Webology (ISSN: 1735-188X)
Volume 19, Number 5, 2022

Survival Models By Non-Parametric And Semi-Parametric Methods For Patients Infected With Coronavirus In Al-Kindi Teaching Hospital

Gheed Rafid Amer[1] , Assistant Prof. Ali Yassin Ghani Al-Badrawi[2]

[1]Department of statistics, Mustansiriyah University, Falastin St., Baghdad, Iraq.

[2]Department of statistics, Mustansiriyah University, Falastin St., Baghdad, Iraq.

Abstract

A study of the non-parametric survival model (Kaplan-Meier) and the semi-parametric Cox Regression model. From the practical side, it was found that the effect of the change of age by (3.483) when the patient's age was transferred from one age group to another on the estimation of the survival function by semi-parametric method using the (Cox Regression) model. From the comparison between the models of survival (nonparametric, semi-parametric) from the mean squares of relative error (RMSE) statistics, it was found that the best model for estimating the survival function is the nonparametric model (Kaplan-Meier). The study came out with several results, the most important of which is that by estimating the survival function by the nonparametric method (Kaplan-Meier), it is possible to obtain the lowest cumulative risk rate for each survival time. This means that the probability of the patient staying in the time period (t) increases and that the risk rate is affected by the change in the patient's age and duration of stay when estimating the survival and cumulative risks by the semi-parametric method (Cox Regerssion).

Keywords

Non-Parametric; Semi-Parametric; Survival models

Introduction

Due to the lack of studies in the survival analysis for emerging corona disease (Covid 19), the study of survival analysis for patients infected with Coronavirus at Al-Kindi Teaching Hospital was chosen because the disease led to the end of life for many people.

Research Aims

1. Study and analysis of survival rates for people infected with the Coronavirus hospitalized in Al-Kindi Teaching Hospital and the most important factors that affect survival time.
2. Comparison of survival models estimated by non-parametric and semi-parametric estimation methods by calculating mean relative error (RMSE).

3. Evaluation of the most important explanatory variables identified in the semi-parametric model affect survival time.

The Probability density function

It is a function used to represent the probability distribution of any random variable. The survival time T is like any continuous random variable that has a domain of positive values only and has a probability density function defined as a single probability target that may fail during the period from (t) to (t + Δt) according to the formula (1):

$$f(t) = \Pr(T = t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(t \leq T \leq t + \Delta t) \qquad (1)$$

where Δt is very little but sufficient time for the event to occur.

The Survival function

The survival function of the random variable survival time (T) represents the probability that the survival time for an item is greater than or equal to the observed survival time (t). The probability of the event we are interested in will occur at a time greater than or equal to the observed survival time t, and it is denoted by the symbol S(t) .It is expressed by the formula (2):

$$S(t) = P(\text{an individual fails at or after } t)$$

$$S(t) = P(T \geq t) \qquad (2)$$

Hazard function h(t)

The risk function represents the probability of death of the patient under study during the period ((t+Δt,t), given that the patient was alive during time t. Therefore, the risk function represented the instantaneous failure rate of the individual live to the observed survival time t and symbolized by the symbol h(t). Its formula (4):

$$h(t) = \lim_{\Delta t \to 0} \frac{P((t \leq T < t + \Delta t)|T \geq t)}{\Delta t} \qquad (3)$$

$$h(t) = \frac{f(t)}{S(t)} \qquad (4)$$

One of the properties of the risk function is that it is a positive function h(t)≥0, and has no upper bound. The importance of the risk function comes from the fact that it expresses the change during the patient's life or represents the risk for each item. The aggregate risk function is defined as the cumulative sum of the risk rates faced by a given individual from the origin of time until the observed survival time t, symbolized by the symbol H(t). It is expressed mathematically as in formulas (5) and (6):

$$H(t) = \int_{0}^{t} h(u)du \qquad (5)$$

$$H(t) = \int_{0}^{t} \frac{f(u)}{S(u)} du = - \int_{0}^{t} \frac{1}{S(u)} \left\{\frac{d}{du} S(u)\right\} du \quad = - \ln S(t) \qquad (6)$$

Nonparametric Method

Nonparametric estimation methods depend on the direct inference of the survival function by arranging and experimenting with data on survival time. And resort to it when finding the appropriate theoretical distribution of the data is impossible. We will discuss one of these methods, which is the Kaplan-Meier method.

Kaplan-Meier method

This method is one of the most widely used nonparametric estimation methods. This is due to its relevance to various life-length data, whether in the medical fields (to measure part of the patients' lives for a certain period) or in the industrial fields (the maintenance officer measures the time until the failure of the product or machine). The estimation of the survival function according to this method is defined by the formula (7):

$$\hat{S}_{K.M(t_i)} = \prod_{t_i \leq t} \left[\frac{n_i - d_i}{n_i}\right] \qquad (7)$$

Since:

$\hat{S}_{K.M(t_i)}$: Estimation of the survival function by the Kaplan–Meier method
n_i: The setting of times to stay in time (t_i).
d_i: The number of people who died at the moment in time(t_i)

Semi- Parametric Method

Most models that contain a set of unknown parameters (β) representing the parametric part, with an anonymous link function representing the non-parametric part, are called semi-parametric models [g(.)]. This model is often used in models where parametric assumptions are ill-defined and inconsistent. Or that the nonparametric model is not fully functional. One of the semi-parametric methods used in this research to find the risk function is the Cox Regression method.

Cox Regression Method

The description of the relationship between the Hazard Rate and a set of explanatory variables can be done with a regression model of the formula (8):

$$ln[h(t)] = ln[h_0(t)] + \sum_{i=1}^{p} \beta_i x_i \qquad (8)$$

Or in the formula (9):

$$ln[-lnS(t;x)] = x'\underline{\beta} + ln[h_0(t)] \qquad (9)$$

It is a (Cox Regression) model, where the parameter vector (β) is estimated (through the partial potential function). The introductory risk rate $[h_0(t)]$ is represented when the explanatory variables $[x_i]$ are equal to zero, or the model is described in terms of (relative risk) in the formula (10):

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \sum_{i=1}^{p} \beta_i x_i \qquad (10)$$

It is a form that does not have a hard limit parameter. Where this parameter becomes part of (h(t)), as for the cumulative risk function, and when it is under the assumption of the regression of the relative position and in the presence of explanatory variables, it is in the formula (11):

$$H(t,x) = H_0(t)\exp\left(\sum_{i=1}^{p} \beta_i x_i\right) \qquad (11)$$

It is possible to use the ranks of the failure time (t) to estimate the parameters of the model in the formula (11), and it is noticeable that the relationship $H_0(t)$ describes the survival time (t) and its absence with in ($\exp\left(\sum_{i=1}^{p} \beta_i x_i\right)$). In the same way, the cumulative survival function is described by the formula (12):

$$S(t,x) = S_0\exp\left(\sum_{i=1}^{p} \beta_i x_i\right) \qquad (12)$$

The parameters (β) indicate the amount of change in the logarithm of the risk rate when (x_i) it increases by one unit, noting that the positive value of the parameters (β) indicates that increasing the value of the explanatory change by one unit leads to an increase in the risk function and thus the situation worsens. As for the negative value of the parameter, it indicates that by increasing the explanatory variable by one unit, the risk decreases, and the case under study improves.

Cox Regression Parameters Estimation

The method of estimating the parameters of the (Cox) model in 1975, the scientist proposed the partial possibility method to estimate the parameters of the (Cox Regression) model, as this method gives an expectation of the values of the dependent variable through the independent variables in the absence of knowledge of the nature of the primary risk function ($h_0(t)$) The partial function is in form (13):

$$l_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_i \beta_i}}{\sum_{j=R(t_i)} e^{x_j \beta_j}} \qquad (13)$$

Where (m) represents the number of failures, and (x_i) represents the variable values of the item whose survival time (t_i), and the risk group ($R(t_i)$), that is, all the items at risk of failure are $R(t_i) = \{j = t_j \geq t_i\}$, and by taking the natural logarithm of the formula (14) it becomes:

$$\log\left(l_p(\beta)\right) = \sum_{i=1}^{m}\left\{x_i\beta_i - \log\sum_{j=R(t_i)} e^{x_j\beta_j}\right\} \qquad (14)$$

After deriving the equation (14) and equating it to zero, we get the model's parameters.

Research data

Data were obtained from the tympanic membranes of diagnosed and diagnosed MERS-CoV patients in Al Kindi Teaching Hospital. These patients were monitored for a period of (60) days, starting from (15-6-2021) until (14-8-2021), and the sample size was (50) patients. The data for this study are the dates of diagnosis of the disease until the date of death, improvement or loss of follow-up of the patient, representing the survival time.

Survival function estimation

The nonparametric method

By (Kaplan-Meier) method, the survival function and cumulative risk were estimated by (SPSS) program, and the results are in Table 1, where the survival function started approximately (98%), but over time, the number of those who died it has increased, and then the survival function has decreased and ended about (15%), and this confirms that the survival function is inversely proportional to time, while the cumulative risk function is directly proportional to the increase in the period, then the incremental risk rate increases to reach (2), which It indicates an increase in the probability of the death of the injured in the period, as the highest risk rate is for the patient who remained under observation for a period of (33) days until he died.

Table 1. Survival and cumulative risk functions Kaplan – Meier.

i	t	S(t)	H(t)	i	t	S(t)	H(t)
1	2	0.97872	0.02151	13	12	0.63495	0.45421
2	3	0.95691	0.04398	14	12	0.63495	0.45421
3	5	0.91029	0.09399	15	14	0.5976	0.51484
4	5	0.91029	0.09399	16	15	0.55776	0.58383
5	6	0.88569	0.12139	17	17	0.51485	0.66387
6	7	0.83508	0.18023	18	19	0.46337	0.76923
7	7	0.83508	0.18023	19	20	0.41188	0.88702
8	8	0.78289	0.24477	20	21	0.3604	1.02055
9	8	0.78289	0.24477	21	28	0.30033	1.20287
10	10	0.75589	0.27986	22	33	0.15017	1.89602
11	11	0.72565	0.32068	23	33	0.15017	1.89602
12	12	0.63495	0.45421				

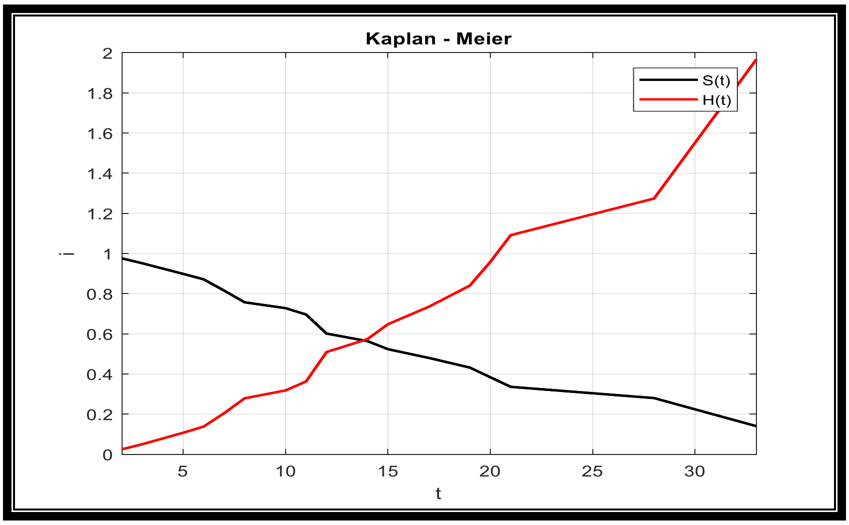

Figure 1. S(t) and H(t) for Kaplan – Meier

Semi-parametric method

The Cox Regression model assumes a risk function for patients infected with Coronavirus, which is affected by three variables, namely, the patient's age, gender, occupation, and survival time. Partial and estimated model parameters as in Table 2.

Table 2. Wald test

Variables	B	SE	Wald	Df	Sig.	Exp(B)
Age	1.248	0.534	5.468	1	0.019	3.483
Sex	-0.487	0.493	0.978	1	0.323	0.614
Job	0.216	0.297	0.526	1	0.468	1.241

The values of the (Wald) test show the values of the test parameters of the model. It indicates through the values of the (sig.) column that the variable of age is the only significant variable in the study variables, because the value of (sig.=0.019) is less than (0.05), and this indicates that the age of the patient has a significant effect on the survival time according to the data of this research, and since the parameter value (1.248), which is a positive amount, indicates that the increase in the explanatory variable the patient’s age, i.e. (transferring from one age group to a later age group) leads to an increase in the risk and that the patient’s condition tends to for the worse by ($e^{1.248} = 3.483$), where the following relationship describes the survival model:

$$\mathbf{S(t)} = S_0 e^{1.248t_i} \qquad (15)$$

As for the survival and risk functions of the estimated model, they are shown in Table (3), from which the survival function started at about (97%). With time, the number of those who died has increased. Then the survival function may range between increasing and decreasing as it decreases and ends at approximately (2%) at the survival time (33) days. This confirms that the survival function is inversely proportional to time. In contrast, the cumulative risk function is directly proportional to time and is affected by the age variable. The probability of the death of the injured in the period, as the highest cumulative risk rate is for the patient who remained under observation for a period of (33) days until he died at the age of (68).

Table 3. Survival and cumulative risk functions Cox Regerssion

i	t	Age	S(t)	H(t)	i	t	Age	S(t)	H(t)
1	2	78	0.97434	0.02599	13	12	69	0.5854	0.53546
2	3	79	0.9319	0.07053	14	12	63	0.64948	0.43158
3	5	60	0.86032	0.15045	15	14	80	0.44909	0.80053
4	5	87	0.89165	0.11468	16	15	60	0.24452	1.40847
5	6	65	0.82309	0.19469	17	17	66	0.26958	1.31088
6	7	69	0.80224	0.22035	18	19	58	0.69461	0.36441
7	7	60	0.7608	0.27339	19	20	69	0.2194	1.51686
8	8	62	0.67634	0.39105	20	21	85	0.15776	1.84666
9	8	57	0.8938	0.11227	21	28	55	0.50164	0.68987
10	10	55	0.88621	0.1208	22	33	55	0.42017	0.86709
11	11	73	0.73352	0.30991	23	33	67	0.02359	3.74707
12	12	63	0.41831	0.87153					

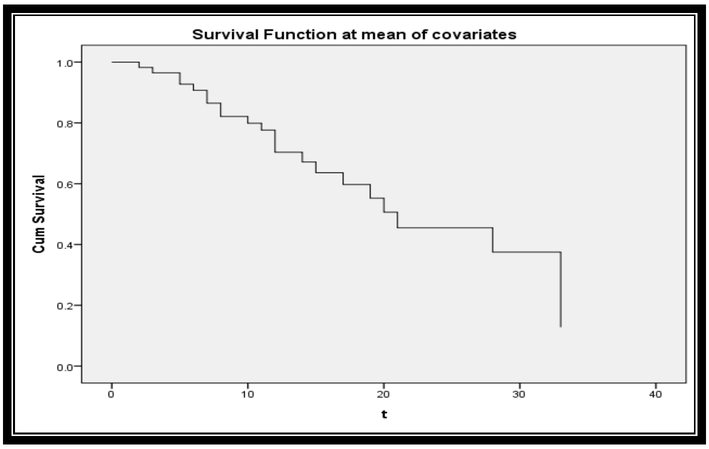

Figure 2. S(t) for Cox Regression

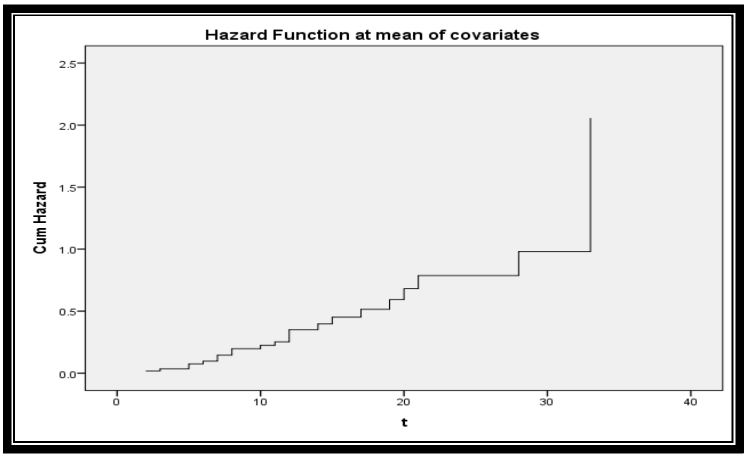

Figure 3. H(t) for Cox Regression

An advantage in estimating the survival function

To know whether the estimation of the survival function by the nonparametric method using the (Kaplan-Meier) is best? or is the estimation of the survival function by the semi-parametric method (Cox Regression) the best? The RMSE statistic was used, as in the table:

Table 4. The sum of squares of errors.

Method	Kaplan – Meier	Cox Regression
RMSE	0.21977	0.2403

Table 4 shows the value of (RMSE) for estimating the survival function by the nonparametric method using (Kaplan-Meier) is less than (RMSE) for semi-parametric survival function estimation (Cox Regression), then the nonparametric method by (Kaplan-Meier) in estimating the survival function is the best.

Conclusions

When estimating the survival function by the nonparametric method (Kaplan-Meier), it is possible to obtain the lowest cumulative risk rate for each survival time, which means the patient's probability of survival in the period (t) increases. We conclude from the mean of the relative error squares that the difference between the survival models estimated by (non-parametric, semi-parametric) is small.

Acknowledgements

Reference

Al-Nasser, A. M. H. (2009). An Introduction to Statistical Reliability.

Arslan, T., Acitas, S., & Senoglu, B. (2017). Generalized Lindley and Power Lindley distributions for modeling the wind speed data. Energy Conversion and Management, 152, 300-311. doi:https://doi.org/10.1016/j.enconman.2017.08.017

Cox, D. R., & Oakes, D. (1984). Analysis of survival data: Chapman and Hail, New York.

D. R. Cox, & E. J. Shell. (1968). A general dimensions of residuals with discussion. Journal of royal Statistical Society, 30, 248-275.

Liao, H.-W. (1998). A simulation study of estimators in stratified proportional hazards models. In.

Machin, D., Cheung, Y. B., & Parmar, M. (2006). Survival analysis: a practical approach: John Wiley & Sons.

Wang, M.-C., & Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. Journal of the American Statistical Association, 94(445), 146-153.