Hacettepe Journal of Mathematics and Statistics
Volume 36 (2) (2007), 157 – 167

COX REGRESSION MODELS WITH NONPROPORTIONAL HAZARDS APPLIED TO LUNG CANCER SURVIVAL DATA

Nihal Ata[*] and M. Tekin Sözer [†]

Received 02 : 08 : 2006 : Accepted 28 : 09 : 2007

Abstract

The Cox regression model, which is widely used for the analysis of treatment and prognostic effects with censored survival data, makes the assumption of constant hazard ratio. In the violation of this assumption, different methods should be used to deal with non-proportionality of hazards. In this study, the stratified Cox regression model and extended Cox regression model, which uses time dependent covariate terms with fixed functions of time are discussed. The results are illustrated by an analysis of lung cancer data in order to compare these methods with respect to Cox regression model in the presence of nonproportional hazards.

Keywords: Cox regression model, Hazard ratio, Non-proportional hazards, Stratified Cox regression model, Extended Cox regression model, Time dependent-covariate, Lung cancer.

2000 AMS Classification: 62 N 01, 62 P 10.

1. Introduction

Survival analysis is a class of statistical methods for studying the occurrence and timing of events and is useful for studying many kinds of events in both the social and natural sciences.

Survival data have some features that are difficult to handle with traditional statistical methods: censoring and time-dependent 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 are proportional over the period of follow-up time. If hazard ratios vary with time, then the assumption of

[*]Hacettepe University, Faculty of Science, Department of Statistics, 06800, Beytepe, Ankara, Turkey. E-mail: `nihalata@hacettepe.edu.tr`

[†]Hacettepe University, Faculty of Science, Department of Actuarial Sciences, 06800 Beytepe, Ankara, Turkey. E-mail: `sozer@hacettepe.edu.tr`

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 paper, we explain the proportional hazards assumption and investigate and discuss the methods which can be used when the hazards are nonproportional.

2. Cox Regression Model

The most common approach to model covariate effects on survival is the Cox regression model, which takes into account the effect of censored observations [4]. Although the model is based on the assumption of proportional hazards, no particular form of probability distribution is assumed for the survival times. The model is therefore referred to as a semi-parametric model.

Let $x_1, x_2, \ldots, x_p$ be the values of p covariates $X_1, X_2, \ldots, X_p$. According to the Cox regression model, the hazard function is given as follows:

$$h(t) = h_0(t) \exp(\sum_{i=1}^{p} \beta_i x_i), \tag{2.1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is a $1 \times p$ vector of regression parameters and $h_0(t)$ is the baseline hazard function at that time. Coefficient vectors of the covariates are estimated using a maximum likelihood (ML) procedure. ML estimates are obtained by maximizing a (partial) likelihood function (L) [3].

3. Assessment of Proportional Hazards Assumption

A key assumption of the Cox regression model is proportional hazards. The proportional hazards assumption means that the hazard ratio is constant over time, or that the hazard for an individual is proportional to the hazard for any other individual [13].

Let $x^* = (x_1^*, x_2^*, \ldots, x_p^*)$ and $x = (x_1, x_2, \ldots, x_p)$ be the covariates of two individuals. The hazard ratio is given as follows:

$$\exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(x_i^* - x_i\right)\right]. \tag{3.1}$$

When the value of the exponential expression for the estimated hazard ratio is a constant that does not depend on time, the proportional hazards assumption is satisfied.

An assessment of the proportional hazards assumption can be done by many numerical or graphical approaches. None of these approaches are known to be better than the others in finding out whether the hazards are proportional or not. Interpreting graphical plots can be arbitrary. The conclusions are highly dependent on the subjectivity of the researcher. Some of these graphical approaches are log-minus-log survival plots of survival functions, a plot of survival curves based on the Cox regression model and Kaplan-Meier estimates for each group, a plot of cumulative baseline hazards in different groups [2], a plot of difference of the log cumulative baseline hazard versus time, a smoothed plot of the ratio of log-cumulative hazard rates versus time, a smoothed plot of scaled Schoenfeld residuals versus time and a plot of the estimated cumulative hazard versus the number of failures [1].

There are various numerical approaches in finding non-proportionality, such as a test including a time dependent covariate in the model [4], a test based on the Schoenfeld partial residuals [12] which is a measure of the difference between the observed and expected value of the covariate at each time [13], a test based on a comparison of different generalized rank estimators of the hazard ratio [6], and a test based on a semi-parametric generalization of the Cox regression model [9, 11].

4. Nonproportional Hazards Models

Since the Cox regression model relies on the hazards being proportional, i.e. on the effect of a given covariate not changing over time, it is very important to verify that the covariates satisfy the assumption of proportionality. If this assumption is violated, the simple Cox regression model is invalid and more complicated analyses such as the stratified Cox regression model or the extended Cox regression model are required.

4.1. The Stratified Cox Regression Model. The stratified Cox regression model is a modification of the Cox regression model by the stratification of a covariate that does not satisfy the proportional hazards assumption. Covariates that are assumed to satisfy the proportional hazards assumption are included in the model, whereas the predictor being stratified is not included.

Let k covariates fail to satisfy the proportional hazards assumption, and p covariates satisfy proportional hazards assumption. The covariates not satisfying the proportional hazards assumption are denoted by $Z_1, Z_2, \ldots, Z_k$, and the covariates satisfying the proportional hazards assumption are denoted by $X_1, X_2, \ldots, X_p$. To form the stratified Cox regression model, a new variable is defined from z variables and denoted by z^*. The stratification variable z^* has k^* categories, where k^* is the total number of combinations (strata) formed after categorizing each of z's. There are interaction and no-interaction models defined in the concept of the stratified Cox regression model [7, 8].

4.1.1. *No-Interaction Model.* The no-interaction model is defined as follows:

$$h_g(t,x) = h_{0g}(t)\exp\left[\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right], \ g = 1, 2, \ldots, k^*, \tag{4.1}$$

where the subscript g represents the strata. The strata are the different categorizations of the stratum variable. The variable z^* is not implicitly included in the model, whereas the x's which are assumed to satisfy the proportional hazards assumption are included in the model. The baseline hazard function, $h_{0g}(t)$, is different for each stratum. However, the coefficients $\beta_1, \beta_2, \ldots, \beta_p$ are the same for each stratum. Since the coefficients of the x's are the same for each stratum, the hazard ratios are same for each stratum.

To obtain estimates of the regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$, a likelihood function L that is obtained by multiplying together the likelihood functions for each stratum is maximized [8].

4.1.2. *Interaction Model.* The interaction model can be formed in two ways according to the number of variables that do not satisfy the proportional hazards assumption.

(i). Let there be one variable that does not satisfy the proportional hazards assumption.

The data set can be stratified into k strata according to the variable that does not satisfy the proportional hazards assumption. In this case, the interaction model is defined as follows:

$$h_g(t,x) = h_{0g}(t)\exp\left[\beta_{1g} x_1 + \beta_{2g} x_2 + \cdots + \beta_{pg} x_p\right], \ g = 1, 2, \ldots, k \tag{4.2}$$

The covariates and the products of each of these covariates with the variable that does not satisfy the proportional hazards assumption can be included in the model. In this case, an alternative interaction model is defined as follows:

$$\begin{aligned} h_g(t,x) = h_{0g}(t)\exp\big[&\beta_1^* x_1 + \beta_2^* x_2 + \cdots + \beta_p^* x_p \\ &+ \beta_{p+1}^*(x_1 \times z) + \cdots + \beta_{2p}^*(x_p \times z)\big], \end{aligned} \tag{4.3}$$

where the coefficients β^* do not have a subscript g representing the stratum [8].

There is a relation between the coefficients of the interaction model and the alternative interaction model. Let $z = (0, 1, 2, \ldots, k-1)$ be the only variable that does not satisfy the proportional hazards assumption and denote its number of strata by $g = 1, 2, \dots, k$. The general relation between the β coefficients of these models are given below:

$$\begin{aligned} &\beta_{1k} = \beta_1^*, \dots, \beta_{pk} = \beta_p^*, \ k = 1, \\ &\beta_{1k} = (k-1)(\beta_1^* + \beta_{p+1}^*), \dots, \beta_{pk} = (k-1)(\beta_p^* + \beta_{2p}^*), \ k > 1. \end{aligned} \tag{4.4}$$

(ii). Let there be two or more variables that do not satisfy proportional hazard.

The new strata variable is represented by $z^* = (0, 1, 2, \dots, k^* - 1)$. Here k^* is the product of the levels of the variables that do not satisfy the proportional hazards assumption. In this case the alternate interaction and interaction models are given by (4.5) and (4.6), respectively.

$$h_g(t, x) = h_{0g}(t) \exp\left[\beta_{1g} x_1 + \cdots + \beta_{pg} x_p\right], \ g = 1, 2, \dots, k^*, \tag{4.5}$$

$$\begin{aligned} h_g(t, x) = h_{0g}(t) \exp\big[&\beta_1 x_1 + \cdots + \beta_p x_p + \beta_{11}(z_1^* \times x_1) + \cdots \\ &+ \beta_{1(k^*-1)}(z_{k^*-1}^* \times x_1) + \cdots + \beta_{p1}(z_1^* \times x_p) + \cdots \\ &+ \beta_{p(k^*-1)}(z_{k^*-1}^* \times x_p)\big]. \end{aligned} \tag{4.6}$$

The no-interaction assumption

The stratified Cox regression model contains regression coefficients that do not vary over the strata. This property of the model is known as the no-interaction assumption. If interaction is allowed for, different coefficients for each of the stratum are obtained.

The test that is used to examine the no-interaction assumption is the likelihood ratio test statistics. For this test statistic, log likelihood functions of the interaction and no-interaction models are used. The interaction model differs from the no-interaction model by containing product terms. Thus, the null hypothesis is that the coefficients of the product terms are equal to zero. The likelihood ratio test statistic shows a Chi-square distribution with $p(k^* - 1)$ degrees of freedom under the null hypothesis [8].

4.2. Extended Cox Regression Model. In the Cox regression model, there can be variables which involve t. Such variables are called *time-dependent* variables. A time-dependent variable is defined as any variable whose value for a given subject may differ over time (t). 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 [5, 10].

In this model, the Cox regression model is extend to a model which contains time-dependent covariates and the product of these covariates with a function of time. Let $x_1, x_2, \dots, x_{p_1}$ be time-independent covariates, $x_1(t), x_2(t), \dots, x_{p_2}(t)$ the time-dependent covariates and set $x(t) = (x_1, x_2, \dots, x_{p_1}, x_1(t), x_2(t), \dots, x_{p_2}(t))$. The extended Cox regression model is defined as follows:

$$h\left(t, x(t)\right) = h_0(t) \exp\left[\sum_{i=1}^{p_1} \beta_i x_i + \sum_{j=1}^{p_2} \delta_j x_j(t)\right], \tag{4.7}$$

where β and δ are the coefficient vectors of the covariates, p_1 is the number of covariates that satisfy the proportional hazards assumption and p_2 the number of covariates not satisfing the proportional hazards assumption. The computations for this model are more complicated than for the Cox regression model, because the risk sets used to form the likelihood function are more complicated with time dependent variables.

Two sets of predictors, $x(t)$ and $x^*(t)$, identify two specifications at time t for the combined set of predictors containing both time-independent and time-dependent variables. The hazard ratio for the extended Cox regression model which is a function of time is given as follows [7,8]:

$$\exp\left[\sum_{i=1}^{p_1}\hat{\beta}_i\left(x_i^{*}-x_i\right)+\sum_{j=1}^{p_2}\hat{\delta}_j\left(x_j^{*}(t)-x_j(t)\right)\right]. \tag{4.8}$$

While investigating the proportional hazards assumption, the extended Cox regression model is used and in this case the model is given as follows:

$$h\left(t,x(t)\right)=h_0(t)\exp\left[\sum_{j=1}^{p}\beta_j x_j+\sum_{j=1}^{p}\delta_j x_j g_j(t)\right], \tag{4.9}$$

where $g_j(t)$ is defined as a function of time. In this model, the critical decision is the form that the functions $g_j(t)$ should take. The possible forms of $g_j(t)$ are given below:

(i) All the $g_j(t)$ can be zero.
(ii) $g_j(t)=t$.
(iii) $g_j(t)=\log t$
(iv) $g_j(t)$ is a Heavyside (step) function. When this function is used, we get constant hazard ratios for different time intervals [8, 10, 13].

Let C be the only covariate. Then the hazard ratios for the extended Cox regression model with one step (Heavyside) function and two step functions are given in Table 1.

Table 1. Hazard Ratios for the Extended Cox Regression with Heavyside Functions

	Step function	Time interval	Hazard ratio
One step function	$g(t)=\begin{cases}1 & \text{if } t\geq t_0\\ 0 & \text{if } t<t_0\end{cases}$	$t\geq t_0$	$\exp\left(\hat{\beta}+\hat{\delta}\right)$
		$t<t_0$	$\exp\left(\hat{\beta}\right)$
Two step functions	$g_1(t)=\begin{cases}1 & \text{if } t\geq t_0\\ 0 & \text{if } t<t_0\end{cases}$	$t\geq t_0$	$\exp(\hat{\delta}_1)$
	$g_2(t)=\begin{cases}1 & \text{if } t<t_0\\ 0 & \text{if } t\geq t_0\end{cases}$	$t<t_0$	$\exp(\hat{\delta}_2)$

More than two step functions can be used, and in this way we get constant hazard ratios within different time intervals. The model does not involve a main effect term and is defined as follows:

$$\begin{aligned}h(t,x(t))=h_0(t)\exp\big[&\delta_1(C\times g_1(t))+\delta_2(C\times g_2(t))+\delta_3(C\times g_3(t))+\cdots\\ &+\delta_k(C\times g_k(t))\big],\end{aligned} \tag{4.10}$$

where $t_1,t_2,\ldots,t_k$ are time intervals and

$$g_1(t)=\begin{cases}1 & \text{if } 0\leq t<t_1\\ 0 & \text{otherwise}\end{cases},\ldots,g_k(t)=\begin{cases}1 & \text{if } t_{k-1}\leq t<t_k\\ 0 & \text{otherwise}\end{cases} \tag{4.11}$$

are step functions. The hazard ratios for this model are as given below:

$$\exp(\hat{\delta}_1)\text{ for }0\leq t<t_1,\ldots,\exp(\hat{\delta}_k)\text{ for }t_{k-1}\leq t<t_k. \tag{4.12}$$

5. An Application to Lung Cancer Data

Patients diagnosed with lung cancer were taken into this study based on Cox regression analysis. In the following analysis, the time after the operation of the recurrence of the illness is the endpoint of interest. This variable is measured in months. There was an 8-year follow-up period for the patients. Patients who were still alive at the end of the follow-up period were treated as censored observations. The complete data set consists of 236 observations, of which 60.2% are censored. The aim of of the analysis was to try and determine prognostic factors that affect the survival time of lung cancer patients by using the statistical analysis software SAS 8.2.

In the following study, qualitative covariates such as

five-level covariate age ($x_1 = 39, 40 - 49, 50 - 59, 60 - 69, 70$),

four-level covariate cigarette consumption (package per year, $x_2 = 5, 6$ - $30, 31$ - $60, 61$),

two-level covariate extended resection ($x_3 = \text{not present}, \text{present}$),

four-level covariate tumour size (mm, $x_4 = 30, 31$ - $40, 41$ - $50, 51$),

four-level covariate tumour stage (x_5 stage 1; stage 2, stage 3, stage 4), and

two-level covariate invasion ($x_6 = \text{not present}, \text{present}$)

were used. Since the calculations are quite difficult, in applications the use of covariates with two or more levels are usually avoided.

The example of real data permits a focused comparison of various competitive techniques related to the Cox regression model which are useful in the presence of nonproportional hazards.

Firstly, the Cox regression model was applied to the data set before investigating the proportional hazards assumption. Cigarette consumption, tumour size and tumour stage were found to be significant at the 95% confidence level. Then the proportional hazards assumption was assessed by a statistical test. This test was accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of the individual failure times. If the proportional hazards assumption is met then the correlation should be near zero [13]. It was found that the extended resection variable does not satisfy the proportional hazards assumption. It is shown by a correlation analysis of the partial residuals with time that the p value obtained for this variable is 0.0041. For other variables (all levels of each covariate) used in the Cox regression model the proportional hazards assumption holds. The same conclusion can be derived from the log-minus-log plots.

The stratified Cox regression model can be described as a no-interaction model and as an interaction model. In the no interaction model, the extended resection variable which causes nonproportional hazards was used as the strata variable and Cox regression analysis carried out. The results obtainrd are given in Table 2.

In the no-interaction model, from the p values, cigarette consumption, tumour size and tumour stage are found to be important risk factors which affect the failure. $x_2(4)$, $x_4(4)$, $x_5(3)$ and $x_5(4)$ are the important levels. Patients whose cigarette consumption is equal to or higher than 61 have 3.6 times the hazard faced by patients whose cigarette consumption is less than or equal to 5. Patients whose tumour size is greater than or equal to 50 mm have 2.2 times the hazard faced by patients whose tumour size is equal to or less than 30 mm. Patients whose tumour stage is 3 have 2.2 times, and patients whose tumour stage is 4 have 6.7 times the hazard faced by patients whose tumour stage is 1.

Table 2. Results for the No-interaction and Interaction Models

Variable	No-interaction model		Interaction model			
			Strata 1		Strata 2	
	exp(β)	p	exp(β)	p	exp(β)	p
Age		0.5659		0.0938		0.4118
$x_1(2)$	0.6640	0.4573	0.5492	0.3392	3.5068	0.4439
$x_1(3)$	1.1599	0.7659	1.0649	0.9128	3.2267	0.3372
$x_1(4)$	0.8531	0.7546	0.6619	0.4706	4.9234	0.2229
$x_1(5)$	1.0626	0.9150	1.7847	0.3790	1.0129	0.9928
Cigarette csp.		0.0130		0.2702		0.8741
$x_2(2)$	1.8010	0.3052	1.9974	0.0214	16739.465	0.9365
$x_2(3)$	1.5173	0.4507	1.4713	0.5056	10228.862	0.9397
$x_2(4)$	3.6366	0.0261	3.8261	0.0315	15671.379	0.9369
Tumour size		0.0041		0.0127		0.7915
$x_4(2)$	0.6765	0.2822	0.6828	0.3726	0.6833	0.6369
$x_4(3)$	1.6295	0.1608	1.7363	0.1843	1.6956	0.5995
$x_4(4)$	2.2147	0.0075	2.5091	0.0098	1.1960	0.8447
Tumour stage		0.0000		0.0000		0.2735
$x_5(2)$	1.2492	0.5128	1.0274	0.9446	2.5197	0.4535
$x_5(3)$	2.2451	0.0078	2.9030	0.0012	1.6923	0.6593
$x_5(4)$	6.7165	0.0000	7.6626	0.0000	13.6475	0.0852
Invasion (x_6)	1.2902	0.3512	1.3868	0.3064	1.4608	0.6465

For the interaction model, the data set is divided into two strata according to the extended resection variable. The first stratum consists of patients who don't have an extended resection and the second stratum consists of patients who do. For the first stratum, from the p values, cigarette consumption, tumour size and tumour stage are found to be important risk factors which affect the failure. Patients whose cigarette consumption is equal to or higher than 61 have 3.8 times the hazard faced by patients whose cigarette consumption is less than or equal to 5. Patients whose tumour size is greater than or equal to 50 have 2.5 times the hazard faced by patients whose tumour size is equal to or less than 30. Patients whose tumour stage is 3 have 2.9 times, and patients whose tumour stage is 4 have 7.6 times the hazard faced by patients whose tumour stage is 1. For the second stratum, when we investigate the p values, none of the covariates is an important risk factor at the 95% confidence level. Results for the alternative interaction model are given by Table 3.

When we investigate the interaction model and the alternative interaction model, the relation between the coefficients of these two models is explored. In the alternative interaction model, we have the same findings as with the interaction model.

To evaluate the no-interaction assumption, a likelihood ratio test that compares the no-interaction model to the (full) interaction model is performed. The null hypothesis is that the no-interaction assumption is satisfied. The test statistic which is given by the difference between the log-likelihood statistics for the no-interaction and interaction models was found to be 14.607. This statistic is approximately chi-square with 14 degrees of freedom under the null hypothesis. Since the null hypothesis could not be rejected, this indicates that the no-interaction model is to be preferred to the interaction models.

Table 3. Results for the Alternative Interaction Model

Variable	$\exp(\beta)$	p	Variable	$\exp(\beta)$	p
Age					
$x_1(2)$	0.5492	0.3392	$x_1(2) * x_3$	6.3858	0.2907
$x_1(3)$	1.0649	0.9128	$x_1(3) * x_3$	3.0299	0.4113
$x_1(4)$	0.6619	0.4706	$x_1(4) * x_3$	7.4388	0.1597
$x_1(5)$	1.7847	0.3790	$x_1(5) * x_3$	0.5676	0.7185
Cigarette consumption					
$x_2(2)$	1.9974	0.2702	$x_2(2) * x_3$	3083.0061	0.9136
$x_2(3)$	1.4713	0.5056	$x_2(3) * x_3$	2557.5131	0.9156
$x_2(4)$	3.8261	0.0315	$x_2(4) * x_3$	1506.8052	0.9213
Tumour size					
$x_4(2)$	0.6828	0.3726	$x_4(2) * x_3$	1.0008	1.0000
$x_4(3)$	1.7363	0.1843	$x_4(3) * x_3$	0.9766	0.9826
$x_4(4)$	2.5091	0.0098	$x_4(4) * x_3$	0.4766	0.4498
Tumour stage					
$x_5(2)$	2.9030	0.0012	$x_5(2) * x_3$	2.4526	0.4876
$x_5(3)$	7.6626	0.0000	$x_5(3) * x_3$	0.5830	0.6629
$x_5(4)$	1.3868	0.3064	$x_5(4) * x_3$	1.7810	0.7128
Invasion (x_6)	1.0274	0.9446	$x_6 * x_3$	1.0534	0.9532

Further evidence of the proportional hazards assumption not being satisfied for the extended resection variable can be seen from a graph of the adjusted survival curves stratified by the extended resection variable. Survival curves of patients who have extended resection and those who do not have extended resection begin to diverge after 8 months. It was seen that two curves diverge greatly after 8 months. This indicates that the hazard ratio for the extended resection variable will be much closer to one early on, but quite different from one later on.

In order to illustrate the extended Cox regression model we use step functions, and the results we obtain are given in Table 4.

When the extended Cox regression model is used with one step function and with two step functions, the same results are obtained. Cigarette consumption, tumour size, tumour stage and extended resection are found to be important risk factors which affect the failure. Patients whose cigarette consumption is equal to or higher than 61 have 3.5 times the hazard faced by patients whose cigarette consumption is less than or equal to 5. Patients whose tumour size is greater than or equal to 50 mm. have 2.2 times the hazard faced by patients whose tumour size is equal to or less than 30 mm. Patients whose tumour stage is 3 have 2.2 times, and patients whose tumour stage is 4 have 6.5 times the hazard faced by patients whose tumour stage is 1. Patients who have extended resection have 4 times the hazard faced by patients who do not have extended resection.

Table 4. Results for Extended Cox Regression Model

	One Step Function		Two Step Functions	
Variable	$\exp(\beta)$	p	$\exp(\beta)$	p
Age				
$x_1(2)$	0.6706	0.4672	0.6706	0.4672
$x_1(3)$	1.1715	0.7503	1.1715	0.7503
$x_1(4)$	0.8808	0.8028	0.8808	0.8028
$x_1(5)$	1.0727	0.9016	1.0727	0.9016
Cigarette consumption				
$x_2(2)$	1.8023	0.3031	1.8023	0.3031
$x_2(3)$	1.5322	0.4389	1.5322	0.4389
$x_2(4)$	3.5645	0.0279	3.5645	0.0279
Tumour size				
$x_4(2)$	0.6644	0.2621	0.6644	0.2621
$x_4(3)$	1.6300	0.1585	1.6300	0.1585
$x_4(4)$	2.1879	0.0082	2.1879	0.0082
Tumour stage				
$x_5(2)$	1.2372	0.5289	1.2372	0.5289
$x_5(3)$	2.1848	0.0098	2.1848	0.0098
$x_5(4)$	6.5723	0.0000	6.5723	0.0000
Invasion (x_6)	1.2992	0.3338	1.2992	0.3338
Extended resection (x_3)	4.0452	0.0154	-	-
$x_3 * g(t)$	0.2040	0.0110	-	-
$x_3 * g_1(t)$	-	-	4.0452	0.0154
$x_3 * g_2(t)$	-	-	0.8252	0.5581

For different time intervals, hazard ratios are given in Table 5. Since hazard ratios are constant within these time intervals, the proportional hazards assumption is satisfied.

Table 5. Step Functions and Hazard Ratios

	Step function	Time interval	Hazard ratio
One step function	$g(t) = \begin{cases} 1 & \text{if } t \geq 8 \text{ months} \\ 0 & \text{if } t < 8 \text{ months} \end{cases}$	$t < 8$ months	4.0451
		$t \geq 8$ months	0.8251
Two step functions	$g_1(t) = \begin{cases} 1 & \text{if } t \geq 8 \text{ months} \\ 0 & \text{if } t < 8 \text{ months} \end{cases}$	$t < 8$ months	4.0451
	$g_2(t) = \begin{cases} 1 & \text{if } t < 8 \text{ months} \\ 0 & \text{if } t \geq 8 \text{ months} \end{cases}$	$t \geq 8$ months	0.8251

In survival analysis, comparisons between a number of possible models can also be made on the Akaike's information criterion (AIC) or -2log likelihood function (-2logL). The values of AIC and –2logL for the Cox regression model, extended Cox regression model and stratified Cox regresyon model are given in Table 6. The AIC values of the prognostic factors can be compared across different models [3].

Table 6. -2logL and AIC Values of the Cox Regression Model, Stratified Cox regression model and Extended Cox Regression Model

		Stratified Cox Regression Model					
	Cox Regression Model	No-Interaction Model	Interaction Model		Alternative Interaction Model		Extended Cox Regression Model
			Strata 1	Strata 2	Strata 1	Strata 2	
AIC	926.710	819.579	659.732	173.240	659.732	832.972	922.340
-2log L	896.710	791.579	631.732	145.240	631.732	776.972	890.340

Our study shows that, according to the AIC, using the stratified Cox regression model (no-interaction model) and extended Cox regression model gives more suitable results for survival data in the presence of nonproportional hazards.

6. Conclusion

Ignoring nonproportional hazards in an analysis can lead us to incorrect results. For this reason, before applying the Cox regression model to survival data, one should first check the proportional hazards assumption. In this study, some of the regression models that can be used in the presence of nonproportional hazards are considered. Stratified Cox regression models – interaction and alternative interaction models – are given, and the relations between the regression coefficients of these models are obtained. Also the Cox regression model is extended to allow time-dependent variables, this being called the extended Cox regression model.

All the regression models within the scope of this paper are applied to real survival data of lung cancer patients. When we take AIC - model selection criteria - into consideration, the no-interaction model and extended Cox regression model are found to be more appropriate models for survival data than the Cox regression model if the proportional hazards assumption does not hold. In these models, only the extended resection variable does not satisfy the proportional hazards assumption and cigarette consumption, tumour size and tumour stage are found to be important risk factors which affect the failure. However these models give almost the same hazard ratios.

In case of nonproportional hazards, using the stratified Cox regression and extended Cox regression models is found to be more appropriate than the simple Cox regression model. In our study this claim is supported by the application.

Acknowledgements

We would like to thank Prof. Dr. Hadi Akay and Dr. Ayten Kayi Cangir for permission to analyze their lung cancer data.

References

[1] Arjas, E. *A graphical method for assessing goodness of fit in Cox's proportional hazards model*, Journal of the American Statistical Association **83**, 204–212, 1988.

[2] Andersen, P. K. Borgan, G. and Keiding, N. *Linear nonparametric tests for comparison of counting process with application to censored survival data*, International Statistical Review **50**, 219–258, 1982.

[3] Collett, D. *Modelling Survival Data in Medical Research*, (Chapman&Hall, London, 1994).

[4] Cox, D. R., *Regression models and life-tables*, Journal of the Royal Statistical Society Series B **34**, 187–220, 1972.

[5] Fisher, L. D. ve Lin D. Y. *Time-dependent covariates in the Cox proportional hazards regression model*, Annual Review of Public Health **20**, 145–157, 1999.

[6] Gill, R. D. and Schumacher, M. *A simple test for the proportional hazards assumption*, Biometrika **74**, 289–300, 1987.

[7] Klein, John P. and Moeschberger, M. L. *Survival Analysis Techniques for Censored and Truncated Data*, (Springer, New York, 1997).

[8] Kleinbaum, D. G. *Survival Analysis: A Self-Learning Text*, (Springer, New York, 1996).

[9] Persson, I. *Essays on the assumption of proportional hazards in Cox regression*, http://www.diva-portal.org., 2002.

[10] Pettitt, A. N. and Daud, I. Bin *Investigating time dependence in Cox's proportional hazards model*, Applied Statistics **39**, 313–329, 1990.

[11] Schemper, M. *Cox analysis of survival data with nonproportional hazards functions*, The Statistician **41**, 455–465, 1992.

[12] Schoenfeld, D. *Partial residuals for the proportional hazards model*, Biometrika **69**, 551–55, 1982.

[13] Therneau, T. M. and Grambsch, P. M. *Modelling Survival Data: Extending the Cox Model*, (Springer, New York, 2000).