IOP Conference Series: Earth and Environmental Science

PAPER • **OPEN ACCESS**

Stratified-extended cox model in survival modeling of non-proportional hazard

To cite this article: D J Ratnaningsih *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **299** 012023

View the article online for updates and enhancements.

This content was downloaded from IP address 111.170.106.60 on 11/10/2019 at 00:41

IOP Conf. Series: Earth and Environmental Science **299** (2019) 012023 doi:10.1088/1755-1315/299/1/012023

Stratified-extended cox model in survival modeling of non-proportional hazard

D J Ratnaningsih[1,2*], A Saefuddin[2], A Kurnia[2] and I W Mangku[3]

[1] Departemen of Statistics, Universitas Terbuka, Banten, Indonesia
[2] Departemen of Statistics, IPB University, Bogor 16680, Indonesia
[3] Departemen of Mathematics, IPB University, Bogor 16680, Indonesia

*E-mail: dewijuliahr@gmail.com

Abstract. Cox proportional hazard model is frequently used in survival analysis. Cox proportional hazard model is time independent covariate while many models involve time as a dependent covariate causing incomplete proportional hazard assumption, known as non-proportional hazard. The proposed model in this paper was a non-proportional hazard involving time-independent and time-dependent covariates. The approaching model was carried out by joining a stratified Cox and extended Cox model termed as Stratified-Extended Cox (SE Cox) model. The simulation of the SE Cox model resulted in small MSE for the parameter estimates. In addition, the goodness of value was more appropriate compared to the existing non-proportional hazard model. Hence, the SE Cox model was applied to evaluate student persistence in Universitas Terbuka, Indonesia.

1. Introduction

The survival model used in many fields is the Cox model or Cox proportional hazard (Cox PH) model. Cox [8] and Cox & Oakes [9] developed the model to predict the hazard rate of an object with covariates at risk. The Cox model assumes that the risk level of an individual is proportional at all times known as Cox proportional hazard [18]. Therefore, the risk comparison in the Cox model is assumed to be constant and independent with respect to (w.r.t.) time.

In reality, some the covariates do not fulfill the assumption of time-independent and constant risk. Hence, time-dependent covariates produced a survival model of non-proportional hazard due to the change in time. In other words, the assumption required model is not fulfilling. The violation of the proportional hazard assumption has resulted in a bias parameter estimate [15].

Many types of research have been carried out for non-proportional hazard involving both times independent and time-dependent covariates. Methods developed to handle survival with non-proportional hazard are Cox stratified [5], [20], [1], and Cox extended [17], [26], [13], [2]. Usually, both models are implemented to analyse in health sciences [10], [29].

The Cox stratified model is a modification of the Cox model by stratifying covariates that do not meet the assumption [5], [20], [1]. Meanwhile, the covariates that fulfill Cox's assumptions are put into the model. The stratified covariates are observed as strata in the former model. Although the Cox stratified model produced the same parameter estimates in strata, the baseline hazard values in each stratum are different.

The extended Cox model is an extension of the Cox model containing time-dependent covariates or the multiplication of independent covariates with time functions [17], [26]. A time-dependent covariate

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

is a covariate in which the value changes over time. Although the baseline hazard values are the same, the parameter estimates are different over time.

In many cases, both types of covariates (time independent covariate and time-dependent covariate) are often occurred, for example in the case of data on persistence students at Universitas Terbuka (UT), Open and Distance Higher Education Program. The time-independent covariates are studied program of interest, gender, age, marital status, employment status, educational background, and domicile. In addition, the time-dependent covariate is the number of credits taken, and the number of registered courses per semester [25].

Both Cox stratified and extended models implemented on student learning persistence resulted in non-significant parameter estimate of influential covariates. Therefore, the use of both model separately is not appropriate to analyse in student persistence data. Hence, the proposed model in this paper combine between stratified Cox and extended Cox called Stratified Extended Cox (SE Cox) model. The model is implemented in analysing UT students' persistence data.

2. Stratified-Extended Cox Model

2.1. Definition model

The Stratified Extended Cox (SE Cox) model is a proposed model for dealing with non proportional hazard problem in Cox model caused by involving two types of covariates: time independent covariate and time dependent covariate. SE Cox model combines two methods. There are stratified Cox [11], [5], [20], [1] and extended Cox [17], [26], [2], [13]. In time independent covariates, Cox assumption violations are solved by creating strata on the model, so time independent covariates that do not meet the Cox model assumptions are excluded while time dependent covariate remains incorporated into the model. The time-dependent covariates identified in this model are internal. Internal covariates are related to the individual, and can only be measured when the individual is alive.

In SE Cox model, it is assumed that there are independent variables of p_1 covariates which are time-independent and as many as p_1 time-dependent covariates. In p_1 time-independent covariate as many as k covariates fulfill proportional hazard assumption and notated with $x_1, x_2, ..., x_k$ with $k < p$. Independent variable/covariate which does not fulfill proportional hazard assumption as many as m resulting from $p - k = m$ is $x_{k+1}, x_{k+2}, ..., x_p$ and notated as $z_1, z_2, ..., z_m$. Covariates which do not fulfill proportional hazard assumption z_i with $i = 1, 2, .., m$ are excluded from the Cox model for stratification with as many as m strata. Meanwhile, covariates that fulfill proportional hazard covariates are included in the model. Other time dependent covariates as many as $x_{p_1+1}, x_{p_1+2}, ..., x_{p_2}$, and notated with $x_1(t_j), x_2(t_j), ..., x_{p_2}(t_j)$ and included in the model. The proposed model to solve the non-proportional hazard problem caused by the two covariates is:

$$h_s(t, \boldsymbol{x}) = h_{0s}(t) \exp\left(\sum_{a=1}^{p_1} \beta_{ai} x_{ai} + \sum_{b=1}^{p_2} \delta_{bi} x_{bi}(t_j) \right). \tag{1}$$

where:

$h_{0s}(t)$ = baseline hazard function in every stratum, with strata (s = 1, 2,.., m).
β_{ai} = a coefficient vector as the fixed effect of a-covariate in i-individual.
x_{ai} = time-independent covariate at the a-time in i- individual.
δ_{bi} = a coefficient vector of time-dependent covariate at b time in i- individual.
$x_{bi}(t_j)$ = time-dependent covariate at b time in i- individual at the time of t_j.

2.2. Parameter estimation of model

Parameter estimation in the SE Cox model is based on likelihood. The estimation method used is similar to the Cox PH model. The estimation of parameter by the Maximum Penalize Likelihood Estimation [7]. The maximum likelihood partial function is expressed by L_p. In the partial function of likelihood only *j* object experiences the events that contribute directly, while $n-j$ for example, is a censored object [11].

Likelihood of each observed object states that the possibility of the event at the *t* time depends on when $T_j \geq t.$ The possibility of *i*-th individual which experiences a *j* event, (P_{ij}) depends on a set of objects which may experience an event at the *t* time, also called a set of risk, and notated as $R(t_i).$ The risk possibility of the *i*-th individual at t_j time is defined as:

$$P_{i(t_j)} = \frac{\exp\left(\sum_{a=1}^{p_1} \beta_{ai} x_{ai} + \sum_{b=1}^{p_2} \delta_{bi} x_{bi}(t_j)\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{a=1}^{p_1} \beta_{aj} x_{aj} + \sum_{b=1}^{p_2} \delta_{bj} x_{bj}(t_j)\right)}. \tag{2}$$

For instance, a model parameter estimated is $\boldsymbol{\gamma} = (\beta_1, \beta_2, ..., \beta_{p_1}, \delta_1, \delta_2, ..., \delta_{p_2}),$ and has *m* time-independent covariate which does not fulfill the Cox PH assumption, then there are as many as *m* strata formed. The function of the partial likelihood of the SE Cox model at the *s*-strata is formulated as follows:

$$L_s(\boldsymbol{\gamma}) = \prod_{i=1}^{n_s} P_{i(t_j)} = \prod_{i=1}^{n_s} \frac{\exp\left(\sum_{a=1}^{p_1} \beta_{ai} x_{ai} + \sum_{b=1}^{p_2} \delta_{bi} x_{bi}(t_j)\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{a=1}^{p_1} \beta_{aj} x_{aj} + \sum_{b=1}^{p_2} \delta_{bj} x_{bj}(t_j)\right)}. \tag{3}$$

Likelihood function of all *m* strata are formulated as follows:

$$L_p(\boldsymbol{\beta}) = \prod_{s=1}^{m} L_s = \prod_{s=1}^{m} \prod_{i=1}^{n_s} P_{i(t_j)} = \prod_{s=1}^{m} \prod_{i=1}^{n_s} \frac{\exp\left(\sum_{a=1}^{p_1} \beta_{ai} x_{ai} + \sum_{b=1}^{p_2} \delta_{bi} x_{bi}(t_j)\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{a=1}^{p_1} \beta_{aj} x_{aj} + \sum_{b=1}^{p_2} \delta_{bj} x_{bj}(t_j)\right)}. \tag{4}$$

by using notation:

$$\psi_i = \exp\left(\sum_{a=1}^{p_1} \beta_{ai} x_{ai} + \sum_{b=1}^{p_2} \delta_{bi} x_{bi}(t_j)\right)$$

can be written as (5)

$$L_p(\boldsymbol{\gamma}) = \prod_{s=1}^{m} \prod_{i=1}^{n_s} \frac{\psi_i}{\sum_{j \in R(t_i)} \psi_j}. \tag{5}$$

Next step is to maximize partial likelihood function by taking partial derivative of $\ln L_p(\boldsymbol{\gamma})$ w.r.t. each parameter of model. $\ln L_p(\boldsymbol{\gamma})$ is formulated in the following equation:

$$\ln L_p(\boldsymbol{\beta}) = \sum_{s=1}^{m}\left[\sum_{i=1}^{n_s}\left(\sum_{a=1}^{p_1}\beta_{ai}x_{ai} + \sum_{b=1}^{p_2}\delta_{bi}x_{bi}(t_j)\right)\right] - \sum_{s=1}^{m}\left[\sum_{i=1}^{n_s}\ln\sum_{j\in R(t_i)}\psi_j\right]. \tag{6}$$

Parameter is estimated by finding first derivative each parameter as follows.

$$\frac{\partial \ln L_p(\boldsymbol{\gamma})}{\partial \beta_a} = 0 \qquad a = 1, 2, ..., p_1 \tag{7}$$

$$\frac{\partial \ln L_p(\boldsymbol{\gamma})}{\partial \delta_b} = 0 \qquad b = 1, 2, ..., p_2. \tag{8}$$

Equation (7) results the following equation:

$$\frac{\partial \ln L_p(\boldsymbol{\gamma})}{\partial \beta_a} = \sum_{s=1}^{m}\sum_{i=1}^{n_s}\sum_{a=1}^{p_1} x_{ai} - \sum_{s=1}^{m}\sum_{i=1}^{n_s}\frac{1}{A}\sum_{j\in R(t_i)} B. \tag{9}$$
$$= 0$$

While equation (8) results the following:

$$\frac{\partial \ln L_p(\boldsymbol{\gamma})}{\partial \delta_b} = \sum_{s=1}^{m}\sum_{i=1}^{n_s}\sum_{b=1}^{p_2} x_{bi}(t_j) - \sum_{s=1}^{m}\sum_{i=1}^{n_s}\frac{1}{A}\sum_{j\in R(t_i)} C$$
$$= 0 \tag{10}$$

where:

$$A = \sum_{j\in R(t_i)} \exp\left(\sum_{a=1}^{p_1}\beta_{aj}x_{aj} + \sum_{b=1}^{p_2}\delta_{bj}x_{bj}(t_j)\right)$$

$$B = \sum_{a=1}^{p_1} x_{aj} \exp\left(\sum_{a=1}^{p_1}\beta_{aj}x_{aj} + \sum_{b=1}^{p_2}\delta_{bj}x_{bj}(t_j)\right)$$

$$C = \sum_{b=1}^{p_2} x_{bj}(t_j) \exp\left(\sum_{a=1}^{p_1}\beta_{aj}x_{aj} + \sum_{b=1}^{p_2}\delta_{bj}x_{bj}(t_j)\right).$$

To solve the equation, Newton Raphson iteration can be used. For instance $L_p(\boldsymbol{\gamma})$ is a partial likelihood function of vector $\boldsymbol{\gamma} = (\beta_1, \beta_2, ..., \beta_{p_1}, \delta_1, \delta_2, ..., \delta_{p_2}).$ Suppose that $\boldsymbol{U}(\boldsymbol{\gamma})$ is the first partial derivative vector of $L_p(\boldsymbol{\gamma})$ and if $\boldsymbol{H}(\boldsymbol{\gamma})$ is Hessian matrix of the derivative of the second partial likelihood of $L_p(\boldsymbol{\gamma}).$ In general the second partial likelihood w.r.t. parameter β and δ is:

$$\frac{\partial^2 \ln L_p(\boldsymbol{\gamma})}{\partial^2 \beta_a^2} = \sum_{s=1}^{m}\sum_{i=1}^{n_s}\frac{1}{A}\left(\begin{array}{l}\sum_{j\in R(t_i)}\sum_{a=1}^{p_1} x_{aj}\left(\sum_{j\in R(t_i)}\sum_{a=1}^{p_1} x_{aj}\,\psi_j\right) \\ -\sum_{j\in R(t_i)}\left(\sum_{a=1}^{p_1} x_{aj}\right)^2 \psi_j\end{array}\right). \tag{11}$$

$$\frac{\partial^2 \ln L_p(\boldsymbol{\gamma})}{\partial^2 \delta_b^2} = \sum_{s=1}^{m}\sum_{i=1}^{n_s} \frac{1}{A} \left(\begin{array}{l} \sum_{j\in R(t_i)} \sum_{b=1}^{p_2} x_{bj}(t_j) \left(\sum_{j\in R(t_i)} \sum_{a=1}^{p_1} x_{aj}\psi_j \right) \\ - \sum_{j\in R(t_i)} \sum_{a=1}^{p_1} x_{aj} \sum_{b=1}^{p_2} x_{bj}(t_j)\psi_j \end{array} \right). \tag{12}$$

Parameter estimation of model is done with Newton Raphson iteration method as follows.

a. Determining on the initial value: $\hat{\boldsymbol{\gamma}}_0 = \mathbf{0}$.
b. Calculating $\hat{\gamma}_1 = \hat{\gamma}_0 - H(\hat{\gamma}_0)^{-1} U(\hat{\gamma}_0)$.
c. Iteration is done until the convergent value is obtained: $\hat{\gamma}_{c+1} \cong \hat{\gamma}_c$.

Hosmer and Lemeshow (2008) state that variance estimation of maximum partial likelihood $\hat{\boldsymbol{\gamma}}$ is the inverse of the Hessian matrix or it is formulated as follows:

$$\hat{Var}(\hat{\boldsymbol{\gamma}}) \cong \boldsymbol{H}(\hat{\boldsymbol{\gamma}})^{-1}. \tag{13}$$

2.3. Simulation study

The design of the proposed model simulation follows 3 steps: (1) model simulation approach, (2) data structure and model assumption, and (3) model goodness of test.

2.3.1. Model Simulation Approach. The SE Cox model is an extension of Cox PH to address both time-independent and internal time-dependent covariates that do not meet proportional hazard assumptions. The proportional hazard assumption states that the hazard function ratios of the two individuals are constant over time or equivalent to the assertion that an individual's hazard function compared another individual's hazard function is proportional [14]. The modified model is stratified at time independent covariate that does not meet proportional hazard. Time-independent covariates that do not meet proportional hazards are removed from the model, while those that do and time-dependent covariates that do not meet proportional hazards assumptions are put into the model.

To check whether the proposed model can be applied to the appropriate conditions and gives a good result compared to the existing model, simulations are performed with various treatment combinations. The stratified-extended Cox simulation model is performed as follows.

1) Generating time-dependent covariate data by using the PermAlgo package developed by [28]. It generates timed increment and censors time generation data from the time-dependent covariate with the data distribution specified by the user. In this case, the time-dependent covariate data is analogous to the number of credits taken by the student, $x_i(t_j)$ ~ Uniform (0,100). The generation of these covariates uses a counting process system based on a partial likelihood probability permutation.
2) Generating time independent covariate data. In this situation time, independent covariate data is analogous with age (x_1) and study the background of the students (x_2). Age, x_1~ N(40, 10) while study background, x_2~ Uniform (1,3). Study background is divided into 2 categories: (1) high school, (2) non-degree program, and (3) bachelor. They belong to the strata in the proposed model. These covariates are generated by the re-counting process in the 1st step.
3) To generate data response, survival time is measured in the semester unit, T_{ji} ~ Uniform (0,10).
4) Model parameter $\beta = \log(1.04)$ and $\delta = \log(0.99)$. For model dummy α=2.
5) Censoring 0%, C_0 ~ Uniform (6, 8); censoring 30%, C_{30} ~ Uniform (2, 5); and censoring 50%, C_{50} ~ Uniform (0, 5).
6) Total samples (n) are 100; 500; 1,000 and 2,000. The selection of the samples is in accordance with the variation of the student population that represents each faculty.

7) Iteration is done 1,000 times.

2.3.2. Data structure and model assumptions. The proposed data structure of the model is as follows. For instance, T_i^* is the survival time, i.e. student survival data of the i-th student in semester unit. C_i^* is the censored time during general observation of the i-th student with $T_i^* < C_i^*$. The survival time (T_i) is min (T_i^*, C_i^*) and $\delta_i = I(\{T_i^* \le C_i^*\})$ is the censored indicator with I=1 when T_i^* is observed and I=0 when T_i^* censored. Therefore the observing pairs of the response data is (T_i, δ_i).

In the first step, time-dependent covariate and time-independent covariate data are created. The time-dependent covariate data in this study is assumed to be internal, i.e. the covariates related to the individual and can only be measured when the individual is alive, meaning he/she is still enrolled as UT student. Time-dependent covariate data is the number of credits taken (NC), $x_i(t_j) \sim$ Uniform (0, 100). Time dependent covariate data is performed by using the counting process assumption based on partial likelihood to count the events from time to time. For example, to calculate the time of student's failure in a group of enrolled students in a particular period. The basic idea in the counting process is to find a certain time interval of a starting and an end time of an event. To generate time dependent covariate, PermAlgo Package is applied [28]. The new pairs of data formed through the counting process are the pairs of data with $(N_i(t), Y_i(t))$, function, where [3], [12]:

$N_i(t)$ = the number of observed events at [0,*t*] interval for *i*-th individual.

$Y_i(t)$ = $\begin{cases} 1, & \text{the } i\text{-th unit is at risk at time } t \\ 0, & \text{others} \end{cases}$

This formula is a special case of the right censored data:

$N_i(t) = I(\{T_i \le t,\ \delta_i = 1\})$.

Furthermore, to generate time independent covariate, age (x_1) is assumed N (40, 10). The second step is to generate data for censoring, by assuming that data are distributed uniformly with a censoring percentage of 0%, 30%, and 50%. Each is distributed with $C_0 \sim$ Uniform (6, 8); $C_{30} \sim$ Uniform (2, 5); and $C_{50} \sim$ Uniform (0, 5). The censoring percentage is selected based on an existing real data condition. The generated data is expected to be appropriate or close to real data. PermAlgo is applied for the first and second steps of data input.

The next step is to insert another time independent covariate, i.e. educational background (x_2). The covariate $x_2 \sim$ Uniform (1, 3) is assumed to be fixed strata because, in real data, one significant covariate that does not meet the proportional hazard assumption is the educational background. To obtain the strata covariate data, the re-counting process is used for PermAlgo generated data. The newly inserted data formed is used to perform the proposed modeling.

2.3.3. Goddness of the model. The measurement of goodness model compared to the result of parameter estimation is reflected in the generated bias value and MSE [23].

$$Bias(\boldsymbol{\gamma}) = \left(\frac{1}{m}\sum_{i=1}^{m}\hat{\boldsymbol{\gamma}}^{(i)}\right) - \hat{\boldsymbol{\gamma}}^{(0)} \quad (14)$$

$$MSE(\boldsymbol{\gamma}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\boldsymbol{\gamma}}^{(0)} - \hat{\boldsymbol{\gamma}}^{(i)})^2 \quad (15)$$

m is the number of iteration, $\hat{\boldsymbol{\gamma}}^{(0)}$ is the parameter of the initial model and $\hat{\boldsymbol{\gamma}}^{(1)}$ is the model parameter resulted from model simulation. Model evaluation is done through the maximum value of log likelihood. Lee [19] states that another goodness measurement for model evaluation is AIC (Akaike Information Criterion) and SBC (Schwarz's Bayesian Criterion). AIC and SBC are functions of a series

of *n* observations of a sum square error, and a number of predictor variables $k \le p+1$ where *k* includes an intercept. AIC value is formulated as follows:

$$AIC = n \ln\left[\frac{SSE}{n}\right] + 2k. \tag{16}$$

SBC values is formulated as follows:

$$SBC = n \ln\left[\frac{SSE}{n}\right] + k \ln n. \tag{17}$$

The best model is selected by examining all three values. A model is considered good if it has a maximum log-likelihood value and minimum AIC and SBC values. The simulation result is applied to UT student resistance data. The data used are UT survival learning time of the students who registered at the first semester 2008 (or 2008.1) until 2015.2. This is the response variable. The time-dependent covariates that have a fixed effect on student survival are educational background, gender, age, marital status, employment status, and domicile. Meanwhile, the time-dependent covariates observed are the number of credits taken and the number of courses enrolled per semester.

3. Result and Discussion

This section describes the goodness and constancy of the proposed model simulation results. Simulations were performed on several combinations of censor treatment, the number of data and iterations. The results are then compared with other existing models. As described in the model simulation, there are 12 combinations of of treatments, each of which is compared to the model's good value which includes: bias, Loglik value, AIC, and BIC. The result of the SE Cox model is compared with the Cox model, extended Cox, stratified Cox, and dummy extended. Previous strata study uses dummy variables. The results of the study are presented in the table of model simulation results.

3.1. Result of the Model Simulations

In this simulation, censor percentage and sample size used are 0%; 30%; 50% and 100; 500; and 2,000 respectively with 1,000 iterations. The result of the third simulation and the different data are presented in Table 1, Table 2, and Table 3.

Table 1 shows the simulation results with a small number of n (n = 100). The parameter bias of the SE Cox model is relatively small compared to the other four models in each censoring percentage tested. So is the mean square error (MSE) value each censoring percentage of the MSE SE Cox models is relatively small compared to other models. However, for model goodness of test of Loglik, BIC, and AIC values, of the five models tested, the stratified Cox model and SE Cox model provide better value than others. The stratified Cox model has a better value for the model's goodness of test than the SE Cox model. However, the differences between these three values tend to be not significant.

Table 2 shows the simulation results of n = 500. Table 2 can be showed that the bias and MSE estimation with 0%, 30%, and 50% censoring in the SE Cox model are the smallest, except for parameter b1. The value of bias and MSE estimation at 0% and 30% censoring in the Cox model are the smallest. Whereas the goodness of values for the five models tested, the best model is the stratified Cox and the SE Cox model.

Table 1 and Table 2 show that the number of tested samples is relatively small. The small samples in this simulation aim to show that the proposed model can be used and is suitable for small data. The test is done because a small number of data is often found in real data. The results of the above analysis show that the proposed model (SE Cox model) provides a reasonable bias and MSE value as well as Loglik, AIC, and BIC as an alternative to non-proportional hazard modeling on survival analysis with small data. The model simulation for large data is represented by n = 2,000. The simulation results of the five models tested on these two large n types with three types of censoring are presented in Table 3.

Table 3 shows that at 0% censoring, the SE Cox model provides a relatively small estimation and MSE estimation values compared to the other four models in estimator b0. Meanwhile, at 30% and 50% censoring, the SE Cox and extended Cox models provide a better estimator and MSE bias values than the others on both estimates (b0 and b1). For the model test scores shown by Loglik, AIC, and BIC values, the SE Cox model ranks second after the stratified Cox model. However, the difference in model goodness of test values for both models is relatively small and not very significant. Thus, at with various types of censoring, the SE Cox model can be used as an alternative to non-proportional hazard modeling involving two types of covariates in the model (time independent covariate and time-dependent covariate).

At 30% and 50% censoring, the SE Cox model provides a better bias and MSE estimations b0 and b1 than other models. When viewed from the model's virtues, as seen in the previous tables, the stratified Cox model has relatively better model goodness of value than the SE Cox model. The three values of model goodness in the three types of censoring being tested have a relatively small and no significant difference. In addition, the use of the stratified Cox model on data involving two types of covariates is less precise and parameter estimation of the model becomes biased. Thus, the SE Cox model can be used as an alternative non-proportional hazard modeling wherein the model there are two types of covariates that cause the proportional hazard assumption was not met.

Table 1. Result of model simulation on various types of censoring with *n*=100

Parameter Bias and Goodness of Model	Censoring 0% n=100				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.26122	
b0	0.01385	0.01268	0.01354	0.01781	0.01227
b1	0.00355	0.00377	0.00393	0.00422	0.00362
MSE : dummy				0.12251	
b0	0.00032	0.00027	0.00031	0.00043	0.00027
b1	0.00003	0.00003	0.00003	0.00003	0.00003
Loglik	-353.681	-354.695	-257.623	-335.878	-258.259
BIC	716.521	718.5485	524.4157	685.4935	525.6757
AIC	711.363	713.3905	519.2457	677.7565	520.5177
Parameter Bias and Goodness of Model	Censoring 30% n=100				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.31378	
b0	0.01091	0.01052	0.01042	0.01661	0.01014
b1	0.01064	0.00322	0.01103	0.0036	0.00319
MSE : dummy				0.15852	
b0	0.00028	0.00025	0.00027	0.00041	0.00024
b1	0.00013	0.00003	0.00014	0.00003	0.00003
Loglik	-297.982	-304.598	-218.107	-287.13	-224.744
BIC	604.718	617.9491	444.9687	587.3913	458.2427
AIC	599.964	613.195	440.2146	580.2602	453.4886
Parameter Bias and Goodness of Model	Censoring 50% n=100				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.34906	
b0	0.01097	0.01057	0.01048	0.01604	0.01016
b1	0.01088	0.00302	0.01148	0.00323	0.00283
MSE : dummy				0.21015	
b0	0.00034	0.00031	0.00033	0.00046	0.0003
b1	0.00014	0.00004	0.00016	0.00004	0.00004
Loglik	-207.397	-212.143	-149.274	-199.279	-154.077
BIC	422.9299	432.4222	306.6842	410.7617	316.2898
AIC	418.7942	428.2865	302.5485	404.5581	312.154

Table 2. Result of model simulation on various types of censoring with n = 500

Parameter Bias and Goodness of Model	Censoring 0% n=500				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.27466	
b0	0.01344	0.01247	0.01335	0.01784	0.01234
b1	0.00333	0.00408	0.00339	0.00443	0.00405
MSE : dummy				0.08564	
b0	0.0002	0.00018	0.0002	0.00034	0.00017
b1	0.00001	0.00002	0.00001	0.00002	0.00002
Loglik	-2550.5	-2553.24	-2048.74	-2457.96	-2051.83
BIC	5113.387	5118.86	4109.867	4934.492	4116.044
AIC	5105.009	5110.482	4101.489	4921.925	4107.666

Parameter Bias and Goodness of Model	Censoring 0% n=500				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.31438	
b0	0.0113	0.01073	0.01123	0.01642	0.01069
b1	0.0104	0.00338	0.01049	0.0037	0.00336
MSE : dummy				0.07726	
b0	0.00019	0.00016	0.00019	0.00032	0.00016
b1	0.00001	0.00002	0.00001	0.00002	0.00002
Loglik	-2126.61	-2159.71	-1716.2	-2070.25	-1749.32
BIC	4265.19	4331.382	3444.362	4158.458	3510.618
AIC	4257.22	4323.411	3436.391	4146.502	3502.647

Parameter Bias and Goodness of Model	Censoring 0% n=500				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.32617	
b0	0.01103	0.01074	0.01095	0.0163	0.01065
b1	0.01049	0.00361	0.01064	0.00385	0.00359
MSE : dummy				0.12135	
b0	0.00016	0.00015	0.00016	0.0003	0.00015
b1	0.00011	0.00002	0.00012	0.00002	0.00002
Loglik	-1507.9	-1531.59	-1206.17	-1463.77	-1229.98
BIC	3027.168	3074.536	2423.709	2944.572	2471.315
AIC	3019.809	3067.176	2416.349	2933.533	2463.955

Table 3. Result of model simulation on various types of censoring with n = 2,000

Parameter Bias and Goodness of Model	Censoring 0% n=2,000				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.27347	
b0	0.01343	0.01252	0.01341	0.01789	0.01249
b1	0.00323	0.00406	0.00325	0.00441	0.00405
MSE : dummy				0.07726	
b0	0.00019	0.00016	0.00019	0.00032	0.00016
b1	0.00001	0.00002	0.00001	0.00002	0.00002
Loglik	-12904.8	-12914.1	-10881.7	-12531.3	-10891.4
BIC	25824.7	25843.37	21778.47	25085.35	21797.86
AIC	25813.55	25832.22	21767.32	25068.62	21786.71

Parameter Bias and Goodness of Model	Censoring 0% n=2,000				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
b0	0.01147	0.01101	0.01146	0.01668	0.01099
b1	0.01042	0.00347	0.01044	0.00377	0.00346
MSE : dummy				0.10421	
bo	0.00014	0.00013	0.00014	0.00028	0.00013
b1	0.00011	0.00001	0.00011	0.00002	0.00001
Loglik	-10715.8	-10849.3	-9062.74	-10491.9	-9196.27
BIC	21446.34	21713.36	18140.22	21005.84	18407.29
AIC	21435.6	21702.61	18129.47	20989.73	18396.55

Parameter Bias and Goodness of Model	Censoring 0% n=2,000				
	Cox	Extended	Stratified	Dummy Extended	Startified Extended
Bias : dummy				0.3332	
b0	0.01101	0.0108	0.01098	0.01623	0.01078
b1	0.01027	0.00344	0.01031	0.00368	0.00343
MSE : dummy				0.11495	
b0	0.00013	0.00012	0.00013	0.00027	0.00012
b1	0.00011	0.00001	0.00011	0.00001	0.00001
Loglik	-7653.89	-7747.75	-6437.74	-7476.89	-6531.71
BIC	15321.9	15509.64	12889.61	14974.97	13077.56
AIC	15311.77	15499.51	12879.48	14959.77	13067.43

3.2. Application to student retention data universitas terbuka

The proposed model in this paper is applied to the data of students survival in the Universitas Terbuka (Indonesia). Data used is the duration of study time of UT students who register in 2008 semester 1 (or 2008.1) until 2015.2. This time data is the response variable measured in semester units. The time-dependent covariates that are considered influential of student survival are educational background, the preferred course of study, gender, age, marital status, employment status, and student domicile.

Meanwhile, the time-dependent covariates observed are the number of credits taken and the number of courses enrolled per semester. Description of UT students' survival data is presented based on demographic characteristics, such as residence, gender, age, marital status, employment status, and academic characteristics, such as formal education background, study program, the number of credits taken, and the number of courses taken.

Table 4 shows that the number of observations is as many as 4,483 students, consisting of 1,574 censored people (35.11%) and 2,909 (64.89%) not censored. Students are considered censored if they are registered or have graduated or enrolled in other study programs. Whereas non-active student is not censored. Students are said to be non-active if they do not register for 4 consecutive semesters [4]. The description of UT students according to the demographic characteristics is also presented in Table 5. Those who do not register live in the districts, female, between 35-45 years old, married, and work. This is emphasized by [27], that in the open and distance education system, the learning process is more complex due to older age, working for status and family. Orr [22] also states the same that students who already pursue their career (work) can not study full time. Kadarko [16] reveals that age is a significant factor of self-study capacity where one has to have orientation and apply strategies in learning from the modules provided and also in perceiving the non-conventional academic environment.

Table 4. UT students based on demographic characteristics

Covariates observed	Categories	Censored Status		Total
		Censored	ot Censored	
Domicile	District	1,265	2,338	3,603
	Cities	309	571	880
Sex	Female	923	1,558	2,481
	Male	651	1,351	2,002
Age	< 35 years	87	294	381
	35-45 years	1,049	1,859	2,908
	> 45 years	438	756	1,194
Marital status	Single	464	1,081	1,545
	Married	1,110	1,828	2,938
Working status	Not working	58	294	352
	Working	1,516	2,615	4,131
Total		1,574	2,909	4,483

The academic characteristics of the students presented in Table 5 shows that non registered students are dominated by those with high school education background with 75 credits attained and the number of courses taken/semester is between 5 to 8 courses. Coggins [6] argues that one of the causes of high drop out in the open and distance education system is the education background of the students and the length of the study period. In general, high school leavers who enroll at UT do not have the idea of independent learning habit and they come from various types of high schools.

Ratnaningsih [25] further found that the students, among others, (1) do not understand the learning system at UT, (2) have no idea of self-study, (3) do not have learning motivation, (4) do not have peers to discuss, (5) have a lack of learning materials, and (6) come from diverse educational backgrounds. In addition to that, they do not develop a regular self-study system as [21], [30] found. Puspitasari and Islam [24] suggest that distance education students from different age groups have different levels of readiness in self-study. The higher the age, the higher the possibility of self-study. Of the 10 study programs observed, the highest percentage of non-active students is those in Agribusiness Study Program and Biology Education, by 70.18% and 70.10%, respectively.

Table 5. UT Students based on academic characteristics

Covariates observed	Categories	Censored Status		Total
		Censored	Not Censored	
Formal Education Level	High School	813	1,911	2,724
	Diploma	752	959	1,711
	Bachelor	9	39	48
Study Program	Agribusiness (Fishery)	25	31	56
	Agribusiness (Animal Husbandry)	17	40	57
	Public Administration (Personnel Administration & Management)	35	3	38
	Public Administration	586	1,195	1,781
	Indonesian Language and Literature Education	291	505	796
	English Language Education	287	442	729
	Biology Education	119	279	398
	Economics Education	171	348	519
	Chemistry Education	30	64	94
	English Literature (Translation)	13	2	15
Number of Credit Taken (System Credit Semester)	SCS < 75	46	2,197	2,243
	$75 \leq SCS \leq 120$	86	402	488
	SCS> 120	1,442	310	1,752
Number of Courses Taken	NCT <5	410	391	801
	$5 \leq NCT \leq 8$	1,149	2,146	3,295
	NCT > 8	15	372	387

Based on the problems, the SE Cox model is proposed to overcome the assumption that the Cox model is not met. The analysis of survival data of UT students with the Cox model yield 6 covariates that do not fulfill the Cox assumption (Table 7), proven by small p-value (*). The goodness of value of the Cox model is seen in AIC, BIC, and Loglik, with 41,147.47; 41,207.49; and -22,987.64 respectively. If the log-log graphic of a survival test is used, the plot of the four covariates is presented in Figures 1 and 2.

Figure 1 shows six unparalleled plots of the covariates. Study program covariate and education (educational background) and age are time-independent covariates. Meanwhile, SCS (the number of credits taken) and the NCT (number of courses taken) and GPA are time-dependent covariates. For modeling purposes, the covariate is used as strata (fixed) in real data and adjusted to the model in the simulation, which is education. Another reason for making education as strata is for efficient estimation of model parameters because the number of categories is small (only 3 categories). The use of multiple strata (such as study program, which has 10 categories) would lead to biases and inefficient models. Therefore, in this model, only education covariate is used for strata.

Table 6. Statistic value using Cox proportional hazard model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Study program	-0.03783	0.96288	0.01309	-2.891	0.00384**
Education	0.4597	1.58361	0.04704	9.773	< 2e-16***
Domicile	-0.01967	0.98052	0.04759	-0.413	0.67932
Age	-0.1728	0.84131	0.041	-4.215	0.000025** *
Gender	0.05666	1.0583	0.03846	1.473	0.14065
Work	0.02466	1.02497	0.06521	0.378	0.70526
Marital status	-0.07177	0.93075	0.0458	-1.567	0.11712
Credit unit	-1.46516	0.23104	0.03403	-43.06	< 2e-16***
Course unit	0.19522	1.21557	0.0397	4.918	8.67e-07***
GPA	-0.8492	0.42776	0.03984	-21.316	< 2e-16***

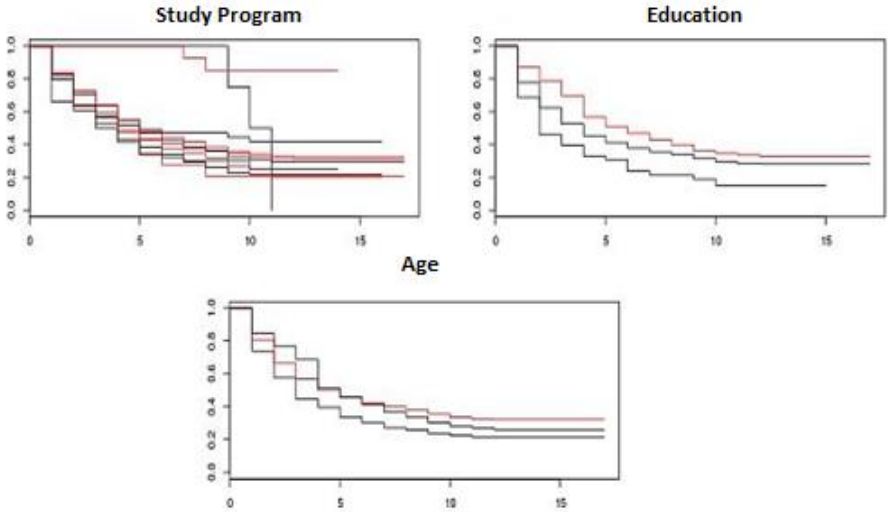

Figure 1. Log-log survival plot for time independent covariate.

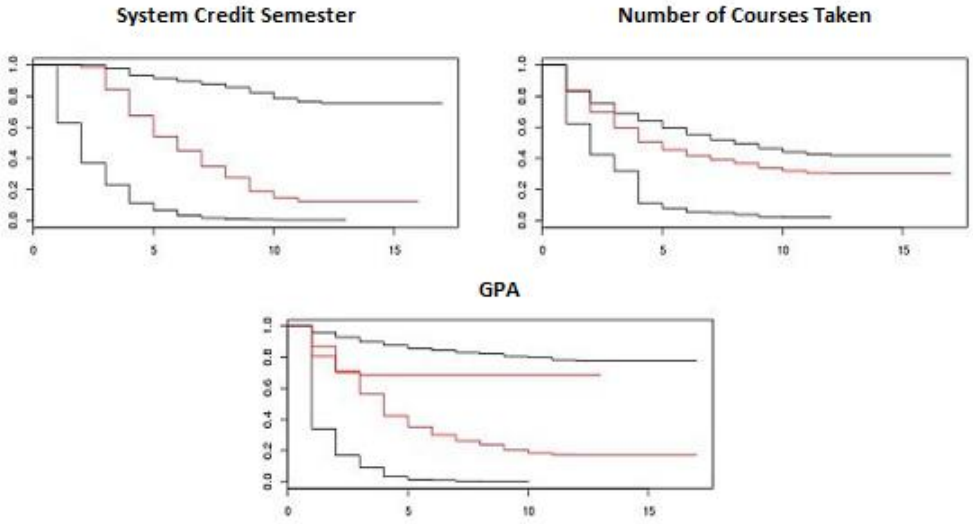

Figure 2. Log-log survival plot for time dependent covariate.

Modeling result using SE Cox model toward study survival of UT students are presented in Table 7.

Table 7. Statistic values using SE Cox model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Study Prog2	0.163092	1.177145	0.240123	0.679	0.49701	
Study Prog3	-1.44033	0.23685	0.607671	-2.37	0.01778	*
Study Prog4	0.317251	1.373347	0.185831	1.707	0.08779	.
Study Prog5	0.280462	1.323741	0.188217	1.49	0.13262	
Study Prog6	0.534342	1.706326	0.188222	2.839	0.00453	**
Study Prog7	0.186211	1.204676	0.192494	0.967	0.33336	
Study Prog8	0.201044	1.222679	0.190682	1.054	0.29173	
Study Prog9	0.893834	2.444484	0.221165	4.041	5.31E-05	***
Study Prog10	-2.94505	0.052599	0.736862	-3.997	6.42E-05	***
Domicile	0.059798	1.061622	0.048089	1.243	0.21368	
Age2	0.053887	1.055366	0.069487	0.776	0.43804	
Age3	-0.03428	0.966302	0.089053	-0.385	0.70029	
Gender	0.089179	1.093277	0.039023	2.285	0.06423	
Work	-0.15036	0.860397	0.066515	-2.261	0.52379	
Merried	-0.07989	0.923219	0.045124	-1.77	0.07666	.
t_SCS	-0.31524	0.729613	0.006408	-49.195	< 2e-16	***
t_NCT	0.139449	1.14964	0.012133	11.493	< 2e-15	***
t_GPA	-1.18319	0.3063	0.037891	-31.226	< 2e-16	***

Table 7 shows that by using the SE Cox model, there are 4 significant covariates that do not fulfill the Cox PH assumption, ie. study program, number of credit taken, number of courses taken, and GPA. Only 4 study programs that do not fulfill the Cox PH assumption, i.e. Study Prog3, Study Prog6, Study Prog9, and Study Prog10, each above 50% censoring percentage. Goodness of value of stratified-extended Cox model in real data is AIC = 36,972.37; BIC = 37,020.18; and Loglik = -20,478.19. The four values are better than the Cox PH Model. Therefore it can be stated that the proposed model (SE Cox model) has a better goodness model compared to the existing models to solve the problem of the non-proportional hazard model due to time-independent covariate and time-dependent covariate in the model.

4. Discussion

The simulation of the treatment combination shows that the bias parameter of the SE Cox model is smaller than other models, so is the goodness model, shown here by AIC, BIC, and Logik. Out of the five values tested on either small or large samples, the SE Cox is better than the other four models. Thus, the proposed alternative model is non-proportional hazard data modeling caused by the presence of 2 types of covariates, ie time independent covariate and time-dependent covariate.

The result of the model application of the real data shows that the SE Cox model can be used to model the UT students' survival data since it consists of time independent covariates and time-dependent covariates. The results show that there are fewer covariates that do not meet the Cox assumption. In addition to that, as far as the goodness of values is a concern, the SE Cox model gives better value than the Cox model. Thus better and more meaningful modeling.

However, in the proposed model does have weaknesses, namely the limitations on the strata. Strata used in this model is fixed and has relatively limited categories (only 3 categories). The strata used are education. The analysis shows that by using the proposed model, a time-independent covariate which does not fulfill the assumption still exists, that is selected study program. Study program covariate is most likely random due to the fact that students are allowed to choose their study programs based on their interests and educational background. The number of categorizations of the study program in real data is significant. If the study program is used as a stratum, then there will be as many as 10 strata. This becomes inefficient and may lead to bias. Therefore, other modeling efforts should be made to accommodate many strata. One possible model approach is to develop a mixed model. The use of mixed models with the study program as a random effect is expected to solve the weaknesses in the model proposed in this paper.

5. Conclusion

The simulation of the treatment combination shows that the bias parameter of the SE Cox model is smaller than other models, so is the goodness model, shown here by AIC, BIC, and Loglik. Out of the five values tested on either small or large samples, the SE Cox is better than the other four models. Thus, the proposed alternative model is non-proportional hazard data modeling caused by the presence of 2 types of covariates, ie time independent covariate and time-dependent covariate.

The result of the model application of the real data shows that the SE Cox model can be used to model the UT students' survival data since it consists of time independent covariates and time-dependent covariates. The results show that there are fewer covariates that do not meet the Cox assumption. In addition to that, as far as the goodness of values is a concern, the SE Cox model gives better value than the Cox model. Thus better and more meaningful modeling.

However, in the proposed model does have weaknesses, namely the limitations on the strata. Strata used in this model is fixed and has relatively limited categories (only 3 categories). The strata used are education. The analysis shows that by using the proposed model, a time-independent covariate which does not fulfill the assumption still exists, that is selected study program. Study program covariate is most likely random due to the fact that students are allowed to choose their study programs based on their interests and educational background. The number of categorizations of the study program in real data is significant. If the study program is used as a stratum, then there will be as many as 10 strata.

This becomes inefficient and may lead to bias. Therefore, other modeling efforts should be made to accommodate many strata. One possible model approach is to develop a mixed model. The use of mixed models with the study program as a random effect is expected to solve the weaknesses in the model proposed in this paper.

References

[1] Abdelaal M M and Zakria S H 2015 *American Journal of Theoretical and Applied Statistics* **4** 504.
[2] Adeleke K A, Abiodun A A, Ipinyomi R A 2015 *Journal of Modern Applied Statistical Methods* **14** 68.
[3] Andersen P K and Gill R 1982 *Annals of Statistics* **10** 1100.
[4] Anonymous 2018 *Universitas Terbuka Catalogue* (Jakarta: Publishing Center of Universitas Terbuka)
[5] Ata N and Sözer M T 2007 *Hacettepe Journal of Mathematics and Statistics* **36** 157.
[6] Coggins C 1989 *Preferred Learning Syles and Their Impact On Completion Of External Degree Programs* In Moore M G and Clar G C (Eds.) *Reading in Distance Learning and Instruction* (University Park, PA: ACSDE)
[7] Collet D 1997 *Modelling Survival Data in Medical Research* Second Edition (London: Chapman & Hall)
[8] Cox D R 1972 *Journal of the Royal Statistical Society* **34** 187.
[9] Cox D R and Oakes D 1984 *Analysis of Survival Data* (London: Chapman & Hall)
[10] Demidenco E 2013 *Mixed Models Theory and Application With R* Second Edition (New Jersey: John Wiley & Sons)
[11] Dupuy J F and Leconte E 2006 *Estimation in a Partially Observed Stratified Cox Model* (accessed 2016 December 05 on https://scholar.google.com/cita-tions?user=cQLxz-X8AAAAJ&hl=en)
[12] Fleming T R and Harrington D P 1991 *Counting Processes and Survival Analysis* (New York: Wiley)
[13] Gellar J E, Colantuoni E, Needhan D M and Crainiceanu C M 2015 *Statistical Modelling* **15** 1.
[14] Guo S 2010 *Statistical Analysis (Pocket Guides to Social Work Researchh Methods)* (Oxford: University Press USA)
[15] Henderson R and Oman J P 1999 *Journal of the Royal Statistical Society* **61** 367.
[16] Kadarko W 2000 *Journal of Open Distance Education* **3** 1.
[17] Kleinbaum D G and Klein M 2012 *Survival Analysis: A Self-Learning Text* 3rd ed Gail M, Krickeberg K, Samet J M, Tsiatis A, Wong W (New York: Springer)
[18] Lee E T 1992 *Statistical Methods for Survival Data Analysis* (New York: John Wiley & Sons Inc)
[19] Lee M C 2014 *International Journal of Computer Science & Information Technology* **6** 103.
[20] Mehrotra D V and Su S C 2012 *Statistics Medicine*. **31** 1849.
[21] Nugraheni E and Pangaribuan N 2006 *Journal of Open Distance Education* **7** 68.
[22] Orr S 2000 *The International Journal of Education Management* **14** 54.
[23] Polat E and Gunay S 2014 *International Journal of Mathematics Trends and Technology* **9** 132.
[24] Puspitasari K A and Islam S 2002 *Readiness of Student Self Study and Potential Student Candidate of Distance Education* (Jakarta: Universitas Terbuka)
[25] Ratnaningsih D J, Saefuddin A and Wijayanto H 2008 *Journal of Open Distance Education* **9** 101.
[26] Saegusa T, Di C and Chen YQ 2014 *Biometrics* **70** 619.
[27] Schuemer R 1993 *Some Psychological Aspects of Distance Education* (Hagen, Germany: Institute for Research into Distance Education)
[28] Sylvestre M P, Edens T, MacKenzie T and Abrahamowicz M 2015 *Permutational Algorithm to Simulate Survival Data* (Package 'PermAlgo')

[29] Therneau T M and Grambsch P M 2000 *Modeling survival data: extending the Cox model. Statistics for biology and health* (New York: Springer Science+Business Media)
[30] Yunus M, Pannen P, Darojat O and Julaeha S 2005 *Journal of Open Distance Education* **6** 1.