American Journal of Theoretical and Applied Statistics
2015; 4(6): 504-512
Published online October 29, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150406.21
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Modeling Survival Data by Using Cox Regression Model

Medhat Mohamed Ahmed Abdelaal*, Sally Hossam Eldin Ahmed Zakria

Statistics and Mathematics Department, Faculty of Commerce, Ain Shams University, Cairo, Egypt

Email address:
medhatal@commerce.asu.edu.eg (M. M. A. Abdelaal), sally.hossam21@yahoo.com(S. H. Eldin Ahmed Zakria)

To cite this article:
Medhat Mohamed Ahmed Abdelaal, Sally Hossam Eldin Ahmed Zakria. Modeling Survival Data by Using Cox Regression Model. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 6, 2015, pp. 504-512. doi: 10.11648/j.ajtas.20150406.21

Abstract: Survival analysis refers to the general set of statistical methods developed specifically to model the timing of events. A popular regression model for the analysis of survival data is the Cox proportional hazards regression model. The Cox regression model is a semi parametric model, making fewer assumptions than typical parametric methods but more assumptions than those nonparametric methods. The main objective of this paper is to construct Cox proportional hazards regression model for examining the covariate effects on the hazard function and to determine the risk factors affecting the outcome of liver transplantation operation for end-stage liver disease. This article will focus on a review of (a) the Cox model and interpretation of its results, (b) assessment of the validity of the PH assumption, and (c) accommodating non-proportional hazards using covariate stratification. Cox PH model showed that the variables: Recipient age, $MELD_3$ Score, Ln_Creatinine, and GRWR are statistically significant and selected as significant factors for risk of death after liver transplantation operation. Also the scaled Schoenfeld residual displayed non-proportionality for variable Recipient Age and this variable needed to be stratified. And the Cox-Snell residual showed the Cox PH model does not fit these data adequately. So the stratified Cox model could be more appropriate to the current study. The stratified Cox model with interaction and with no interaction were applied and showed that the no-interaction model is acceptable at 0.05 level of significance and the variables $MELD_3$ Score, Ln_Creatinine are statistically significant and selected as significant factors for risk of death after liver transplantation operation at 0.05 level of significance.

Keywords: Survival Analysis, Censoring, Cox Proportional Hazard Regression Model, Cox- Snell Residual, Stratified Cox Regression Model

1. Introduction

The most common approach to model covariate effects on survival is the Cox proportional hazard model, which can handle censored and/or truncated observations [1]. Regression analysis is generally used for identifying the risk factors. But due to the presence of censoring in survival data, ordinary regression models cannot be used. Also simple logistic regression analysis has the limitation of only allowing a view of survival probability over the entire study period as a single time interval and it assume that every patient is at risk over the entire study period. This is not valid for studies with long follow up or other situations where patients have variable time at risk. For this purpose, in survival analysis, Cox's regression model is widely applicable.

The distinguishing feature of Cox PH model is its ability to estimate the relationship between the hazard rate and explanatory variables without having to make any assumptions about the shape of the baseline hazard function. Hence the Cox model is sometimes referred to as a semi-parametric model.

The Cox regression model is a statistical theory of counting processes that unifies and extends nonparametric censored survival analysis. The approach integrates the benefits of nonparametric and parametric approaches to statistical inferences [2].

The Cox proportional hazards regression model relates covariates to the hazard function as follows:

$$h(t|x) = h_0(t)c(\beta' x) \qquad (1)$$

Where $h_0(t)$ is called the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the model are zero, $\beta' = (\beta_1, \beta_2 \dots \dots \beta_p)$ is a parameter vector of regression coefficients, $x = (x_1, x_2 \dots . x_p)'$ is the value of the vector of explanatory variables for a particular individual, and c(·) is a fixed, known scalar function [3].

This is a semi-parametric model where the baseline hazard $h_0(\mathrm{t})$ is estimated non-parametrically, while the covariate effect is constrained by the parametric representation $c(\beta' x)$. Where, $\mathrm{c}(\cdot)$ take an exponential form:

$$c(\beta' x) = e^{(\beta' x)} = e^{\left(\sum_{j=1}^{p} \beta_j x_{ji}\right)} \quad (2)$$

Which assures that the hazard is non-negative and assumes that covariate effects on the hazard are multiplicative. So

$$h(t|x) = h_0(t)c(\beta' x) = h_0(t)e^{(\beta' x)} = h_0(t)e^{\left(\sum_{j=1}^{p} \beta_j x_{ji}\right)} \quad (3)$$

Proportional hazards

The Cox model is called a proportional hazards model since the ratio of hazard rates of two individuals with covariate values x_1 and x_2, at time *t* is:

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)e^{(\beta' x_1)}}{h_0(t)e^{(\beta' x_2)}} = \frac{e^{(\beta' x_1)}}{e^{(\beta' x_2)}} = e^{[\beta'(x_1 - x_2)]} \quad (4)$$

The hazard ratio is time-independent as, the ratio does not depend on t.

Since the hazard function at t given covariate *x is* $h(t|x) = h_0(t)e^{(\beta' x)}$. The survival function, the cumulative hazard function and probability density function can be derived as follows:

$$H(t|x) = \int_0^t h(u \mid x)du = \int_0^t h_0(u)e^{(\beta' x)}du = H_0(t)e^{(\beta' x)} \quad (5)$$

$$S(t|x) = e^{-[H(t \mid x)]} = e^{-\left[H_0(t)e^{(\beta' x)}\right]} \quad (6)$$

$$f(t|x) = h_0(t)e^{(\beta' x)}\, e^{-\left[H_0(t)e^{(\beta' x)}\right]} \quad (7)$$

2. Parametric PH Models

Parametric models need some special assumptions about $h_0(\mathrm{t})$, such as the exponential and Weibull distributions. But the advantage of Cox model is the fact that such assumptions can be avoided.

The Parametric PH model is the parametric versions of the Cox proportional hazards model. It is given in similar form to the Cox PH models. The main difference between the two kinds of models is that the baseline hazard function is assumed to follow a specific distribution when a fully parametric PH model is fitted to the data, while the Cox model has not such assumption. A number of different parametric PH models can be derived by choosing different hazard functions [4]. The commonly used models are exponential, Weibull, or Gompertz models.

Weibull PH model:

The Weibull model allows for hazard rates to be non-constant but monotonic that either increase or decrease exponentially with time.

Under the Weibull PH model, the hazard function of a particular patient with covariates $(x_1, x_2 .. x_p)$ is given by:

$$h(t \mid x) = \lambda\gamma\, t^{(\gamma-1)}e^{(\beta_1 x_1 + \beta_2 x_2 \ldots\ldots + \beta_p x_p)} = \lambda\gamma\, t^{(\gamma-1)}e^{(\beta' x)} \quad (8)$$

Where λ is the scale parameter and γ is the shape parameter

Exponential PH model:

The hazard function under this model is to assume that it is constant over time. Under the exponential PH model, the hazard function of a particular patient is given by:

$$h(t \mid z) = \lambda e^{(\beta_1 x_1 + \beta_2 x_2 \ldots\ldots + \beta_p x_p)} = \lambda e^{(\beta' x)} \quad (9)$$

3. Likelihood Estimation for the Cox PH Model

Derivation of an estimator of β cannot be based on an ordinary likelihood function since $h_o(\mathrm{t})$ is not specified parametrically in the Cox model. Instead, partial likelihood function has been proposed by Cox [5] for the estimation of regression parameters which is a function depending on β only.

Cox partial likelihood

Let $\mathrm{t}_1\ \mathrm{t}_2 \ldots\ldots\ \mathrm{t}_n$ be the observed survival time for n individuals, and the ordered death time of r individuals be $\mathrm{t}_{(1)}\ \mathrm{t}_{(2)} \ldots\ \mathrm{t}_{(r)}$ .The set of individuals who are at risk at t_j is denoted by $\mathrm{R}(\mathrm{t}_j)$. So that $\mathrm{R}(\mathrm{t}_j)$ is the group of individuals who are alive and uncensored at a time just prior to t_j. The conditional probability that the i^{th} individual dies at t_j given that one individual from the risk set on $\mathrm{R}(\mathrm{t}_j)$ dies at t_j is:

$$= \frac{h_i(t_j)}{\sum_{k \in R(t_j)} h_k(t_j)} = \frac{h_0(t_j)e^{(\beta' x_j)}}{\sum_{k \in R(t_j)} h_0(t_j)e^{(\beta' x_k)}} = \frac{e^{(\beta' x_j)}}{\sum_{k \in R(t_j)} e^{(\beta' x_k)}} \quad (10)$$

By taking the product of these conditional probabilities over r death times gives:

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{(\beta' x_j)}}{\sum_{k \in R(t_j)} e^{(\beta' x_k)}} \quad (11)$$

Then the partial likelihood function for the Cox PH model is given by:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{(\beta' x_i)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right]^{\delta_i} \quad (12)$$

Where $R(t_i)$ is the risk set at time t_i and δ_j is the event indicator which is zero if the i^{th} survival time is right censored and unity otherwise. This is the partial likelihood defined by Cox. The Cox methodology uses the partial likelihood to yield estimates of β that are consistent and efficient regardless of the form of $h_o(t)$. The partial likelihood is valid when there are no ties in the dataset.

4. The Score Function and Information Matrix

The regression coefficients β are estimated with $\hat{\beta}$ that maximize the partial likelihood. Assuming no ties, the log

partial likelihood is:

$$l(\beta) = \log \mathrm{L}(\beta) = log \prod_{i=1}^{n} \left[\frac{e^{(\beta' x_i)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right]^{\delta_i} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_i \left[\beta' x_i - log\left(\sum_{k \in R(t_i)} e^{(\beta' x_k)} \right) \right] \quad (13)$$

Then the score function which is the first partial derivative:

$$U_h(\beta) = \frac{\partial l(\beta)}{\partial \beta_h} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_i \left[x_{ih} - \left(\frac{\sum_{k \in R(t_i)} x_{kh}\, e^{(\beta' x_k)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right) \right] \quad (14)$$

For $h = 1, 2 \ldots p$. The maximum partial likelihood estimate $\widehat{\beta}$ can be obtained uniquely by solving the partial likelihood equation:

$$U_h(\beta) = 0$$

Whereas the second derivative of the partial likelihood is given by:

$$\frac{\partial^2 l(\beta)}{\partial \beta_g \partial \beta_h} = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_i \left[\left[\frac{\sum_{k \in R(t_i)} x_{kh}\, x_{kg}\, e^{(\beta' x_k)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right] - \left[\frac{\sum_{k \in R(t_i)} x_{kh}\, e^{(\beta' x_k)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right] \left[\frac{\sum_{k \in R(t_i)} x_{kg}\, e^{(\beta' x_k)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right] \right] \quad (15)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_g \partial \beta_h} = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_i \left[\left[\frac{\sum_{k \in R(t_i)} x_{kh}\, x_{kg}\, e^{(\beta' x_k)}}{\sum_{k \in R(t_i)} e^{(\beta' x_k)}} \right] - \left[\frac{\left(\sum_{k \in R(t_i)} x_{kh}\, e^{(\beta' x_k)} \right) \left(\sum_{k \in R(t_i)} x_{kg}\, e^{(\beta' x_k)} \right)}{\left[\sum_{k \in R(t_i)} e^{(\beta' x_k)} \right]^2} \right] \right] \quad (16)$$

This matrix for $g = 1, 2 \ldots p$. and $h = 1, 2 \ldots p.$, is a sum over $i = 1, 2 \ldots n.$ of weighted covariance matrices for the x vector in the populations at risk at the time t_i.

The negative of the second partial derivatives provide the observed information matrix which estimates the covariance matrix of the estimated regression coefficients [6].

Consequently,

$$I(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta_g \partial \beta_h} \quad (17)$$

$I(\beta)$ is defined as the observed information matrix .

The score vector $U(\beta_0)$ evaluated at the true value of β will be asymptotically distributed as a multivariate normal with mean vector zero and covariance matrix which can be unbiased estimated by $I(\beta_0)$.

$$U(\beta_0) \sim N(0, I(\beta_0)) \quad (18)$$

The estimate $\hat{\beta}$ will also be asymptoticly normal

$$\hat{\beta} \sim N(\beta_0, I^{-1}(\beta_0)) \quad (19)$$

5. Model Checking in Cox Regression Model

After the model has been fitted, the adequacy of the fitted model needs to be assessed which is usually performed using model residuals.

5.1. Cox-Snell Residuals

The Cox-Snell residual is given by Cox and Snell, which is used for assessing the fitness of PH model [7]. The Cox-Snell residual for the i[th] individual is defined as:

$$r_{ci} = exp\left(\widehat{\beta}' x_i\right) \widehat{H}_0(t_i) \quad (20)$$

Where $\widehat{H}_0(\mathrm{t_i})$ is an estimate of the baseline cumulative hazard function at time t_i. In practice the Nelson – Aalen estimate is generally used. . If the final PH model is correct and the $\hat{\beta}$ are close to the true values of the β, then r_{ci} should resemble a censored sample from a unit exponential distribution. Therefore, a plot of the Nelson-Aalen cumulative hazard estimate of residuals $\widehat{H}(r_{ci})$ versus residuals r_{ci} should be a straight line through the origin with a slope of 1, if the fitted model is correct.

5.2. Proportional Hazard Assumption Checking

The main assumption of the Cox proportional hazards model is proportional hazards, which mean that the hazard ratio is constant over time. There are several methods for verifying that a model satisfies the assumption of proportionality (Graphical method, Scaled Schoenfeld residuals, Adding time dependent covariate) [8].

- Graphical method

According to Cox regression model the survival function for i[th] individual is given by:

$$S_i(t) = [S_0(t)]^{exp\{\beta' x_i\}}$$

Where $x = (x_1, x_2 \ldots\ldots x_p)'$ is the values of the vector of explanatory variables for a particular individual. By taking the logarithm twice, we get:

$$ln[-lnS_i(t)] = \beta' x_{\mathrm{i}} + ln[-ln\, S_0(t)] \quad (21)$$

Then the difference in log-log curves corresponding to two different individuals with variables $x_1 = (x_{11}, x_{12}, \ldots\ldots x_{1p})$ and $x_2 = \left(x_{21}, x_{22}, \ldots\ldots x_{2p}\right)$ which does not depend on t is given by:

$$ln[-lnS_i(t, \mathrm{x_1})] - ln[-lnS_i(t, \mathrm{x_2})] = \sum_{i=1}^{p} \beta(x_{1i} - x_{2i}) \quad (22)$$

This provides the basis for assessing the validity of PH assumption. By plotting estimated $-log(-log(survival))$ versus survival time for two groups we would see *parallel curves* if the hazards are proportional [9]. This method does not work well for categorical predictors with many levels because the graph becomes cluttered.

- Scaled Schoenfeld residuals

Scaled Schoenfeld residuals are defined as the product of the inverse of the estimated variance-covariance matrix of the k[th] Schoenfeld residual and the k[th] Schoenfeld residual [10]. The scaled Schoenfeld residual can be used to assess time trends and lack of proportionality.

$$r^*_{pji} = (V^{-1}) r_{pji} \quad (23)$$

Where r^*_{pji} is the Scaled Schoenfeld residual and r_{pji} is the Schoenfeld residual.

Under the null hypothesis, we expect to see a constant function over time. When the proportional hazards assumption holds, straight horizontal line with zero slope is expected.

6. The Stratified Cox Regression Model

The stratified Cox regression model is a modification of the Cox regression model that allows for control by 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, however the predictors being stratified is not included. There are interaction and no-interaction models defined in the stratified Cox regression model [11].

6.1. No-Interaction Model

In the stratified model with no interaction, the strata divide the individuals into *K* disjoint groups, each having a distinct baseline hazard $h_{0k}(t)$ but a common value for the regression parameter which means that the coefficients $\beta_1, \beta_2 \ldots\ldots \beta_p$ are the same for each stratum. The hazard function for the failure time of an individual in stratum k takes the form:

$$h_k(t|x) = h_{0k}(t) exp(\beta' x) \quad (24)$$

Where k denotes the particular stratum $(k = 1, \ldots, K)$, , β is a vector of unknown regression parameters, and $h_{0k}(t)$are K unknown baseline hazard functions. The subscript k in the equation indicates that each stratum has its own baseline hazard function while the β' are the same across strata.

Under the stratified model, it can be seen that individuals within the k [th] stratum share the same baseline hazard function $h_{0k}(t)$ which implies that the proportional hazards for two individuals in the same stratum still holds:

$$\frac{h_k(t|x_1)}{h_k(t|x_2)} = exp(x_1 - x_2)\acute{\beta} \quad (25)$$

On the other hand, individuals from different groups can have non-proportional hazards as their baseline hazards functions may differ.

$$\frac{h_{0k}(t)}{h_{0k}(t)}, \text{comparing strata } k \text{ to } \acute{k}.$$

Since these functions are unrestricted, any relationship of this hazard ratio over time is possible.

The partial likelihood for the stratified Cox model is the product of partial likelihoods in each stratum:

$$L(\beta) = \prod_{k=1}^{K} L_k(\beta) \quad (26)$$

6.2. Interaction Model

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:

$$h_k(t|x) = h_{0k}(t)\, exp\left[\beta_{1k}x_1 + \beta_{2k}x_2 , \ldots + \beta_{pk}x_p\right] \quad (27)$$

In this interaction model, each regression coefficient has the subscript$k$, which denotes the kth stratum and indicates that the regression coefficients are different for different strata. So if there is no interaction the stratified Cox regression model will contains regression coefficients that do not vary over the strata. If interaction is allowed for, different coefficients for each of the stratum are obtained. Likelihood ratio test statistics is used to examine the no-interaction assumption.

$$LR = -2log\, L_{(no\ int.)} - \left[-2log\, L_{(\ int.)}\right] \quad (28)$$

The likelihood ratio (LR) test compares log likelihood statistics for the interaction model and the no-interaction model.

7. Data Set

This part of the study includes shedding light on the case study and the collected data description. The study involved 308 patients performed liver transplantation operation consecutively admitted to Egypt Air hospital and Ain Shams University hospital from January 2007 to May 2013. Among 308 patients 4 were excluded due to their age were less than 18 years and 2 for re-transplantation. So the study included 302 patients. In the following analysis, the time after the operation till death is the endpoint of interest, this variable is measured in months. There was a 2-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 302 observations, of which 81.45% are censored. The results of Cox PH regression model is obtained by using (STATA) statistical packages.

8. Description of the Variables

- *Survival time:* time to death or censoring time, measured in months.
- *Death status:* the event indicator, equal to 1 for those died during the period of the study and 0 for those who were not died or censored.
- *Recipient Age:* a dummy variable, equal to 0 if the patient age less than 50 and 1 if his age greater than or equal 50.
- *Recipient Sex:* a dummy variable, equal to 1 for male and 0 for female.

- *Donor Age:* a dummy variable, equal to 0 if the patient age less than 30 and 1 if his age greater than or equal 30.
- *Donor Sex:* a dummy variable, equal to 1 for male and 0 for female.
- *BMI:* Body Mass Index KG/m^2
- *CTP score:* stands for Child-Turcotte-Pugh score. A categorical variable, with codes 1 for class A, 2 for class B and 3 class C. Since the variable CTP has three levels, it is included in the model using the subgroup CTP_1 as the reference group.
- *MELD Score:* stands for Model for End stage Liver Disease. A categorical variable, with codes 1 for MELD score from $\{6 \text{ to } 12\}$, 2 for MELD score from $\{13 \text{ to } 18\}$and 3 for MELD score from $\{19 \text{ or higher}\}$. Since the variable MELD has three levels, it is included in the model using the subgroup $MELD_1$ as the reference group.
- *HCC:* stands for Hepatocellular Carcinoma .A dummy variable, equal to 1 for patients suffer from HCC and 0 if they did not.
- *Ascites:* a dummy variable, equal to 1 if the patients suffer from Ascites and 0 if they did not.
- *Encephalopathy:* a dummy variable, equal to 1 if the patients suffer from Encephalopathy and 0 if they did not.
- *Ln-Total Bilirubin:* mg/dl, this variable is transformed by taking the logarithm to decrease the influence of extreme values and to fit normal distribution.
- *Ln-Creatinine:* mg/dl, this variable is transformed by taking the logarithm to decrease the influence of extreme values.
- *Albumin:* a dummy variable, equal to 1 if the Albumin level is less than or equal 2.6 (≤ 2.6 mg/dl) and 0 if the Albumin level is higher than 2.6 (> 2.6mg/dl).
- *Inverse-INR:* this variable is transformed by taking the inverse to decrease the influence of extreme values.
- *Sodium:* a categorical variable, with codes 1 for Sodium level $\{\leq 130 \text{ mg/dl}\}$, 2 for Sodium level from $\{131 \text{ to } 135 \text{ mg/dl}\}$ and 3 for Sodium level from $\{136\text{mg/dl or higher}\}$. Since the variable Sodium has three levels, it is included in the model using the subgroup $Sodium_1$ as the reference group.
- *Ln-Calcium:* mg/dl, this variable is transformed by taking the logarithm to decrease the influence of extreme values.
- *Ln-Potassium:* mg/dl, this variable is transformed by taking the logarithm to decrease the influence of extreme values.
- *GRWR:* Graft to recipient weight ratio showed in percentage.

9. Analysis and Results

Table (1). *Univariate Cox PH regression analysis.*

Covariates	β	Hazard Ratio	95% CI LL	95% CI UL	p-value
Recipient Age <50		1			
Recipient Age >=50	0.505	1.657	0.964	2.846	0.067*
Recipient Sex Female		1			
Recipient Sex Male	-0.289	0.748	0.354	1.581	0.448
Donor Age < 30		1			
Donor Age >= 30	0.049	1.0507	0.616	1.789	0.855
Donor Sex Female		1			
Donor Sex Male	0.242	1.274	0.643	2.526	0.475
BMI	0.0312	1.031	0.966	1.101	0.352
HCC No		1			
HCC YES	0.131	1.14	0.667	1.95	0.63
CTP Score A		1			
CTP Score B	-0.094	0.909	0.117	7.04	0.928
CTP Score C	0.34	1.405	0.193	10.202	0.736
$MELD_1$ (6 to 12)		1			
$MELD_2$ (13 to 18	1.256	3.513	0.823	14.986	0.089*
$MELD_3$ (19 or higher)	2.218	9.1916	2.204	38.316	0.002*
Ascites No		1			
Ascites Yes	0.086	1.09	0.563	2.111	0.797
Encphalopathy No		1			
Encphalopathy Yes	0.218	1.244	0.725	2.134	0.426
Ln_TBilirubin	0.0057	1.0057	0.668	1.512	0.978
Ln_Creatinine	0.807	2.242	1.35	3.723	0.002*
Albumin < 2.6		1			
Albumin >= 2.6	0.2505	1.284	0.751	2.195	0.300
INV_INR	0.805	2.238	0.383	13.07	0.371
$Sodium_1$ (<= 130)		1			
$Sodium_2$ (131- 135)	-0.271	0.762	0.367	1.579	0.465
$Sodium_3$ (>136)	-0.425	0.653	0.33	1.293	0.200*
Ln- Calcium	-0.801	0.448	0.011	18.137	0.671
Ln-Potassium	0.642	1.901	0.395	9.153	0.423
GRWR	-1.77	0.1702	0.031	0.906	0.038*

To determine the variables to be included in the final model, the univariate Cox PH regression analysis is applied first to identify the impact of individual variable on time to event before proceeding more complicated model selection. Variables are identified as significant using a 0.2 significance level in the univariate analysis.

Table (1) presents the univariate Cox PH analysis .The first column is showing the coefficients β the parameter estimate, in the second column the hazard ratio for a one unit change in the predictor, then the 95% confidence interval and finally the p-value.

According to the univariate Cox PH analysis, that the covariates Recipient Age ($p = 0.067$) , $MELD_2$ ($p = 0.089$), $MELD_3$ ($p = 0.002$), Ln creatinine ($p = 0.002$), Sodium $_3$ ($p = 0.2$), and the GRWR ($p = 0.038$) are statistically significant and selected as significant factors for risk of death after liver transplantation operation.

- Multivariate Cox PH regression analysis

We then conducted full multivariate Cox PH analysis (by using stepwise selection process) including all the potential risk factors that had a p-value of less than or equal 0.2 in univariate Cox PH analysis.

To select the best subgroup of variables in our model, the approach of stepwise was applied. The stepwise selection process consists of a series of alternating forward selection and backward elimination steps. The former step adds variables into the model, and the latter step removes variables from the model. The threshold for variable selection into the model is setting with $p\ value$ at 0.2 (SLENTRY = 0.2), while the threshold for variable removing from the model is setting with $p\ value$ at 0.1 (SLSTAY = 0.1). It means only variables with $p\ value$ less than 0.2 will be tested in the model, and to keep it in the model, its $p\ value$ should be less than 0.1. The results from the stepwise proportional hazard regression are displayed as below.

Table (2). *Multivariate Cox PH regression analysis.*

Covariates	β	Hazard Ratio	95% CL LL	95% CL UL	p-value
Recipient Age <50		1			
Recipient Age >=50	0.748	2.113	1.163	3.841	0.014*
MELD1 (6 to 12)		1			
MELD2 (13 to 18	1.159	3.189	0.739	13.75	0.101
MELD3 (19 or higher)	2. 057	7.827	1.851	33.08	0.005*
Ln_Creatinine	0.5107	1.666	1.005	2.762	0.048*
Sodium3	-0.185	0.8305	0.463	1.487	0.532
GRWR	-1.827	0.16	0.022	1.173	0.071*

To further optimize the Cox model, the variable with the highest $p - value$ and over threshold of significance are removed from the predictive model one by one until all the rest variables are shown significant impact on the prediction of hazard rate. From Table (2) the variable Sodium$_3$ is the one with highest p-value= 0.532, so it is removed. The result is shown as below Table (3).

Table (3). *Elimination of variable with high p- value by Stepwise.*

Covariates	β	Hazard Ratio	95% CL LL	95% CL UL	p- value
Recipient Age <50		1			
Recipient Age >=50	0.596	1.816	1.04	3.154	0.034*
MELD1 (6 to 12)		1			
MELD2 (13 to 18	1.214	3.368	0.787	14.417	0.102
MELD3 (19 or higher)	2.18	8.853	2.11	37.129	0.003*
Ln_Creatinine	0.5151	1.673	0.99	2.805	0.051*
GRWR	-1.381	0.251	0.046	1.366	0.089*

And then, the same strategy is applied for the following analysis. The variable $MELD_2$ has highest $p - value = 0.102$, so removed in the following step as shown in Table (4).

Table (4). *The final Cox PH model.*

Covariates	β	Hazard Ratio	95% CL LL	95% CL UL	p- value
Recipient Age <50		1			
Recipient Age >=50	0.604	1.831	1.0547	3.178	0.032*
MELD1 (6 to 12)		1			
MELD3 (19 or higher)	1.160	3.190	1.834	5.548	0.0001*
Ln_Creatinine	0.518	1.678	1.003	2.807	0.048*
GRWR	-1.423	0.2408	0.043	1.334	0.090*

The final model is obtained as in Table (4), the data showed that most of the predictors are significant in the model with their p-value less than 0.05 except for the GRWR. After we built a multivariate model of main effects, we then check all the interactions between predictors. To test the interaction among variables, the list of all raw variables and all possible combinations of interactions are included for proportional hazard regression analysis however none of the

interactions are significant. Eventually, the final model is generated including the variables Recipient age, $MELD_3$, Ln_Creatinine, and GRWR.

The final multivariate Cox PH model is then given by:

$$h_i(t) = h_0(t)\, Exp\,((0.604\, Recipient\ Age + 1.160\, MELD_3 + 0.518\ \mathrm{Ln}\,.\,\mathrm{creatinine} - 1.423\, GRWR)$$

The final multivariate Cox PH model concluded that:

- The higher: the Recipient age, the MELD score and the Creatinine level the more the risk of death after LDLT.
- The Larger the transplanted liver graft the lower the risk of death after LDLT.

After fitting Cox PH model, we can plot the survivor, cumulative hazard, and the estimated hazard functions, as shown in figure (1).

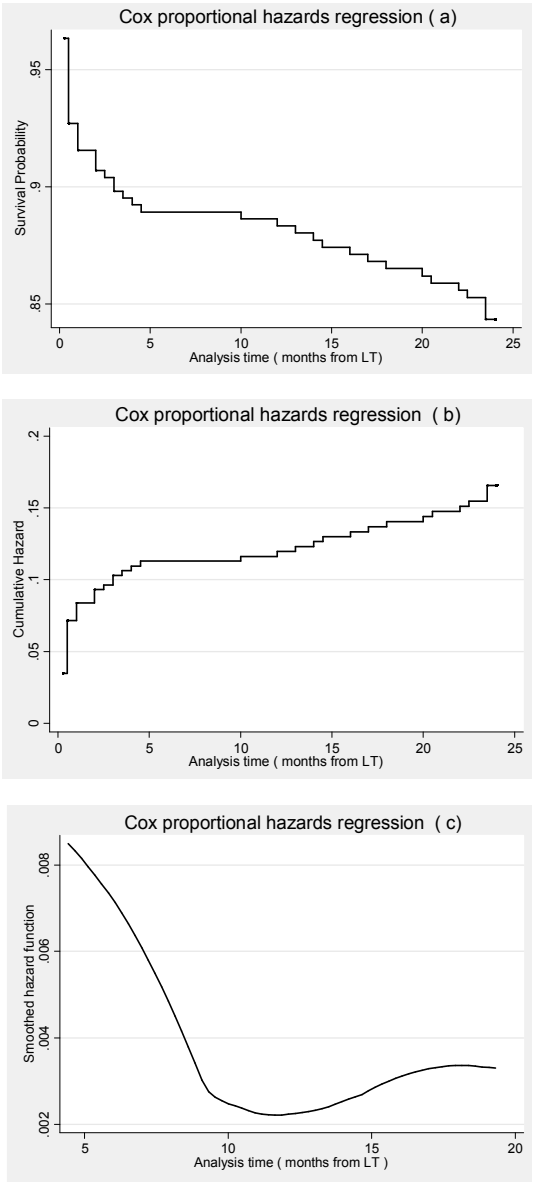


Figure (1). *Cox PH model: (a) survivorship function; (b) cumulative hazard function; (c) estimated hazard function.*

It is obvious from figure (1-c) that the hazard function is not monotonic, as it first very high during the early weeks after the LT operation, then decreases and tends to stabilize during the first year from LT, and after the 1st year it begins to increase slightly.

10. Model checking

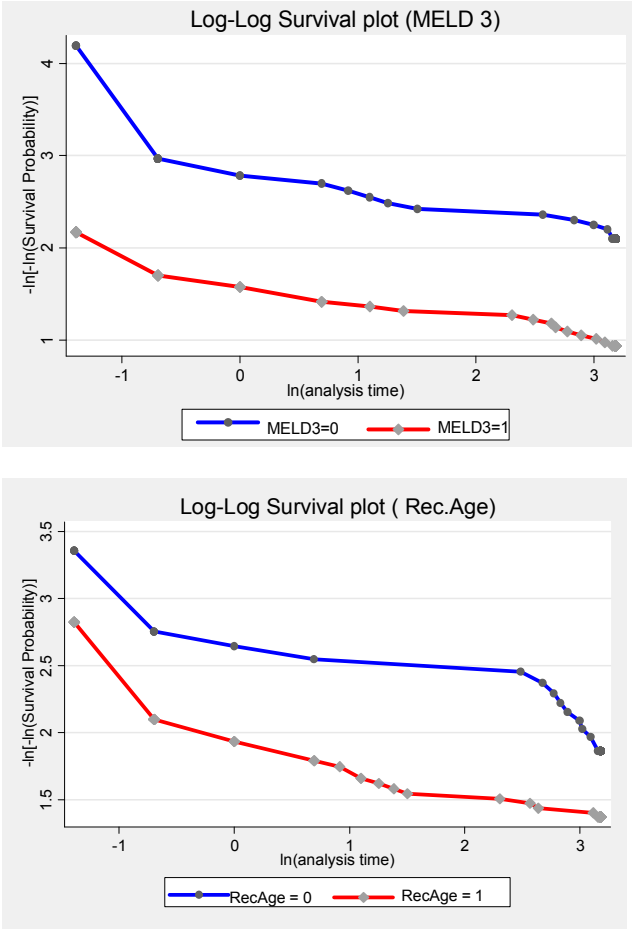
Adequacy of a fitted model needs to be assessed after a model has been constructed.

10.1. The PH Assumption Checking

The final model is based on a major assumption that the hazards between groups are proportional. To test the assumption of proportionality, the scaled Schoenfeld residuals and log-log survival plot have been used.

- Log- Log Survival Plot

Figure (2) shows $-log(-log(survival))$ plot of the variable Recipient Age and $MELD_3$. For Recipient Age the plotted lines are not parallel however for $MELD_3$ the plotted lines are parallel. Although using graphs to assess the validity of the assumption is subjective, it can be a helpful tool.



Figur (2). *Log-Log Survival plot For Rec.Age and* $MELD_3$.

- Scaled Schoenfeld residuals

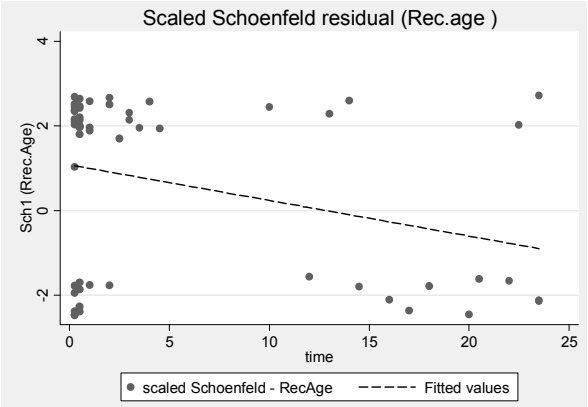
Scaled Schoenfeld residuals is based on the principle that,

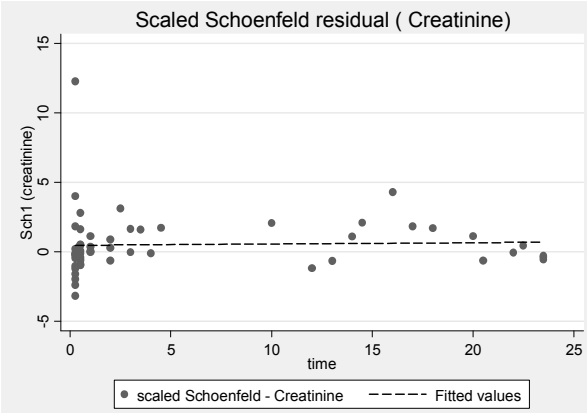
for a given regressor, the assumption restricts $\beta(t_k) = \beta$ for all t_k. This implies that a plot of $\beta(t_k)$ versus time will have a slope of zero. The null hypothesis is having zero slope, which is equivalent to testing that the log hazard-ratio function is constant over time.

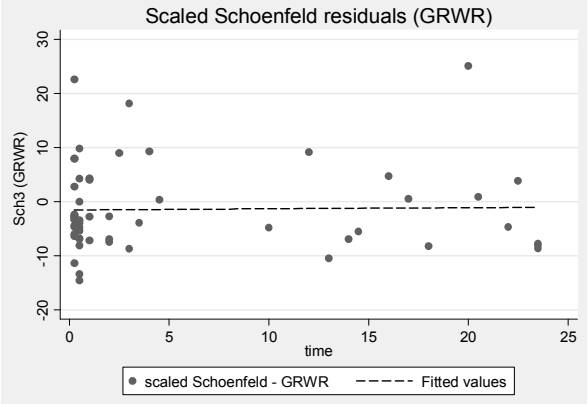
Table (5). *Test PH assumption by using Scaled Schoenfeld residuals.*

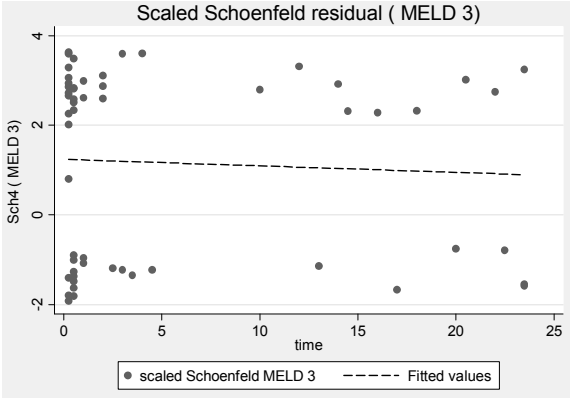
	rho	*chi2*	*df*	*Prob > chi2*
Recipient Age	-0.321	5.7	1	0.017
$MELD_3$	-0.0558	0.17	1	0.68
Creatinine	0.0348	0.08	1	0.774
GRWR	0.0191	0.03	1	0.859
Global test		5. 72	4	0.221

Table (5) shows both covariate-specific and global tests. It is obvious that Recipient age variable violates the proportional-hazards assumption $p\ value$ =0.017 which is less than 0.05.







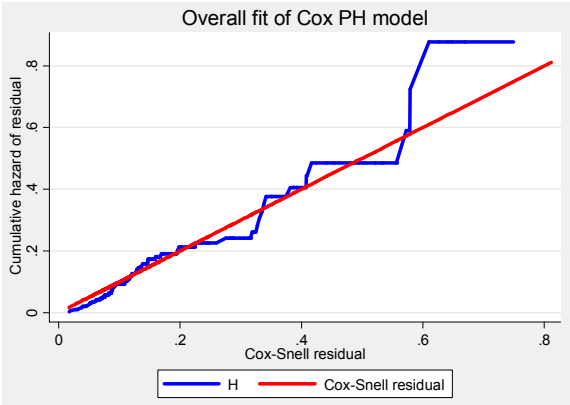


Figur (3). *Scaled Schoenfeld residuals to test PH assumption.*

Also figure (3) supports the results obtained before, variables Creatinine, $MELD_3$, and GRWR , having zero slope and does not violate the proportional-hazards assumption . However, Recipient age variable violates the proportional-hazards assumption as it does not have a zero slope.

10.2. Cox-Snell Residuals

We assess goodness of fit by using Cox–Snell residual plot.



Figur (4). *The Cox-Snell residual Plot for Cox PH model.*

In Figure (4), the blue line is the estimation of Cox-Snell residuals while the red line is the origin with a slope equals to 1. The plot suggests that the Cox PH model does not fit the straight line adequately. There is some evidence of a systematic deviation from the straight line which gives us some concern about the adequacy of the fitted model. As the scaled Schoenfeld residuals and $-log(-log(survival))$ showed that the Cox PH model displayed non-proportionality for variable Recipient Age, and the Cox-Snell residuals suggests that the Cox PH model does not fit the data adequately, so the Stratified Cox regression model is more adequate to be used.

11. The Stratified Cox Regression Model

As Schoenfeld residuals showed that the Cox PH model displayed non-proportionality for variable Recipient Age, which means that there is an interaction between this variable

and time, so the Stratified Cox regression model is more adequate to be used. Here we applied the stratified Cox regression with no interaction and with interaction model.

Table (6). *Results for the No-interaction and Interaction Models.*

Variables	No interaction model			Interaction model					
				Strata 1 Recipient Age < 50			Strata 2 Recipient Age ≥ 50		
	Coef.	**H.R.**	***P* value**	**Coef.**	**H.R.**	***P* value**	**Coef.**	**H.R.**	***P* value**
$MELD_3$	1.15	3.178	0.0001	0.970	2.63	0.034	1.271	3.567	0.0001
Ln.creat.	0.521	1.684	0.050	0.2846	1.329	0.0669	0.5606	1.751	0.045
GRWR	-1.40	0.246	0.107	-1.497	0.223	0.274	-1.347	0.259	0.232

Log likelihood for no interaction model = -256.31462
Log likelihood for interaction model = -256.072857

It is clear that in the no interaction model there is different baseline hazard function for each stratum however the coefficients are the same. However in the interaction model different baseline hazard function and different coefficients are obtained for each stratum.

Table (6) shows the application of the stratified Cox regression with no interaction and with interaction model, to determine which model is more appropriate statistically the likelihood ratio test is used, which compares log-likelihood statistics for the interaction model and the no-interaction model.

Under the null hypothesis H_0: the no interaction model is correct and statistically appropriate.

The Likelihood ratio statistics is calculated as follows:

$$\left[-2\,log\;\; likelihood_{(\,no\;interaction\;model)}\right] - \left[-2\,log\;\; likelihood_{(\,interaction\;model)}\right]$$

$$\text{The Likelihood ratio statistics} = 512.6292 - 512.1457 = 0.48348$$

This value (0.48348) is not significant at the 0.05 level of significance for 2 degrees of freedom.

Thus, it appears that despite the numerical difference between corresponding coefficients in $Recepient\ Age < 50$ and $Recepient\ age \geq 50$ models, there is no statistically significant difference. We can therefore conclude for these data that the stratified Cox model with no-interaction model is acceptable (at 0.05 level of significance).

12. Conclusions

The Cox PH model was used to examine the covariate effects on the hazard function and to determine the risk factors affecting the outcome of liver transplantation operation for end-stage liver disease.

The final multivariate Cox PH model is then given by:

$$h_i(t) = h_0(t)\,Exp\,((0.604\,Recipient\ Age + 1.160\,MELD_3 + 0.518\,\text{Ln . creatinine} - 1.423\,GRWR)$$

Cox PH model showed that the variables: Recipient age, $MELD_3$, Ln_Creatinine, and GRWR are statistically significant and selected as significant factors for risk of death after liver transplantation operation.

The scaled Schoenfeld residuals showed that the Cox PH model displayed non-proportionality for variable Recipient Age and the Stratified Cox regression model is more adequate to be used.

Also the Cox-Snell residual showed that there is some evidence of a systematic deviation from the straight line which gives us some concern about the adequacy of the fitted model. So we concluded that the Cox PH model does not fit these data adequately. The stratified Cox model with interaction and with no interaction were applied and showed that the no-interaction model is acceptable at 0.05 level of significance and the variables $MELD_3$ Score, Ln_Creatinine are statistically significant and selected as significant factors for risk of death after liver transplantation operation at 0.05 level of significance.

References

[1] Therneau, T. M., Grambsch, P. M. (2000). Modeling Survival Data, Extending the Cox Model. Springer, New York.

[2] Gill, Richard D. (1984). Understanding Cox's Regression Model: A Martingale Approach. Journal of American Statistical Association, 79: 441-447.

[3] David W. Hosmer, Jr., and Stanley Lemeshow (1999). Applied survival analysis: regression modeling of time to event data. Wiley, New York.

[4] Lawless, J. F. (2003). Statistical Models and Methods for Lifetime Data. Second Edition, Wiley, New York.

[5] Kalbfleisch, J.D. & Prentice, R.L. (2002). The Statistical Analysis of Failure Time Data. Wiley, New York.

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

[7] Cox, D.R. and Oakes, D., (1984) Analysis of Survival Data. Chapman and Hall, London.

[8] Collett D. (1994). Modeling survival data in Medical research. Chapman & Hall, London.

[9] Klembaum, D. G. (1996). Survival Analysis: A Self learning text. Springer, New York.

[10] Grambsch, P. and Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrrika, 81: 515–526.

[11] Lee, E. T., and Wang, J. W. (2003). Statistical Methods for Survival Data Analysis. Third Edition, Wiley, New York.