: Term plot of partial residuals against Age.

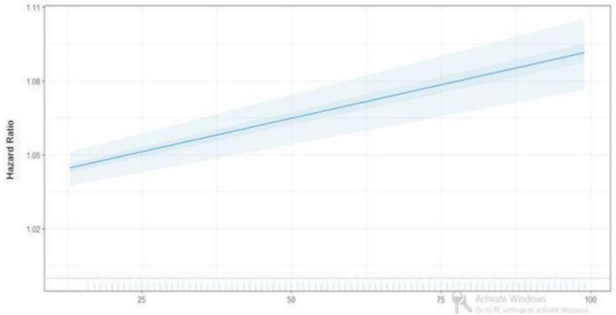

Figure-7: Bootstrap simulated Hazard ratios obtained from Cox PH model.

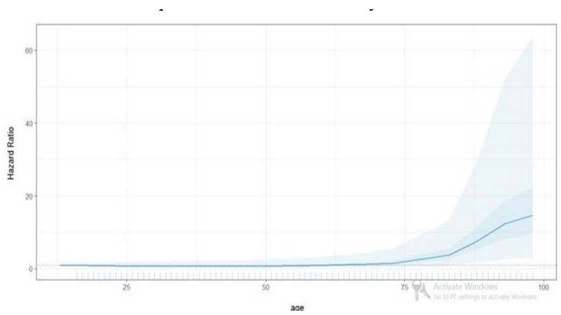

Figure-8: Bootstrap simulated Hazard ratios obtained from non-PH Cox model.

Result from the study of fitting Proportional Hazard model shows that all variables have significant effect with the exception of Descending and Sigmoid level which is insignificant. Similarly, the Likelihood ratio, Wald test and Score test are significant with values 1739, 1741 and 1775 respectively on 9 degree of freedom. In addition, the Schoenfeld residuals test for testing Proportional Hazard assumption shows that all correlation coefficients are significant except for the Sex male covariate.

However, fitting a non-Proportional Hazard model with Penalized B-Spline for continuous and categorical variables shows that “Age” has a significant effect on hazard which indicates non-linearity. Similarly, the Likelihood ratio test gives a significant result with value 1825 on 12 degree of freedom. Also, result of analysis of deviance depicts Non-proportionality.

Bootstrap simulation shows that the data set does not support proportional hazard model assumptions. The graph indicates that the hazard ratio increases with increase in patient’s age. Similarly, the bootstrap simulation of Non-PH model reveals that hazard ratio increases non-linearly with an increase in the ages of patients.

Conclusion

This study establishes the non-linearity of continuous variable “age” on hazard. In addition, Schoenfeld residual test and plot of estimated effect shows that Proportional Hazard assumptions may have been violated. Similarly, term plot of the age covariate indicates that the trend of the age covariate is non-linear and this claim is being supported by analysis of deviance.

Hence, the purpose of the study to ascertain the inherent trend in Cox PH model shows that the continuous variable “age” exhibited a non-linear trend. Therefore, assessment of Cox PH model is essential to avoid wrong statistical inference.

References

1. Fisher, L. D., & Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, 20(1), 145-157.

2. Auton, T. (2001). Modelling Survival Data: Extending the Cox Model. *JSTOR*, 50(4), 558-559.

3. Abrahamowicz, M., & MacKenzie, T. A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in medicine*, 26(2), 392-408.

4. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-202.

5. Ata, N., & Sözer, M. T. (2007). Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, 36(2), 157-167.

6. Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420), 942-951.

7. Friedman J.H. and Silverman B.W. (1989). Flexible Parsimonious Smoothing and Additive modelling. *American Statistical Association and the American Society for Quality control*, 31(1), 3-21.

8. Amzat K.T and Adeosun S.A (2014). On a Sequential Probit Model of Infant Mortality in Nigeria. *International Journal of Mathematics and Statistics Invention*, 2(3), 89-94.

9. Heinzl, H., & Kaider, A. (1997). Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer methods and programs in biomedicine*, 54(3), 201-208.

10. Hess, K. R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in medicine*, 14(15), 1707-1723.

11. Eisen, E. A., Agalliu, I., Thurston, S. W., Coull, B. A., & Checkoway, H. (2004). Smoothing in occupational cohort studies: an illustration based on penalised splines. *Occupational and environmental medicine*, 61(10), 854-860.

12. Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11(2), 89-121.

13. Strasak, A. M., Lang, S., Kneib, T., Brant, L. J., Klenk, J., Hilbe, W., ... & VHM & PP Study Group (2009). Use of penalized splines in extended Cox-type additive hazard regression to flexibly estimate the effect of time-varying serum uric acid on risk of cancer incidence: a prospective, population-based study in 78,850 men. *Annals of epidemiology*, 19(1), 15-24.

14. Wand, M. P. & Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2), 179-198.

15. Malloy, E. J., Spiegelman, D., & Eisen, E. A. (2009). Comparing measures of model selection for penalized splines in Cox models. *Computational statistics & data analysis*, 53(7), 2605-2616.