Review Article

iMedPub Journals
www.imedpub.com

Insights in Biomedicine
ISSN 2572-5610

2019
Vol.4 No.1:4

DOI: 10.36648/2572-5610.4.1.55

Estimation of Right-Censored Data with Partially Linear Models: A Comparison for Different Censorship Solution Methods

Yilmaz E*

Faculty of Science, Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey

*Corresponding author: Dr. Ersin Yilmaz

yilmazersin13@hotmail.com

Faculty of Science, Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey.

Tel: +90 252 211 10

Abstract

Recently, because there is information inflation, collected data generally include missing or censored values. Therefore, it is hard modelling and analyzing datasets accurately especially in medical and clinical studies. In this paper, to handle the right-censored data which is the most common kind of censored data, three different solution methods are introduced. After that to see their effects on the modelling process, the semiparametric regression model is used based on smoothing spline.

Keywords: Censored data; Survival analysis; Partially linear models; Imputation methods; Solution of censorship

Received: December 28, 2019; **Accepted:** January 30, 2019; **Published:** February 02, 2019

Citation: Yilmaz E (2019) Estimation of Right-Censored Data with Partially Linear Models: A Comparison for Different Censorship Solution Methods. Insights Biomed Vol.4 No.1:4

Introduction

In biomedical applications mainly in clinical trials the two important issues arise when studying time to event data. We will assume the event to be death. It can be any event of interest.

1. Some individuals are still alive at the end of the study or analysis so the event of interest namely death hasn't occurred. Therefore, we have right censored data.
2. Length of follow-up varies due to staggered entry. So, we can't observe the event for those individuals with insufficient follow-up time.

Right-censored data is a common phenomenon that emerges in various applied fields. In statistics literature, this kind of data is encountered especially in medical studies and generally, datasets are formed by incomplete observations. It can be clearly said that one of the most important problems which distort the data quality is censored observations. As known, in practice, datasets are commonly problematic. Generally, datasets are formed by missing or censored observations because of many different reasons such as a death of patients abruptly, withdraw from the study, equipment malfunctions and so on.

In terms of regression analysis, classical methodology is inapplicable for such an incomplete data. As known, in analyzing survival data, there are three common approaches to overcome the censorship which are Kaplan-Meier weights that proposed by Miller and improved by Stute, synthetic data transformation which is proposed by various authors such as Buckley and James, Koul et al. and Leurgans, imputation methods which have extensive literature but for this paper, only *k*NN imputation method is proposed by Batista and Monard and Troyanskaya et al. [1-8]. studied the imputation of microarray data. In the literature, mentioned three methods are always studied separately. In the literature, mentioned three methods are always studied separately. Qualitative comparisons were made between the weighted log-rank statistics and weighted Kaplan-Meier (WKM) statistics. A statement of null asymptotic distribution theory is given, and the choice of weight function is discussed in detail. Small-sample simulation studies indicates that statistics compare favorably with the log-rank procedure even under the proportional hazards alternative and which can perform better than it under the crossing hazards alternative [9-11].

It is aimed to compare mentioned methods and arising the important differences and properties of them when data is modelled by semi-parametric regression method based on smoothing splines. In addition to that, the interactive web application is presented to provide simplicity for modelling the survival data with semi-parametric models. In statistics, semiparametric regression includes regression models which combines nonparametric and parametric models. They are mostly used in situations where the fully nonparametric model may not perform well or when the researchers wants to use a parametric model but the functional form with respect to a regressors subset or the density errors is unknown.

Smoothing splines

A semi-parametric regression model can be written as follows for the uncensored observations:

$$Y_i = x_i^T \beta + f(t_i) + \varepsilon_i, i = 1, \ldots, n \quad (1)$$

where Y_i's are the values of response variable, $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ and t_i's represent the explanatory variables (x_i, t_i) which are either independent and identically distributed (i.i.d.) fixed design points, $\beta = (\beta_1, \ldots \beta_p)^T$ is a vector of regression coefficients, f (.) is an unknown function from R^1 and $\varepsilon_1 \ldots, \varepsilon_n$ are independent random errors with mean zero and finite variance $\sigma_i^2 = E(\varepsilon_1^2)$.

In this paper, our interest is estimating the model (1) when Y'_i is observed incompletely and right-censored by a random censoring variable C_i. Here, it should be noted that all through this paper, it has been assumed that explanatory variables (x_i, t_i) are completely observed. Consequently, incompletely observed variables (Y_i, X_i, t_i) replace with (Z_i, δ_i, X_i, t_i). Here, binary dataset (Z_i, δ_i) can be expressed as follows:

$$Z_i = \min(Y_i, C_i) \text{ and } \delta_i = I(Y_i \le C_i) \quad (2)$$

Where $Z_{(i)}$ is the updated response variable with respect to censored data with unknown distribution and δ'_i are the values of censoring indicator function which contains censoring information. If i^{th} observation is censored, then $\delta_i = 0$ or $\delta_i = 1$. Thus, model is rewritten as follows:

$$Z_i = x_i^T \beta + f(t_i) + \varepsilon_i, i = 1, \ldots, n \quad (3)$$

To see details about estimation of model (3) Green and Silverman can be inspected.

Censorship solutions

In order to involve the censorship in the estimation process, three common methods are used which are Kaplan-Meier weights, synthetic data transformation and *k*NN imputation method. These methods turn data into an appropriate form for modelling procedure.

Kaplan-Meier weights: To overcome the right-censored response observations Kaplan-Meier (K-M) weights are expressed which is discussed by Stute with details. K-M weights can be calculated based on the Kaplan-Meier estimator $\hat{F}$ of the distribution function F of lifetime's Y'_i s at each value $Z_{(i)}$ given by:

$$w_{(i)} = \hat{F}(Z_{(i)}) - \hat{F}(Z_{(i-1)}) = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(i)}} \quad (4)$$

Where δ_i denotes the value of censoring indicator associated with ordered values $Z_{(i)}$

Synthetic data transformation: Synthetic data transformation is a common method in the literature to handle with censored data and various authors are proposed different methods such as; Leurgans, Buckley and James and Koul et al. [5-7]. In this paper, synthetic response values are obtained based on the method proposed by Koul et al. [6]. In this sense, data transformation is realized by:

$$Z_{iG} = \frac{\delta_i Z_i}{1 - G(Z_i)} = \frac{\delta_i Z_i}{\bar{G}(Z_i)} \quad (5)$$

where G is the distribution function of censoring variable as mentioned before?

kNN imputation method: Imputation is a class of methods that focuses to fill the censored observations with estimated ones. It can be realized with using true relationships in the dataset that can be helpful in estimating censored observations. In this paper, kNN imputation is used. In this context, it can be said that using imputation methods for handling the censored data has some differences from the mentioned two methods. Some differences from other two methods and the important properties of *k*NN imputation can be ordered as follows

1. Method is free from distribution. This feature provides an important advantage when dealing with data that does not fit any distribution family.
2. Right-censored data points are completed with actual observations, not synthetic or constructed values.
3. Different from synthetic data transformation and K-M weights, *k*NN method utilizes the explanatory variables to supply additional information in completing censored data points.
4. One of the most important properties of kNN imputation is fully nonparametric method and it does not include any assumptions about the relationship between observation pairs (Xi, Yi) *or* $(Xi, Zi), i = 1, \ldots, n$.

Method can work with discrete and continuous variables. It uses most frequently used data point among *k*-nearest neighbours. For continuous attributes, it uses mean value of *k*-nearest neighbors.

Discussion and Conclusion

Researchers deal with datasets from different populations have different distributions. Mentioned solution methods for censorship have some advantages of their own according to properties of data. If dataset is suitable for Kaplan-Meier estimator, Kaplan-Meier weights and Synthetic data transformation can be useful. On the other hand, kNN method can work ultimately free from assumptions which is its important advantage. As a result is up to the researcher.

References

1 Miller RG (1976) Least squares regression with censored data. Biometrika 63: 449-464.

2 Stute W (1993) Consistent estimation under random censorship when co-variables are present. J Multivariate Analy 45: 89-103.

3 Stute W (1995) The central limit theorem under random censorship. Annals Stat 2: 422-439.

4 Stute W (1999) Non-linear censored regression. Statistica Sinica 9: 1089-1102.

5 Buckley J, James I (1979) Linear regression with censored data. Biometrika 66: 429-436.

6 Koul H, Susarla V, Ryzin VJ (1981) Regression analysis with randomly right-censored data. Annals Stat 1: 1276-1285.

7 Leurgans S (1987) Linear models, random censoring and synthetic data. Biometrika 74: 301-309.

8 Batista G, Monard M (2003) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17: 519-533.

9 Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformat 17: 520-525.

10 Green PJ, Silverman BW (1994) Non-parametric regression and generalized linear model. Chapman Hall 1: 1-2.

11 Kaplan EL, Meier P (1958) Non-parametric estimation from incomplete observations. Journal of The Am Stat Assoc 53: 457-481.