University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2011 *Paper* 95

Subsample ignorable likelihood for accelerated failure time models with missing predictors

Nanhua Zhang* Roderick J. Little†

*University ofSouth Florida, nzhang1@health.usf.edu
†University of Michigan, rlittle@umich.edu

Subsample ignorable likelihood for accelerated failure time models with missing predictors

By NANHUA ZHANG [1,*] AND RODERICK J. LITTLE [2,]**

[1]Department of Epidemiology & Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612-3085, U.S.A.

[2]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.

[*] email: nzhang1@health.usf.edu

[**] email: rlittle@umich.edu

Abstract: Missing values in predictors are a common problem in survival analysis. In this paper, we review complete-case analysis and maximum likelihood estimation for accelerated failure time models with missing predictors, and apply a new method called subsample ignorable likelihood (Little and Zhang 2011) to this class of models. The approach applies a likelihood-based method to a subsample of observations that are complete on a subset of the covariates, chosen based on assumptions about the missing data mechanism. We give conditions on the missing data mechanism under which the subsample ignorable likelihood method is consistent, while both complete-case analysis and ignorable maximum likelihood are inconsistent. We illustrate the properties of the proposed method by simulation and apply the method to a real dataset.

1. Introduction

The accelerated failure time model (AFT model; Kalbfleisch and Prentice, 2002) is a common form of regression analysis when the outcome is a (possibly censored) survival time, such as the time to develop a disease or death. The model is specified by

$$\log(T_i) = x_i^T \beta + \sigma \varepsilon_i, \ \ \varepsilon_i \overset{iid}{\sim} S_0(.), \ i = 1, 2, ..., n,$$

in which T_i's are the actual failure times, x_i's are vectors of covariates, β is the vector of regression coefficients, σ is the scale parameter, and $S_0(.)$ is a known baseline survival distribution. We obtain the log-normal accelerated time model if S_0 is the standard normal distribution, the log-logistic AFT model if S_0 is the logistic distribution, and the Weibull AFT model if S_0 is the extreme-value distribution (See Table 1). The actual failure time, T_i, is not observed if the study terminates before the failure happens. Let δ_i denote the censoring indicator, equal to 1 if failure is observed, and 0 if failure is censored. Given censored survival data $\{(x_i, t_i, \delta_i), i = 1, ..n\}$, where t_i is the random observed time (failure or censoring), the likelihood function can be written as

$$L(\beta, \sigma \mid \cdot) = \prod_{i=1}^{n} L(\beta, \sigma \mid x_i, t_i, \delta_i)$$

where

$$L(\beta, \sigma \mid x_i, t_i, \delta_i) = \left(\frac{1}{\sigma t_i} f_0(u_i) \right)^{\delta_i} S_0(u_i)^{1-\delta_i}, u_i = \frac{\log(t_i) - x_i^T \beta}{\sigma}.$$

Analysis of disease registry and mortality data are often complicated by incomplete covariate data, because a variable is not measured or the subject does not respond to certain questions. We consider the accelerated failure time model with missing covariates. Common current approaches are:

(a) Complete-case (CC) analysis, which excludes subjects with missing covariate data;

(b) Ignorable likelihood (IL) methods, which base the inference on the observed likelihood for a model that does not include a distribution for the missing data mechanism of the missing covariates. The censoring mechanism for the outcome is strictly speaking non-ignorable but known, and is incorporated by including the censoring indicator in the likelihood – see for example Little and Rubin, 2002, Chapter 15. Examples of IL methods include ignorable maximum likelihood (IML; Meng and Schenker, 1999; Cho and Schenker, 1999; Lipsitz and Ibrahim, 1996a; Lipsitz and Ibrahim, 1996b), Bayesian inferences (Chen, Ibrahim, and Lipsitz, 2002; Bedrick, Christensen, and Johnson, 2000), and multiple imputation (Giorgi et al., 2008; White and Royston, 2009),

(c) Non-ignorable modeling methods, which jointly model the variables and the missing data mechanism for the covariates (Hemming and Hutton, 2010; Herring, Ibrahim, and Lipsitz, 2002). This approach is less common in practice, because it is difficult to specify the model for the missing data mechanism correctly, and problems with identifying the parameters (Little and Rubin (2002), Ch. 15).

Ignorable likelihood methods have the advantage of retaining all data, but they assume that the missing data are missing at random (MAR), in the sense that the missingness of the covariates does not depend on the missing values, after conditioning on the observed data (Rubin, 1976; Little and Rubin, 2002). Complete-case analysis involves a loss of information but has the advantage of yielding valid inference when the missingness depends only on the covariates, but not on the failure time. Little and Zhang (2011) provide a formal justification based on partial likelihood ideas.

In this article, we apply the subsample ignorable likelihood method proposed in Little and Zhang (2011) to the accelerated failure time model (SILAFT). The method mitigates the information loss of CC analysis while retaining the property of allowing missingness of some covariate to depend on their underlying values, a nonignorable mechanism where IL methods are subject to bias. The key idea is to partition the covariates into three sets – one set (say Z) fully observed, one set (say W) for which the missingness is assumed to depend on covariates (including W) but not on the failure time, and one set (say X) for which the missingness are assumed MAR in the subsample of cases with W fully observed. The proposed SILAFT methods apply an IL method to the subsample of case with W fully observed. Particular forms of SILAFT methods include ignorable maximum likelihood, Bayesian inference, and multiple imputation. Conditions formalized in section 4 indicate that SILAFT gives valid estimates in some circumstances where both CC and IL methods are biased.

Section 2 presents a motivating problem based on data from the National Longitudinal Mortality Study (NLMS), where the AFT model is applied to study the relationship between mortality and education and income, adjusting for race, gender, and

marital status. In this application, gender and age are fully observed, but the other variables have missing values; it was thought that missingness of education, race, and marital status was at random, but missingness of household income was likely to depend on income. In this example, Z consists of age and gender, W consists of income and X consists of education, race and marital status. The SILAFT methods apply an IL method to the subsample of cases with income observed.

Section 3 presents the proposed SILAFT method and describes conditions on the missing data mechanism under which it gives consistent estimates, but both IL and CC analyses are biased. We illustrate the properties of the SILAFT methods and alternatives in Section 4, using simulation studies. In Section 5, we apply the method to the motivating data from the NLMS (Sorlie et al., 1995). We conclude with some discussion in Section 6.

2. Motivating problem: social inequalities in mortality

Social inequalities, as measures in variables such as education and income, have been shown to be related to mortality (Antonovsky, 1967; Black et al., 1982; Hann et al., 1987; Sorlie et al., 1995). However, social inequalities are usually viewed as causally irrelevant "confounding variables" rather than risk factors of mortality (Rothman 1986). Accordingly, Link and Phelan (1995) proposed that socioeconomic status is a "fundamental cause" of disparity in mortality. In this paper, we use a dataset from the National Longitudinal Mortality Study (NLMS) (Sorlie et al., 1995) to study the relationship between income and education-related social inequality and mortality. We

use 579,566 subjects who were at least 25 years old at baseline survey from the total of 988,346 participants from the NLMS study and the following variables are extracted:

(a) Time to event outcome: The length of follow-up period (in years) and the death indicator (0=Alive, 1=Dead);

(b) Two socioeconomic status measures at baseline: Adjusted household income and education;

(c) Other covariates: Age at baseline, gender, race and marital status.

The AFT model is used to study the effect of income and education on time to death. Some of the variables have missing values – see Table 2 for the number of missing values for each variable. CC analysis suffers from a loss of all observations that contain missing values. IL methods capture the partial information from the incomplete cases that is lost by CC analysis but assume that the missing values are MAR. It is reasonable to assume MAR for the missingness of education, race, and marital status, but the missingness of household income is thought to depend on the underlying value income – often individuals with high or low values of income are less likely to respond to income than others (David et al., 1986, Lillard et al., 1986, Yan et al., 2010). If these assumptions are correct, the IL methods yield biased estimates of the AFT model. This motivates SILAFT, which allows assumptions of missingness at random for some variables (Education, Race, and Marital status) and assumptions of missingness not at random for others (Adjusted household income), in a sense defined precisely in Section 4.

Table 2 About Here

3. Subsample Ignorable Likelihood AFT models

We consider the missing data pattern in Table 3, which includes a set of completely observed covariates *Z* and two sets of covariates with missing values, namely *W* and *X*.

Table 3 About Here

The columns R_{w_i} and R_{x_i} represents vectors of response indicators for w_i and x_i, the values of *W* and *X* for unit *i*, with entries 1 if a variable is observed and 0 if a variable is missing. To describe missing data patterns for a set of variables (say *v*), it is convenient to write $u_v = (1,...,1)$ to denote a vector of 1's of the same length as the vector *v*, and $\overline{u}_v$ to denote a vector of 0's and 1's of the same length as *v* for which at least one entry is zero. In Table 3, $R_{w_i} = u_w, R_x = \overline{u}_x$ for the complete cases in Pattern 1, $R_{w_i} = u_w, R_x = \overline{u}_x$ for the cases in Pattern 2, where *W* is fully observed and *X* has at least one missing value, and $R_{w_i} = \overline{u}_w$ for the cases in Pattern 3, where *W* has at least one missing value. The pattern of missing values will typically vary for cases within these three sets, but we do not need to distinguish them for the present discussion. Interest concerns the parameters ϕ of the distribution of (t,δ) given (*Z*, *W*, *X*), say $p((t_i,\delta_i) \mid z_i, w_i, x_i, \phi)$. We propose SILAFT, which discards data in Pattern 3 and applies an IL method to the subsample of cases in Patterns 1 and 2 with both *Z* and *W* observed. The division of covariates into *W* and *X* for SILAFT is determined by assumptions about the missing data mechanism. Specifically, the method is valid under the following two assumptions:

(a) <u>Covariate missingness of *W*</u>: the probability that *W* is fully observed depends only on the covariates and not (t,δ), that is:

$$p\left(R_{w_i} = u_w \mid z_i, w_i, x_i, (t_i, \delta_i), \psi_w)\right) = p\left(R_{w_i} = u_w \mid z_i, w_i, x_i, \psi_w)\right) \text{ for all } (t_i, \delta_i) \qquad (1)$$

(b) <u>Subsample MAR of *X*</u>: Missingness of *X* is MAR within the subsample of cases for which *W* is fully observed, that is:

$$\begin{aligned} &p(R_{x_i} \mid z_i, w_i, x_i, (t_i, \delta_i), R_{w_i} = u_w) = \\ &\qquad p(R_{x_i} \mid z_i, w_i, (t_i, \delta_i), x_{\text{obs},i}, R_{w_i} = u_w) \quad \text{for all } x_{\text{mis},i}, \end{aligned} \qquad (2)$$

The validity of SILAFT under (1) and (2) follows from similar arguments to those in Little and Zhang (2011). We first consider the conditional likelihood for a set of parameters ζ based on the joint distribution of $X, (t, \delta), R_X$ given *W* and *Z* and $R_{w_i} = u_w$, that is, restricted to cases *i* with *W* fully observed:

$$L_{\text{cc,w}}(\zeta) = \prod_{i=1}^{m+r} p\left((t_i, \delta_i), x_{\text{obs},i}, R_{x_i} \mid w_i, z_i, R_{w_i} = u_w; \zeta\right),$$

where $\zeta = (\theta, \psi)$. By a direct application of Rubin's (1976) theory, under the subsample MAR condition (6), this likelihood factorizes as

$$L_{\text{cc,w}}(\zeta) = \prod_{i=1}^{m+r} p\left((t_i, \delta_i), x_{\text{obs},i} \mid w_i, z_i, (t_i, \delta_i), R_{w_i} = u_w; \theta\right) \times \prod_{i=1}^{m+r} p\left(R_{x_i} \mid w_i, x_{\text{obs},i}, (t_i, \delta_i), z_i, R_{w_i} = u_w; \psi\right),$$

where the second component on the right side does not involve θ, and the first component on the right side, namely

$$L_{\text{ign,w}}(\theta) = \prod_{i=1}^{m+r} p\left(x_{\text{obs},i}, y_{\text{obs},i} \mid w_i, z_i, R_{w_i} = u_w; \theta\right),$$

is the likelihood for the subsample with w_i observed, ignoring the distribution of the missing data indicators R_{x_i}. Thus inference about θ, the parameter of the distribution (*X*, (t, δ)) given (*W*, *Z*), based on $L_{\text{ign,w}}(\theta)$ is valid. Now factorize

$$p\left(x_i,(t_i,\delta_i)\mid w_i,z_i,R_{w_i}=u_w;\theta\right)=$$
$$p\left((t_i,\delta_i)\mid x_i,w_i,z_i,R_{w_i}=u_w;\theta\right)\times p\left(x_i\mid w_i,z_i,R_{w_i}=u_w;\theta\right).$$

By assumption (1), $p\left((t_i,\delta_i)\mid x_i,w_i,z_i,R_{w_i}=u_w;\theta\right)=p\left((t_i,\delta_i)\mid x_i,w_i,z_i,\phi\right)$, where $\phi=\phi(\theta)$ is the parameter of the regression of interest, and the conditioning on the cases with W observed is removed. Thus, under assumptions (1) and (2), we can base inferences about θ on $L_{\text{ign,w}}(\theta)$, and then derive likelihood inferences about $\phi=\phi(\theta)$ as in Section 3.

The missing data mechanism defined by conditions (1) and (2) is suitable in empirical studies where it is natural to assume covariate-dependent missingness for some covariates and subsample MAR missingness for others. For example, in the motivating example concerning the time to mortality on socioeconomic variables in Section 2.2, Income may be covariate-dependent and the Education and Race may be subsample MAR. Generally, SILAFT methods are based on a partial likelihood (Cox 1972) with the component $L_{\text{ign,w}}(\theta)$ discarded from the analysis, and hence involve a loss of efficiency relative to full likelihood methods. However, they are more efficient than CC analysis, and avoid the need to specify the form of the missing data mechanism beyond assumptions (1) and (2).

Assumptions (1) and (2) differ from the assumptions under which IL and CC methods are valid. Specifically, IL inference assumes the data are MAR, that is:

$$p(R_{w_i},R_{x_i}\mid z_i,w_i,x_i,(t_i,\delta_i),\psi)=p(R_{w_i},R_{x_i}\mid z_i,w_{\text{obs},i},x_{\text{obs},i},(t_i,\delta_i),\psi)\quad\text{for all } w_{\text{mis},i},x_{\text{mis},i}. \tag{3}$$

where missingness of both w_i and (x_i, y_i) can depend on missing components of w_i. CC analysis yields valid inferences if the probability that an observation is complete does not depend on the outcomes, that is:

$$p\left(R_{w_i} = u_w, R_{x_i} = u_x \mid z_i, w_i, x_i, (t_i, \delta_i), \psi)\right) = \\ p\left(R_{w_i} = u_w, R_{x_i} = u_x \mid z_i, w_i, x_i, \psi)\right) \text{ for all } (t_i, \delta_i). \tag{4}$$

This differs from the assumption (2) in that missingness of x_i in (2) can depend on (t_i, δ_i). If this is not the case, then CC yields valid inferences but is less efficient than SILAFT, since SILAFT uses the data in Pattern 2, which are discarded by CC.

4. Simulation Study

As a numerical illustration of this theory, we simulate data for the pattern of Table 3, under a variety of missing data mechanisms. For each of 1000 replications, 1000 observations $(z_i, w_i, x_i, (t_i, \delta_i))$, i = 1,…,1000 on Z, W, X and (t, δ) were generated as follows:

$$z_i \sim N(0,1), w_i \sim Bernoulli(0.5), x_i \sim N(0,1), i = 1,...,1000,$$

and

$$\left(y_i \mid z_i, w_i, x_i\right) \sim_{\text{ind}} LN(1 + z_i + w_i + x_i, 1),$$

where $y_i = \log\left(T_i\right)$ and LN denotes log-normal distribution. T_i is censored at 30, which produces roughly 15% of censoring.

Missing values of W and X were then generated from the following two logistic models:

$$\text{logit}\left(P(R_{w_i}=0\mid z_i,w_i,x_i,(t_i,\delta_i))\right)=\alpha_0^{(w)}+\alpha_z^{(w)}z_i+\alpha_w^{(w)}w_i+\alpha_x^{(w)}x_i+\alpha_t^{(w)}t_i$$
$$\text{logit}\left(P(R_{x_i}=0\mid R_{w_i}=1,z_i,w_i,x_i,(t_i,\delta_i))\right)=\alpha_0^{(x)}+\alpha_z^{(x)}z_i+\alpha_w^{(x)}w_i+\alpha_x^{(x)}x_i+\alpha_t^{(x)}t_i$$

with x_i fully observed when w_i is missing.

For the missing data generation schemes above, CC analysis is valid if both $\alpha_t^{(w)}$ and $\alpha_t^{(x)}$ are zero; IL is valid if $\alpha_w^{(w)}$, $\alpha_x^{(w)}$ and $\alpha_x^{(x)}$ are zero; SILAFT is valid if $\alpha_t^{(w)}$ and $\alpha_x^{(x)}$ are zero. Four missing data mechanisms were created using different sets of values for the regression coefficients such that, in mechanism (I) all three methods (CC, IL and SILAFT) are consistent, while in mechanisms (II), (III) and (IV), just one of the three methods is valid. The simulation setup is summarized in Table 4.

Table 4 About Here

These missing data mechanisms all generate approximately 30% and 20% values missing in *W* and *X*, respectively.

Four specific versions of the methods are applied to estimate the regression coefficients:

(1) CC: Complete-case analysis, using;

(2) IML: ignorable ML for the whole dataset;

(3) SILAFT: IML for the subsample with *W* observed;

(4) BD: least squares estimates from the regression before deletion (BD), as a benchmark method.

For each method, Table 5 summarizes the root mean squared errors (RMSEs) of estimates of all the regression coefficients, and Tables 6 reports respectively the empirical bias, RMSE and coverage probability of estimates of the individual regression

coefficients. Results in bold type reflect situations where the method is consistent based on the theory of Section 4, and hence should do well. The results are based on 1000 repetitions in each simulation.

Tables 5 and 6 About Here

In general, the simulation results are in line with theoretical expectations. All methods are valid in mechanism I. In mechanism II, CC is valid but IL and SILAFT are inconsistent; IL is consistent in mechanism III but CC and SILAFT are biased. In mechanism IV, SILAFT is consistent but CC and IL are inconsistent, and in this case SILAFT has small empirical bias and generally performs best, except for some individual coefficients where the gain in efficiency of IL compensates for the bias of that method. We now describe results in a bit more detail.

For mechanism I, all three methods yield consistent estimates, IL is best since it makes full use of the data, CC is the worst since it discards the most information, and SILAFT lies between CC and IL, since it retains some incomplete cases and drops others.

For mechanism II, CC is valid and in general has the lowest RMSEs, while both IL and SILAFT are biased. However, IL yield comparable or even smaller RMSEs than CC for β_z and β_w, reflecting gains in efficiency that compensate for bias in these parameter estimates.

For mechanism III, IL is the only valid method among the three, and is clearly the best method. Both CC and SILAFT lead to biased estimates, as shown in Table 5, with SILAFT being better than CC since it is incorporates features of IL as a method.

In mechanism IV, SILAFT is valid while CC and IL are biased. The RMSEs from SILAFT are generally the smallest, except that IL yields a smaller RMSE than SILAFT for β_w and β_x.

In some of these situations, supporters of IL may note that it competes well with other methods, despite its theoretical inconsistency and the quite sizeable sample size. This suggests a degree of robustness for IL, which has the virtue of retaining all the data.

5. Application to motivating example

We now apply the proposed method to the data from the NLMS study that were presented in Section 2. We fit log-linear models of the follow-up period (in years) on the adjusted household income (in 1000 dollars per year) and education, adjusting for race, gender, marital status, and baseline age (in years). Adjusted household income data are categorical in NLMS, and we use the median of the corresponding category as a proxy to the true adjusted household income. Education is dichotomized to be greater than high school and high school or less.

Age and gender are fully observed, whereas adjusted household income, education, race, and marital status are subject to missing data, with the percentage shown in Table 1. We assume covariate missingness for adjusted household income, given evidence that people with high or low income are more likely to fail to report it, and we assume subsample missingness at random for other covariates.

With those plausible assumptions, SILAFT on the subsample with adjusted household income observed yields consistent estimates of the regression, whereas IL on the whole sample may be biased. CC analysis is also valid since there is little evidence to

believe that missingness of covariates depends on the follow-up period; however, SILAFT is preferred over CC analysis since it uses more information in the incomplete cases than does CC analysis.

Table 7 About Here

The results of CC analysis, IL and SILAFT are shown in Table 7. All three methods yield similar estimates because the missing proportions of the variables are small. The IL method gives smaller standard errors than CC because it uses more sample than CC. SILAFT is a hybrid of CC and IL, yielding standard errors of SILAFT that lie between CC and IL. There is positive effect of adjusted household income and education, with survival time increasing as adjusted household income and education increases. Race and gender are significant, with white and female having significantly longer survival time than black and male, respectively. Marriage seems to have a protective effect, with married people more likely to live longer.

6. Discussion

We propose subsample ignorable likelihood for accelerated failure time model (SILAFT), which applies an analysis that assumes MAR to a subsample of the data that is complete on a subset of covariates. The methods work for a class of missing data mechanisms, defined in eq. (1) and (2), where both IL and CC fail to give consistent estimates. It is easy to implement, since existing software for ignorable likelihood methods is all that is required. This extends the class of models for data MNAR that can be handled by a selective use of MAR data methods and allows combinations of MAR and MNAR data mechanisms for difference variables in the data set.

The general rationale of SILAFT is partial likelihood (Cox, 1972). This involves a loss of efficiency relative to full modeling, but it is much simpler, since the latter requires specifying a precise form of the missing data mechanism via a model for the missing data indicator, which is vulnerable to model misspecification. An important topic is how much efficiency is lost by SILAFT relative to full likelihood methods. SILAFT involves minimal loss when the fraction of cases in the subsample with the MNAR subset W observed is relatively high, and hence the method is most beneficial relative to CC analysis when the fraction of information in the pattern with W complete but other variables incomplete is relatively high. We present the subsample ignorable likelihood idea in the accelerated failure time model setting, but the general idea of subsample ignorable likelihood can be applied to other models of failure time, such as the Cox proportional hazard regression model.

The validity of the SILAFT methods rests on the assumptions (1) and (2), concerning which variables are considered covariate-dependent MNAR and which are considered subsample MAR. The choice requires an understanding about the missing data mechanism in the particular context. It is aided by learning more about the missing data mechanism, e.g. by recording reasons why particular values are missing. In cases where a choice cannot be made, an alternative strategy is simply to see whether key results are robust of alternative methods. Thus, one might apply CC analysis, IL and SILAFT for the subsample judiciously chosen on the basis of assumptions (1) and (2), to assess sensitivity of key inferences to alternative assumptions about the missing data mechanism.

Acknowledgements.

This paper uses data supplied by the National Heart, Lung, and Blood Institute, NIH, DHHS from the National Longitudinal Mortality Study. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the National Heart, Lung, and Blood Institute, the Bureau of the Census, or the National Center for Health Statistics.

References

Antonovsky A (1967). Social Class, Life Expectancy, and Overall Mortality. *Milbank Memorial Fund Quarterly* 45: 31-73.

Bedrick EJ, Christensen R, Johnson WO (2000). Bayesian accelerated failure time analysis with application to veterinary epidemiology. Statistics in Medicine 19: 221-237.

Black D, Morris JN, Smith C, Townsend P (1982). *Inequalities in Health: The Black Report*. Middlesex, England: Penguin.

Cho M, Schenker N (1999). Fitting the log-F accelerated failure time model with incomplete covariate data. Biometrics 55: 826-833.

David M , Little RJA, Samuhel, ME, Triest RK (1986) Alternative Methods for CPS Income Imputation. *J. Am. Statist. Assoc.* 86:29-41.

Giogi R, Belot A, Gaudart J, Launoy G (2008). The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Statistics in Medicine* 27: 6310-31.

Haan M, Kaplan GA, Camacho T (1987). Poverty and Health: Prospective Evidence from the Alameda County Study. *American Journal of Epidemiology* 125: 989-98.

Hemming K, Hutton JL (2010). Bayesian sensitivity models for missing covariates in the analysis of survival data. Journal of Evaluation in Clinical Practice 18: 238-246.

Herring AH, Ibrahim JG, Lipsitz SR (2002). Maximum likelihood estimation in random effects cure rate models with nonignorable missing covariates. Biostatistics 3: 387-405.

Kalbfleisch JD, Prentice RL (2002). The Statistical Analysis of Failure Time Data (2nd Ed.). Wiley, New York.

Lillard L, Smith JP, Welch F (1986). What do We Really Know About Wages: The Importance of Nonreporting and Census Imputation. *Journal of Political Economy* 94: 489-506.

Link BG, Phelan JC (1995). Social Conditions as Fundamental Causes of Disease. *Journal of Health and Social Behavior* (extra issue): 80-94.

Lipsitz SR, Ibrahim JG (1996a). A conditional model for incomplete covariates in parametric regression models. *Biometrika* 83: 916-922.

Lipsitz SR, Ibrahim JG (1996b). Using the EM-algorithm for survival data with incomplete categorical covariates. *Life Data Analysis* 2: 5-14.

Little RJA, Rubin DB (2002). Statistical Analysis with missing data (2nd Edition). Hoboken, NJ: John Wiley.

Little RJA, Zhang N (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society, Series C* 60: 591-605.

Meng X, Schenker N (1999). Maximum likelihood estimation for linear regression models with right censored outcome and missing predictors. *Computational Statistics and Data Analysis* 29: 471-483.

Rothman K (1986). *Modern Epidemiology*. Boston, MA: Little, Brown, and Compnay.

Rubin DB (1976). Inference and missing data. *Biometrika* 63: 581-592.

Sorlie PD, Backlund E, Keller JB (1995). U.S. Mortality by Economic, Demographic, and Social Characteristics: The National Longitudinal Mortality Study. *American Journal of Public Health* 85: 949-56.

White IR, Royston P (2009). Imputing missing covariate values for the Cox model. Statistics in Medicine 28: 1982-98.

Yan T, Curtin R, Jans M (2010). Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics* 26: 145-164.

Table 1. Baseline survival distribution

Baseline Distribution	$f_0(u)$	$S_0(u)$
Normal	$(2\pi)^{-1} e^{-0.5u^2}$	$1-\Phi(u)$
Logistic	$e^u / (1+e^u)^2$	$(1+e^u)^{-1}$
Extreme value	$\log(2)e^u e^{-\log(2)e^u}$	$e^{-\log(2)e^u}$

Table 2: Missingness in the National Longitudinal Mortality Study (NLMS)

Variables	# of subject missing (n = 579,566)	# of subject missing in the subsample with income observed (n = 559,517)
Income	20,049	0
Education	2,229	185
Race	2124	1997
Gender	0	0
Marital Status	2610	502
Age at baseline	0	0

Table 3: General Missing Data Structure for Section 3

Pattern	Observation, i	z_i	w_i	x_i	(t_i, δ_i)	R_{w_i}	R_{x_i}
1	$i = 1,\ldots,m$	√	√	√	√	u_w	u_x
2	$i = m+1,\ldots,m+r$	√	√	x	√	u_w	$\bar{u}_x$
3	$i = m+r+1,\ldots,n$	√	x	?	√	$\bar{u}_w$	u_x or $\bar{u}_x$

Key: √ denotes observed, x denotes at least one entry missing, ? denotes observed or missing

Table 4: Missing data mechanisms generated in the simulations

Mechanisms	$\alpha_0^{(w)}$	$\alpha_z^{(w)}$	$\alpha_w^{(w)}$	$\alpha_x^{(w)}$	$\alpha_t^{(w)}$	$\alpha_0^{(x)}$	$\alpha_z^{(x)}$	$\alpha_w^{(x)}$	$\alpha_x^{(x)}$	$\alpha_t^{(x)}$
I: All valid	-1	1	0	0	0	-1	1	0	0	0
II: CC valid	-1.7	1	1	1	0	-1.7	1	1	1	0
III: IL valid	-4	1	0	0	0.25	-2.5	1	1	0	0.25
IV: SILAFT valid	-1.5	1	1	0	0	-3.5	1	1	0	0.25

Missing value of W and X are generated based on the following logistic models:

$$\text{logit}\left(P(R_{w_i}=0 \mid z_i, w_i, x_i, (t_i,\delta_i))\right) = \alpha_0^{(w)} + \alpha_z^{(w)} z_i + \alpha_w^{(w)} w_i + \alpha_x^{(w)} x_i + \alpha_t^{(w)} t_i \quad .$$

$$\text{logit}\left(P(R_{x_i}=0 \mid R_{w_i}=1, z_i, w_i, x_i, (t_i,\delta_i))\right) = \alpha_0^{(x)} + \alpha_z^{(x)} z_i + \alpha_w^{(x)} w_i + \alpha_x^{(x)} x_i + \alpha_t^{(x)} t_i$$

In particular, for the four missing data mechanisms:
I: Missingness of $W = f(Z)$, Missingness of $X = f(Z|W$ observed), all four methods are valid;
II: Missingness of $W = f(Z, W, X)$, Missingness of $X = f(Z, W, X|W$ observed), only CC valid;
III: Missingness of $W = f(Z)$, Missingness of $X = f(Z, W|W$ observed), only IL valid;
IV: Missingness of $W = f(Z, W, (t,\delta))$, Missingness of $X = f(Z, W, (t,\delta), W$ observed), only SILAFT valid.

Table 5. Summary RMSEs*1000 of Estimated Regression Coefficients for Before Deletion (BD), Complete Cases (CC), Ignorable Likelihood (IL) and Subsample AFT model, under Four Missing Data Mechanisms

	I	II	III	IV
BD	**92**	**96**	**95**	**91**
CC	**133**	**125**	564	441
IL	**109**	140	**117**	138
SILAFT	**125**	157	420	**119**

*Four missing data mechanisms:
I: Missingness of $W = f(Z)$, Missingness of $X = f(Z|W$ observed), all four methods are valid;
II: Missingness of $W = f(Z, W, X)$, Missingness of $X = f(Z, W, X|W$ observed), only CC valid;
III: Missingness of $W = f(Z)$, Missingness of $X = f(Z, W, (t, \delta)|W$ observed), only IML valid;
IV: Missingness of $W = f(Z, W)$, Missingness of $X = f(Z, W, (t, \delta)|W$ observed), only SILAFT valid.

RMSE estimates $1000 * \sqrt{E\left(\left\|\beta_r - \beta_{TRUE}\right\|^2\right)}$, with r denoting the rth repetition.

Bold values are for methods consistent for the mechanism generating the data

Table 6. RMSE, Empirical Bias, and 95% confidence coverage for Individual Regression Coefficients under Four Missing Data Mechanisms (1000 replications)

RMSE*1000																
	Mechanism I				Mechanism II				Mechanism III				Mechanism IV			
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**42**	**37**	**65**	**35**	**46**	**33**	**69**	**35**	**47**	**34**	**67**	**35**	**45**	**34**	**64**	**33**
CC	**65**	**57**	**90**	**45**	**56**	**46**	**89**	**47**	371	264	262	206	263	208	234	165
IL	**51**	**40**	**79**	**37**	93	45	82	48	**60**	**37**	**85**	**39**	87	58	82	38
SILAFT	**61**	**53**	**85**	**43**	93	60	99	50	255	218	180	178	**54**	**49**	**83**	**45**
Bias*1000																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**-3**	**-3**	**3**	**1**	**-7**	**1**	**14**	**-1**	**-4**	**1**	**4**	**1**	**-2**	**4**	**1**	**0**
CC	**-7**	**-4**	**4**	**0**	**-4**	**3**	**9**	**-2**	-367	-259	-250	-201	-258	-201	-219	-159
IL	**-3**	**-2**	**3**	**0**	79	24	25	29	**-3**	**1**	**3**	**2**	70	44	8	1
SILAFT	**-8**	**-4**	**6**	**0**	76	42	54	16	-249	-214	-164	-173	**0**	**4**	**-4**	**-1**
95% Confidence coverage																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**95.7**	**94.1**	**94.8**	**94.8**	**95.4**	**95.0**	**95.4**	**94.6**	**95.9**	**94.1**	**94.3**	**95.2**	**94.6**	**93.5**	**95.4**	**95.4**
CC	**93.9**	**94.1**	**94.1**	**96.0**	**95.6**	**94.8**	**95.2**	**95.5**	0.0	0.0	13.2	0.2	0.4	1.1	4.0	19.6
IL	**95.5**	**93.6**	**95.5**	**94.9**	63.5	90.6	94.3	87.5	**94.4**	**95.6**	**93.8**	**95.0**	71.4	81.8	94.5	94.2
SILAFT	**95.5**	**93.3**	**94.1**	**94.1**	74.0	84.4	91.3	91.3	0.3	0.0	42.2	1.9	**95.0**	**93.9**	**96.2**	**94.7**

Table 7. Estimates of AFT models: National Longitudinal Mortality Study

	CC			IL			SILAFT		
Parameter	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Intercept	7.02	.0206	<.0001	7.03	.0203	<.0001	7.02	.0206	<.0001
Education: > HS vs. HS or less	.15	.0082	<.0001	.15	.0081	0.299	.15	.0082	<.0001
Adjusted Income	.08	.0018	<.0001	.08	.0018	0.0005	.08	.0018	<.0001
Race: Black vs. White	-.20	.0112	<.0001	-.19	.0111	0.0173	-.20	.0112	<.0001
Race: Other vs. White	.15	.0247	<.0001	.15	.0244	0.2138	.15	.0246	<.0001
Gender: Female vs. Male	.59	.0069	<.0001	.59	.0068	<.0001	.59	.0069	<.0001
Marital Status: Married vs. Other	.21	.0075	<.0001	.21	.0074	0.1715	.21	.0075	<.0001
Age at baseline	-.08	.0003	<.0001	-.07	.0003	<.0001	-.07	.0003	<.0001