Linear Regression Model with a Randomly Censored Predictor: Estimation Procedures

Folefac D Atem[1]* and Roland A Matsouaka[2]

[1]Department of Biostatistics, University of Texas Health Science Center, USA

[2]Department of Biostatistics and Bioinformatics & Duke Clinical Research Institute, Duke University, USA

*Corresponding author: Folefac D Atem, Department of Biostatistics, University of Texas Health Science Center, Houston, TX, USA, Tel: 214-648-1080; Email: Folefac.D.Atem@uth.tmc.edu

Abstract

We consider linear regression model estimation where the covariate of interest is randomly censored. Under a non-informative censoring mechanism, one may obtain valid estimates by deleting censored observations. However, this comes at a cost of lost in formation and decreased efficiency, especially under head censoring. Other methods for dealing with censored covariates such a signoring censoring or replacing censored observations with a fixed number, often lead to severely biased results and are of limited practicality. Parametric methods based on maximum likely hood estimation as well as semi parametric and non parametric methods have been successfully used in linear regression estimation with censored covariates where censoring is due to a limit of detection. In this paper, we adapt some of these methods to handle randomly censored covariates and compare them under different scenario store cently developed semiparametric and non parametric methods for randomly censored covariates. Specifically, we consider both dependent and independent randomly censored mechanisms as well as the impact of using an on parametric algorithm on the distribution of the randomly censored covariate. Through extensive simulation studies we compare the performance of these methods under different cenarios. Finally, we illustrate and compare the methods using the Framingham Health Study data to assess the association between low\density lipo protein (LDL) in off spring and parental ageaton set of a clinically-diagnosed cardiovascular event.

Introduction

Modeling continuous outcome data using linear regression usually assumes in theory that the values of the covariates are fully observed. However, in practice and especially for any large dataset, it is unlikely that complete information will be available for all study participants. The issue of censored data is ubiquitous and affects many studies and permeates a wide range of research areas, including medicine, economics, and social sciences. Multiple reasons lead to in complete observations in a data set including on response, attrition, and absence of even to interest. Usually during the design, implementation, and data collection phases of a study, efforts are made to minimize the occurrence of incomplete data whenever possible and, when unavoidable, to understand the reasons for such a discrepancy in order to handle the available data adequately and run appropriate statistical analyses. Although there is an extensive literature on missing data [1-3] and censored outcomes, [4-6] only a small number of paper shave explored scenario in which the covariate is censored [7-10].

Arguably, the in adequacy of linear regression models on censored outcome variables has sparked an interest in alternative methods, and subsequently has led to major developments of regression models for survival analysis for decades. Extensive literature has been published regarding censored outcomes, especially in studies of time- to -even to us comes where censoring is due to loss of follow-up, dropout, or study termination [4-6].

While there is a vast literature on censored outcomes and different related method shave been discussed extensively, only a limited number of papers have focused on the issue of censored covariates. Ignoring or using a wrong approach to account for the censored nature of a covariate in regression model estimation can lead to analytical issues and spurious results [9-11]. It is important that censored covariates be recognized, acknowledged, and handled appropriately to produce reliable results [12-20]. However, the vast majority of literature on censored covariates has focused on censoring due limit of detection or type 1 censored covariates where observations of the covariate below such a limit cannot be measured or detected, but recorded at or less than the limit of detection value [11,12,21-26]. Only a handful of publications have investigated the implications of randomly censored covariates where some observations of the covariate are censored at varying censoring time points [8,20,27-29]. Censored covariate measurements

arise when, for some participants in a study, the as curtained information of interest has not yet occurred (or will not occur) at the time of assessment.

This is due to a time lag between the time when a covariate is measured (usually) and the occurrence (or non-occurrence) of an event of particular interest that needs to happen for such a measure to be available and assessed. For instance, Clayton [8] investigated family an aggregation in chronic disease incidence and modeled the possible influence that parental age at onset of a given disease might have on an individual's risk of succumbing to a particular disease. Using the Framingham Heart Study an ongoing multi generation all and mark study designed to identify factors and characteristics that contribute to the development of cardiovascular disease (CVD) and other diseases through long-term, active surve alliance and monitoring Atem & Mat souaka [6] studied the impact that age at onset of clinically-diagnosed cardiovascular events in parents may have on the onset of cardiovascular events among offspring.

In both cases, even if important factors have been thoroughly measured for both parents and their offspring, it is unlikely that all parents have had or will have developed the disease of interest at the time of investigation. This means that the variable" parental age at onset of a given disease" is guaranteed to be censored, i.e., not fully observed. Therefore, it is extremely important in any statistical analysis to account for the fact that the variable of interest is censored for some participants.

In theory, there are many ways to address the issue of censored covariates in data analyses. From a practical point of view, however, the most important questions are: When and under what conditions can one safely consider the problem of censored covariates to be trivial? How can current methods be applied and under what conditions can one expect (asymptotically) unbiased and meaningful results? In general, in appropriate handling of censored covariates may affect the type I error [11], yield biased results, hinder the power to detect any meaningful treatment differences, or lead to loss of efficiency in estimating the coefficient parameters of are aggression model [10].

Complete case analysis, where by observations with censored covariate values are discarded (on purpose or through a soft ware default option), is the most commonly used method. When the sample size of the data is large, the censoring mechanism is independent of the outcome, and the proportion of censored data is relatively small, complete case analysis of the data can be employed safely [1,2] since the impact t of censored observations on the analysis of the data may be negligible. In that case the complete-case analysis yields valid (consistent and a asymptotically unbiased) estimates for regression coefficient parameters.

However, under moderate to heavy censoring of data there might be a substantial loss of efficiency due to the reduction in the sample size and the significant loss of information on other fully observed covariates and on the outcome measures of the deleted observations [9]. Further-more, when the censoring mechanism is informative, using a complete-case analysis can lead to biased results which are in part exacerbated by the losses of information and efficiency Case analysis can lead to biased results which are in part exacter bated by the losses of information and efficiency increase stricting the analysis so Lely to truly observed covariate measures may introduce some imbalance in the data set in a way that misrepresents the population under study.

When the issue of censored covariate is not ignored all together, simple substitution methods (or to be more accurate adhoc fill- in methods) where censored observations are replaced by the overall mean or median of the observed variable or, alternatively, by a constant are frequently used because they are simple, easy to understand, are easy to implement. Unfortunately, they may lead to substantial biased estimates and in accurate conclusions [13,21,24,25,30-32]. Sever a non-trivial statistical methods have been developed specifically to input censored covariates and estimate regression coefficient parameters in a model with a censored covariate [33,34] Some of these methods, known as parametric methods, use maximum likelihood estimation (MLE) under the assumption that the covariate follows a specific distribution [7,10,16,18,34-38].

For example, when such a distribution as assumption is plausible, Richardson & Ciampi [17] proposed using MLE and input censored observations with E $(X|X<\xi)$ in the context where measurements of the covariate, X, are left-censored due to limit of detection ξ. However, this approach has some limitations, especially when the censored covariate is correlated with other covariates [34,35,37,38,39].

As we know, an MLE method relies on a parametric distribution assumption of the censored covariate, i.e., the postulated distribution is assumed to be true and correctly specified. It is less reliable when the distribution assumption is in corrector when the data set is so small that it becomes question able whether the assumed distribution fits the data well. In that case, a semiparametric model that makes weaker parametric distribution as assumptions or, even better, a nonparametric method that does not assume any specific distribution model at all is preferable [15,34,35].

As we previously mentioned, most of these methods highlighted above have been developed to account for type I censoring or limit of detection and are typically developed for left-censored covariates. Nevertheless, it is fairly straight forward to adapt the methods for a limit of detection data or type I censored covariate to a right-censored covariate. However, to the best of our knowledge, no such parametric approach employed for type I censored covariates has been extended to handle a randomly censored covariates. In addition, barring a few papers on dependent (randomly) censoring mechanism [20],

the vast majority of published methods for type I censoring rely on the assumption that the censoring mechanism is independent of the outcome in interest [12,15,19,32].

Our primary objective in this paper is to adapt a parametric method proposed in linear regression models with a type I censored covariate to the case in which a covariate is randomly censored covariate. We use simulation studies to compare this newly developed method to the methods proposed by Sampene & Folefac [20] in which randomly censored covariate values are replaced by a nonparametric and a semiparametric estimations of E $(X|X{\geq}\tau)$ or E $(X|X{\geq}\tau,Y)$, where τ denotes the maximum observation time for the variable X, and the outcome of interest Y. For this purpose, we will consider both dependent and independent censoring mechanism, which occurred depending on whether such a censoring mechanism depends or not on the outcome of interest. Furthermore, we will also compare the fore mentioned non parametric estimation method to the commonly used deletion or complete-case analysis and give recommendations on the methods of estimation based on our simulation results. We begin in Section [21] by presenting parametric c and non-parametric methods used in the censored covariate literature. We then introduce the methods proposed by Sampene & Folefac [20] to handle randomly censored covariates. In Section [18] we run simulation studies to compare each of the discussed methods as well as the complete- case analysis method. Finally, we apply these methods in Section [16] to the Framingham Offspring Study to assess the influence of parental age at onset of cardiovascular disease on the systolic blood pressure of their offspring.

Notation and Methods

We consider n study participants independently sampled from a referenced population. Let Yi, Xi and Ci be, respectively, the continuous outcome variable, the potentially censored covariate (from which we are interested in making inferences), and the right censoring variable, where i ind exes subjects. Due to the right censoring in the covariate X, for each participant ti, we observe the vector

$[[(Y_i,V_i,\delta_i)]]^T$where V_i min$(X_i,)$,i1,2,...,n.

The linear regression model is given by

$Y=\beta_0+\beta_1 X +\varepsilon$

Where the parameter coefficients β_0 and $[[\beta]]_1$ are the intercept and slope, respectively. The random error$\varepsilon$is assumed to be independent of X and follows a normal distribution with mean 0and variance$[[\sigma]]^2$ i.e. $\varepsilon \sim N(0,\sigma^2)$. We consider two different cases of censoring mechanisms. In the first case, we assume that the censoring mechanism is non- informative, i.e., C is independent of the outcome Y. For the second case, we assume that the censoring mechanism depends on the outcome Y, in

a sense that there is some known point$[[c]]_0 \in R$ such that the random variable C follows a distribution characterized by the distribution function$[[G]]_0$ when $Y<c_0$and by the distribution function $G_1$when $Y{\geq}c_0$. For simplicity and demonstrative purposes, we limit our discussion to cases with no additional covariates. If, in practice, the data a than d contain a set of additional fully- observed (i.e. non-censored) covariates, Z, the method discussed here could easily be extended to accommodate such covariates.

Parametric Method

Maximum likelihood estimation

A parametric method assumes an underlying distribution of the population from which the data at than d were sampled and uses the maximum likelihood estimation method to draw inference. Suppose that the censoring C is independent of Y; the simplies$[[X]]_i$and$\varepsilon_i$are independent. Therefore, the distribution of Y is a product of the distributions of$[[X]]_i$andε_i. The likelihood L of Y is made up of two components; one based on the uncensored observed $([[X]]_1,...,X_m)$and the other on the right censored$[[(X]]_{(m+1)},...,X_n)$:

$L=\prod_{(i=1)}^m [[f(Y_i,X_i)]]^\delta \prod_{(i=m+1)}^n [[f(Y_i,X_i>c_i)]]^{(1-\delta)}$

With $\delta=1$ if$[[X]]_i$ is observed and $\delta=0$ if$[[X]]_i$ is censored, I ,...,n. The maximum likelihood estimate of the unknown regression parameter corresponding to the censored covariate X is derived from the log-likely hood function

$\log(L)=\delta\sum_{(i=1)}^m \log(l_{i1}) +(1-\delta) \sum_{(i=m+1)}^n \log(l_{i2})$

Where $\log(l_{i1})=\log[f(Y_i,X_i)]$ and$[[\log]](l_{i2})=\log[f(Y_i,X_i>c_i)]$. Suppose that X follows a normal distribution with mean $\mu_x^{(2)}$and variance$[[\sigma]]_x^2$, we have

$\log(l_{i1})=1/(2\pi\sigma\sigma_x) \exp(([[-\varepsilon]]_i^2)/(2\sigma^2)-([[x_i-\mu]]_x)^2/(2\sigma^2))$

For the second component, consider $\Phi(x)$ the cumulative Gaussian distribution function and define $\varepsilon_{(i\mu_x)}=y_i-\beta_0-\beta_1 \mu_x$ and $Q=\sqrt{(\sigma_x^{(-2)}+\beta^2 \sigma^{(-2)})}$. we show in the Appendix that

$\blacksquare (@\varepsilon)_i=y_i-\beta_0-\beta_1 x_i$

$f(Y_i,X_i>c_i)=1/\sqrt{(2\pi(\sigma^2+\beta_1^2 \sigma_x^2))} \Phi\{Q(\mu_x-(\beta_1 \varepsilon_{([[i\mu]]_x)})/(2\sigma^2)-c_i)\}\exp\{-\varepsilon_{([[i\mu]]_x)}/(2\sigma^2) (1-(\beta_1^2)/(2\sigma^2 Q^2)) \}$.

When censoring is dependent as described above, Ci and Y are dependent. The likelihood can be expressed as

$L=\prod_{(i=1)}^m [[f(Y_i,X_i)]]^\delta \prod_{(i=m+1)}^k [[G_0 (Y_i,X_i>c_i)]] \prod_{(i=k+1)}^m [[G_1 (Y_i,X_i>c_i)]]$

Where the data is made of fully observed $x\_1,...,x\_m$ and right censored $x\_(m+1),...,x\_n$. The censored component is divided into $x\_(m+1),...,x\_(k\ )$ and $〖x〗\_(k+1),...,x\_n$ components with associated distribution $〖G〗\_0$ and $〖G〗\_(1\ )$ respectively, as in

equation (2).

Overview of the nonparametric methods

As stated in the introduction, most of the methods described in published literature that examine the issue of covariates are subject to the limit of detection. Prior to the late1990's, the most common approach to handling such censored covariates was the complete-case analysis method. Alternatively, several naïve adhoc alternative methods have been proposed, including substitution methods, which consist of replacing censored covariate values with either a function of the limit of detection, ξ, e.g. $ξ, √2\ ξ, 2ξ$ [19] or the mean of the observed covariate measures (mean substitution) [37] as well as dichotomizing the potentially censored covariate into a binary covariate [18,13] In evitable, each of these adhoc methods leads to a biased estimation of β1.

For instance, Helsel investigated the use of the senaive substitution methods and concluded that they are inefficient and have no mathematically plausible backing [14]. The extent of the inefficiency depends on the extent and the severity of censoring (i.e., the distance between the limit of detection or random censoring value and the natural limit for X) of censoring. Finally, Ate metal [8,27] explored additional non-parametric methods, based on multiple imputation approach, but concluded that these methods were not efficient when applied to the cases of dependent censoring. Recently, Sampene and A team proposed two conditional multiple imputation methods for estimation and inference [20]. The under pinning's of these methods involve replacing the randomly censored values X_i by estimates of $E(X\_i |X\_i>τ)$, for $i=m+1,...,n$. In the absence of additional covariates, the former is determined via a Kaplan-Meier estimator and performs well when the correlation between X and Y is weak. When the correlation between X and Y is strong, similar to case of missing covariate the outcome of interest is included in the imputation $E(X\_i│X\_i>τ,Yi)$ [1].

This conditional imputation involving outcome Y, unlike the imputation not involving Y used estimates from the Cox proportional hazard hence the name Cox Multiple Imputation. To estimate the corresponding variance of $〖β〗\_1$ for inference, Sampene and A team suggested using either a conditional multiple imputation or a conditional single imputation along with a boot strap re sampling procedure to correct for the under estimation of the variance in here n t to the single imputation [20]. In doing so, we showed that these improvements to the complete-case analysis method result invalid inferences regardless of whether the censoring mechanism is dependent or independent of the outcome. Furthermore, using simulation studies, they demonstrated that the multiple imputation method

dissimilar to the conditional single imputation with boot strap re sampling.

In then extsection, we will run simulation studies to compare the complete-case analysis, parametric, mean imputation, naïveadhoc, conditional single imputation, conditional multiple imputation and Coxmultiple imputation methods. It is worth mentioning that Coleetal & Nieetal (2009). Have explored the parametric approach for type1 censoring. They showed that this approach is very efficient for limit of detection data. However, it is worth exploring how well this parametric approach works compared to others nonparametric approaches when censoring is random [35].

Monte carlo simulations

Data generation and simulation setup: We assumed that the true linear regression model is given by $Y=β\_0+β\_1 X+ε$, with $(β\_0,β\_1)=(1,0.5)$ and $ε \sim N(0,1)$. The variable X as well as the censoring variable distribution were generated from a two-parameter We I bull distribution

$$f(u)=θ/λ\ (u/λ)^{(θ-1)} \exp[-(u/λ)^θ]$$

Where θ is the shape parameter so known as the we I bull slope, with $θ>0$ and $λ, λ>0$, is the scale parameter. More precisely, we generated K=2000 samples of size n=100 and n=500 respectively, and choose $ε \sim N(0,1)$ in each case. For independent censoring mechanism, we considered the following distribution

1. $X \sim Weibull((3/4,1)/4)$ and $C \sim Weibull(1,q)$ for q=1.50 and0.35

2. $X \sim Weibull(1,1/4)$ and $C \sim Weibull(5/4,q)$ with q=1 and 0.4

3. $X \sim Weibull(2,1/4)$ and $C \sim Weibull(9/4,q)$ with q=0.50 and 0.30.

The selected values of q allowed us to obtain, respectively, 20% and 40% censoring. Under dependent censoring, we defined the corresponding mechanism such that $C=C\_1$ if Y>1.02 and $C=C\_(2\ )$ if Y≤1.02. We also considered the following data generating distributions to obtain 20% and 40% censoring: $X \sim Weibull(2,1/4), C1 \sim Weibull(9/4,q)$ and $C2 \sim Weibull(10/4,q)$ with q=0.50 and 0.30, respectively.

Simulation results

Table 1-4 summarize the results of the four sets of simulations performed for light censoring (20%) and heavy censoring (40%) in terms of the following quantities:

A. $Bias=1/K \sum\_(k=1)^K▒〖β ̂\_1k-β\_1 〗$, which is the (overall) deviation of a parameter estimate from the true parameter $β\_1=0.5$, where $β\_1k$ is the estimate from the k-th generated data set;

B. Empirical standard error, $SE(β ̂\_1)$, of the estimate $〖β〗\_1$ overall K simulation data sets

C. Simulation error, i.e., the average of model-based standard errors;

D. Mean squared error (MSE), which is the expectation of the squared aviation of a parameter estimate from the truth. It is equal to $[Bias]^2 + SE(\hat{\beta}_1)$;

E. Cover age probability which is the proportion of simulated samples for which the 95% confidence interval $\beta\_1k \pm Z_{(1-\alpha/2)} SE(\hat{\beta}\_1k)$ includes the true coefficient $\beta\_1 = 0.5$, for $k = 1,...,K$.

Table1: Case1: X~*Weibull* (3/4,1/4)and censoring sin dependent of Y.

Light Censoring					
N=100	**Bias**	$SE(\tilde{\beta}_1)$	**Simulation Error**	**MSE**	**Coverage Probability**
Actual data (No Censoring)	0.0012	0.2527	0.2536	0.0639	0.97
Complete-case	0.0044	0.4138	0.4163	0.1712	0.955
Mean Substitution	0.0044	0.42	0.4163	0.1712	0.962
Maximum Likelihood	0.051	0.3345	0.3377	0.1145	0.945
Conditional Single Imputation	0.0361	0.2901	0.3138	0.0855	0.969
Conditional Multiple Imputation	0.0014	0.4011	0.4111	0.0842	0.96
Cox Based Multiple Imputation	0.0013	0.4201	0.4211	0.1765	0.966
N=500					
Actual data (No Censoring)	0.0009	0.1103	0.1118	0.0122	0.971
Complete-case	0.0021	0.1795	0.1799	0.0322	0.955
Mean Substitution	0.0021	0.182	0.1799	0.0331	0.957
Maximum Likelihood	0.0361	0.1469	0.153	0.0229	0.945
Conditional Single Imputation	0.0132	0.1297	0.1524	0.017	0.97
Conditional Multiple Imputation	0.0011	0.1811	0.1815	0.0324	0.96
Cox Based Multiple Imputation	0.0012	0.1705	0.1743	0.029	0.966
Heavy Censoring					
N=100					
Actual data (No Censoring)	0.0012	0.2527	0.2536	0.0639	0.97
Complete-case	-0.0048	0.8683	0.9052	0.754	0.943
Mean Substitution	-0.0048	0.8852	0.9052	0.7836	0.946
Maximum Likelihood	0.2007	0.5229	0.5421	0.3137	0.9
Conditional Single Imputation	0.1	0.4542	0.5657	0.2163	0.93
Conditional Multiple Imputation	0.0545	0.8101	0.7887	0.6592	0.97
Cox Based Multiple Imputation	0.0029	0.87	0.8911	0.7569	0.965
N=500					
Actual data (No Censoring)	0.0009	0.1103	0.1118	0.0122	0.971
Complete-case	-0.0125	0.3744	0.3845	0.1403	0.948
Mean Substitution	-0.0126	0.382	0.3845	0.1461	0.952

Maximum Likelihood	0.0949	0.2334	0.2369	0.6348	0.911
Conditional Single Imputation	0.0613	0.1812	0.2188	0.0366	0.95
Conditional Multiple Imputation	0.006	0.3891	0.3691	0.1475	0.961
Cox Based Multiple Imputation	0.0021	0.3786	0.3888	0.1433	0.966

Table 2: Case2: X~Weibull (1,1/4)and censoring is independent of Y.

Light Censoring					
N=100	Bias	$SE(\hat{\beta}_1)$	Simulation Error	MSE	Coverage Probability
Actual data (No Censoring)	0.0086	0.3934	0.3969	0.1548	0.968
Complete-case	0.0232	0.5419	0.5717	0.2942	0.942
Mean Substitution	0.0232	0.5446	0.5717	0.2971	0.942
Maximum Likelihood	0.051	0.4605	0.4833	0.2147	0.939
Conditional Single Imputation	0.0185	0.4399	0.4641	0.1939	0.966
Conditional Multiple Imputation	0.017	0.5818	0.5956	0.3388	0.96
Cox Based Multiple Imputation	0.0131	0.5463	0.5554	0.2986	0.962
N=500					
Actual data(No Censoring)	0.0048	0.1757	0.1731	0.0309	0.973
Complete-case	0.0073	0.2426	0.2315	0.0589	0.958
Mean Substitution	0.0073	0.2436	0.2314	0.0594	0.959
Maximum Likelihood	0.0338	0.2075	0.2047	0.0442	0.944
Conditional Single Imputation	0.0058	0.1927	0.2801	0.0372	0.972
Conditional Multiple Imputation	0.0092	0.2481	0.2566	0.0616	0.961
Cox Based Multiple Imputation	0.006	0.245	0.2456	0.6001	0.962
Heavy Censoring					
N=100					
Actual data (No Censoring)	0.0086	0.3934	0.3969	0.1548	0.968
Complete-case	0.0334	0.8828	0.9124	0.7805	0.941
Mean Substitution	0.0334	0.889	0.9124	0.7914	0.941
Maximum Likelihood	0.0914	0.5967	0.6065	0.3644	0.892
Conditional Single Imputation	0.0663	0.5466	0.6072	0.3032	0.96
Conditional Multiple Imputation	0.0311	0.86	0.834	0.7406	0.961
Cox Based Multiple Imputation	0.02	0.9001	0.9004	0.8106	0.959
N = 500					
Actual data (No Censoring)	0.0048	0.1757	0.1731	0.0309	0.973
Complete-case	0.0108	0.3893	0.3865	0.1517	0.956

Mean Substitution	0.0108	0.3922	0.3864	0.1539	0.956
Maximum Likelihood	0.09	0.2704	0.2767	0.0812	0.935
Conditional Single Imputation	0.0367	0.2319	0.2564	0.0551	0.96
Conditional Multiple Imputation	0.0121	0.4	0.3811	0.1601	0.961
Cox Based Multiple Imputation	0.0061	0.3869	0.3905	0.1497	0.966

Table 1 & 2 show that when the distribution of X is highly skewed (Figure 1), the parametric approach results in large bias and MSE as compared to the conditional multiple imputation approach. Although the complete case is unbiased, deleting observations reduces the sample size, which results in an increased standard error and larger MSE as compared to both the maximum likelihood and the conditional multiple imputation methods. Despite being unbiased, both the complete case and the mean substitution methods are inefficient with higher MSE as compared to the maximum likelihood approach, conditional multiple imputation and the Cox multiple imputation approach. The single conditional imputation is unbiased and is more efficient than the mean imputation with smaller MSE because its underestimates the standard error when imputed values are used as true values with no uncertainty. All approaches resulted in acceptable cover age probabilities.
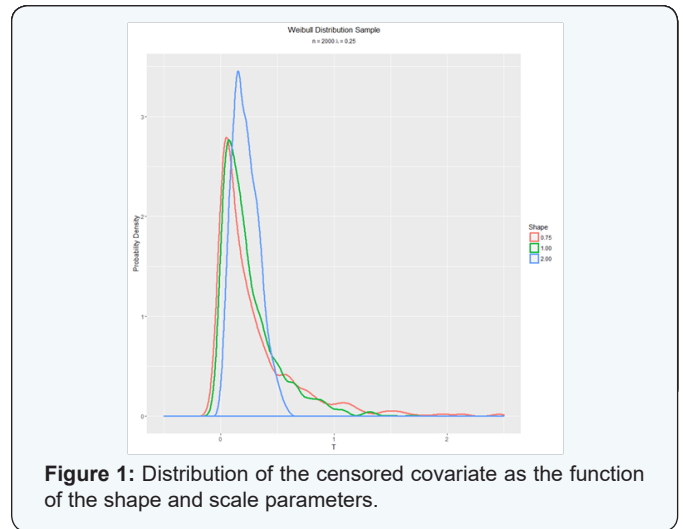

Figure 1: Distribution of the censored covariate as the function of the shape and scale parameters.

Table 3: Case3: X~Weibull (2,1/4)and censoring is independent of Y.

Light Censoring					
N=100	**Bias**		**Simulation Error**	**MSE**	**Coverage Probability**
Actual data (No Censoring)	0.0264	0.8322	0.8258	0.6932	0.958
Complete-case	0.0465	1.037	1.0286	1.0775	0.948
Mean Substitution	0.0465	1.0366	1.0286	1.0767	0.948
Maximum Likelihood	0.0294	0.8894	0.8827	0.7919	0.952
Conditional Single Imputation	0.0361	0.8855	1.4533	0.7854	0.943
Conditional Multiple Imputation	0.0451	1.0349	1.0381	1.0731	0.959
Cox Based Multiple Imputation	0.043	1.0391	1.0396	1.0816	0.959
N=500					
Actual data (No Censoring)	0.0008	0.3795	0.3765	0.144	0.97
Complete-case	0.0064	0.4698	0.4703	0.2208	0.949
Mean Substitution	0.0064	0.4698	0.4703	0.2207	0.948
Maximum Likelihood	0.0085	0.4085	0.4132	0.1604	0.97
Conditional Single Imputation	0.0044	0.4005	0.6782	0.1604	0.97
Conditional Multiple Imputation	0.0019	0.5101	0.5139	0.2602	0.961

Cox Based Multiple Imputation	0.0021	0.4704	0.4777	0.2212	0.969
Heavy Censoring					
N=100					
Actual data(No Censoring)	0.0264	0.8322	0.8258	0.6932	0.958
Complete-case	0.0516	1.4517	1.4531	2.1074	0.943
Mean Substitution	0.0516	1.4555	1.4531	2.1142	0.945
Maximum Likelihood	0.0204	1.0124	1.07	1.0254	0.941
Conditional Single Imputation	0.0371	1.0079	1.3468	1.0172	0.942
Conditional Multiple Imputation	0.0429	1.5321	1.5386	2.3492	0.959
Cox Based Multiple Imputation	0.0916	1.4573	1.4857	2.2157	0.959
N=500					
Actual data(No Censoring)	0.0008	0.3795	0.3765	0.144	0.97
Complete-case	0.0067	0.6537	0.6514	0.4274	0.946
Mean Substitution	0.0067	0.6538	0.6515	0.4275	0.947
Maximum Likelihood	0.0191	0.4639	0.4585	0.2156	0.951
Conditional Single Imputation	0.0153	0.4442	0.7648	0.1975	0.97
Conditional Multiple Imputation	0.006	0.6779	0.6768	0.4631	0.959
Cox Based Multiple Imputation	0.0061	0.6537	0.6617	0.4274	0.973

Table 4: Case4: X~Weibull (2,1/4)and censoring depends on Y.

Light Censoring					
N=100	Bias	$SE(\hat{\beta}_1)$	Simulation Error	MSE	Coverage Probability
Actual data (No Censoring)	-0.0022	0.8335	0.8484	0.6947	0.949
Complete-case	0.0157	1.0285	1.0365	1.0581	0.944
Mean Substitution	0.0157	1.0298	1.0365	1.0607	0.945
Maximum Likelihood	0.0109	0.8881	0.9122	0.7888	0.948
Conditional Single Imputation	0.0045	0.8847	1.1293	0.7827	0.959
Conditional Multiple Imputation	0.003	1.0421	1.0471	1.0859	0.949
Cox Based Multiple Imputation	0.0049	1.4071	1.4771	1.98	0.966
N=500					
Actualdata (No Censoring)	0.0012	0.3793	0.3843	0.1439	0.968
Complete-case	-0.0067	0.4684	0.4685	0.2194	0.947
Mean Substitution	-0.0067	0.4688	0.4685	0.2198	0.948
Maximum Likelihood	-0.0063	0.407	0.4105	0.1657	0.953
Conditional Single Imputation	-0.0039	0.3992	0.7403	0.1594	0.966

Conditional Multiple Imputation	0.0025	0.45	0.4796	0.2025	0.956
Cox Based Multiple Imputation	0.0032	0.5062	0.5111	0.2562	0.967
Heavy Censoring					
N=100					
Actualdata (No Censoring)	-0.0022	0.8335	0.8484	0.6947	0.949
Complete-case	0.0156	1.4648	1.4558	2.412	0.944
Mean Substitution	0.0156	1.4671	1.4558	2.1526	0.945
Maximum Likelihood	0.0328	1.0092	1.009	1.0196	0.944
Conditional Single Imputation	0.0349	1.0006	1.2314	1.0024	0.933
Conditional Multiple Imputation	0.0201	1.4001	1.4091	1.9607	0.966
Cox Based Multiple Imputation	0.045	1.0955	1.0998	1.2116	0.966
N=500					
Actualdata (No Censoring)	0.0012	0.3793	0.3843	0.1439	0.968
Complete-case	0.0102	0.6578	0.6611	0.4328	0.947
Mean Substitution	0.0102	0.6596	0.6611	0.4352	0.948
Maximum Likelihood	0.0164	0.4672	0.463	0.2185	0.952
Conditional Single Imputation	0.0301	0.4436	0.6156	0.1977	0.966
Conditional Multiple Imputation	0.0134	0.6867	0.6867	0.4717	0.966
Cox Based Multiple Imputation	0.0141	0.6771	0.6846	0.4587	0.966

Tables 3 [16] show that, when the distribution of X is close to normal (Figure 1), the maximum likelihood approach results in smaller bias and standard error. The log likely hood [40] is derived under the normal distribution as assumption. Therefore, the distribution of X in (Table 3 & 4), Is close to the true distribution from which the maximum likelihood method is based, provides a better and a more efficient parameter estimate than the multiple imputation methods. The standard error and MSE is smaller than that of both the conditional multiple and Cox multiple imputation approaches. The other imputation methods are less efficient and more biased. Overall, it is worth mentioning that the Cox multiple imputation is more efficient than the conditional imputation when the data is well powered. As the sample size increases, this Cox multiple imputation is very efficient, which might be due to the fact that this approach uses one additional parameter Y in the imputation model as compared to the Kaplan Meier based conditional imputation that does not involve the outcome in the imputation model.

Illustrative example

Association between parentage of cardiac events and low density lipoprotein (LDL) in offspring: According to the American Heart Association cardiovascular disease (CVD) is a multi-faceted disease that affects the heart or blood vessels. CVD includes hypertensive, rheumatic, congenital, and vulvar heart diseases as well as cardio my apathies, heart arrhythmias, carditis, aortic aneurysms, peripheral artery disease, venous thrombosis, coronary death, myocardial infarction, coronary in-sufficiency, angina, is chemic stroke, hemorrhagic stroke, transient is chemic attack ,peripheral artery disease, and heart failure. It is the global leading cause of death, accounting for over 30% of all deaths worldwide-approximately 17.3 million deaths per year. In the United States, someone dies from CVD every 39 seconds, with most of those deaths being attributed to coronary heart disease [40-42]. Though the death rate due to CVD has decreased slowly over the last 30 years, death from heart disease remains the leading cause of death in the United

States, and caring for patients with poor cardiovascular health continues to be one of the largest burdens on the health care system today. From 1990to2009, CVD ranked first in the number of days for which patient received hospital care [43], yet 72% of Americans do not consider themselves at risk for heart disease [40].

Associations have long been established between CVD and a wide variety of risk factors, including non-modifiable variable such as family history [44-50]. Blood levels of low density lipoproteins (LDL), one of the five major groups of lipo proteins categorized by density, are regarded as a strong predictor of CVD. To illustrate the methods proposed in this paper, we study the association between LDL in offspring and age at onset of clinically diagnosed cardiovascular even tin parents, using data from the Framingham Heart Study data base and looking at both the Original and Offspring cohorts [51].

The Framingham Heart Study (FHS) is an ongoing prospective study of the etiology of cardiovascular disease, among other prevalent diseases. The study began in 1948 and enrolled 5, 209 participants (55% women) aged 28 and 62 years old residing in Framingham, Massachusetts a spart of the original cohort who have been followed up to the present. In 1971, the Framingham Offspring Study was established with a sample of 5, 124 men and women aged 5to70 years old who were either (genetic or adoptive) offspring or spouses of offspring of the original cohort [52-54]. Study participants are examined routinely to update their health status information and potential risk factors. Standard clinical examinations included physician interview, physical examination, and laboratory tests, and continue to the present. Participants in the original cohort have been followed biannually; there were 40 participants during the 32nd Exam visit held in 2012- 2014. In the offspring cohort, participants have been followed approximately every four years. The Offspring Examination Cycle 9 covered the years 2011 to 2014 and had 2430 participants.

In this example, we performed two separate analyses, one for each parent, to evaluate the relationship between ages of CVD in parents and log(LDL)in offspring. Data gathered from the original cohort (Exam12 (1971-1974); 3, 261participants) and the offspring cohort (Exam1 (1971-1975); 5, 124 participants) were used. We deleted all missing data and restricted the LDL to physician recorded values; this reduced the sample size ton= 2622 (1,401 mothers and 1,221fathers).Of the 1401 mothers in the final data set, 907 of them (i.e., 35.26%) experienced a cardiovascular event whereas 909 (i.e., 74.45%) out of 1,221 fathers experienced a cardiovascular event. The median age of CVD was 66years and 63years for mothers and fathers, respectively.

Results of the data analyses are provided in (Table 5 & 6). The results for the complete- case analysis, mean substitution, maximum likelihood, conditional single imputation and the

conditional multiple imputation are consistent with the simulation results. With a larger sample, the assumption of normality for the censored covariate is met and the parametric method provides better estimates along with smaller standard error. On the other hands, the results from the adhoc substitution methods are inconsistent with the simulation results. This is because there is no scientific base for such substitutions.

Table 5: Relationship between Maternal age of onset of CVD and LDL in off spring.

Method	Estimate	SE	P-value
Complete-case (64.74%ofthedata)	0.0034	0.00012	0.0002
Mean Substitution	0.0044	0.0012	0.0002
Maximum Likelihood	0.1999	0.0797	0.0123
Conditional Single Imputation	0.0022	0.0009	0.0276
Conditional Multiple Imputation	0.0023	0.001	0.0284
Cox Based Multiple Imputation	0.002	0.0009	0.0286

Table 6: Relationship between Paternal age of onset of CVD and LDL in offspring.

Method	Estimate	SE	P-value
Complete-case(74.45%ofthedata)	0.0024	0.0013	0.0675
Mean Substitution	0.0024	0.0013	0.0675
Maximum Likelihood	0.766	0.148	<.0001
Conditional Single Imputation	0.0018	0.001	0.0742
Conditional Multiple Imputation	0.0018	0.0011	0.0787
Cox Based Multiple Imputation	0.0017	0.001	0.0768

Conclusion

Most of the literature on censored covariates deals with the issue of limits of detection, the point at which observations below this limit cannot be measured or detected and are instead recorded at the limit of detection value. [13,21,25] In this paper, we considered the estimation of linear regression models when the covariate of interest is randomly censored. We evaluated non parametric conditional imputation methods based on the Kaplan-Meier estimate to impute a censored covariate. We compared this nonparametric approach based on Kaplan-Meier to the regression from the full data (without censoring), the complete-case analysis, a naive adhoc substitution (replacing censored values by the mean of the observed covariate values) and the maximum likelihood approach.

Parametric estimators were determined via maximum likelihood estimation method based on an under lying distribution as assumption of the censored covariate. Throughout our simulations, we demonstrate that then alive adhoc substitution method provides biased estimation of the regression parameter of the censored covariate. As Helsel pointed out, these substitution methods area kin to fabricating

data; they don't have any theoretical basis and should thus be discouraged [23].

The complete case analysis method is the widely used approach for handling censored predictor sanities easy to implement. The obvious spit fall of the complete case approach is that it potentially sacrifices information by discarding observations. Although, this method leads to unbiased estimates under independent censoring, it can result in a substantial loss of power, especially under moderate to high percentages of censored observations. Under dependent censoring, complete-case analysis may lead to model misspecification due to selection bias if a group of subjects with similar characteristics do not experience the even to fin terestor leave study before its completion.

The mean substitution approach is easy and looks reliable but a detailed analysis of this approach shows it has many short comings. We cannot always guarantee that the mean of the complete case will be greater than the time at censoring. One basic assumption of censored data is that, if the event is to occur, it can only happen after the censored time. Furthermore, this approach does not make use of the available information, that is, the time at censoring. Using parametric methods requires prior knowledge or postulating a distribution model for the censored covariate. When the postulated parametric distribution of the censored covariate corresponds to the true distribution, the maximum likelihood estimation method and then on parametric method via Kaplan Meier estimation all provides consistent estimates, under independent censoring. Under dependent censoring, if the distribution of X and C are similar, these methods are efficient; however, if the distribution of X and C are dissimilar the MLE approach will be highly inefficient (as shown in section 2.1). Therefore, we propose the use of Kaplan Meier non parametric imputation in absence of prior knowledge of the distribution of censored covariate when such a distribution cannot be accurately as certained. On the other hand, if the sample size is large and the distribution of X is a member of the exponential family, the MLE approach can be suitable.

References

1. http://www.cdc.gov/nchs/data/hus/hus15.pdf

2. https://www.heart.org/idc/groups/heart-public/@wcm/@adt/documents/downloadable/ucm_441512.pdf

3. Albert PS, Harel O, Perkins N, and Browne B (2010). Use of multiple assays subject to detection limits with regression modeling in assessing the relationship between exposure and outcome. Epidemiology (Cambridge,Mass.),21(Suppl4):S35-S43.

4. Arunajadai SG, Rauh VA (2012) Handling covariates subject to limits of detection in regression. Environmental and ecological statistics 19(3):369-391,.

5. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA (2016) Survival analysis with a randomly censored covariate. Technical report.

6. Atem F Qian J Maye J E Johnson KA Betensky RA (2015) Linear regression with a randomly censored covariate: application to an Alzheimer's study. Technical report.

7. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA (2016) Multiple imputation of a randomly censored covariate improves logistic regression analysis. Journal of Applied Statistics Pp1-11.

8. Austin P, Hoch JS (2004) Estimating linear regression models in the presence of a censored independent variable. Statistics in medicine 23(3): 411-429.

9. Bernhardt PW, Wang HJ, Zhang D (2013) Statistical methods for generalized line are models with covariates subject to detection limits. Statistics in Biosciences Pp. 1-22.

10. Bernhardt PW, Wang HJ, Zhang D (2014) Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. Comput Stat Data Anal 69: 81-91.

11. Clayton DG(1978) A model for association in bivariate life tables and its application nine pidemiological studies off a milial tendency in chronic disease incidence. Biometrika 65(1):141-151.

12. Cole SR, Chu H, Nie L, Schisterman EF (2009)Estimating the odds ratio when exposure has a limit of detection. Int J Epidemiol 38(6): 1674-1680.

13. Angeloand GD, Weissfeld L (2008) An index approach forth eco x model with left censored covariates. Statistics in medicine 27(22):4502-4514.

14. Dawber TR, Moore FE, Mann GV (2015) Coronary heart disease in the Framingham study. International journal of epidemiology 44(6): 1767-1780.

15. Donahue RP, Abbott RD, Bloom E, Reed DM, Yano K (1987) Central obesity and coronary heart disease in men. Lancet 329(8537): 821-824.

16. Helsel DR (2005) More than obvious: better methods for inter preting non detect data. Environmental science&technology.39(20):419A-423A.

17. Helsel DR (2006) Fabricating data:how substituting values for non detects can ruin results, and what can be done about it. Chemosphere 65(11): 2434-2439.

18. Helsel DR (2005) Non detects and data analysis. Statistics for censored environmental data. Wiley-Interscience.

19. Hornung RW, Reed LD (1990) Estimation of average concentration in the presence of non detectable values. Applied occupational and environmental hygiene 5(1): 46-51.

20. Howard BV, Robbins DC, Sievers ML, Lee ET, Rhoades D, et al. (2000) Ldl cholesterol as strong predictor of coronary heart disease in diabetic individuals with insulin resistance and low Ldl the strong heart study. Arteriosclerosis, thrombosis, and vascular biology 20(3): 830-835.

21. Kannel WB, Dawber TR, Thomas, Namara Mc (1965) Comparison of serum lipids in the prediction of coronary heart disease. Framingham study indicates that cholesterol level and blood pressure are major factors in coronary heart disease; effect of obesity and cigarette smoking also noted. R I Med J 48: 243-250.

22. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP (1979) An investigation of coronary heart disease in families the Framingham offspring study. Am J Epidemiol 110(3): 281-290.

23. Keil U (2000) Coronary artery disease: the role of lipids, hyper tension and smoking. Basic Res Cardiol 95(1): I52-158.

24. Keysetal A (1970) Coronary heart disease in seven countries. Circulation 41(4): 186-195.

25. Klein JP, Moeschberger ML (2005) Survival analysis: techniques for censored and truncated data. Springer Science&Business Media.

26. Kleinbaum DG, Klein M (1996) A Self-Learning Text. (3ʳᵈ Edn)Survival analysis. Springer.

27. Kong S, Nan B (2016) Semiparametric approach to regression with a covariate subject to a detection limit. Biometrika.

28. Lagakos SW (1979) General right censoring and its impact on the analysis of survival data. Biometrics 35(1): 139-156.

29. Langohr K, Gómez G, Muga R (2004) A parametric survival model with an interval-censore covariate. Statistics in medicine, 23(20): 3159-3175.

30. Lee M, Kong L, Weissfeld L (2012) Multiple imputation for left-censored biomarker data based on Gibbs sampling method. Statistics in medicine 31(17): 1838-1848.

31. Lee S Park S Park J(2003) The proportional hazards regression with a censored covariate. Statistics & probability letters 61(3): 309-319.

32. Lipsitz S, Parzen M, Natarajan S, Ibrahim J (2004) Fitzmaurice G Generalized linear models with acoarsened covariate. Journal of the Royal Statistical Society: SeriesC (Applied Statistics) 53(2): 279-292.

33. Little RJ (1992) Regression with missing X's: are view. Journal of the American Statistical Association, 87(420): 1227-1237.

34. Little RJ, Rubin DB (2014) Statistical analysis with missing data. John Wiley&Sons.

35. Lynn HS (2001) Maximum like lihood inference for left-censored HIVRNA data. Stat Med 20(1): 33-45.

36. Mahmood SS, Levy D, Vasan RS, Wang TJ (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet 383(9921): 999-1008.

37. Manson JE, Colditz GA, Stampfer MJ, Willett WC, Rosner B, et al. (1990) A prospective study of obesity and risk of coronary heart disease in women. N Engl J Med 322(13): 882-889.

38. May RC, Ibrahim JG, Chu H (2011) Maximum likelihood estimationin generalized linear models with multiple covariates subject to detection limits. Stat Med 30(20): 2551-2561.

39. Neaton JD, Wentworth D (1992) Serum cholesterol, blood pressure, cigarette smoking, and death from coronary heart disease over all findings and differences by age for 316099white men. Arch Intern Med 152(1): 56-64.

40. Nie L, Chu H, Liu C, Cole SR, Vexler A, et al. (2010) Linear regression with an independent variable subject to a detection limit. Epidemiology 21(Suppl. 4): S17-S24.

41. Richardson DB, Ciampi A (2003) Effects of exposure measurement error when an exposure variable is constrained by a lower limit. Am J Epidemiol 157(4): 355-363.

42. Rigobon R, Stoker TM (2007) Estimation with censored regressors: Basic issues*. International Economic Review 48(4):1441-1467.

43. Rigobon R, Stoker TM (2009) Bias from censored regressors. Journal of Business& Economic Statistics, 27(3):340-353.

44. Sampene E, Atem FD (2015) Imputing a randomly censored covariate in aline arregression model. Technical report.

45. Sattar A, Sinha SK, Morris N (2012) A parametric survival model when a covariate is subject to left-censoring. J Biom Biostat 3(2).

46. Sattar A, Sinha SK, Wang XF, Li Y (2015) Frailty models for pneumonia to death with a left-censored covariate. Statistics in medicine 34(4):2266-2280.

47. Schisterman EF, Little RJ (2010) Opening the black box of biomarker measurement error. Epidemiology 21(Suppl.4): S1-3.

48. Schisterman EF, Vexler A, Whitcomb BW, Liu A (2006) The limitations due to exposure detection limits for regression models. Am J Epidemiol 163(4): 374-383.

49. Schwalm JD, McKee M, Huffman MD, Yusuf S (2016) Resource effective strategies to prevent and treat cardiovascular disease. Circulation 133(8):742-755.

50. Vexler A, Tao G, Chen X (2002). A tool kit for clinical statisticians to fix problems based on biomarker measurements subject to instrumental limitations: From repeated measurement techniques to a hybrid pooled-unspooled design. Advanced Protocols in Oxidative Stress III, pages439-460.

51. Wang HJ, Feng X (2012) Multiple imputation form-regression with censored covariates. Journal of the American Statistical Association 107(497): 194-204.

52. Wu H, Chen Q, Ware LB, Koyama T (2012) A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: an application to acute lung injury. J Appl Stat 39(8): 1733-1747.

Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats
  (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission
https://juniperpublishers.com/online-submission.php