Modeling survival durations of patients undergoing aortic valve replacement

Liberato Camilleri, Lawrence Grech, Alex Manche'
Department of Statistics and Operations Research
University of Malta
Msida (MSD 06) Malta
E-mail: liberato.camilleri@um.edu.mt

KEYWORDS
Kaplan Meier, Nelson Aalen, Cox regression models, AFT and PH parametric survival models.

ABSTRACT

Survival analysis is a useful statistical tool for problems that deal with survival data. This data is used in order to analyze the predicted duration for a certain event to occur. Initial survival analysis was linked explicitly with events related to death. However, this is no longer the case and nowadays survival analysis is used in almost all research areas to model duration of device failure or relapse duration to drug, smoking and alcohol addiction. This paper presents several approaches to model survival durations of patients undergoing aortic valve replacement. These survival models will be used to relate survival durations for censored data to several pre- and post-operative patient related variables to identify risks factors.

1. Introduction

Throughout the centuries, survival analysis was used solely to investigate mortality rates; however, in the last fifty years, applications of survival analysis have been extended to various fields. Survival analysis is now used in marketing to model shelf-life duration or consumption duration of food products; in industry to model the lifetime duration of electronic devices; in criminology to model prison durations of offenders; in health insurance to model cure durations from certain diseases, and in sociology to model marriage durations before divorce. Moreover, survival analysis can be used to estimate survival durations and life expectancy, amongst other applications.

In 1958, Kaplan and Meier presented the product limit estimator to estimate the survival function from life duration data. This non-parametric statistic accommodates censored data to estimate survival probabilities and hazard rates. Initially Kaplan and Meier submitted separate papers with similar results but John Tukey, the editor of the Journal of the American Statistical Association, convinced them to combine their efforts and produce a single paper. An alternative non-parametric approach is the Nelson-Aalen estimator, which can be used to estimate the cumulative hazard rate function for censored data. This estimator was originally introduced by Nelson but later on Aalen extended its use by investigating its small and large sample properties using martingale methods. No distributional assumptions are required for both Kaplan Meier and Nelson Aalen estimators.

The seminal paper entitled 'Regression models and life tables' proposed by Cox (1972) introduced the proportional hazard (PH) model. The semi-parametric model specifies that the conditional hazard function of failure time given a set of predictors is the product of an unknown baseline hazard function, which is a function of time (parametric part) and an exponential function of the linear combination of the predictors (non-parametric part). The Cox model can be used to compare the relative forces of two lives, given that they have the same baseline hazard. In this approach, Cox estimated the regression parameters by maximizing the partial log-likelihood function. Breslow (1972) suggested an alternative approach in which the cumulative baseline hazard and the regression parameters are estimated simultaneously. The Breslow estimator, which yields both the estimator for the cumulative baseline hazard function as well as Cox's estimator of the regression parameters has been used extensively in research and is implemented in several statistical software packages.

The Cox regression model is based on the assumption that the effects of the covariates being predicted remain constant over time. This limitation can be a problem when the shape and nature of the hazard functions are unknown. On the other hand, parametric models are based on the assumption that the lifetime distribution belongs to a given family of parametric distributions. This parametric approach links the survival duration to a set of predictors using a specified probability distribution for the hazard function. For a constant hazard function the exponential distribution is used; for a monotonic increasing/decreasing hazard function the Weibull or Gompertz distribution is used; and for a humped hazard function the log-normal or log-logistic distribution is used. The accelerated failure time (AFT) parametric models relax the assumption of proportional hazards and assume that the logarithm of the survival time is a linear function of the predictors. In other words, in PH models the predictors act multiplicatively on the hazard, while in AFT models the predictors act multiplicatively on time.

2. Non-Parametric Survival Techniques

Let $t_1 < t_2 < ... < t_r$ be the ordered times of observed deaths of N lives. Moreover, let d_i be the number of deaths observed at time t_i for $1 \le i \le r$; let c_i be the number of censored observation in the time interval $[t_i, t_{i+1})$ and let n_i be the number of lives still alive before t_i. Moreover, let h_i be the hazard rate at time t_i, which is the probability of instantaneous death at t_i.

$$h_i = P(T = t_i \,|\, T \geq t_i) \tag{1}$$

If the observed deaths and survivals are independent, it can be shown that the likelihood function is the product of the likelihood of all deaths and the likelihood of all censored lives surviving until the times at which observations are censored. Moreover, they showed that likelihood function can be expressed as the product of independent binomial likelihoods.

$$L = \prod_{i=1}^{r} h_i^{d_i} \left(1 - h_i\right)^{n_i - d_i} \tag{2}$$

By differentiating the log-likelihood function with respect to h_i and setting the result equal to 0 yields:

$$\hat{h}_i = \frac{d_i}{n_i} \tag{3}$$

By assuming non-informative censoring, the Kaplan-Meier estimate of the survival function $S(t)$ is given by:

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \left(1 - \hat{h}_i\right) = \prod_{t_i \leq t} \left(\frac{n_i - d_i}{n_i}\right) \tag{4}$$

Moreover, var$[S_{KM}(t)]$ is given by:

$$\text{var}[\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 \sum_{t_i \leq t} \frac{d_i}{n_i (n_i - d_i)} \tag{5}$$

In the Nelson-Aalen approach, $\hat{h}_i$ values are used to estimate the integrated hazard function $\Lambda(t)$ given by:

$$\Lambda(t) = \int_0^t u_s ds + \sum_{t_i \leq t} h_i \tag{6}$$

By assuming non-informative censoring and by considering discrete hazards h_i occurring at times $t_i \leq t$, the Nelson-Aalen estimate of the survival function $S(t)$ is given by:

$$\hat{S}_{NA}(t) = e^{-\hat{\Lambda}(t)} \text{ where } \hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{7}$$

Moreover, var$[\hat{\Lambda}(t)]$ is given by:

$$\text{var}[\hat{\Lambda}(t)] = \sum_{t_i \leq t} \frac{d_i (n_i - d_i)}{n_i^3} \tag{8}$$

When h_i is very small, $e^{-h_i} \approx 1 - h_i$, hence

$$\hat{S}_{NA}(t) = e^{-\hat{\Lambda}(t)} = \prod_{t_j \leq t} e^{-\hat{h}_i} \approx \prod_{t_j \leq t} (1 - \hat{h}_i) = \hat{S}_{KM}(t) \tag{9}$$

3. Semi-Parametric Survival Techniques

One of the most widely used in survival models is the proportional hazards model, proposed by Cox (1972). The semi-parametric survival model is made up of the two components. The parametric part is the baseline hazard function, $h_0(t)$ which defines how risk varies with time. The non-parametric part is the exponential function, $\exp(\mathbf{x}_i'\boldsymbol{\beta})$, of a linear combination of the predictors (risk factors). The proportional hazard model is given by:

$$h(t \,|\, \mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i'\boldsymbol{\beta}) \text{ where } \mathbf{x}_i'\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j x_{ij} \tag{10}$$

With an increase in the j^{th} covariate, the hazard rate increases when $\beta_j > 0$, and decreases when $\beta_j < 0$. The absolute force of mortality of a life cannot be estimated without estimating the baseline hazard; however, if one wishes to compare the relative forces of mortality of two lives with similar baseline hazards then

$$\frac{h(t \,|\, \mathbf{x}_1)}{h(t \,|\, \mathbf{x}_2)} = \frac{h_0(t) \exp \sum_{j=1}^{p} \beta_j x_{1j}}{h_0(t) \exp \sum_{j=1}^{p} \beta_j x_{2j}} = \exp \sum_{j=1}^{p} \beta_j (x_{1j} - x_{2j})$$

This implies that the hazards of the two lives will remain proportional over time. Moreover, the logarithm of the hazard ratio increases by β_j for every 1 unit increase in $x_{1j} - x_{2j}$.

Let $R(t_i)$ denote the set of lives that are at risk before time t_i. If $d_i = 1$ for $1 \leq i \leq r$ then the partial likelihood function is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{k \in R(t_i)} \exp(\mathbf{x}_k'\boldsymbol{\beta})} \tag{11}$$

Each observed lifetime contributes to the probability that the life observed to die should have been the one out of the $R(t_i)$ lives at risk to die, conditional on the fact that one death was observed at time t_i. So the contribution from the first death to the partial likelihood is the force of mortality for the first life to die divided by the total force of mortality for the lives in the risk group just before the event occurred. The partial likelihood function considers solely observed deaths and the contribution of the censored observation enters indirectly in the total force of mortality, which is the denominator of the partial likelihood function. The baseline hazard function disappears from the partial likelihood function because it cancels out. Ties occur when some observations are censored exactly at an observed death or there may be more than one death at each observed lifetime ($d_i > 1$). The first case is dealt with by assuming that censoring always occur after the death was observed. When two or more lives die at the same time t_i their contribution to the partial likelihood should be included in the risk group $R(t_i)$. The modified partial likelihood function is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\left[\sum_{k \in R(t_i)} \exp(\mathbf{x}_k'\boldsymbol{\beta})\right]^{d_i}} \tag{12}$$

When maximizing the partial likelihood function, estimation of the parameters will be based on the order, rather than the time, in which the deaths occurred. Nevertheless, the model seeks to identify the factors that influence mortality rates and hence increase or reduce the chance of a premature death. Maximization of the partial likelihood yields the maximum likelihood estimates of the parameters and provides the link between of the observed covariates and hazard rates. When

the Cox model includes several covariates, the process of maximizing the partial likelihood function may be very cumbersome and cannot be achieved directly. However, it can be maximized using an iterative numerical technique such as the Newton-Raphson method.

The partial likelihood estimator for $\boldsymbol{\beta}$ is unbiased and has an asymptotic multivariate normal distribution. Moreover, the asymptotic variance matrix can be estimated by the inverse of the observed information matrix from which the standard errors of the parameter estimates can be computed.

$$\text{var}(\hat{\beta}_j) = -\frac{1}{\partial^2 \log L / \partial \beta_j^2} \text{ evaluated at } \hat{\beta}_j$$

4. Parametric Survival Techniques

To adjust the survival functions for the effects of the covariates, two models are used which include the accelerated failure-time (AFT) model and the proportional hazards (PH) model. In the PH model, the concomitant predictors (covariates) have a multiplicative effect on the hazard function

$$h(t) = h_0(t)\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \qquad (13)$$

where $\beta_0, \beta_1, ..., \beta_p$ are regression parameters; $x_1, x_2, ..., x_p$ are predictors and $h_0(t)$ is the baseline hazard function. In the PH model, the covariates have a multiplicative effect on the hazard function. The PH models accommodated by STATA include the Exponential, Gompertz and Weibull distributions.

Table 1: PH models accommodated by STATA

Distribution	Survival Function	Parametrization
Exponential	$\exp\left(-\lambda_j t_j\right)$	$\lambda_j = \exp(\mathbf{x}_i'\boldsymbol{\beta})$
Weibull	$\exp\left(-\lambda_j t_j^{\alpha}\right)$	$\lambda_j = \exp(\mathbf{x}_i'\boldsymbol{\beta})$
Gompertz	$\exp[-\lambda_j \gamma^{-1}(e^{\gamma t_j} - 1)]$	$\lambda_j = \exp(\mathbf{x}_i'\boldsymbol{\beta})$

In the AFT model, the logarithm of the survival time is expressed as a linear function of the covariates.

$$\log t = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \qquad (14)$$

The AFT model changes the time scale by a factor of $\exp[-(\mathbf{x}_i'\boldsymbol{\beta})]$. Depending on whether this factor is greater or less than 1, time is either accelerated or decelerated. The AFT models accommodated by STATA include the Exponential, Weibull, Log-normal and Log-logistic distributions.

Table 2: AFT models accommodated by STATA

Distribution	Survival Function	Parametrization
Exponential	$\exp\left(-\lambda_j t_j\right)$	$\lambda_j = \exp[-(\mathbf{x}_i'\boldsymbol{\beta})]$
Weibull	$\exp\left(-\lambda_j t_j^{\alpha}\right)$	$\lambda_j = \exp[-\alpha(\mathbf{x}_i'\boldsymbol{\beta})]$
Log-normal	$1\text{-}\Phi[(\log(t_j) - \mu_j)/\sigma]$	$\mu_j = \mathbf{x}_i'\boldsymbol{\beta}$
Log-logistic	$[1 + (\lambda_j t_j)^{1/\gamma}]^{-1}$	$\lambda_j = \exp[-(\mathbf{x}_i'\boldsymbol{\beta})]$

To determine which model provides the best fit, the researcher can either use the Akaike Information Criterion (AIC), proposed by Akaike (1974) or the Bayesian Information Criterion (BIC) proposed by Schwartz (1978). The AIC penalizes the log-likelihood by the number of estimated parameters (p), while BIC penalizes the log-likelihood by the sample size (N) and the number of estimated parameters (p).

$$AIC = -2(\text{log-likelihood}) + 2p \qquad (15)$$

$$BIC = -2(\text{log-likelihood})(1) + p \log N \qquad (16)$$

The model which provides the smallest information criterion provides the best fit for a particular dataset.

5. Application

The dataset consists of 480 patients who underwent an aortic valve replacement at the cardiothoracic centre in a Maltese hospital. This data was collected by a cardio-vascular surgeon over a period of 16 years, ranging between 2003 and 2019. Most of the patients who underwent this treatment were aged over 60 years, which is expected since the prevalence of heart disease increases drastically with age. After surgery, all patients had follow-up appointments. The time of death of patients who died before the end of the investigation period (2019) was recorded and the survival duration was computed. Patients who were still alive after the end of the investigation period were right censored.

The dataset includes a number of patient-related explanatory variables, together with other information related to the patients' health conditions in pre-operative and the post-operative periods. In this study, the dependent variable is **Time**, which is a continuous variable measuring the survival duration between the surgery and the time of death/end of the investigation period. The categorical variable **Status** indicates whether the patient was dead or alive at the end of the investigation period and will be used as a censoring variable. The continuous variable **BMI** provides the ratio of the patient's weight (kilograms) to the patient's height squared (m^2). The **Parsonnet Score** has a metric scale and measures the risk of death of a patient after undergoing heart surgery, where the larger the score the higher is the risk. The continuous variables **HDU** and **ITU** record the duration (days) of the patient's recovery in the High Dependency Unit and the Intensive Therapy Unit respectively. The categorical variables **Diabetes**, **Hypertension** and **Dialysis** indicate whether the patient was diabetic, had high blood pressure and was on dialysis. The categorical variable **Transfusion** indicates whether the patient required/not required blood transfusion directly from another individual. The continuous variable **Ventilation** measures the duration (hours) that the patient spent on a life-assisting mechanical ventilator following the surgery. The categorical variable **Creatinine** indicates the presence/absence of waste product in the blood that normally passes through the kidneys and is eliminated through urine. The continuous variable **Bleeding** measures the blood volume (millilitres) that was provided to the patient during or after surgery. The categorical variable **IABP** indicates whether or not the patient required an intra-aortic balloon pump during heart surgery.

Table 3: Descriptive statistics of continuous variables

Variable	Mean	St. Deviation
Time	4012.16	1576.43
BMI	29.44	4.330
Parsonnet Score	6.24	5.122
ITU	1.04	0.331
HDU	1.19	3.603
Ventilation	5.24	6.766
Bleeding	565.66	268.434

Table 4: Frequency table (categorical variables)

Variable	Frequency	Percentage
Diabetes	440	29.7%
Hypertension	753	50.9%
Transfusion	518	35.0%
Dialysis	25	1.7%
Creatinine	47	3.2%
IABP	39	2.6%

Of the 1480 patients participating in the study, 22.8% died before the end of the investigation period, while the rest 77.2% were right censored. Table 3 displays the means and standard deviations of each continuous risk factor. The mean Parsonnet score (6.24) indicates that the risk of mortality is fair and that there is a 5% predicted mortality rate. All the patients undergoing heart surgery spend one night in ITU and are retained in this unit if health condition is critical. If the patients' health condition is not life-threatening, they are transferred to the HDU for a convalescence period. The mean duration of patients requiring support of a ventilator was 5.24 hours and the mean blood volume transfused was 565.66 millilitres; however, these values were considerably larger for high risk patients. The mean BMI (29.44 kg/m^2) is larger than average indicating that the majority of the patients were overweight or obese.

Table 4 displays the frequency and percentage of each categorical risk factor. 29.7% of the patients were diabetic; 50.9% suffered from high blood pressure; 1.7% were on dialysis, 2.6% required the use of an intra-aortic balloon pump during surgery; 35% required blood transfusion and 3.2% of the patients had the presence of creatinine.

6. Results of non-parametric survival methods

The Kaplan-Meier estimates of the survival probabilities were computed using the facilities of STATA. Figures 1 and 2 show the Kaplan Meier survival distributions and 95% confidence intervals when patients are grouped by diabetes and dialysis condition. The Logrank test was used to compare survival distributions of groups of patients clustered by categorical risk factors. The Logrank test identifies two significant risk factors, which include dialysis [$X^2(1) = 54.51$, $p < 0.001$] and diabetes [$X^2(1) = 11.51$, $p = 0.007$]. Hypertension, creatinine, transfusion and IABP were not found to be significant risk factors. Figures 3 and 4 show the Nelson Aalen cumulative hazard functions and 95% confidence intervals when patients are grouped by diabetes and dialysis condition.

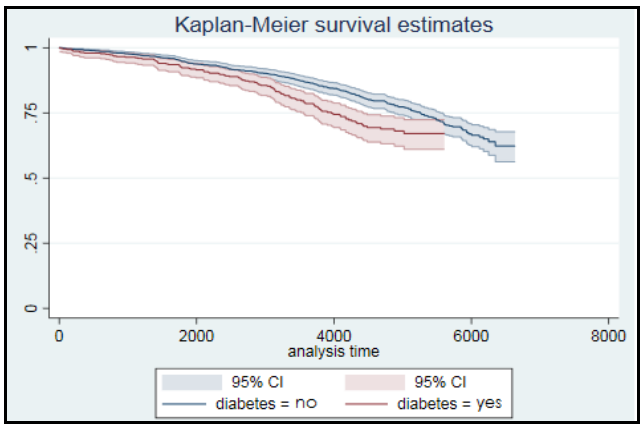

Figure 1: Survival function for patients grouped by diabetes

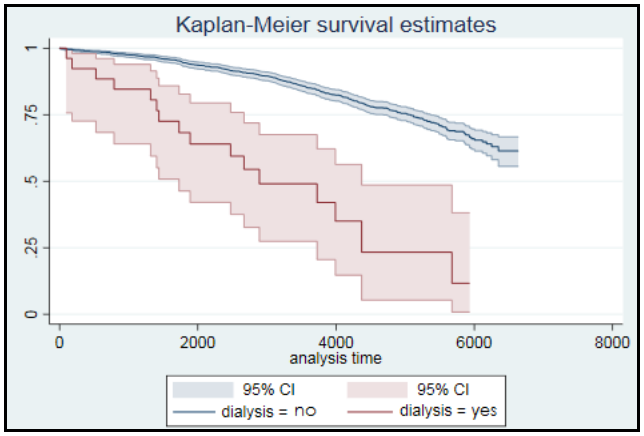

Figure 2: Survival function for patients grouped by dialysis

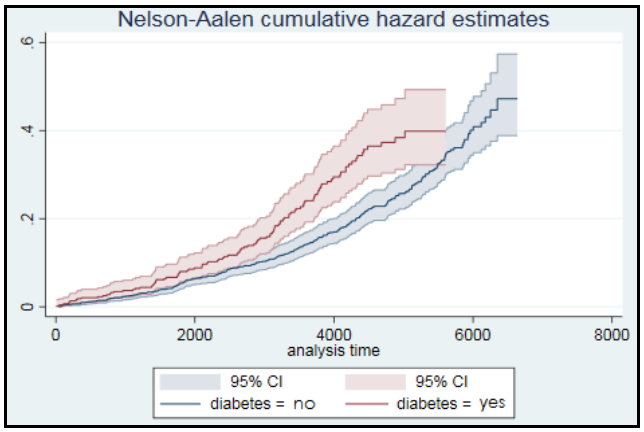

Figure 3: Cumulative hazard functions (grouped by diabetes)

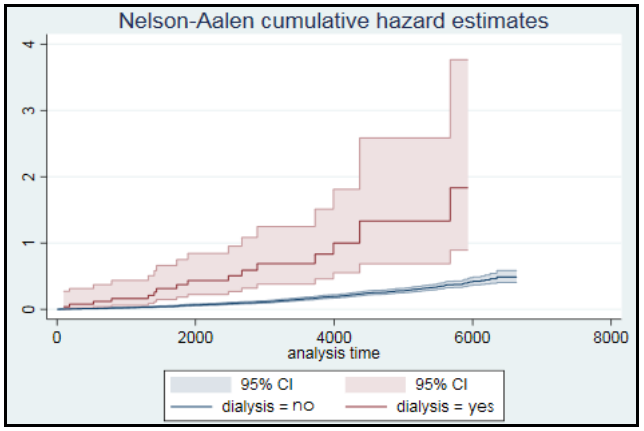

Figure 4: Cumulative hazard functions (grouped by dialysis)

7. Results of semi-parametric survival methods

The `estat phtest` was used to test the proportionality hazard assumption for the Cox regression model that includes 12 risk factors. This test confirmed that the proportionality hazard assumption is satisfied [$X^2(12) = 10.38$, $p = 0.583$].

Table 5: Hazard ratios of Cox model

Parameter	HR	S.E.	Z	$P > \|z\|$
BMI	1.001	0.0124	0.11	0.910
Diabetes	1.304	0.1726	2.00	0.045
Hypertension	1.439	0.4953	1.06	0.290
Parsonnet	1.091	0.0105	9.03	0.000
ITU	0.847	0.1606	-0.87	0.382
HDU	1.051	0.0119	4.36	0.000
Ventilation	0.992	0.0112	-0.72	0.474
Bleeding	1.000	0.0002	0.23	0.815
IABP	1.470	0.5076	1.12	0.265
Dialysis	4.837	1.2884	5.92	0.000
Creatinine	0.994	0.0114	-0.50	0.615
Transfusion	0.942	0.1322	-0.42	0.675
Log-Likelihood				-2222.45
LR test				$X^2(12) = 135.53$, $p < 0.001$

Table 5 displays the hazard ratios and standard errors of the 12 risk factors. To identify the parsimonious model a backward elimination procedure was used. Table 6 shows that this model includes four significant risk factors, where the Parsonnet score is the best predictor of survival duration because it has the lowest p-value. It is followed by dialysis condition, treatment duration in the High Dependency Unit and diabetes condition.

Table 6: Hazard ratios of parsimonious Cox model

Parameter	HR	S.E.	Z	$P > \|z\|$
Diabetes	1.514	0.1835	2.26	0.024
Parsonnet	1.087	0.0103	8.85	0.000
HDU	1.047	0.0098	4.89	0.000
Dialysis	4.268	1.1079	5.59	0.000
Log-Likelihood				-2229.31
LR test				$X^2(4) = 121.81$, $p < 0.001$

The hazards of death of patients who are on dialysis or are diabetic are respectively 4.268 and 1.514 times than patients who do not have these conditions. Moreover, for every extra day of treatment in the High Dependency Unit, the hazard of death increases by 4.7% and for every 1 unit increase in the Parsonnet score the risk of death increases by 8.7%, given that other effects are kept constant.

8. Results of parametric survival methods

Since the proportionality hazard assumption was satisfied for this data set, PH survival models were fitted. These include the Exponential distribution leading to a constant hazard; the Gompertz leading to an exponential hazard, and the Weibull distribution leading to a monotonic increasing or decreasing hazard. Table 7 displays the log-likelihood, AIC and BIC values for the parametric survival models assuming these three survival distributions. The parametric survival model assuming a Gompertz distribution provides the best fit since it yields the smallest AIC and BIC values. The estimate of the ancillary parameter γ is 0.00034.

Table 7: AIC and BIC values of the PH survival models

Distribution	Log-likelihood	*p*	AIC	BIC
Exponential	-962.3	4	1932.6	1953.8
Gompertz	-917.0	5	1844.0	1870.5
Weibull	-927.1	5	1864.2	1890.7

Table 8 displays the hazard ratios and standard errors of the 4 significant risk factors of the Gompertz survival model and the results resemble those of the Cox regression model.

Table 8: Hazard ratios of Gompertz survival model

Parameter	HR	S.E.	Z	$P > \|z\|$
Diabetes	1.430	0.1641	2.18	0.029
Parsonnet	1.094	0.0102	9.63	0.000
HDU	1.051	0.0109	4.79	0.000
Dialysis	4.182	1.1508	5.43	0.000
Log-Likelihood				-917.0
LR test				$X^2(4) = 110.45$, $p < 0.001$

9. Conclusion

The Cox regression model is based on the assumption that the effects of the covariates being predicted remain constant over time. This limitation can be a problem when the shape and nature of the hazard functions are unknown. On the other hand, parametric models are based on the assumption that the survival distribution has a known parametric form. The two modeling approaches yielded similar results because the proportionality hazard assumption was satisfied.

References

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

Kaplan, E. L., & Meier, P (1958). Nonparametric estimation from incomplete observations, Journal of the American Statistical Association 53(282), 457-481.

AUTHOR BIOGRAPHY

LIBERATO CAMILLERI studied Mathematics and Statistics at the University of Malta. He received his PhD degree in Applied Statistics from Lancaster University. His research specialization areas are related to statistical models, which include Generalized Linear models, Latent Class models, Multilevel models and Mixture models. He is an associate professor and Head of the Statistics department at the University of Malta.