Error-Contaminated Survival Data under Additive Hazards Models

Ying Yan and Grace Y. Yi

Department of Statistics and Actuarial Science
University of Waterloo

January 10, 2014

Outline

- Asymptotic Bias Analysis
- Valid Methods that Corrects for Bias
- Model Misspecification and Model Checking
- ACTG 175 Study

Data Structure

- Data Structure: $\{S_i, \delta_i, Z_i(t), i = 1, \cdots, n\}$
  - Observed event time: $S_i = \min(T_i, C_i)$
    - failure time T_i
    - censoring time C_i
  - Censoring indicator: δ_i
  - Covariates: $Z_i(t) = (X_i^T, V_i^T(t))^T$
    - X_i is error-prone
    - $V_i(t)$ is precisely measured and external

- Counting Process Formulation:
  - Counting process: $N_i(t)$
  - At-risk process: $Y_i(t) = I(S_i \geq t)$
  - Martingale: $M_i(t; \beta, \Lambda_0) = N_i(t) - \int_0^t Y_i(u)\{\lambda_0(u)du + \beta^T Z_i(u)du\}$

AIDS Clinical Trials Group (ACTG) 175 Study (Hammer et al. 1996, *New England Journal of Medicine*)

- A double-blind randomized clinical trial
  - Four types of HIV treatments
  - Inaccurate baseline measurements on CD4 counts
  - Around 2100 subjects in this study
  - Censoring proportion: 75.6%
- Our interest: association between treatments and survival

Additive Hazards (AH) Models

1 Lin and Ying (1994, *Biometrika*):

$$\lambda(t|Z_i) = \lambda_0(t) + \beta^T Z_i(t)$$

- $\lambda_0(t)$ is left unspecified

2 McKeague and Sasieni (1994, *Biometrika*):

$$\lambda(t; Z_i) = \alpha_0^T(t) Z_{i,1}(t) + \beta^T Z_{i,2}(t)$$

- $\alpha_0(t)$ is a vector of unknown functions

3 Aalen (1980; 1989, *Stat. in Med.*):

$$\lambda(t; Z) = \beta^T(t) Z_i(t),$$

- $\beta(t)$ is a vector of unknown functions

Compare AH with Proportional Hazards (PH) Model (Cox 1972, *JRSSB*)

- Different focus of PH and AH:
  - PH: the *ratio* of the two hazard functions
  - AH: the *difference* of the two hazard functions
- How to choose between PH and AH?
  - PH: useful when the *relative* covariate effects are of interest
  - AH: useful when the *absolute* covariate effects are the focus
- Link of AH and PH:
  - Special cases of Box-Cox transformation models (e.g., Zeng, Yin and Ibrahim 2006, *JASA*)

Inference for AH model by Lin and Ying (1994)

- Pseudo score function:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - \bar{Z}(t)\} \{dN_i(t) - Y_i(t)\beta^T Z_i(t)dt\},$$

- Solving $U(\beta) = 0$ leads to the pseudo score estimator:

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^{\tau} Y_i(t)\{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - \bar{Z}(t)\} dN_i(t)\right]$$

  - Advantage of AH: Explicit expression of $\hat{\beta}$

Inference for AH model with Measurement Error

- Underlying data: $\{S_i, \delta_i, X_i, V_i(t), i = 1, \cdots, n\}$; X_i is unobservable
- Observed data: $\{S_i, \delta_i, W_{ir}, V_i(t), i = 1, \cdots, n; r = 1, \cdots, n_i\}$
- Measurement error model:

$$W_{ir} = X_i + \epsilon_{ir}, i = 1, \cdots, n; r = 1, \cdots, n_i$$

  - ϵ_{ir} is mean 0 and independent of X_i, $V_i(t)$, S_i, δ_i

- Goal: Inference of β in AH model

$$\lambda(t|X_i, V_i(t)) = \lambda_0(t) + \beta_x^T X_i + \beta_v^T V_i(t)$$

using observed data

Open questions

- Questions:
  - What is the effect that ignores measurement error when using AH models?
  - Is it reasonable to assume measurement error distribution?
  - What happens if we wrongly using AH models?
  - Is it possible check survival model (including AH and PH models) assumptions with mismeasured covariates?

- We aim at providing answers to these questions.

Section 1: Asymptotic Bias Analysis

Observed Hazard Function

- Observed hazard function:

$$\lambda^*(t; W_i, V_i(t)) = \lim_{\Delta t \to 0} \frac{\Pr(T_i \le t + \Delta t | T_i \ge t, W_i, V_i(t))}{\Delta t}$$

- We show

$$\begin{aligned} & \lambda^*(t; W_i, V_i(t)) \\ = \quad & \lambda_0(t) + \beta_x^T E\{X_i | T_i \ge t, W_i, V_i(t)\} + \beta_v^T V_i(t) \end{aligned}$$

- Naive hazard function

$$\lambda^{**}(t; W_i, V_i(t)) = \lambda_0(t) + \beta_x^T \bar{W}_{i\cdot} + \beta_v^T V_i(t)$$

Naive Estimator

- Naive estimator based on $\lambda^{**}(t; W_i, V_i(t))$:

$$\hat{\beta}_{nv} = \left[\sum_{i=1}^{n} \int_0^{\tau} Y_i(t)\{\hat{Z}_i(t) - \tilde{Z}(t)\}^{\otimes 2} dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^{\tau} \{\hat{Z}_i(t) - \tilde{Z}(t)\} dN_i(t)\right]$$

  - $\hat{Z}_i(t) = (\bar{W}_{i\cdot}{}^T, V_i^T(t))^T$
  - $\tilde{Z}(t) = \sum_{i=1}^{n} Y_i(t)\hat{Z}_i(t) / \sum_{i=1}^{n} Y_i(t)$

Asymptotic Bias Analysis of the Naive Estimator

- We obtain

$$\beta^*_{nv} = (B_1 + B_2)^{-1} B_1 \beta,$$

  - β^*_{nv}: limit of $\hat{\beta}_{nv}$
  - Magnitude of B_2 is determined by
    - degree of measurement error
    - numbers n_i of replicates
    - observation process for survival information

Asymptotic Bias Analysis: A simple numerical study

- Data Distribution:
  - $X_i \sim UNIF(-1, 1)$
  - $\lambda(t; X_i) = \lambda_0(t) + X_i\beta$, with $\lambda_0(t) = 1$ and $\beta = 1$
  - $C_i \sim UNIF(0, 4.2)$
  - $W_{ir} = X_i + \epsilon_{ir}$
  - $\epsilon_{ir} \sim N(0, \sigma^2)$, $n_i = 1, 2, 4, 8$ and σ^2 varies from 0 to $Var(X_i)$
- Measure of bias: Asymptotic relative bias

$$(\beta^*_{nv} - \beta)/\beta$$

Asymptotic Bias Analysis

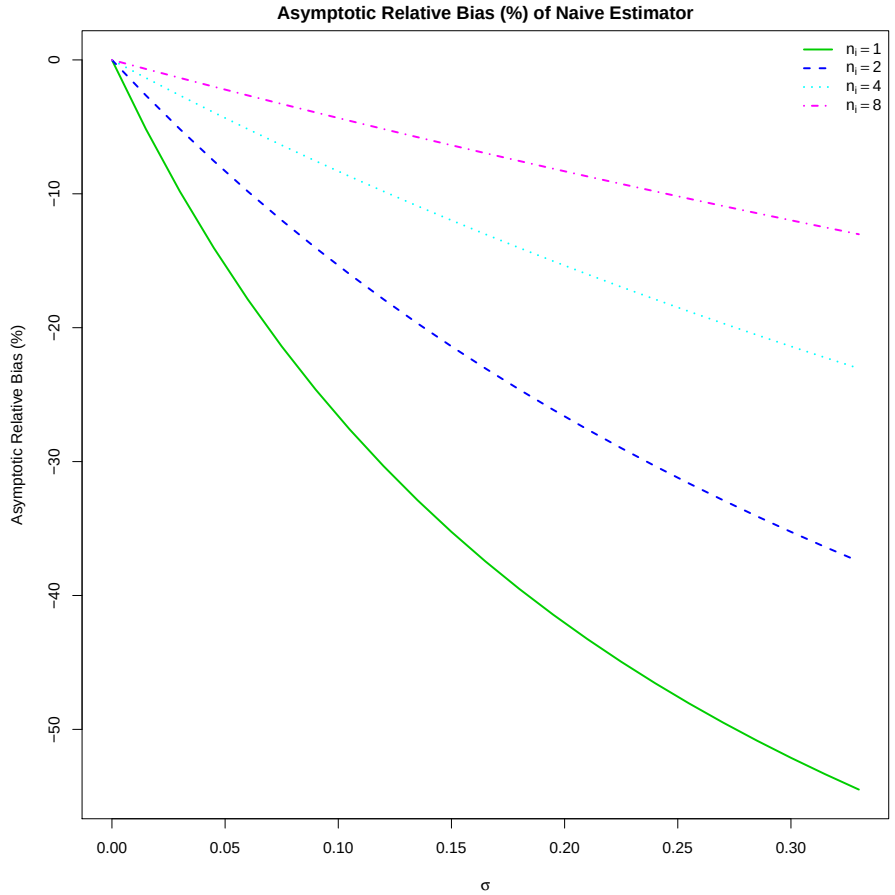

Figure : Asymptotic Bias Analysis

Conclusion

- Measurement error attenuates $\hat{\beta}_{nv}$ to 0
- Degree of attenuation can be reduced by increasing the number n_i of replicated measurements
- Demand on consistent methods to adjust for measurement error

Section 2: Valid Methods that Corrects for Bias

Method 1

- $U_{nv}(\beta)$: naive pseudo score function based on observed data
- Goal: Construct a corrected version of $U_{nv}(\beta)$ with unbiased property
- We obtain

$$U_c(\beta) = U_{nv}(\beta) + \int_0^\tau \left\{ 1 - \frac{1}{\sum_{j=1}^n Y_j(t)} \right\} \sum_{i=1}^n \{ Y_i(t) \hat{\Sigma}_1 \beta / n_i \} dt$$

  - $\hat{\Sigma}_1$ reflects degree of measurement error
  - $U_c(\beta)$ is unbiased: $E\{U_c(\beta)\} = 0$

Method 1: Corrected Pseudo Score Estimator

- Corrected pseudo score estimator of β:

$$\begin{aligned}\hat{\beta}_c &= \left(\sum_{i=1}^n \int_0^\tau Y_i(t)(\hat{Z}_i(t)-\tilde{Z}(t))^{\otimes 2}dt - \int_0^\tau \left\{1-\frac{1}{\sum_{j=1}^n Y_j(t)}\right\}\sum_{i=1}^n \{Y_i(t)\hat{\Sigma}_1/n_i\}dt\right)^{-1} \\ &\quad \times\left(\sum_{i=1}^n \int_0^\tau \{\hat{Z}_i(t)-\tilde{Z}(t)\}dN_i(t)\right)\end{aligned}$$

  - Explicit expression
  - No distributional assumption on measurement error
  - $\hat{\beta}_c$ is consistent of β and asymptotic normal

- Corrected pseudo score estimator of cumulative hazard:

$$\hat{\Lambda}_0(t;\hat{\beta}_c) = \int_0^t \left\{\frac{\sum_{i=1}^n dN_i(u)}{\sum_{j=1}^n Y_j(u)}\right\} - \int_0^t \hat{\beta}_c^T \tilde{Z}(u)du$$

  - $\hat{\Lambda}_0(t;\hat{\beta}_c)$ is uniformly consistent of $\Lambda_0(t)$ and converges weakly to a mean-zero Gaussian process

Method 2

- In the absence of measurement error:

$$\sum_{i=1}^{n} \int_0^{\tau} dM_i(t;\beta,\Lambda_0) = 0,$$

$$\sum_{i=1}^{n} \int_0^{\tau} Z_i(t) dM_i(t;\beta,\Lambda_0) = 0.$$

- In the presence of measurement error:

$$\sum_{i=1}^{n} \int_0^{\tau} d\tilde{M}_i(t;\beta,\Lambda_0) = 0,$$

$$\sum_{i=1}^{n} \int_0^{\tau} \hat{Z}_i(t) d\tilde{M}_i(t;\beta,\Lambda_0) + \sum_{i=1}^{n} \int_0^{\tau} Y_i(t) \hat{\Sigma}_1 \beta / n_i dt = 0.$$

Proposed Method 2: Estimating Equation Estimator

- Estimating equation estimator of β:

$$\begin{aligned}\hat{\beta}_e &= \left(\sum_{i=1}^n \int_0^\tau Y_i(t)(\hat{Z}_i(t)-\tilde{Z}(t))^{\otimes 2}dt - \sum_{i=1}^n \{Y_i(t)\hat{\Sigma}_1/n_i\}dt\right)^{-1} \\ &\quad \times\left(\sum_{i=1}^n \int_0^\tau \{\hat{Z}_i(t)-\tilde{Z}(t)\}dN_i(t)\right)\end{aligned}$$

  - Explicit expression
  - No distributional assumption on measurement error
  - $\hat{\beta}_e$ is consistent of β and asymptotic normal

- Estimating equation estimator of cumulative hazard:

$$\hat{\Lambda}_0(t;\hat{\beta}_e) = \int_0^t \left\{\frac{\sum_{i=1}^n dN_i(u)}{\sum_{j=1}^n Y_j(u)}\right\} - \int_0^t \hat{\beta}_e^T \tilde{Z}(u)du$$

  - $\hat{\Lambda}_0(t;\hat{\beta}_e)$ is uniformly consistent of $\Lambda_0(t)$ and converges weakly to a mean-zero Gaussian process

Section 3: Model Misspecification and Model Checking

Consequence of Misspecifying AH model

- True model:
$$\lambda(t; Z_i(t)) = \lambda_{cox}(t) \exp\{\alpha^T Z_i(t)\}$$
- Working model:
$$\lambda(t; Z_i(t)) = \lambda_0(t) + \beta^T Z_i(t)$$
- We show that
$$\beta \approx R\alpha$$
- R depends on survival, censoring and covariate distrbutions

Conclusion

- β usually differs from α
- Their interpretations are different
- Misspecification of AH model distorts inference procedure
  - β usually differs from α
  - Different interpretation of parameters
- Currently no tool for checking survival models (including AH and PH models) in the presence of measurement error
- Our aim: develop a valid tool to check AH model with covariate measurement error

Model Checking Statistic

- We propose an overall goodness-of-fit test statistic

$$S_c = \sup_{t\in[0,\tau]} \sum (\hat{\Sigma}_c^{-1})_{jj}^{1/2} |n^{-1/2} U_{c,(j)}(\hat{\beta}_c, t)|,$$

  - $U_c(\hat{\beta}_c, t)$: A mean-zero process based on observed data

- Reason: an abnormally large value of S_c can indicate that the AH model and/or the error model are incorrectly specified

- Resampling techniques used to obtain the quantiles of S_c

Simulation Study: Case 1

- True model:

$$\lambda(t; X_i, V_i) = \lambda_0(t) + X_i\beta_x + V_i\beta_v$$

- $W_{ir} = X_i + \epsilon_{ir}$

Table : Empirical size of the proposed test statistic

σ	Method	No Censoring	30% Censoring
	S_{true}	0.057	0.054
0.25	S_{nv}	0.158	0.056
	S_c	0.047	0.062
0.75	S_{nv}	0.595	0.065
	S_c	0.047	0.052

Simulation Study: Case 2

- True model:

$$\lambda(t|Z_i) = \lambda_0(t)\exp(X_i\alpha_x + V_i\alpha_v)$$

- $W_{ir} = X_i + \epsilon_{ir}$

Table : Empirical power of the proposed test statistic

σ	Method	No Censoring	30% Censoring
	S_{true}	0.920	0.791
0.25	S_c	0.872	0.740
0.75	S_c	0.462	0.402

Section 4: ACTG 175 Study

Models

- AH model:

$$\lambda(t; Z_i) = \lambda_0(t) + X_i\beta_x + V_i\beta_v$$

  - X_i unobservable true baseline CD4 counts
  - V_i: treatment indicator

- Error Model:

$$W_{ir} = X_i + \epsilon_{ir}$$

  - W_{ir}: replicated measurements of CD4 count

- Reliability ratio:

$$\hat{\Sigma}_{xx}/(\hat{\Sigma}_{xx} + \hat{\Sigma}_0) = 69.3\%$$

  - A considerable degree of measurement error

- Goodness of Fit: p-value of S_c is 0.859
  - no evidence against AH model or error model

Result

Table : Analyse of the ACTG 175 dataset using different methods

Data	Method	CD4			Treatment		
		EST	MVE	MCI	EST	MVE	MCI
Data Subsets with Replicates	$\hat{\beta}_{nv}$	-4.67	2.15	$(-5.58, -3.77)$	-2.12	1.18	$(-2.80, -1.45)$
	$\hat{\beta}_{szs}$	-5.76	3.36	$(-6.90, -4.63)$	-2.16	1.19	$(-2.84, -1.49)$
	$\hat{\beta}_{rc}$	-5.71	3.20	$(-6.82, -4.60)$	-2.14	1.18	$(-2.81, -1.47)$
	$\hat{\beta}_{c}$	-5.78	3.40	$(-6.93, -4.64)$	-2.16	1.19	$(-2.84, -1.49)$
	$\hat{\beta}_{e}$	-5.79	3.40	$(-6.93, -4.64)$	-2.16	1.19	$(-2.84, -1.49)$
Full Data	$\hat{\beta}_{nv}$	-4.72	2.13	$(-5.62, -3.81)$	-2.15	1.16	$(-2.81, -1.48)$
	$\hat{\beta}_{rc}$	-5.77	3.19	$(-6.88, -4.67)$	-2.16	1.16	$(-2.83, -1.50)$
	$\hat{\beta}_{c}$	-5.85	3.41	$(-7.00, -4.71)$	-2.18	1.17	$(-2.86, -1.51)$
	$\hat{\beta}_{e}$	-5.85	3.41	$(-7.00, -4.71)$	-2.18	1.17	$(-2.86, -1.51)$

Acknowledgment

- This talk is based on our paper
  - Yan, Y. and Yi, G. Y. (2014?), A Class of Functional Methods for Error-Contaminated Survival Data under Additive Hazards Models with Replicate Measurements, *JASA*, Revised.

- We thank Dr. Michael Hughes for providing the ACTG 175 dataset.

- This research was supported by the Natural Sciences and Engineering Research Council of Canada.

- Yan was partially funded by the Trainee Award of The CANNeCTIN Biostatistics and Methodological Innovation Program.