**AN ADDITIVE SUBDISTRIBUTION HAZARDS MODEL**

**FOR COMPETING RISKS DATA**

**Molydah S1, Danardono2**

1 Post Graduate Student of Department of Mathematics,

Universitas Gadjah Mada

2 Department of Mathematics, Universitas Gadjah Mada Yogyakarta

**e-mail**: *molydah@mail.ugm.ac.id*

**DOI: 10.14710/medstat.XX.X.XX-XX**

|  |  |
| --- | --- |
| **Article Info:**  Received:  Accepted:  Available Online:  **Keywords:**  *Additive hazard model; competing risk; cumulative incidence function; sub-distribution hazard*. | **Abstract:** Competing risk failure time data occur frequently in medical investigations. Several methods have been proposed for the analysis of these data. The classic approach is to model all cause-specific hazards and then estimate the cumulative incidence curve based on these cause-specific hazards. Unfortunately, the cause-specific hazard function does not have a direct interpretation in terms of survival probabilities for the particular failure type. In this paper, we consider a more flexible model for the sub-distribution. It is a combination of the additive model and the Cox model and allows one to perform a more detailed study of covariate effects. One advantage of this approach is that our regression modeling allows for non-proportional hazards. This leads to a new simple goodness-of-fit procedure for the proportional sub-distribution hazards assumption that is very easy to use. We applied this method to melanoma data and estimated the cumulative death rate for those who died from melanoma after surgical removal of the tumor. It was found that two covariates had a time-varying effect and two other covariates had a constant effect in predicting the cumulative incidence curve in patients who died of melanoma following tumor removal surgery. |

1. **INTRODUCTION**

Survival analysis is a statistical method where the outcome variable that is considered is the time until an event occurs or survival time. In survival data, the problem that often arises is the presence of censored observations (Allison, 2010). Censored observations occur when the survival time of the individual being observed is not known with certainty. Censored is the basic concept that distinguishes survival analysis from conventional statistical methods, causing univariate, bivariate, and multivariate analysis to be invalid for analyzing survival data, so special statistical methods are needed, one of which is the Cox regression model, also known as the Cox Proportional Hazard model (Cox, 1972). In survival analysis the most popular regression method used is the semiparametric regression method, this is because semiparametric regression does not require assumptions about the survival time distribution, but the results of the parameter estimates are close to the parametric regression method. The semiparametric regression method that is often used in survival analysis is the Cox regression model. The application of the Cox model for some situations is sometimes inappropriate, one of which is when competing risks occur. In general, competing risks arise when an individual can experience more than one type of event, and the occurrence of these events precludes the occurrence of other types of events. Often, the timing of an event in competing risk is influenced by one or several independent variables (covariate).

The cumulative event curve (CIF) is the probability that a certain type of event occurs at or before a certain point in time and is an appropriate summary curve in analyzing competing risk. The Kaplan-Meier (KM) method has become a widely used tool for estimating survival functions and cumulative occurrence functions. This method is conceptually easy to understand and easy to compute, however, if there is more than one type of event (or failure), and if these events are dependent, the KM method is biased. This bias arises because the KM method assumes that all events are independent, therefore the KM method is not appropriate for estimating the cumulative event curve.

In biomedical studies, it is important to study the effect of covariates on the cumulative occurrence function of a particular failure. The standard approach is to model a cause-specific hazard for all causes. The Cox proportional hazard model is the most commonly used regression model for all causes. This approach is valid when all cause-specific hazards are modeled correctly because the cumulative incidence curve for a particular cause is a function of all cause-specific hazards. However, this method has the disadvantage that it is difficult to identify which specific covariate has a time-varying effect on the cumulative incidence curve. Overcoming this problem, Fine and Gray (1999) proposed a model based on sub-distribution that can estimate the effect of covariates on the cumulative incidence curve. In this study, the authors focus on a more flexible and general model for the hazard sub-distribution function in estimating the effect of covariates on CIF. This model is a combination of the additive model and the Cox model, making it possible to carry out a more detailed study of covariate effects. The method will be applied to the censored competing risk data. Sun and Liu (2006) extended the Cox proportional model to a more general additive hazard model for hazard sub-distributions with independent time covariates and used the IPCW technique for censored data. In addition, the advantages of the model allow for non-proportional hazard in this case leading to the goodness-of-fit test procedure.

1. **LITERATURE REVIEW**
   1. **Survival Analysis**

Survival analysis is one of the statistical methods used to analyze data where the variable considered is the time until an event occurs. Time can be expressed in units of days, weeks, months, or even years, which are obtained from the initial observation of an individual until an event occurs for that individual (Kleinbaum and Klein, 2012). The purpose of survival analysis is to determine the relationship between the time of occurrence and the independent variables measured at the time of the study. In addition, it is also used to identify the factors that significantly influence an event. In survival data, one of the problems that arise is incomplete observations, which are generally grouped into censored data and truncated data. Censored data is data that cannot be observed in its entirety because the research subject is missing or for other reasons so that the data cannot be retrieved, or until the end of the study, the subject has not experienced a certain event. According to Collet (2003), censored data types are divided into three types, namely:

1. The left censorship occurs when the initial time of the observation subject is not observed, but the full failure time event can be observed before the research study ends.
2. The right censorship occurs when the subjects included in the observation can be observed in full but, until the end of the study, had not experienced an incident.
3. An interval sensor is a sensor whose survival time is within an interval of certain.

Another problem with survival data is that the truncated data is divided into left- and right-truncated data. Left-truncated data occurs when the individual's incident time is less than the left-truncated time, while right-truncated data occurs when the individual's incident time is more than the right-truncated time. Individuals who experienced a left or right slit were not included in the observation.

* 1. **Survival Function and Hazard Function**

In survival analysis, there are two basic functions, namely the survival function and the hazard function. The survival function is a basic function used to describe the phenomenon of its occurrence. The survival function is denoted by S(*t*), which is the opportunity for an individual to survive longer than time *t* (Kleinbaum and Klein, 2012). The survival function is a probability, so its value is always in the interval [0, 1]. The survival function is defined as follows:

|  |  |
| --- | --- |
|  | (1) |

If is a continuous random variable, then the survival function is the complement of the cumulative distribution function, where the cumulative distribution is defined as , so the survival function can be written as follows:

|  |  |
| --- | --- |
|  | (2) |

Furthermore, the value of can be obtained through the probability density function as follows:

|  |  |
| --- | --- |
|  | (3) |

Theoretically, the survival function is a non-increasing function with respect to time *t* with the following characteristics:

1. The survival function is a decreasing monotone function.
2. For = 0 then , meaning that the chance for a research unit to survive at is 1.
3. For maka shows that as time goes by, the chance for a research unit to survive will get smaller, so that if is very large, then the chance for a research unit to survive will be close to zero.

An important measure in survival analysis besides the survival function is the hazard function. The hazard function often referred to as the hazard rate, is denoted by The hazard function can be defined as the rate of occurrence of an event if it is known that an object of research survives until time . Systematically, it can be written as follows:

|  |  |
| --- | --- |
|  | (4) |

* 1. **Estimation of Survival Function and Hazard Function**

The Kaplan-Meier estimator or Product Limit Estimator is a nonparametric estimator that is often used to estimate survival functions. The Kaplan-Meier estimator is given as follows (Klein and Moeschberger, 2005):

|  |  |
| --- | --- |
|  | (5) |

where is the number of individuals who experience the event at time and is the number of individuals at risk at time where is the survival time observed in object, for . The Kaplan-Meier estimator is a function of the ladder that goes down when there is an event.

To estimate the survival function using the Kaplan-Meier estimator , first calculate the standard error or variance of the survival function. The variance of the Kaplan-Meier estimator can be found using the Greenwoods formula (Klein and Moeschberger, 2005):

|  |  |
| --- | --- |
|  | (6) |

or you can use the following formula as an alternative:

|  |  |
| --- | --- |
|  | (7) |

* 1. **Competing Risk Model**

Competing risk is a situation where an individual may experience more than one type of event, and the occurrence of one type of event precludes the occurrence of another. If failure were the different causes of death, then only death from the first cause would be observed and considered an event of interest, whereas death from any other cause would be considered a competing risk. The existence of competitive risks implies that the usual methods of survival must be applied with caution, and the consequences of their use must be recognized. One method that is often used to estimate the probability of survival over a certain period of time is the Kaplan-Meier method.

In the case of competing risks, the Kaplan-Meier method considers failures from other causes or competing risks as censored observations, but this causes bias. Prentice, et al (1978) suggest an approach that can be used in calculating competing risk, namely the cumulative event function (CIF), by using this technique, the probability of each event that occurs is partitioned into the probability for each type of event. In general, the standard approach that is often used in estimating and modeling the cumulative event function is by estimating and modeling the cause specific hazard cause as shown in Figure 1.

|  |
| --- |
|  |
| **Figure 2**. Figure Illustrating the Competing Risks Model |

* 1. **Additive Sub-distribution Hazard Model**

In the case of competing risks, assuming two types of failures the cumulative incidence function for cause 1 given a set of covariates *x* is given by

|  |  |
| --- | --- |
|  |  |

where is the failure time, ε indicates the cause of failure. To estimate , following Gray (1988) and Fine and Gray (1999), we consider the modeling of the hazard function of instead of the cause-specific hazard function. A major advantage of this is that one can directly estimate without simultaneously estimating sub-distributions corresponding to other failure Specifically, define

|  |  |
| --- | --- |
|  |  |

Then }, where =. It the following, it is assumed that has the form

|  |  |
| --- | --- |
|  | (8) |

where is an unknown *q*-vector of time-varying components representing effects of covariates ***X*** on , is a *p*-vector of unknown regression parameters denoting the effects of covariates **Z** on , and is an unspecified baseline hazard function. The above model assumes that covariates affect in two ways, additive covariate effects described by an additive model and multiplicative covariate effects characterized by the Cox model. The additive model allows time-varying covariate effects, while the Cox model allows only a common dependence through the baseline.

* 1. **Estimation Procedures**

be independent replicates of {}. Then the observed data are }. Define and . To estimate the unknown parameters , and , following Martinussen and Scheike (2002) and using IPCW techniques, we propose to employ the following score functions

For right censored competing risks data, the counting process, , is observed only up to the censoring time for censored individuals. Now consider the inverse censoring weighted response | that has the mean equal to the cumulative incidence function, since

|  |  |
| --- | --- |
|  |  |

where is a prespecified constant such that For a given , solving the second and third score equations gives a weighted Aelen estimator

|  |  |
| --- | --- |
|  | (9) |

for , and a Breslow estimator

|  |  |
| --- | --- |
|  | (10) |

for , where and **.** By inserting (2) and (3) in the first score equation for , we obtain

|  |  |
| --- | --- |
|  | (11) |

where Let denote the solution to (4). Once is obtained, one can estimate the cumulative baseline hazard function and the additive components by the Breslow estimator and the weighted Aalen estimator , respectively. If there is only a single cause of failure, (2) —(4) reduce to the score equations of Martinussen and Scheike (2002). To investigate asymptotic properties of , and , let and be defined as ,,**Q** and with replaced by . In the Appendix, it is shown that has an asymptotic normal distribution with mean zero and a covariance matrix that can be consistently estimated by

|  |  |
| --- | --- |
|  | (12) |

It thus follows that has an asymptotic normal distribution with mean zero and covariance matrix that can be consistently estimated by

|  |  |
| --- | --- |
|  | (13) |

Furthermore, that converges weakly to a zero-mean Gaussian process whose covariance function at can be consistently estimated by

|  |  |
| --- | --- |
|  | (14) |

Similarly, converges weakly to a zero-mean Gaussian process whose covariance function at can be consistently estimated by

|  |  |
| --- | --- |
|  | (15) |

* 1. **Prediction of Cumulative Incidence Functions**

One of the main goals in survival analysis is to predict certain survival probabilities for future subjects. To predict under (8) for a patient with a set of covariates and , one can first estimate the cumulative distribution hazard by

|  |  |
| --- | --- |
|  | (16) |

The predicted cumulative incidence is then given by . Furthermore, an estimate for the 100*p*th percentile of can be obtained by solving the equation , where is such that . Using the functional -method, one can show that for a known, monotone, absolutely continuous transformation converges weakly to a zero-mean Gaussian process whose covariance function at can be consistently estimated by

|  |  |
| --- | --- |
|  | (17) |

The above transformation is usually chosen to stabilize the variance and to ensure that pointwise and simultaneous confidence intervals for the probability are bounded between and . One commonly used choice is Sometimes one is interested in constructing confidence bands for , or . This may be analytically difficult since the limiting Gaussian processes for , , and } do not have independent increments. To this end, we propose to use the following simulation approach to approximate these limiting distri butions as in Lin, Fleming and Wei (1994) and Scheike and Zhang (2003). Let be be a simple random sample of size from the standard normal distribution and independent of the observed data. Then one can construct the simultaneous confidence bands for , or or tp by replacing and with and , respectively, and repeatedly generating normal random samples given the observed data. Note that since is nonnegative, one may want to use the log transformation for the construction of its confidence bands.

* 1. **Test for Model Identification**

This section considers the goodness of fit test of the model and the of time-varying covariates. For these, we develop some asymptotically procedures. To evaluate the goodness of fit of the covariates included in the multiplicative part of the model, following Lin, Wei and Ying (1993) and Wei (1984), consider the cumulative score processes. The observed score process is given by , t), and its asymptotic distribution is equivalent distribution of

|  |  |
| --- | --- |
|  |  |

where is the minus of the derivative of . Note that if multiplicative part of the model is appropriate, the components of the score process should behave as under the null. This suggests that we can use the following test statistics

|  |  |
| --- | --- |
|  | (18) |

where denotes the *j*th component of . The percentiles of this test statistic can be estimated empirically using a number of simulated processes as discussed in the previous section, or in Lin, Wei and Ying (1993). Now consider testing if covariate *j*, included in the additive part of the model, is significant. For this, we suggest the test statistic

|  |  |
| --- | --- |
|  | (19) |

where is the *j*th component of and is the estimate of the variance of . Sometimes one may also be interested in testing if an additive component has indeed a time-varying effect. To this end, we propose the test statistic

|  |  |
| --- | --- |
|  | (20) |

Note that evaluates the departure of from the null, while measures the departure between and the estimate of the constant effect under the null, . Also note that the asymptotic distribution of is equivalent to the asymptotic distribution of where is defined in (7). Then the percentiles of the above two test statistics can be simulated as before. The proposed tests are simple to implement and are omnibus. Additionally, one can plot the estimated cumulative regression function and use the plots to visually examine whether a covariate has a time-varying effect on the cumulative incidence function.

1. **MATERIAL AND METHOD** 
   1. **Data**

The data used in this case study is secondary data entitled Melanoma obtained from the MASS package in program R (Venables & Ripley, 2002). Melanoma data consists of measurements performed on patients with malignant melanoma, a type of skin cancer. Each patient underwent surgical removal of the tumor at the Department of Plastic Surgery, Odense University Hospital, Denmark during the period 1962 to 1977. The surgery consisted of the complete removal of the tumor together with approximately 2.5 cm of the surrounding skin. Among the measurements taken were the thickness of the tumor and see whether there was ulceration or not. This measure is considered an important prognostic variable in patients with thick tumors and/or an increased likelihood of death from melanoma.

* 1. **Variables and Methods**

The case of melanoma patients, there are three possible events, namely patients who died from melanoma, patients who died not because of melanoma, and patients who were still alive until the end of the study. Each patient will only experience one event out of three possibilities. The incident of concern is the patient who died of melanoma. If the patient dies not due to melanoma then the patient is considered a competing risk individual. Each patient is represented by one row in the data set. Of the 205 patients under observation, 134 patients were still alive, while 71 patients had died, of which 57 patients died of melanoma and 14 patients died of causes unrelated to melanoma. The following variables are used in the data:

1. Dependent variable is a time variable (in days), which is the duration of time the patient survives during surgery until he dies or is censored, and an event variable which is the type of event experienced by the patient.
2. Independent variable (covariate) is a variable that is thought to influence the response variable, including the variables sex, age, thickness, and ulcer. The sex variable is the sex variable of the patient. Age variable (in years), is the patient's age variable at the time of surgery. Thick variable, is the patient's tumor thickness variable (in mm) at the time of surgery. The ulcer variable is an ulcer indicator variable.

The following is a description of the data for categorical variables:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1**. Table of frequencies and coding of categorical variables | | | |
| Variables | Category | Code | Total |
| Event | Censored | 0 | 134 |
| Death with Melanoma | 1 | 57 |
| Death witout melanoma | 2 | 14 |
| Sex | Male | 0 | 79 |
| Famale | 1 | 126 |
| Ulcer | Absent | 0 | 90 |
| Present | 1 | 115 |

In this simulation study overall of 205 patients with four risk factors will entered into the regression model to predict the likelihood of the cumulative incidence function of patients dying of melanoma using flexible models. This model is available in the **timereg** package (Scheike & Martinussen, 2006) and **cmprsk** packages (Gray, 2022) in the R program. In this case, the researchers only focused on patients who died after undergoing tumor removal surgery due to melanoma.

1. **RESULTS AND DISCUSSION**

In the analysis, to visually examine which risk factors had a time-varying effect, we first fitted the data with the additive model, with covariates including sex, age, thick and ulcer, and for all cases. The test gave p-value of 0,335, 0,736, 0,003, and 0,005 for sex, age, thick, and ulcer, respectively, which indicates that thick and ulcer variables had time-varying effects while, sex and age had a constant effects on the cumulative incidence function of patients who died of melanoma after perform tumor removal surgery.

To fit (8) given the above results, let X include thick and ulcer, and Z be sex and age. That is, assume that thick and ulcer affected the hazard of melanoma time-dependently and sex and age affected the hazard constantly. For the goodness-of-fit assessment of the proposed model, we first considered the test for sex and age variables and obtained p-value 0.210 and 0,694 , respectively, that sex and age variables could be included in the multiplicative part. Second, we performed and tests for thick and ulcer variables. The test gave p-values similar to those obtained before and confirmed that thick and ulcer had time-varying effects. Based on the additive model, the test gave p-values of 0.000 and 0,000 for thick and ulcer, respectively, which indicates that both risk factors had significant on the cumulative incidence function in patients who died from melanoma. The above results (1) is an appropriate model for this data set. Based on a multiplicative model, the test gave p-values of 0.018 and 0,265 for sex and age, respectively, this indicates that sex had a significant on the cumulative incidence function of patients who died from melanoma, while the variable age is not a significant on the cumulative incidence function of patients who died from melanoma after perform tumor removal surgery.

In the next process, we consider predictions for two different patients defined by the new data assignment below. Patient type I: (sex = 1), 52 years old, thick= 0,62 and (ulcer = present), and patient type II: (sex = 0), 52 years old, thick= 0,62 and (ulcer = absent). Predictions based on the model can be seen in the following figure:

|  |
| --- |
| Patient type II  Patient type I |
| **Figure 2**. Prediction of the Cumulative Incidence Curves |

1. **CONCLUSION**

The additive sub-distribution hazard regression model is a flexible model for regression analysis of competing risk failure time data. In applying additive sub-distribution hazard regression analysis to melanoma patient data, it was found that the thick and ulcer variables had a time-varying effect on the cumulative incidence function of patients who died of melanoma, while the sex and age variables were not constant on the cumulative incidence function of patients who died because of melanoma after perform tumor removal surgery.

**REFERENCES**

Allison, Paul. D., 2010, Survival Analysis, United Kingdom: Emerald Group Publishing

Ltd.

Cox, D. R., 1972 Regression Models and Life-Table, Jurnal of the Royal Statistical

Society: Series B (Methodologi), nomor 2, volume 34, halaman 187-220.

Collet, D., 2003, Modelling Survival Data in Medical Research (Second Edition), London:

Chapman dan Hall/CRS.

Fine, J. P., dan Gray. R. J., 1999, A Proportional Hazards Model for the Subdistribution of a Competing Risk, Journal of the American Statistical Association, No. 446, Vol 94, p 496-509.

Gray, R. J., 1988, A Class of K-sample Test for Comparing the Cumulative Incidence of a Competing Risk. Ann, Statist, Vol 16, p1141-1154.

Gray B., 2022, \_cmprsk: Subdistribution Analysis of Competing Risks\_. R package version 2.2-11, <https://CRAN.R-project.org/package=cmprsk>

Lin, D. Y., Wei, L. J., and Ying, Z., 1993, Checking the Cox model with cumulative sums martingale-based residuals, Biometrika, Vol. 80, p. 283-298.

Kleinbaum, D. G., dan Klein, M., 2015, Survival Analysis A Self Learning Text (Second Edition), Springer.

Klein, P., dan Moeschberger, L., 2005, Survival Analysis: Techniques for Censored and Truncated Data (Second Edition), Newyork: Springer.

Kleinbaum, D. G., dan Klein, M., 2015, Survival Analysis A Self Learning Text (Third Edition), Newyork: Springer.

Martinussen, T., dan Scheike, T. H., 2002, A Flexible Additive Multiplicative Hazard Model, Journal of Biometrika, Vol 89, p 283-298.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Jr., Flournoy, V. T., Breslow, N. E., 1978, The Analysis of Failure Times in the Presence of Competing Risks, Journal of Biometrics, No 4, Vol 34, p 541-554.

Scheike, T. H & Martinussen, T, 2006, Dynamic Regression models for survival data. Springer, New-York

Sun, L., Liu, J., Sun, J., Zhang, M. J., 2006, Modeling The Subdistribution of a Competing Risks, Journal of the Statistica Sinica, No. 4, Vol. 16, p 1367-1385.

Venables, W. N. & Ripley, B. D. ,2002, Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Wei, L. J., 1984, Testing goodness of fit for proportional hazards model with censored observations, J. Amer. Statist. Assoc, Vol. 79, p 649-652.