Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1261–1270. doi:10.1111/biom.12922.

Threshold regression to accommodate a censored covariate

Jing Qian^{1,*}, Sy Han Chiou², Jacqueline E. Maye^{3,4}, Folefac Atem⁵, Keith A. Johnson³, and Rebecca A. Betensky²

¹Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, U.S.A.

²Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

³Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts, U.S.A.

⁴Department of Clinical and Health Psychology, University of Florida, Gainesville, Florida, U.S.A.

⁵Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, Houston, Texas, U.S.A.

SUMMARY:

In several common study designs, regression modeling is complicated by the presence of censored covariates. Examples of such covariates include maternal age of onset of dementia that may be right censored in an Alzheimer's amyloid imaging study of healthy subjects, metabolite measurements that are subject to limit of detection censoring in a case-control study of cardiovascular disease, and progressive biomarkers whose baseline values are of interest, but are measured post-baseline in longitudinal neuropsychological studies of Alzheimer's disease. We propose threshold regression approaches for linear regression models with a covariate that is subject to random censoring. Threshold regression methods allow for immediate testing of the significance of the effect of a censored covariate. In addition, they provide for unbiased estimation of the regression coefficient of the censored covariate. We derive the asymptotic properties of the resulting estimators under mild regularity conditions. Simulations demonstrate that the proposed estimators have good finite-sample performance, and often offer improved efficiency over existing methods. We also derive a principled method for selection of the threshold. We illustrate the approach in application to an Alzheimer's disease study that investigated brain amyloid levels in older individuals, as measured through positron emission tomography scans, as a function of maternal age of dementia onset, with adjustment for other covariates. We have developed an R package, *censCov*, for implementation of our method, available at CRAN.

* qian@umass.edu.

¹⁰ **Supplementary Materials**

Web Appendix A referenced in Sections 3 and 7, Web Appendix B referenced in Section 3, Web Appendix C referenced in Section 6, and Web Appendices D and E referenced in Section 7, are available with this paper at the *Biometrics* website on Wiley Online Library. An R package, *censCov*, for implementation of the method, is available at <https://CRAN.R-project.org/package=censCov>.

Keywords

Alzheimer's disease; Bias correction; Censored predictor; Cox proportional hazards model; Kaplan-Meier estimator; Limit of detection

1. Introduction

In several common study designs, regression modeling is complicated by the presence of censored covariates. For example, maternal history of dementia is a major risk factor for Alzheimer's disease and dementia (Edland et al., 1996). To investigate this, one substudy of Harvard Aging Brain Study (HABS) aimed to assess whether beta-amyloid deposition in aging adults is associated with maternal age at dementia onset. However, since not all mothers had experienced onset of dementia, their ages of onset are right censored by their ages at the offspring interview, or their ages at death if this occurred prior to the interview. Among the 141 participants in the study, 99 had censored maternal age of dementia onset.

While there is a large literature on the treatment of censored outcome variables and of missing covariates, there is a sparse literature on the treatment of censored covariates, which is a structured form of missingness. Use of the observed values of the covariates will likely produce biased estimates of the regression coefficients (Rigobon and Stoker, 2009) and inflated type I error (Austin and Hoch, 2004). A simple approach is based on discarding all subjects who have a censored covariate and fitting the regression model including only those with completely observed covariates; this is termed *complete-case regression* (Little, 1992). In the context of the brain imaging study, this is valid as long as age at censoring is independent of offspring beta-amyloid, given age at maternal dementia and other covariates. Even if this assumption holds, the approach is potentially highly inefficient even with moderate censoring. Most available methods address the simple case of type I censoring, in which the covariate is censored by a fixed limit of detection. Some methods assume parametric distributions for the censored covariate and replace censored observations with the expectation of the covariate given that it lies below the limit of detection (Lynn, 2001; Richardson and Ciampi, 2003; Schisterman et al., 2006), while others apply maximum likelihood estimation that accounts for the censoring (Rigobon and Stoker, 2007; May et al., 2011). All of these methods require correct specification of the distribution of the covariate and the substitution method leads to underestimation of standard errors. Wang and Feng (2012) developed distribution-free approaches using multiple imputation based on M-regression models. Two papers evaluated the considerable bias of the estimated coefficient of a thresholded censored covariate (Austin and Hoch, 2004; Rigobon and Stoker, 2009), but did not derive a bias correction or investigate the properties of associated hypothesis tests. When both the outcome variable and the covariates are subject to right censoring, Dabrowska (1995) proposed an estimate of the conditional survival function of the outcome variable in the nonparametric regression context. Tsimikas et al. (2012) proposed an estimating function approach for a generalized linear model with a censored covariate. The method requires assumption of a parametric distribution for the censored covariate. Atem et al. (2017) developed a multiple imputation method that assumed a Cox proportional hazards model for the distribution of the censored covariate given other covariates in the linear

regression model. This method is computationally intensive and requires modeling for the censored covariate given other covariates.

There is a clear need for statistical methods that accommodate randomly censored covariates, such as maternal age of onset of dementia. In this paper, we propose three methods for estimation and inference for linear regression models with a covariate that is subject to random censoring. The *deletion threshold regression* approach selects a threshold and constructs a derived binary covariate that indicates whether the covariate of interest is known to lie above or below the threshold; we delete observations that are indeterminate due to censoring. The *complete threshold regression* approach constructs a derived binary covariate that indicates whether the *observed* value of the covariate lies above or below the threshold. The *residual threshold regression* approach extends the threshold methods to cases of dependence between the censored covariate and other covariates. We illustrate how the proposed threshold regression methods can be used to directly test the association of the censored covariate with the outcome and how this association can be consistently estimated. We have developed an R package, censCov (Qian and Chiou, 2017) for implementation of our methods.

In Section 2, we introduce notation and the model. In Section 3, we describe the deletion threshold method and derive procedures for hypothesis testing and estimation of the regression coefficient of the potentially censored covariate. We also derive the asymptotic properties of the proposed estimator and selection of the threshold to optimize test power. In Section 4, we describe the complete threshold method and derive the associated inferential procedures. In Section 5, we extend the threshold regression methods to accommodate dependence between the censored covariate and other covariates. In Section 6 we propose a *reverse survival regression* approach that reverses the roles of the outcome and covariate and employs a survival model. This is an alternative approach that provides a test of association, but not an estimate. We perform comprehensive simulation studies in Section 7, and apply the methods to the Alzheimer's imaging study in Section 8. We conclude in Section 9.

2. Notation and Model

We consider the following linear regression model

$$Y = \alpha_0 + \alpha_1 X + \alpha_2^T \mathbf{Z} + \epsilon, \quad (1)$$

where X is the covariate of interest that is subject to right censoring, $\mathbf{Z}$ is a $p \times 1$ vector of additional covariates that are completely observed, and ϵ is the random error term. Let C be the variable that potentially censors X . We assume that the random error, ϵ , has mean 0 and finite variance σ^2 , and that ϵ is independent of $X, C, \mathbf{Z}$, i.e., $\epsilon \perp (X, C, \mathbf{Z})^T$. Let $I(A) = 1$ if the event A occurs and 0 otherwise. The observed data are $\{Y_i, U_i, \mathbf{Z}_i, \delta_i\}$ for $i = 1, 2, \dots, n$, where $U_i = \min(X_i, C_i)$, $\delta_i = I(X_i \leq C_i)$ and n is the sample size. Our primary scientific interest is in the parameter α_1 which captures the association between Y and X : we aim to test $H_0: \alpha_1 = 0$ and consistently estimate α_1 .

3. Deletion threshold regression

Let t^* be a scalar that lies within the support of the distribution of X . In many cases, t^* will be a natural threshold for X . To accommodate the right censoring of X , we define a derived binary covariate, X^* , as

$$X^* = \begin{cases} 1, & \text{if } X > t^*, C > t^*; \\ 0, & \text{if } X \leq t^*, X < C. \end{cases}$$

If $C < X$ and $C \leq t^*$, then we consider the observation to be noninformative with respect to the ordering of X and t^* and we delete it from our analysis. Assuming the linear regression model (1) and that $\epsilon \perp (X, C, \mathbf{Z}^T)^T$, from which it follows that $\partial \beta / \partial C \perp (X^*, C, \mathbf{Z}^T)^T$, it additionally follows that $E(Y|X^* = 0, \mathbf{Z}) = \alpha_0 + \alpha_1 E(X|\mathbf{Z}, X^* = 0) + \alpha_2^T \mathbf{Z}$ and $E(Y|X^* = 1, \mathbf{Z}) = \alpha_0 + \alpha_1 E(X|\mathbf{Z}, X^* = 1) + \alpha_2^T \mathbf{Z}$. These equations justify fitting a *deletion threshold regression* model, which is a linear model for $E(Y|X^*, \mathbf{Z})$:

$$E(Y|X^*, \mathbf{Z}) = \beta_0(t^*, \mathbf{Z}) + \beta_1(t^*, \mathbf{Z})X^* + \beta_2^T \mathbf{Z}. \quad (2)$$

This implies that $\beta_0(t^*, \mathbf{Z}) + \beta_2^T \mathbf{Z} = \alpha_0 + \alpha_1 E(X|\mathbf{Z}, X^* = 0) + \alpha_2^T \mathbf{Z}$ and $\beta_0(t^*, \mathbf{Z}) + \beta_1(t^*, \mathbf{Z}) + \beta_2^T \mathbf{Z} = \alpha_0 + \alpha_1 E(X|\mathbf{Z}, X^* = 1) + \alpha_2^T \mathbf{Z}$, which suggests that $\beta_1(t^*, \mathbf{Z}) = \alpha_1 \{E(X|\mathbf{Z}, X^* = 1) - E(X|\mathbf{Z}, X^* = 0)\}$. Under independence of X and $\mathbf{Z}$ given X^* ,

$$\beta_1(t^*) = \beta_1(t^*, \mathbf{Z}) = \alpha_1 \{E(X|X^* = 1) - E(X|X^* = 0)\} \equiv \alpha_1 \mu(t^*). \quad (3)$$

Since $\mu(t^*) > 0$, it follows that a test of $H'_0: \beta_1(t^*) = 0$ is a valid test of $H_0: \alpha_1 = 0$; this is true even if C and X are dependent as we made no use of the assumption of independent censoring. The choice of t^* impacts the power of the hypothesis test $H'_0: \beta_1(t^*) = 0$ we address this in Section 3.4. We address the violation of independence of X and $\mathbf{Z}$ given X^* in Section 5.

3.1 Consistent estimation of α_1

To estimate α_1 under independence of X and $\mathbf{Z}$ given X^* , we first obtain a consistent estimator of $\beta_1(t^*)$ by fitting model (2) using maximum likelihood or quasi-likelihood estimating equations (McCullagh and Nelder, 1989), depending on whether or not ϵ is normally distributed. We then estimate the bias-correction term, $\mu(t^*)$, in equation (3), which is the difference between two conditional means of X . The first conditional mean, $E(X|X^* = 0)$, is easily estimated using its empirical counterpart, i.e., $E(X|X^* = 0)$. The second conditional mean, $E(X|X^* = 1)$, requires some approximation. Let $S_X(\cdot)$ and $f_X(\cdot)$ denote

the survival function and probability density function of X , respectively, and note that $\int_t^\infty S_X(u)du = \int_t^\infty u f_X(u)du - tS_X(t)$. Therefore,

$$E(X|X^* = 1) = E(X|X > t^*) = \frac{\int_{t^*}^\infty u f_X(u)du}{S_X(t^*)} = \frac{\int_{t^*}^\infty S_X(u)du}{S_X(t^*)} + t^*. \quad (4)$$

Under independent censoring, i.e., $C \perp X$, we use the Kaplan-Meier estimator $\hat{S}_X(x)$ for $S_X(x)$. Then $\int_{t^*}^\infty S_X(u)du$ can be estimated by $\int_{t^*}^\infty \hat{S}_X(u)du$ using the trapezoidal rule. To compute the partial area under the Kaplan-Meier curve $\int_{t^*}^\infty \hat{S}_X(u)du$ more accurately when observations from the tail are sparse, we employ an approximation for the tail (Section 3.2).

Let $X_{(1)} < \dots < X_{(\kappa)}$ be the observed, uncensored, ordered failure times in the sample, and $X_{(0)} = 0$. An estimator for α_1 is thus given by $\hat{\alpha}_1 = \hat{\beta}_1(t^*)/\hat{\mu}(t^*)$ where $\hat{\mu}(t^*)$ is equal to

$$\frac{\sum_{j=1}^k I\{X_{(j)} > t^*\} [X_{(j)} - \{X_{(j-1)} \vee t^*\}] [\hat{S}\{X_{(j-1)} \vee t^*\} + \hat{S}\{X_{(j)} \vee t^*\}]}{2\hat{S}(t^*)} + t^* - \frac{\sum_{i=1}^n \delta_i U_i I(U_i \leq t^*)}{\sum_{i=1}^n \delta_i I(U_i \leq t^*)}.$$

If C and X are dependent, but independent conditional on $\mathbf{Z}$, we use the Breslow estimator from the Cox model for X given $\mathbf{Z}$ in place of the Kaplan-Meier estimator. In the presence of competing risks, we estimate one minus the cumulative incidence function (Fine and Gray, 1999) in place of the Kaplan-Meier estimator. We employ the nonparametric bootstrap (Efron and Tibshirani, 1994) for variance estimation of $\hat{\alpha}_1$. We justify this in Section 3.3.

3.2 Tail estimation of $S_X(x)$

For accurate estimation of the mean $E(X)$ or conditional mean $E(X|X > t^*)$ using the Kaplan-Meier curve $\hat{S}_X(u)$ especially under heavy censoring, we consider different strategies for tail corrections. One strategy is to treat the largest observation of X as an observed failure even if it is censored (Efron, 1967). This will not overcome underestimation of the conditional mean $E(X|X > t^*)$ if the distribution of X has a much heavier right tail than the censoring distribution C , and thus the estimate of α_1 may be still biased upward. A second approach is to parametrically approximate the tail of $S_X(x)$ (Gong and Fang, 2012); this relies on the accuracy of the parametric assumption. A third approach is to increase the largest observed time to the upper limit of the support of X and set it to be an event. This relies on knowledge of the support of X . In our Alzheimer's disease example X is age of onset of dementia, and so a natural upper limit is available. We evaluate this strategy and assess its sensitivity to choice of upper limit in Section 7. In certain scenarios an external estimate of $S_X(u)$ may be available. In the absence of knowledge of knowledge of X , we

consider a hybrid estimator that is a weighted average of a threshold regression estimator that uses an imperfect estimate of $S_X(x)$ and the complete-case regression estimator. This can reduce the bias of the threshold estimator and potentially achieve lower variance (see Web Appendix A).

3.3 Asymptotic properties

In Web Appendix B, we provide a detailed proof of the following theorem stating that the proposed estimator, $\hat{\alpha}_1$, is consistent and asymptotically normal:

THEOREM 1: Suppose that the linear regression model (1) and the threshold regression model (2) hold. Under independence of the random error, ϵ , and $(X, C, \mathbf{Z}^T)^T$, independence of X and $\mathbf{Z}$ given X^* , and the regularity conditions $C_1 - C_6$ listed in the Web Appendix B, $\hat{\alpha}_1$ converges to α_1 almost surely, and $n^{1/2}(\hat{\alpha}_1 - \alpha_1)$ is asymptotically normal with mean 0.

3.4 Threshold selection

Equality (3) establishes that a test of $H'_0: \beta_1(t^*) = 0$ is a valid test for $H_0: \alpha_1 = 0$. The threshold t^* clearly impacts the power of the test of H'_0 , and thus its selection is an important consideration. Under model (2), in conjunction with our assumptions regarding ϵ , the hypothesis test $H'_0: \beta_1(t^*) = 0$ is essentially a two-sample test comparing the means of two normal or asymptotically normal distributions with equal variances, with power function

$$\Phi\left(-z_{1-\alpha/2} + \frac{|\mu_1(t^*) - \mu_2(t^*)|}{\sigma\sqrt{1/n_1(t^*) + 1/n_2(t^*)}}\right),$$

where $\Phi(\cdot)$ and z_α are the cumulative distribution

function and the $100 \times \alpha$ th percentile of the standard normal, respectively, σ^2 is the variance of ϵ , $\mu_1(t^*) = E(Y|X^* = 1)$, $\mu_2(t^*) = E(Y|X^* = 0)$, and $n_1(t^*)$ and $n_2(t^*)$ are the numbers of observations with $X^* = 1$ and $X^* = 0$, respectively. It follows from model (2) and the independence of X and $\mathbf{Z}$ given X^* that $\mu_1(t^*) - \mu_2(t^*) = \beta_1(t^*)$. Since $\beta_1(t^*) = \alpha_1\mu(t^*)$ by (3), and the power is increasing in the objective function $\psi_1(t^*) = \hat{\mu}(t^*)/\sqrt{1/n_1(t^*) + 1/n_2(t^*)}$ should be selected to maximize $\psi_1(t^*)$. Importantly, this does not require a correction for maximal selection since $\mu(t^*)$ is unrelated to the association between X and Y and depends only on the distributions of X and C .

Remark 1: When t^* is selected based on the sample, the extension of Theorem 1 is straightforward. Let t_0^* be the maximizer of the objective function $\psi_1(t^*)$, and $\Psi_1(t^*)$ be the limit function of $n^{1/2}\psi_1(t^*)$. Using empirical processes theory, we can show that t_0^* is a consistent estimator of t_0 , which is the maximizer of $\Psi_1(t^*)$. Thus, the conclusion in Theorem 1 still holds. More details are provided in the Web Appendix B.

4. Complete threshold regression

To avoid deletion of indeterminate observations, we propose an alternative, *complete threshold regression* method that retains all observations. This approach conditions on the

derived binary covariate U^* that indicates whether $U = X \wedge C \leq t^*$ or $U > t^*$. As U is completely observed, there are no indeterminate observations and thus no deletions. As in Section 3, we fit the derived model:

$$E(Y|U^*, \mathbf{Z}) = \gamma_0(t^*, \mathbf{Z}) + \gamma_1(t^*, \mathbf{Z})U^* + \gamma_2^T \mathbf{Z}. \quad (5)$$

It follows that $\gamma_1(t^*, \mathbf{Z}) = \alpha_1 \{E(X|\mathbf{Z}, U^* = 1) - E(X|\mathbf{Z}, U^* = 0)\}$, which under independence of X and $\mathbf{Z}$ given U^* yields

$$\gamma_1(t^*) = \alpha_1 \{E(X|U > t^*) - E(X|U \leq t^*)\} \equiv \alpha_1 \nu(t^*). \quad (6)$$

Since $\nu(t^*) > 0$, it follows that a test of $H'_0: \gamma_1(t^*) = 0$ is a valid test of $H_0: \alpha_1 = 0$. As for the deletion method, the test is valid even if C and X are dependent. The choice of t^* impacts the power of the hypothesis test $H'_0: \beta_1(t^*) = 0$ (see Section 4.2).

In the special case when the primary goal is to estimate α_2 , the association of a set of completely observed covariates $\mathbf{Z}$ with Y , while adjusting for a censored covariate X , the complete threshold regression may result in substantial gains in efficiency relative to deletion threshold regression and complete-case regression. This is due to its retention of all observations and the equality of γ_2 in model (5) with α_2 in model (1), when $X \perp \mathbf{Z}$ given X^* . A downside is that this approach may sacrifice efficiency in the testing and estimation of the association of X due to its potential misclassification of indeterminate observations.

4.1 Consistent estimation of α_1

A consistent estimate of $\gamma_1(t^*)$ is obtained by fitting the threshold regression model (5). Given (4) and $E(X) = E(X|U \leq t^*)\Pr(U \leq t^*) + E(X|U > t^*)\Pr(U > t^*)$,

$$\nu(t^*) = \left[\left[\frac{\int_{t^*}^{\infty} S_X(u) du}{S_X(t^*)} + t^* \right] - \int_0^{\infty} S_X(u) du \right] \{ \Pr(U \leq t^*) \}^{-1}.$$

As in Section 3.1,

$$\hat{\alpha}_1 = \hat{\gamma}_1(t^*)/\hat{\nu}(t^*). \quad (7)$$

where $\hat{\nu}(t^*)$ can be estimated as $\hat{\mu}(t^*)$ as described in Section 3.1, depending on the censoring mechanism. Under independent censoring (i.e., $C \perp X$), and using the Kaplan-Meier estimator $\hat{S}_X(x)$ for $S_X(x)$, we propose:

$$\hat{\nu}(t^*) = \left[\frac{\sum_{j=1}^k I\{X_{(j)} > t^*\} [X_{(j)} - \{X_{(j-1)} \vee t^*\}] [\hat{S}\{X_{(j-1)} \vee t^*\} + \hat{S}\{X_{(j)} \vee t^*\}]}{2\hat{S}(t^*)} + t^* - \frac{\sum_{j=1}^k [X_{(j)} - \{X_{(j-1)} \vee t^*\}] [\hat{S}\{X_{(j-1)} \vee t^*\} + \hat{S}\{X_{(j)} \vee t^*\}]}{2} \right] \frac{n}{\sum_{i=1}^n I(U_i \leq t^*)}.$$

Estimation of $\nu(t^*)$ under conditionally independent censoring follows as in Section 3.1.

4.2 Threshold selection

We have justified that a test of $H'_0: \gamma_1(t^*) = 0$ is a valid test of $H_0: \alpha_1 = 0$ based on equation (6). Since $\gamma_1(t^*)$ depends on t^* , we apply our approach for optimal threshold selection from Section 3.4 and select t^* to maximize the power of $H'_0: \gamma_1(t^*) = 0$. Namely, t^* is selected to maximize the objective function $\psi_2(t^*) = \hat{\nu}(t^*)/\sqrt{1/n'_1(t^*) + 1/n'_2(t^*)}$ where $n'_1(t^*)$ and $n'_2(t^*)$ are number of observations satisfying $U \leq t^*$ and $U > t^*$, respectively.

4.3 Asymptotic properties

Using similar arguments to those used for Theorem 1, we can establish the strong consistency and asymptotic normality of the proposed estimator $\hat{\alpha}_1$ in (7). Likewise, we estimate the variance for $\hat{\alpha}_1$ in (7) using the nonparametric bootstrap; this is justified as in Section 3.3.

5. Residual threshold regression

In Section 3 we assumed that X and Z are independent given X^* . Within the context of the family history of dementia study, this means that maternal age at dementia onset (X) and offspring's education level (Z) are independent within the set of younger ages of onset, and separately within the set of older ages at onset, while over all ages of onset there is an association. Independence of X and Z given X^* can be tested via Cox models for X with covariates Z in the subsets $X^*=0$ and $X^*=1$. If dependence is found, we propose the following two-step *residual threshold regression* to estimate α_1 . First, we obtain $\hat{\alpha}_2$ using complete-case analysis. Second, using the residual $Y - \hat{\alpha}_2^T Z$ as the new response variable and X as the covariate, we fit the proposed deletion threshold or complete threshold regression to obtain $\hat{\alpha}_1$. The optimal threshold for this method is the same as for the deletion (complete) threshold method. As the new response variable involves the estimated $\hat{\alpha}_2$ we cannot use the analytical variance estimate of $\hat{\beta}_1$ from fitting (2). Instead, we use the bootstrap procedure to estimate the variance of $\hat{\beta}_1$ as well as $\hat{\alpha}_1$. We study this in simulation in Section 7.5.

The asymptotic properties of $\hat{\alpha}_1$ based on the residual threshold regression procedure follow as an extension of Theorem 1. It is straightforward to show that $\hat{\alpha}_2$ the consistent complete-

case estimator of α_2 , is a plug-in estimator in a map from the distribution of $\{Y, \mathbf{Z}, U\}$ to α_2 , and this map is Hadamard-differentiable. Adding this step to the proof of Theorem 1, the consistency and asymptotic normality of $\hat{\alpha}_1$ follow.

In scenarios in which the complete-case analysis fails (e.g., Section 7.4), the residual method, which relies on complete-case analysis, will not work. In this case, when $\mathbf{Z}$ is categorical with K categories, say, $\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(K)}$, we fit the model (2) for each subset of the data with fixed level of $\mathbf{Z}$, say $\mathbf{Z}^{(\kappa)}$, to obtain $\hat{\beta}_1(t^*, \mathbf{Z}^{(\kappa)})$. When $\mathbf{Z}$ is continuous, we discretize it into K distinct categories and take the same approach. To test $H_0: \alpha_1 = 0$, we can test $H'_0: \hat{\beta}_1(t^*, \mathbf{Z}^{(\kappa)}) = 0$ for all $\kappa=1, \dots, K$ and estimate it consistently as

$$\hat{\alpha}_1 = \frac{1}{K} \sum_{\kappa=1}^K \frac{\hat{\beta}_1(t^*, \mathbf{Z}^{(\kappa)})}{\hat{E}(X|\mathbf{Z}^{(\kappa)}, X^* = 1) - \hat{E}(X|\mathbf{Z}^{(\kappa)}, X^* = 0)}.$$

6. Reverse survival regression

Reverse survival regression is an alternative approach for testing the association between Y and X , adjusting for $\mathbf{Z}$. This involves reversal of the roles of Y and X . Specifically, we fit the Cox proportional hazards model with outcome X and covariates Y and $\mathbf{Z}$. This approach naturally handles the censoring of X and is implemented using standard statistical software. Critically, as we justify in Proposition 1, it can be used to test $H_0: \alpha_1 = 0$ in model (1) (see Web Appendix C).

PROPOSITION 1: Under the assumptions of model (1) that the random error e has mean 0, variance σ^2 , probability density function $f_e(\cdot)$ and is independent of $(X, C, \mathbf{Z}^T)$, and the assumption that C is independent of X given Y and $\mathbf{Z}$, the test of $H_0: \alpha_1 = 0$ based on the Cox proportional hazards model that reverses the roles of Y and X yields a valid test for $H_0: \alpha_1 = 0$, where α_1 is the effect of covariate X in model (1).

7. Simulation Studies

We conducted extensive simulation studies to investigate the finite sample performance of the proposed threshold regression methods. We also compared them with complete-case regression and reverse survival regression. The true model in our simulation is given by (1) with $\alpha_0 = 0.5$, $\alpha_1 = 0.5$, $\alpha_2 = -0.5$. Sample sizes of $n = 200$ and 500 were considered. We generated X_i from an exponential distribution $\text{Exp}(3)$, Z_i from $\text{Unif}(1, 6)$, ϵ_i from $\text{Normal}(0, 0.75^2)$, and then obtained Y_i from model (1). Both the scenarios in which $X \perp Z$ and $X \not\perp Z$ (Section 7.5) were considered. We generated the censoring time C_i from $\text{Exp}(\lambda)$ with $\lambda = 0.75, 2.0$, or 4.45 , which yielded light, moderate or heavy censoring with censoring rate 20%, 40%, or 60%, respectively. We evaluated the proposed strategy of estimating the tail of $S_X(x)$ to obtain more accurate estimation of the mean survival, and assessed its sensitivity to different values of τ_x , the upper limit of the support of X . We chose $\tau_x = 1.75$ in our simulation, which corresponds to the 99.5th percentile of X , and also considered $\tau_x = 1.5$ and $\tau_x = 2.0$. We additionally evaluated the strategy of treating the largest observation of X

as an observed failure even if it is censored. We implemented these methods using the R package censCov that we developed (Qian and Chiou, 2017).

7.1 Type I error of hypothesis tests based on $\beta_1(t^*)$ or $\gamma_1(t^*)$

One important usage of the threshold regression methods is a hypothesis test of $H'_0: \beta_1(t^*) = 0$ or $H'_0: \gamma_1(t^*) = 0$ which we have shown is a valid test for $H_0: \alpha_1 = 0$. By setting $\alpha_1 = 0$ in simulations, we evaluated the type I error of the hypothesis test $H'_0: \beta_1(t^*) = 0$ or $H'_0: \gamma_1(t^*) = 0$ at the thresholds optimizing the test powers for the two threshold regression methods. The optimal threshold for each simulated dataset was obtained by maximizing objective function $\psi_1(t^*)$ or $\psi_2(t^*)$ (Sections 3.4 and 4.2). We also calculated the type I error of the hypothesis test $H_0: \alpha_1 = 0$ based on complete-case regression and reverse survival regression. As seen in Table 1 ($n=200$) and Web Table 3 ($n=500$), the two threshold regression methods generally yield accurate estimates of type I error, whether we increased the largest observed time to different τ_x 's or set it to be an event. The complete-case regression and the reverse survival regression also give fairly accurate estimates of type I error. When the sample size is 200 and censoring rate is 60%, we observed some inflation of the type I error for complete-case regression (5.74%).

7.2 Power analysis of the hypothesis test based on $\beta_1(t^*)$ or $\gamma_1(t^*)$

Figure 1 summarizes the simulation results for the power of hypothesis tests based on $\beta_1(t^*)$ or $\gamma_1(t^*)$. For a sample size of 500 and under light, moderate or heavy censoring, we chose $\tau_x=1.75$ and a sequence of threshold values between 0.01 and 0.80 with a grid size 0.01. We then calculated the average power and the average value of the proposed objective function $\psi_1(t^*)$ or $\psi_2(t^*)$ at each threshold value for the deletion threshold method or the complete threshold method, based on 1000 replications. Plots (a), (b) and (c) in Figure 1 correspond to light, moderate and heavy censoring, respectively. The grey solid line and grey dashed line depict the power of the tests under the deletion threshold method and the complete threshold method, respectively. The black solid line and black dashed line represent the proposed objective functions under the deletion threshold and complete threshold methods, respectively. The dotted horizontal line is the power of test based on complete-case regression.

Figure 1 demonstrates empirically that the proposed objective functions $\hat{\psi}_1(t^*)$ or $\hat{\psi}_2(t^*)$ are approximately proportional to the power curves for the threshold methods; this is expected based on our theoretical justification in Sections 3.4 and 4.2. Thus, the objective functions $\hat{\psi}_1(t^*)$ and $\hat{\psi}_2(t^*)$ can be used to identify the optimal threshold leading to the highest power. It is seen in Figure 1 that the deletion threshold method results in higher power than then complete threshold method at every threshold value. Also, under moderate to heavy censoring, the optimal power from both threshold methods, and especially from the deletion threshold method, is much higher than the power based on complete-case regression.

Table 1 ($n=200$) and Web Table 3 ($n=500$) list the power estimates for the hypothesis tests based on $\beta_1(t^*)$ or $\gamma_1(t^*)$ at the optimal threshold, and compares them to their counterparts based on complete-case and reverse survival regression. For the two threshold methods, the

power is insensitive to whether we increase the largest observed time to different τ_x 's or set it to be an event. The power comparisons among methods recapitulate those of Figure 1. It is notable that the reverse survival regression yields higher power than the two threshold methods, and thus is potentially useful for testing association, but is limited in that it does not provide an estimator of α_1 .

7.3 Estimation of α_1

In our simulations, we have evaluated the bias and standard error for $\hat{\alpha}_1$ as well as the bias for $\hat{\beta}_1(t^*)$ or $\hat{\gamma}_1(t^*)$ which are biased estimators of α_1 . Table 2 and Web Table 4 summarize the simulation results (1000 repetitions) for threshold regression methods under light, moderate and heavy censoring. By increasing the largest observed time to τ_x and setting it to be an event, we found that the estimate $\hat{\alpha}_1$ is more accurate in term of both bias and variance, compared to treating the largest observation of X as an observed failure. The advantage is more obvious as the censoring rate increases. The estimation results are robust to the choice of different τ_x 's. The coverage probability of the 95% confidence interval for $\hat{\alpha}_1$ based on threshold regression is reasonably accurate.

The estimates $\hat{\beta}_1(t^*)$ and $\hat{\gamma}_1(t^*)$ are obviously biased, while the estimate $\hat{\alpha}_1$ is virtually unbiased. The bootstrap resampling-based standard error estimate of $\hat{\alpha}_1$ displays close agreement with the empirical standard error. The proposed estimate $\hat{\alpha}_1$ based on deletion threshold method is as efficient as that based on complete-case regression under light censoring, and is more efficient under moderate or heavy censoring. The estimate $\hat{\alpha}_1$ based on complete threshold method is more efficient than that based on complete-case regression under heavy censoring. For the standard error of $\hat{\alpha}_2$ both threshold regression methods are more efficient than the complete-case regression, especially under moderate or heavy censoring, since complete-case regression excludes observations with censored X . Comparing the two threshold methods, the deletion threshold regression yields a more efficient estimate of α_1 , while complete threshold regression results in a more efficient estimate of α_2 . This suggests that keeping observations with $C \leq t^*$, $X > C$ in threshold regression may introduce additional noise in the estimation of α_1 , but may be helpful for the estimation of α_2 since all the observations of Z are used in the regression. When α_1 is of primary interest, we recommend use of the deletion threshold regression method.

In Web Appendix A, we report the simulation results of the proposed hybrid estimators (Section 3.2). This approach is useful in the absence of external information about X .

7.4 Simulation under failure of complete-case regression

We also considered a simulation set-up in which complete-case regression fails, but threshold regression methods are still able to provide valid tests. The true model is the same as model (1) above, with the same regression coefficients. The sample size is $n = 100$. The covariate X_i is Bernoulli, and takes values 2 and 4 with probability 0.6 and 0.4, respectively. The censoring time C_i follows a continuous uniform distribution $Unif(1.0, 3.9)$. The

resulting censoring rate is 60%. The distributions of Z_j and ϵ_j are the same as before. As $X_j = 4$ is always censored, the complete-case regression fails since there is no variation in X_j among the complete cases. However, the deletion threshold regression and the complete threshold regression are still able to yield valid tests for $H_0: \alpha_1 = 0$, with type 1 errors of 0.052 and 0.045, and powers of 0.910 and 0.597, respectively.

7.5 Simulation when X and Z are dependent

We next considered a case in which X and Z are dependent through a normal copula with correlation coefficient $\rho = 0.3, 0.5, 0.7$ or 0.9 . We fit the residual threshold regression as proposed in Section 5, and compared it with the deletion approach, which is appropriate for $\rho = 0$. We did not compare with the complete threshold approach because it is less efficient than the deletion approach. We considered sample sizes $n=200$ and 500 , and increased the largest observed time to τ_x or set it to be an event. First, we evaluated the type 1 error of a test of $H_0: \beta_1(t^*) = 0$. As seen in Web Table 5, the type 1 error of both the deletion and residual threshold regressions are close to the nominal level 0.05.

Table 3 summarizes the estimation results for $\hat{\alpha}_1$ when $n = 200$ and $\tau_x = 1.75$. When $\rho = 0$, the estimation results based on deletion threshold and residual threshold are reassuringly close. Under weak to moderate correlation between X and Z ($\rho = 0.3, 0.5$), the residual threshold and deletion threshold regression exhibit similar performance. However, with strong correlation between X and Z and heavy censoring, the bias of $\hat{\alpha}_1$ from deletion threshold regression is substantial. In contrast, the residual threshold regression exhibits small bias and accurate coverage probability, and is more efficient than the complete-case regression under moderate to heavy censoring. Web Tables 6, 7 and 8 provide additional simulation results under different n and τ_x , from which the same conclusions as in Table 3 can be made.

8. Applications to Alzheimer's disease study

Family history of dementia is a major risk factor for Alzheimers disease and dementia (Jarvik et al., 2005), particularly when the family history is maternal (Edland et al., 1996) and when the age of dementia onset is younger (Silverman et al., 2005). One substudy of the HABS aimed to assess whether beta-amyloid deposition in older adults is associated with maternal age at dementia onset (Maye et al., 2016). All participants were evaluated with interviews, cognitive testing and informant interviews, and judged to be either cognitively normal (clinical dementia rating (CDR) 0) or mildly impaired (CDR 0.5). Among the 141 participants, 99 (70%) have censored maternal age of dementia onset.

We fit a multiple regression model for beta-amyloid, with covariates including maternal age of dementia onset (X) and additional potential predictors (Z): CDR (0 vs 0.5), gender of subject, years of education and subject's age (>81 or not). We tested whether $X \perp Z|X^*$ via a Cox model for X with covariates Z in the subsets $X^*=0$ and $X^*=1$ and found that it is violated for gender and age. Using our proposed threshold selection procedures, we identified age of 83 years as the optimal threshold for all methods. For comparison, we also report the results of the complete-case and the p -value from reverse survival regression.

As seen in Table 4, all methods yielded a negative association between maternal age of dementia onset and beta-amyloid in offspring, which suggests that increased beta-amyloid deposition is associated with younger maternal age at dementia onset. The p -values are comparable across methods, with the exception of the complete threshold (larger) and reverse survival methods (much smaller). There is a larger standard error for the complete case analysis than for the threshold methods. These findings are all consistent with those from our simulations. The slight differences between the deletion and residual threshold methods may reflect a small influence of conditional dependence of age of onset and gender and age. The same negative association was also found via the multiple imputation method in Atem et al. (2017) ($\hat{\alpha}_1 = -0.0049$, p -value = 0.01 as reported in Table 5 therein).

9. Discussion

We have proposed three threshold regression methods to handle a covariate that is subject to random censoring in a linear regression model. The proposed approaches have the advantage of not requiring strong modeling assumptions. They not only allow for efficient hypothesis testing of the censored covariate effect, but also for consistent estimation its regression coefficient. We have additionally derived a straightforward procedure for identification of the optimal threshold. Compared to the complete-case regression, the threshold regression methods result in more efficient estimation of the regression coefficient of the censored covariate and yield more powerful hypothesis tests, especially under heavy censoring. Also, if the censored covariate truly follows a threshold model, then threshold regression methods are expected to offer considerable improvement over complete-case analysis. Deletion threshold regression tends to yield more efficient estimation of the coefficient of the censored covariate than complete threshold regression. However, complete threshold regression may lead to more efficient estimates of regression coefficients of other completely observed covariates. When the censored covariate is correlated with additional covariates, the residual threshold regression method is preferred. The reverse survival regression method appears to provide the most powerful test, but cannot be used for estimation of the regression coefficient.

One extension of interest is to the more complex scenario of multiple censored covariates. We consider model (1) with multiple censored covariates $\mathbf{X} = (X_1, \dots, X_q)^T$, where X_j ($j = 1, \dots, q$) is subject to right censoring by C_j and the C_j 's are mutually independent. Given thresholds $(r^*, \dots, r^*)$, and the associated derived binary covariates $\mathbf{X}^* = (X_1^*, \dots, X_q^*)^T$, we define $\mathbf{X}_{-j}^* = (X_1^*, \dots, X_{j-1}^*, X_{j+1}^*, \dots, X_q^*)^T$. If $X_j \perp \mathbf{X}_{-j}^* | \mathbf{X}_j^*$, then our results for testing and consistent estimation for a single censored covariate extend directly to a_{1j} , $j = 1, \dots, q$. Otherwise, an iterative application of the residual threshold method is promising. The complete regression method may be better suited to the case of multiple censored covariates than the deletion and complete-case methods, both of which may require excessive deletion. A test based on reverse survival regression would require fitting a complicated frailty model. Another extension of interest is to the setting of generalized linear models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the editor, an associate editor, and referees for their helpful comments, which led to an improved article. This work was supported by US National Institutes of Health grants (R21AG053695, R01NS094610, P50AG005134, P01AG036694, T32NS048005) and the Harvard NeuroDiscovery Center.

References

- Atem FD, Qian J, Maye JE, Johnson KA, and Betensky RA (2017). Linear regression with a randomly censored covariate: Application to an Alzheimer's study. *Journal of the Royal Statistical Society: Series C* 66, 313–328.
- Austin PC and Hoch JJ (2004). Estimating linear regression models in the presence of a censored independent variable. *Statistical Methodology* 23, 411–429.
- Dabrowska DM (1995). Nonparametric regression with censored covariates. *Journal of Multivariate Analysis* 54, 253–283.
- Edland SD, Silverman JM, Peskind ER, Tsuang D, Wijsman E, and Morris JC (1996). Increased risk of dementia in mothers of Alzheimer's disease cases: evidence for maternal inheritance. *Neurology* 47 (1), 254–256. [PubMed: 8710088]
- Efron B (1967). The two sample problem with censored data In Cam LML and Neyman J, editors, *Proceedings of the fifth Berkeley symposium mathematical statistics probability*, Volume 4, pages 831–853. University of California Press, Berkeley.
- Efron B and Tibshirani RJ (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York.
- Fine JP and Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94, 496–509.
- Gong Q and Fang L (2012). Asymptotic properties of mean survival estimate based on the Kaplan-Meier curve with an extrapolated tail. *Pharmaceutical Statistics* 11, 135–140. [PubMed: 22323424]
- Jarvik LF, Rue AL, Gokhman I, Harrison T, Holt L, Steh B, et al. (2005). Middle-aged children of Alzheimer parents, a pilot study: stable neurocognitive performance at 20-year follow-up. *Journal of Geriatric Psychiatry and Neurology* 18 (4), 187–191. [PubMed: 16306237]
- Little RA (1992). With missing X's: A review. *Journal of American Statistical Association* 87, 1227–1237.
- Lynn HS (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine* 20, 33–45. [PubMed: 11135346]
- May RC, Ibrahim JG, and Chu H (2011). Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine* 30, 2551–2561. [PubMed: 21710558]
- Maye JE, Betensky RA, Gidyczin CM, Locascio J, Becker JA, Pepin L, et al. (2016). Maternal dementia age at onset in relation to amyloid burden in non-demented elderly offspring. *Neurobiology of Aging* 40, 61–67. [PubMed: 26973104]
- McCullagh P and Nelder J (1989). *Generalized Linear Models*. Chapman and Hall, London, second edition.
- Qian J and Chiou S (2017). *censCov: Linear regression with a randomly censored covariate*. R package version 1.0–0.
- Richardson DB and Ciampi A (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology* 157 (4), 355–363. [PubMed: 12578806]
- Rigobon R and Stoker T (2009). Bias from censored regressors. *Journal of Business and Economic Statistics* 27, 340–353.

- Rigobon R and Stoker TM (2007). Estimation with censored regressors: Basic issues. *International Economic Review*, 48, 1441–1467.
- Schisterman EF, Vexler A, Whitcomb BW, and Liu A (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology* 163 (4), 374–383. [PubMed: 16394206]
- Silverman JM, Ciresi G, Smith CJ, Marin DB, and Schnaider-Beeri M (2005). Variability of familial risk of Alzheimer disease across the late life span. *Archives of General Psychiatry* 62 (5), 565–573. [PubMed: 15867110]
- Tsimikas JV, Bantis LE, and Georgiou SD (2012). Inference in generalized linear regression models with a censored covariate. *Computational Statistics & Data Analysis* 56, 1854–1868.
- Wang H and Feng X (2012). Multiple imputation for M-regression with censored covariates. *Journal of American Statistical Association* 107, 194–204.

Figure 1.

Power of the hypothesis test ($H_0: \beta_1(t^*) = 0$ or $H_0: \gamma_1(t^*) = 0$) with various thresholds on the censored covariate. The grey solid line and grey dashed line represent empirical power of tests under deletion threshold regression and complete threshold regression, respectively. The black solid line and black dashed line represent the proposed objective functions $\psi_1(t^*)$ and $\psi_2(t^*)$ to select optimal thresholds under deletion threshold regression and complete threshold regression, respectively. The dotted horizontal line is the power of test based on complete-case linear regression.

Figure 2.

Identifying the optimal threshold of maternal age of dementia onset in the family history of dementia study. Plot (a): using the proposed objective functions $\psi_1(t^*)$ for deletion threshold regression. plot (b): using the proposed objective functions $\psi_2(t^*)$ for complete threshold regression. The dashed curves are raw values of the proposed objective functions, while the solid ones are smoothed curves using local fitting (e.g., `loess()` function in R).

Table 1

Type I error and power comparison among two threshold regression methods at the optimal thresholds, complete-case regression and reverse survival regression (n=200).

	light censoring (20%)				moderate censoring (40%)			heavy censoring (60%)		
	τ_x	t^*	$\beta_1(\gamma_1)$	α_1	t^*	$\beta_1(\gamma_1)$	α_1	t^*	$\beta_1(\gamma_1)$	α_1
Type I error (%)										
CC				5.14			5.36			5.74
DT	1.75	0.419	5.30	5.26	0.307	5.26	4.78	0.214	4.90	3.98
DT	1.50	0.414	5.26	5.12	0.299	5.50	5.08	0.207	4.86	3.74
DT	2.00	0.425	5.10	5.30	0.315	5.64	5.06	0.233	5.20	4.26
DT	obs	0.412	5.04	4.80	0.289	5.32	4.48	0.185	5.62	3.98
CT	1.75	0.438	4.92	5.18	0.340	5.28	5.12	0.279	4.88	4.72
CT	1.50	0.434	4.98	5.06	0.328	5.26	5.12	0.260	4.48	4.46
CT	2.00	0.445	5.18	5.16	0.348	5.10	5.10	0.298	4.60	4.50
CT	obs	0.430	5.06	4.94	0.312	5.28	4.58	0.208	5.26	3.72
RS				4.68			5.60			5.30
Power (%)										
CC				60.4			28.0			13.0
DT	1.75	0.417	58.3	58.7	0.303	48.7	47.8	0.215	35.9	32.6
DT	1.50	0.412	58.8	58.5	0.298	48.8	48.8	0.207	35.7	32.6
DT	2.00	0.423	59.7	59.7	0.311	48.6	48.0	0.225	35.1	31.4
DT	obs	0.411	58.4	58.2	0.287	48.3	46.6	0.184	34.7	31.2
CT	1.75	0.438	52.1	52.3	0.334	34.1	34.5	0.280	17.1	16.1
CT	1.50	0.432	52.0	51.3	0.322	34.5	34.5	0.263	18.1	17.0
CT	2.00	0.444	52.6	52.3	0.345	34.3	34.2	0.297	16.8	16.8
CT	obs	0.430	52.6	51.3	0.307	34.8	32.7	0.207	17.4	14.3
RS				78.6			64.3			46.3

NOTE: τ_x , the guess of the upper support of X ; t^* , average of the optimal threshold values over simulation replications; $\beta_1(\gamma_1)$, hypothesis test based on β_1 or γ_1 ; α_1 , hypothesis test based on β_1 or γ_1 ; CC, complete-case regression; DT, deletion threshold regression method; CT, complete threshold regression method; RS, reverse survival regression; obs, treating the largest observation of X as an observed failure.

Table 2

Bias, standard error and coverage probability under threshold regression and complete-case regression (n = 200)

	τ_x	t^*	bias $\hat{\beta}_1(\hat{\gamma}_1)$	bias $\hat{\alpha}_1$	SD $\hat{\alpha}_1$	SE $\hat{\alpha}_1$	CP $\hat{\alpha}_1$	SD $\hat{\alpha}_2$	% del	% $\leq t^*$	% $> t^*$
light censoring rate of 20%											
CC				0.003	0.227	0.228	95.2	0.042	20.0		
DT	1.75	0.417	-0.201	0.012	0.236	0.234	94.5	0.041	15.1	62.5	22.4
DT	1.50	0.412	-0.203	0.015	0.235	0.236	94.8	0.041	15.0	62.2	22.7
DT	2.00	0.423	-0.198	0.007	0.230	0.231	94.3	0.041	15.2	62.9	22.7
DT	obs	0.411	-0.202	0.020	0.238	0.239	95.1	0.041	15.0	62.1	22.9
CT	1.75	0.438	-0.224	0.010	0.253	0.254	94.9	0.038	0.0	78.9	21.1
CT	1.50	0.432	-0.226	0.013	0.254	0.257	94.6	0.038	0.0	78.4	21.6
CT	2.00	0.444	-0.223	0.001	0.254	0.251	94.5	0.038	0.0	79.3	20.7
CT	obs	0.430	-0.227	0.017	0.259	0.260	94.8	0.038	0.0	78.2	21.8
moderate censoring rate of 40%											
CC				-0.009	0.340	0.355	95.1	0.048	40.0		
DT	1.75	0.303	-0.233	0.014	0.272	0.273	95.1	0.045	30.0	46.8	23.2
DT	1.50	0.298	-0.236	0.028	0.284	0.281	94.9	0.045	29.8	46.4	23.8
DT	2.00	0.311	-0.229	0.000	0.264	0.266	95.6	0.045	30.3	47.2	22.4
DT	obs	0.287	-0.240	0.053	0.292	0.299	95.1	0.045	29.3	45.6	25.1
CT	1.75	0.334	-0.284	0.004	0.342	0.335	94.5	0.038	0.0	79.0	21.0
CT	1.50	0.322	-0.287	0.026	0.348	0.347	95.5	0.038	0.0	77.9	22.1
CT	2.00	0.345	-0.278	-0.009	0.321	0.324	95.0	0.038	0.0	79.7	20.3
CT	obs	0.307	-0.293	0.050	0.363	0.372	95.5	0.038	0.0	76.4	23.6
heavy censoring rate of 60%											
CC				-0.005	0.668	0.650	95.1	0.061	59.7		
DT	1.75	0.215	-0.258	-0.017	0.311	0.320	96.5	0.052	45.9	32.1	21.9
DT	1.50	0.207	-0.262	0.012	0.324	0.338	96.5	0.052	45.3	31.7	23.0
DT	2.00	0.225	-0.256	-0.054	0.291	0.302	95.6	0.053	46.7	32.6	20.7
DT	obs	0.184	-0.272	0.131	0.395	0.427	95.9	0.051	43.1	30.3	26.6
CT	1.75	0.280	-0.331	-0.038	0.462	0.465	94.2	0.037	0.0	83.6	16.4
CT	1.50	0.263	-0.324	0.016	0.467	0.494	96.0	0.038	0.0	81.9	18.1
CT	2.00	0.297	-0.323	-0.091	0.435	0.429	93.5	0.037	0.0	84.9	15.1
CT	obs	0.207	-0.359	0.133	0.594	0.648	96.8	0.038	0.0	76.1	23.9

NOTE: See the footnote of Table 1 for τ_x , t^* , CC, DT, CT, obs.

SD, standard deviation; SE, mean standard error; CP, coverage (%) of 95% Wald confidence interval; % del, proportion of observations deleted; % $\leq t^*$, proportion of observations below the threshold.

Table 3

Summary statistics for $\hat{\alpha}_1$ under threshold regression and complete-case regression when X and Z are dependent ($n = 200$, $\tau_x = 1:75$, $\alpha_1 = 0:5$).

ρ	light censoring						moderate censoring						heavy censoring					
		bias	SD	SE	CP	PW	bias	SD	SE	CP	PW	bias	SD	SE	CP	PW		
0	CC	0.003	0.228	0.228	95.0	60.0	0.009	0.350	0.353	94.1	30.7	-0.010	0.682	0.650	94.3	12.5		
	DT	0.021	0.228	0.235	95.7	60.6	0.020	0.264	0.275	95.6	49.2	-0.023	0.323	0.318	94.8	36.1		
	RT	0.021	0.228	0.234	95.8	60.5	0.020	0.264	0.274	95.6	49.9	-0.022	0.323	0.318	94.9	36.0		
0.3	CC	0.006	0.231	0.235	95.4	59.1	-0.010	0.368	0.364	95.6	27.6	-0.004	0.732	0.663	92.5	15.0		
	DT	0.007	0.245	0.241	95.5	56.8	0.029	0.283	0.282	95.5	44.4	-0.039	0.338	0.328	95.1	31.8		
	RT	0.021	0.245	0.241	95.6	58.5	0.018	0.284	0.284	95.5	46.2	-0.023	0.343	0.333	95.2	32.5		
0.5	CC	0.007	0.245	0.251	96.1	54.0	-0.010	0.387	0.387	96.1	23.7	-0.021	0.699	0.695	95.2	12.0		
	DT	-0.022	0.249	0.252	95.7	48.0	-0.030	0.298	0.298	94.6	34.7	-0.085	0.355	0.345	93.5	24.2		
	RT	0.019	0.251	0.255	96.0	53.8	0.016	0.302	0.305	95.4	39.3	-0.037	0.368	0.361	94.0	28.0		
0.7	CC	0.013	0.281	0.285	95.6	45.1	-0.008	0.439	0.436	95.7	20.1	0.023	0.812	0.791	94.6	11.0		
	DT	-0.077	0.273	0.276	94.9	34.0	-0.091	0.339	0.332	93.8	23.1	-0.143	0.392	0.388	92.8	17.2		
	RT	0.019	0.281	0.286	95.1	45.4	0.017	0.351	0.351	95.7	32.1	-0.014	0.431	0.434	95.0	22.1		
0.9	CC	0.020	0.373	0.380	95.7	28.1	-0.006	0.576	0.598	95.5	12.4	-0.031	1.157	1.090	94.1	7.5		
	DT	-0.195	0.328	0.326	90.9	16.6	-0.248	0.415	0.410	89.6	9.0	-0.333	0.530	0.496	86.3	8.0		
	RT	0.022	0.361	0.374	95.0	27.8	0.016	0.477	0.497	95.7	16.4	-0.027	0.713	0.679	94.7	11.4		

NOTE: See the footnote of Table 1 for CC, DT. See the footnote of Table 2 for bias, SD, SE, CP.

PW, power (%) of the hypothesis test based on $\hat{\beta}_1$ RT, residual threshold regression estimation procedure when X and Z are dependent as described in Section 5.

Table 4

Linear regression analysis for family history of dementia study

	τ_x	$\hat{\alpha}_1$	(SE)	<i>p</i> -value	95% CI	% del
<i>the variable of interest: Maternal age of demential onset (in years)</i>						
CC		−0.0098	(0.0050)	0.057	(−0.0199, 0.0003)	70.21
DT	105	−0.0048	(0.0026)	0.058	(−0.0099, 0.0002)	39.72
DT	100(obs)	−0.0047	(0.0026)	0.067	(−0.0097, 0.0003)	39.72
RT	105	−0.0050	(0.0029)	0.090	(−0.0108, 0.0008)	39.72
RT	100(obs)	−0.0054	(0.0033)	0.097	(−0.0118, 0.0010)	39.72
CT	105	−0.0038	(0.0036)	0.299	(−0.0108, 0.0033)	0
CT	100(obs)	−0.0036	(0.0036)	0.311	(−0.0107, 0.0034)	0
RS				0.001		0

NOTE: See the footnote of Table 1 for τ_x , CC, DT, CT, RS, obs. See the footnote of Table 3 for RT.

SE, standard error based on nonparametric bootstrap variance estimate with 1000 size; 95% CI, 95% Wald confidence interval; % del, proportion of observations deleted.

Estimate and SE not reported for reverse survival regression due to non-interpretability of the estimate.

Adjusted covariates in the model include CDR (0 vs 0.5), gender of subject, years of education and subject's age (>81 or not).