Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Flexible parametric survival models: An application to gastric cancer data

Nihal Ata Tutkun [1,] *, Müge Yeldan [2], Handan İlhan [2]

[1]*Department of Statistics, Faculty of Science, Hacettepe University, Ankara, Turkey*
[2]*Department of Actuarial Sciences, Faculty of Science, Hacettepe University, Ankara, Turkey*

ARTICLE INFO

Article history:
Received 21 December 2016
Received in revised form
12 February 2017
Accepted 18 February 2017

Keywords:
Cancer
Parametric
Proportional hazards
Spline

ABSTRACT

Flexible parametric survival models using cubic splines become popular in survival data analysis. The property of allowing converging hazard functions leads them to be the alternatives to Cox proportional hazards model and parametric survival models. In this study, flexible parametric survival models are applied to the data set of 106 gastric cancer patients. According to this data set, metastasis and muscle contraction are found as important risk factors on survival.

1. Introduction

Cox (1972) proportional hazards (PH) model is widely used in the analysis of time-to-event data with censoring and covariates. However, it has some disadvantages such as the estimation of baseline hazard function and the assumption of PH. The baseline hazard function is treated as a high-dimensional nuisance parameter and consequently parametric survival models may be appropriate to estimate it. When the PH assumption is violated, parametric survival models give more precise estimates and lead to some benefits. Another, important but less widely used model in survival analysis is proportional odds (PO) model. This model becomes a candidate model under non-proportional hazards. In this study, flexible parametric survival models suggested by Royston and Parmar (2002) based initially on the assumption of either PH or PO scaling of covariate effects is examined. The class of such models is based on transformation of the survival function by a link function proposed by Younes and Lachin (1997) using the parameterized link function of Aranda-Ordaz (1981). Royston and Parmar (2002) used restricted cubic splines to model baseline hazard directly. The restricted cubic splines offer greater flexibility in shape of the hazard function when compared with standard parametric models (Nelson et al., 2007). The purpose of this study is to present the flexible parametric survival models of Royston and Parmar (2002) and apply the models to the real data set in cancer. The comparison of the models is also discussed.

2. Method

2.1. Survival models

Cox PH regression model which was discussed by Cox (1972) is used for analyzing censored survival data. The general Cox PH model is defined through the hazard function h(t) as

$$h(t) = h_0(t)exp(\beta'z)$$

where z is a covariate vector and $h_0(t)$ is the baseline hazard function. The model may be written in integrated form as

$$H(t) = \left(\int_0^t h_0(u)du\right) exp(\beta'z) = H_0(t)exp(\beta'z)$$

where, H(t) is the cumulative hazard function. The model has an assumption of proportional hazards, however it has no distributional assumption. The use of Cox PH regression model may have two difficulties such as:

Baseline hazard function: In the Cox PH model, the baseline hazard function, which may be estimated by the method of Kalbfleisch and Prentice (1980) among others, is treated as a high-dimensional nuisance parameter and is highly erratic. However, the behavior of the hazard function is of potential medical interest because it is directly related to the time-course of an illness. To estimate it informatively (that is, smoothly), some type of parametric model may be appropriate (Royston, 2001).

* Corresponding Author.
Email Address: nihalata@hacettepe.edu.tr (N. A. Tutkun)

https://doi.org/10.21833/ijaas.2017.04.014

Non-proportional hazards: A second issue is how to deal with non-proportional hazards which may occur, for example, when modelling prognostic factors in studies with medium or long follow-up times. Although the Cox PH model may be extended to allow for non-proportional hazards, for example by incorporating time varying regression coefficients there is no 'natural', widely accepted approach, and obtaining a satisfactory model can be complicated. There are further concerns about the complexity involved in the practical interpretation of the coefficients and in the robustness of such models (Royston, 2001).

An alternative approach to modelling survival data is the PO model which was first described in a semi-parametric framework by Bennett (1983) and was further developed by several authors, including importantly by Dinse and Lagakos (1983), Crowder et al. (1991), Collett (1994), Rossini and Tsiatis (1996), Yang and Prentice (1999), and Kirmani and Gupta (2001). The model assumes that the effect of the covariates is to increase or decrease the odds of dying by a given duration by a proportionate amount:

$$O(t;z)=\frac{1-S(t;z)}{S(t;z)}=\frac{1-S_0(t)}{S_0(t)}\exp(\beta'z)=O_0(t)\exp(\beta'z)$$

where, z is a covariate vector, $S_0(t)$ is the baseline survival function that is taken from a suitable distribution, and $\exp(\beta'z)$is a multiplier reflecting the proportionate increase in the odds associated with covariate values z (Rodriguez, 2007). Here, the covariates act multiplicatively on the odds of survival beyond t. The model is a linear model for the log-odds ratio. As for the PH model, a non-parametric estimate of the baseline hazard function can be obtained. The model is then fitted by estimating β–parameters in the linear component of the model and the baseline survival function from the data (Collett, 1994). In PO model there is an assumption that the odds ratio is constant over time. The PO model with its property of convergent hazard functions is of considerable value in modelling survival data with non-proportional hazards.

To allow non-proportional hazards, Royston (2001), Royston and Parmar (2002), and Lambert and Rosyton (2009) developed flexible parametric models based initially on the assumption of either PH or PO scaling of covariate effects. Generically, the class of such models is based on transformation of the survival function by a link function g(.)

$$g[S(t;z)]=g[S_0(t)]+\beta'z$$

where, $S_0(t)=S(t;0)$ is the baseline survival function and β is a vector of parameters to be estimated for covariates z. Within this framework, Younes and Lachin (1997) used the parameterized link function of Aranda-Ordaz (1981) where θ = 1 corresponds to the proportional odds model and θ→ 0 to the proportional hazards model (Royston and Parmar, 2002). For estimation, Younes and Lachin (1997) used B-splines, Shen (1998) used sieve maximum likelihood and monotone splines and Royston and Parmar (2002) used natural cubic splines to model $g[S_0(t)]$ within the Aranda-Ordaz family of link function.

2.2. Parametric survival models with splines

Splines are flexible mathematical functions defined by piecewise polynomials and used in regression type models for non-linear effects. The points at which the polynomials join are called knots. Several techniques have been developed for exploring the functional forms. The most common splines used in practice are cubic splines. However, splines can be of any degree, n. Also, function is forced to have continuous 0th, 1st and 2nd derivatives (Lambert and Royston, 2009). In survival analysis, the splines are used in different aspects in several studies such as Hastie and Tibshirani (1990a,b), Gamerman and West (1987), O'Sullivan (1988), Durrelman and Simon (1989), Sleeper and Harrington (1990), Zucker and Karr (1990), Kooperberg and Stone (1992), Gray (1992, 1994), Rosenberg (1995), Nelson et al. (2007), Govindarajulu et al. (2009), Lambert et al. (2010), Andersson et al. (2011), Bantis et al. (2012), and Rutherford et al. (2015).

The advantages of parametric survival models can be given as: the survival and hazard functions can be derived and manipulated, they do not have an assumption of proportional hazards, they allow prediction at any time point for any set of covariates and they can use of restricted cubic splines for hazard function. Besides, the key point of parametric survival models is the assumption of the survival time distribution. This creates the need of more flexible models. One of the alternatives is modelling the baseline cumulative odds or hazard function as natural cubic spline function of log time (Royston and Parmar, 2002).

The one of the most common distribution of survival time is Weibull distribution. Suppose that T is random variable having a Weibull distribution with characteristic life μ and shape parameter p. Then the log cumulative hazard function is given by,

$$lnH(t)=ln\left[\left(\frac{t}{\mu}\right)^{p}\right]=px-pln\mu=\frac{x-ln\mu}{\sigma}=\gamma_0+\gamma_1 x$$

which is linear in x=lnt. If the distribution of T departs from Weibull then lnH(t) will be related to x by a non-linear function $s\equiv s(x,\gamma)$ where survival function S(t) is exp(-exps) (Royston and Parmar, 2002).

The distribution of the survival time may be different than most common ones, and then more flexible parametric models can be used. The approach taken by Royston and Parmar (2002) is to model the logarithm of the baseline cumulative odds or hazard function as a natural cubic spline function of log time, so the general function $s(x,\gamma)$ is approximated by a spline. The PH spline model with fixed covariate vector z may be written (Royston and Parmar, 2002)

$$g[S(t;z)] = ln\{-lnS(t;z)\} = lnH(t;z) = lnH_0(t) + \beta'z = s(x;\gamma) + \beta'z$$

whereas for the PO spline model,

$$g[S(t;z)] = ln\{S(t;z)^{-1} - 1\} = lnO(t;z) = lnO_0(t) + \beta'x = s(x;\gamma) + \beta'z$$

The restricted cubic splines are defined as cubic splines constrained to be linear beyond boundary knots first knot and final knot. Restricted cubic splines with K knots can be fit by creating K-1 derived variables. For knots k_1, ..., k_K, a restricted cubic spline function can be written as

$$s(x) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{K-1} z_{K-1}.$$

A restricted cubic spline function of lnt, with knots k_0, can be written as $s\{\ln(t)\, I\, \gamma,\ k_0\}$. This is then used for baseline log cumulative hazard in PH model. The estimation of the models is done by using full likelihood which is discussed by Royston and Parmar (2002) and Lambert and Rosyton (2009) in detail.

3. Application: Gastric cancer

Gastric cancer is the third most common cause of cancer-related death in the world and it remains difficult to cure. In this study, we consider patients who were diagnosed with gastric cancer. The data used here is taken from Eroglu et al. (1997), but some of changes were made in the original data set, covariates and categories of the covariates to apply flexible parametric models. In the following analysis death is the endpoint of interest. The survival time (min=1, max=67) is measured in months and the mean survival time is obtained as 42.33±2.94. Patients who were still alive at the end of the follow-up period were treated as censored observations. The complete data set consists of 106 patients, of which 36.8% are censored. In the concept of analysis, prognostic factors which affect the survival time of gastric cancer patients are tried to be determined by flexible parametric survival models.

The mean age is obtained as 56.68 ± 1.2. The information about the covariates used in the following study is given in Table 1.

Table 1: The covariates of the gastric cancer data

Covariates		Number of total observations	Number of failed observations	Number of censored observations
Chemotherapy	No	11 (%10.4)	1 (%2.6)	10 (%14.9)
	Yes	95 (%89.6)	38 (%97.4)	57 (%85.1)
Pathological stage	1	14 (%13.2)	2 (%5.1)	12 (%17.9)
	2	23 (%21.7)	6 (%15.4)	17 (%25.4)
	3	69 (%65.1)	31 (%79.5)	38 (%56.7)
Sex	Female	33 (%31.1)	12 (%30.8)	21 (%31.3)
	Male	73 (%68.9)	27 (%69.2)	46 (%68.7)
Metastasis	No	74 (%69.8)	16 (%41.0)	58 (%89.6)
	Yes	32 (%30.2)	23 (%59.0)	9 (%13.4)
Smoking	No	54 (%50.9)	20 (%51.3)	34 (%50.7)
	Yes	52 (%49.1)	19 (%48.7)	33 (%49.3)
Alcohol	No	94 (%88.7)	35 (%89.7)	59 (%88.1)
	Yes	12 (%11.3)	4 (%10.3)	8 (%11.9)
Ulcer treatment	No	61 (%57.5)	25 (64.1)	36 (53.7)
	Yes	45 (%42.5)	14 (35.9)	31 (46.3)
Family history	No	79 (%74.5)	27 (%69.2)	52 (%77.6)
	Yes	27 (%25.5)	12 (%30.8)	15 (%22.4)
Tumor stage	1	27 (%25.5)	5 (%12.8)	22 (%32.8)
	2	79 (%74.5)	34 (%87.2)	45 (%67.2)
Muscle contraction	No	102 (%96.2)	37 (%94.9)	65 (%97.0)
	Yes	4 (%3.8)	2 (%5.1)	2 (%3.0)
Radiotherapy	No	34 (%32.1)	14 (%35.9)	20 (%29.9)
	Yes	72 (%67.9)	25 (%64.1)	47 (%70.1)

In the non-parametric approach to survival analysis we provide the estimates of Kaplan-Meier (KM) survival function and the log-rank test. The log-rank test allows for testing the equality of survival functions different groups. We observe that the equality of the survival functions of metastasis (p=0.000), family history (p=0.032) and muscle contraction (p=0.000) are rejected wheras it is not rejected for the others at a 95% confidence level.

Firstly, Cox PH model is applied to data set and AIC is obtained as 315.672. According to this model, metastasis and also muscle contraction are found significant at 95% confidence level. Patients with metastasis have 2.95 times the hazard for the patients without metastasis. Patients with muscle contraction have 8.358 times the hazard for the patients without muscle contraction.

The PH assumption of Cox PH model is shown by a correlation analysis of Schoenfeld (1982) residuals with time. All the covariates except metastasis (p=0.0001) hold the PH assumption. Therefore PH assumption does not satisfy for this data set and the results of Cox PH model becomes suspicious. Then flexible parametric survival models are applied to the data set as an alternative survival model. Here the Royston (2001), Royston and Parmar (2002) and Lambert and Rosyton (2009) approach is used. They use cubic splines and suggest selecting the df for the spline part of the model by minimizing the Akaike Information Criterion (AIC). The AIC may also be used to select the scale for the model. Table 2 shows the AICs for the gastric cancer data for PH and PO models with between 1 and 6 df (0 and 5 knots).

Table 2: The AIC values for several spline survival models for the gastric cancer data

Number of knots	0	1	2	3	4	5
AIC(PH model)	190.9251	190.0635	187.7287*	188.5921	188.5103	190.8589
AIC(PO model)	191.3755	192.1880	189.8785	190.4991	190.2563	192.6175

The model minimizing the AIC is the PH model with 2 knots for which AIC = 187.7287. The PH model with 2 knots is significant (p=0.0008) and the results are given by Table 3.

Table 3: The Results of PH model with spline

Covariate		$\hat{\beta}$	Std. Error	p-value	Hazard ratio	95% Confidence interval for hazard ratio	
Age		-0.004	0.019	0.837	0.996	-0.041	0.033
Chemotherapy	No Yes	0.821	1.233	0.506	2.272	-1.596	3.237
Pathological stage	1 2 3	0.296 0.519	1.078 1.234	0.784 0.674	1.344 1.680	-1.819 -1.900	2.410 2.938
Sex	Female Male	-0.427	0.529	0.419	0.652	-1.463	0.608
Metastasis	No Yes	1.129	0.394	0.004*	3.092	0.356	1.902
Smoking	No Yes	0.275	0.464	0.553	1.317	-0.634	1.184
Ulcer treatment	No Yes	-0.478	0.417	0.252	0.620	-1.294	0.339
Family history	1 2	0.753	0.402	0.061	2.124	-0.035	1.541
Alcohol	No Yes	0.207	0.631	0.743	1.229	-1.030	1.444
Tumor stage	1 2	0.560	0.831	0.500	1.751	-1.068	2.189
Muscle contraction	No Yes	2.449	1.011	0.015*	11.572	0.467	4.430
Radiotherapy	No Yes	-0.461	0.420	0.273	0.631	-1.285	0.363

According to the PH model with splines (k=2), metastasis and muscle contraction are found as important risk factors which effect the death risk of the patient. Patients with metastasis have 3.01 times the hazard for the patients without metastasis. Patients with muscle contraction have 11.57 times the hazard for the patients without muscle contraction. The significance of the covariates is same with Cox PH model; however the difference in hazard ratios and model fit is distinctly obvious. This difference is so important that it cannot be ignored in survival data sets.

4. Conclusion

Survival data is generally analyzed by semi-parametric and parametric survival models. The semi-parametric model named as Cox PH model allows the distribution of the survival time to be unknown. The baseline hazard is not necessary to estimate hazard ratio. The model may become invalid in case of non-proportional hazards and it is less consistent with theoretical survival function. Parametric survival models are well known models, since it completely specify hazard and survival functions and may possibly predict time-quantile. However, there is an assumption on underlying distribution. The disadvantage of these models brings about the suggestion of flexible parametric survival models. These models use becomes popular especially in flexibility by increasing the degrees of freedom of the spline functions in estimating hazard function. In this study, Cox PH model and also flexible survival models with splines (PH with spline and PO with spline) are used in analyzing gastric cancer data.

The models are tried to determine the possible effects of age, chemotherapy, pathological stage, sex, metastasis, smoking, ulcer treatment, family history, alcohol, tumor stage, muscle contraction and radiotherapy on survival. The best model is obtained as PH model with splines within the models taken into consideration in this study. Metastasis and muscle contraction are found as important risk factors in this cancer data.

Acknowledgment

We would like to thank Dr. Aydan Eroğlu for permission to use the data set.

References

Andersson T, Dickman P, Eloranta S, and Lambert P (2011). Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC Medical Research Methodology, 11(1). https://doi.org/10.1186/1471-2288-11-96

Aranda-Ordaz FJ (1981). On two families of transformations to additivity for binary response data. Biometrika, 68(2): 357–363.

Bantis L, Tsimikas JV, and Georgiou SD (2012). Survival estimation through the cumulative hazard function with monotone natural cubic splines. Lifetime Data Analysis, 18(3): 364–396.

Bennett S (1983). Analysis of survival data by the proportional odds model. Statistics in Medicine, 2(2): 273–277.

Collett D (1994).Modeling survival data in medical research. Chapman and Hall, London, UK.

Cox DR (1972). Regression models and life-tables. Journal of the Royal Statistical Society, 34(2): 187-220.

Crowder MJ, Kimber AC, Smith RL, and Sweeting TJ (1991). Statistical analysis of reliability data. Chapman and Hall, London, UK.

Dinse GE and Lagakos SW (1983). Regression analysis of tumor prevalence data. Journal of the Royal Statistical Society, 32(3): 236-248.

Durrelman S and Simon R (1989). Flexible regression models with cubic splines. Statistics in Medicine, 8(5): 551-561.

Eroglu A, Altınok M, Ozgen K, and Sertkaya D (1997). A multivariate analysis of clinical and pathological variable in survival after resection of gastric cancer. Turkiye Klinikleri Journal of Case Reports, 15(1): 15-20.

Gamerman D and West M (1987). An application of dynamic survival models in unemployment studies. The Statistician, 36(2/3): 269-274.

Govindarajulu US, Malloy EJ, Ganguli B, Spiegelman D, and Eisen EA (2009). The comparison of alternative smoothing methods for fitting non-linear exposure–response relationships with Cox models in a simulation study. The International Journal of Biostatistics, 5(1): 1-21.

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

Gray RJ (1994). Spline-based tests in survival analysis. Biometrics, 50(3): 640-652.

Hastie T and Tibshirani R (1990a). Exploring the nature of covariate effects in the proportional hazards model. Biometrics, 46(4): 1005-1016.

Hastie TJ and Tibshirani RJ (1990b). Generalized additive models. CRC press, Boca Raton, USA.

Kalbfleisch JD and Prentice RL (1980). The statistical analysis of failure time data. A John Wiley and Sons, New York, USA.

Kirmani SNUA and Gupta RC (2001). On the proportional odds model in survival analysis. Annals of the Institute of Statistical Mathematics, 53(2): 203-216.

Kooperberg C and Stone CJ (1992). Logspline density estimation for censored data. Journal of Computational and Graphical Statistics, 1(4): 301-328.

Lambert PC and Royston P (2009). Further development of flexible parametric models for survival analysis. Stata Journal, 9(2): 265–290.

Lambert PC, Dickman PW, Nelson CP, and Royston P (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models. Statistics in Medicine, 29(7-8): 885–895.

Nelson CP, Lambert PC, Squire IB, and Jones DR (2007). Flexible parametric models for relative survival, with application in coronary heart disease. Statistics in Medicine, 26(30): 5486-5498.

O'Sullivan F (1988). Nonparametric estimation of relative risk using splines and cross-validation. SIAM Journal on Scientific and Statistical Computing, 9(3): 531-542.

Rodriguez G (2007). Lecture notes on generalized linear models. Princeton University, New Jersey, USA. Available online at: http://data.princeton.edu/wws509/notes/c7

Rosenberg PS (1995). Hazard function estimation using B-splines. Biometrics, 51(3):874–887.

Rossini AJ and Tsiatis AA (1996). A semiparametric proportional odds model for the analysis of current status data. Journal of the American Statistical Association, 91(434): 713–721.

Royston P (2001). Flexible parametric alternatives to the Cox model, and the more. The Stata Journal, 1(1): 1-28.

Royston P and Parmar MK (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine, 21(15): 2175– 2197.

Rutherford MJ, Crowther MJ, and Lambert PC (2015). Using restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: A simulation study. Journal of Statistical Computational and Simulation, 85(4): 777-793.

Schoenfeld D (1982). Partial residuals for the proportional hazards model. Biometrika, 69: 51-55.

Shen XT (1998). Proportional odds regression and sieve maximum likelihood estimation. Biometrika, 85(1): 165–177.

Sleeper LA and Harrington DP (1990). Regression splines in the cox model with application to covariate effects in liver disease. Journal of the American Statistical Association, 85(412): 941-949.

Yang S and Prentice RL (1999). Semiparametric inference in the proportional odds regression model. Journal of the American Statistical Association, 94(445): 125–136.

Younes N and Lachin J (1997). Link-based models for survival data with interval and continuous time censoring. Biometrics, 53(4): 1199–1211.

Zucker DM and Karr AF (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. The Annals of Statistics, 18(1): 329-353.