Fast Accelerated Failure Time Modeling for Case-Cohort Data

Sangwook Kang

Department of Statistics
University of Connecticut

Jun 29, 2012

Joint work with Sy Han Chiou and Jun Yan at UConn
INTRODUCTION: CASE-COHORT DESIGN

- Cohort studies could be prohibitively expensive
- Effort and cost arise from the assembling of covariate measurement
- Low disease rate: redundant covariate information on controls
- Case-cohort study design: to reduce cost while achieving the same goals
CASE-COHORT DESIGN

COHORT
CASE-COHORT DESIGN

COHORT

subject with disease (case)  subject without disease (control)
CASE-COHORT DESIGN

COHORT

SRS: SUBCOHORT

subject with disease (case) subject without disease (control)

SRS: SUB COHORT
CASE-COHORT DESIGN

COHORT

SRS: SUBCOHORT

subject with disease (case)  subject without disease (control)

SRS: SUB COHORT
GENERALIZED CASE-COHORT DESIGN

- Disease of interest is not rare or the number of cases is not small
- Obtaining covariate information on all the cases might not be feasible
- Some of the disease outcomes might not be rare when multiple disease outcomes are considered (Breslow and Wellner, 2007)
- Obtain covariate information on subsets of cases and subcohort members
GENERALIZED CASE-COHORT DESIGN

COHORT

- Red: Subject with disease type 1
- Blue: Subject with disease type 2
- Purple: Subject with both disease type 1 and 2
- White: Subject without disease (control)
GENERALIZED CASE-COHORT DESIGN

**COHORT**

- Subject with disease type 1
- Subject with disease type 2
- Subject with both disease type 1 and 2

**SRS SUBCOHORT**

- Subject without disease (control)
GENERALIZED CASE-COHORT DESIGN

Sample case with disease type 1

Subject with disease type 1
Subject without disease (control)
Subject with disease type 2
Subject with both disease type 1 and 2
GENERALIZED CASE-COHORT DESIGN

Sample case with disease type 2

COHORT

Sample case with disease type 1

SRS

SRS

SUBCOHORT

Subject with disease type 1

Subject without disease (control)

Subject with disease type 2

Subject with both disease type 1 and 2
GENERALIZED CASE-COHORT DESIGN

Sample case with disease type 2

Sample case with disease type 1

Subject with disease type 1
Subject with disease type 2
Subject with both disease type 1 and 2

Subject without disease (control)

Generalized case cohort study sample
Semiparametric AFT model

- A log-linear model for the failure times with unspecified error distribution is

\[ T = X^\top \beta + \epsilon, \]

- \( T \): log-transformed failure time,
- \( X \): a \( p \times 1 \) covariate vector,
- \( \beta \): \( p \times 1 \) vector of regression parameters, and
- \( \epsilon \): i.i.d. random variable with an unspecified distribution.

- Nice interpretation: directly relates the failure time to covariates
- Attractive alternative to the popular Cox model
- The AFT model has not been as widely used as the Cox model due to computation difficulties.
EXISTING METHODS

- Failure time data from case-cohort studies
  - Prentice (1986), Self and Prentice (1988), Kulich and Lin (2000), Chen (2001), Kong et al. (2004), and many more
  - Modeling either hazard function or survival function
  - AFT models: Nan et al. (2006), Yu et al. (2007), Kong and Cai (2009)

- Rank based estimators for AFT models
  - Prentice (1978), Tsiatis (1990), Ying (1993)
  - Nice theoretical properties but rarely used in practice
EXISTING METHODS

- Rank based method (Jin et al. 2003)
  - The estimating equation with Gehan’s weight is the gradient of an objective function.
  - The objective function can be solved by linear programming.
  - Adapted to case-cohort data by Kong and Cai (2009)

- Induced smoothing method
  - Brown et al., 2005; Brown and Wang, 2007; Johnson and Strawderman, 2009
  - A smoothed version (add a noise to the parameter and then take the expectation of the estimating equation).
  - Asymptotically equivalent and computationally efficient
A rank based estimating equation with Gehan’s weight for full cohort data is

\[ U_n(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (X_i - X_j) I[e_j(\beta) \geq e_i(\beta)] = 0, \quad (1) \]

- \( T_i, C_i, X_i \), \( i = 1, \ldots, n \): \( n \) independent copies of \( \{ T, C, X \} \),
- \( \Delta_i = I[T_i < C_i] \),
- \( I[\cdot] \): the indicator function,
- \( e_i(\beta) = Y_i - X_i^\top \beta \), and
- \( Y_i = \min(T_i, C_i) \)
For case-cohort data, we consider the following weight-adjusted estimating equation (Kong and Cai, 2009)

$$U_n^c(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} h_j \Delta_i (X_i - X_j) I[e_j(\beta) \geq e_i(\beta)] = 0.$$  \hspace{1cm}(2)

- $h_i = \Delta_i + (1 - \Delta_i) \frac{\xi_i}{p_n}$,
- $\xi_i$: the sub-cohort indicator,
- $p_n = \tilde{n}/n$: sub-cohort inclusion probability, and
- $\tilde{n}$: subcohort sample size
By adapting the idea of Brown and Wang (2007), the smoothed version of (2) is

$$\tilde{U}_n(\beta) = E(U_n(\beta + n^{-1/2} Z)) = \sum_{i=1}^{n} \sum_{j=1}^{n} h_j \Delta_i (X_i - X_j) \Phi \left[ \frac{e_j(\beta) - e_i(\beta)}{r^2_{ij}} \right],$$

(3)

- $r^2_{ij} = n^{-1} (X_i - X_j)^\top (X_i - X_j)$; $Z \sim N(0, 1)$, and
- $\Phi(\cdot)$: standard normal cumulative distribution function.

$\tilde{\beta}_n$: solution to (3).
Estimation of the asymptotic variance of $\tilde{\beta}_n$: need to incorporate extra complexity caused by the data structure

Four variance estimators

- Multiplier Bootstrap
- Induced Smoothing
- Huang’s (2002) Approach
- Zeng and Lin’s (2008) Approach
Multiplier Bootstrap

- Multiplier bootstrap estimator adapted from Jin et al (2006) to case-cohort data.

- $\eta_i$: $i = 1, \ldots, n$, be i.i.d. positive random variables with $E(\eta_i) = Var(\eta_i) = 1$. Define

$$U^*_n(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \eta_i \eta_j h_j \Delta_i (X_i - X_j) \Phi \left[ \frac{e_j(\beta) - e_i(\beta)}{r_{ij}^2} \right].$$

- By repeating this process a large number times, the variance matrix can be estimated using the sample variance matrix of the bootstrap sample.

- This approach is very computing intensive for larger sample sizes or more covariates.
Induced Smoothing

- Following the argument in Zeng and Lin (2008) uniformly in a neighborhood of $\beta_0$, equation (3) can be expressed as

$$n^{-1/2} \sum_{i=1}^{n} h_i S_i(\beta_0) + An^{1/2}(\beta - \beta_0) + o_p(1 + n^{1/2}\|\beta - \beta_0\|),$$

- $S_i(\beta_0)$: a zero-mean random vector, and $A$: asymptotic slope matrix of $n^{-1/2} U_n(\beta_0)$.

- The asymptotic variance can be estimated in a sandwich form; $\Sigma = A^{-1} V (A^{-1})^\top$, where $V = V_1 + V_2$ is the variance of $n^{-1/2} \sum_{i=1}^{n} h_i S_i(\beta_0)$.

- The slope matrix $A$ can be estimated directly by

$$A_n = \frac{1}{n} \frac{\partial}{\partial \beta^\top} U_n(\hat{\beta}_n).$$

- Matrix $V$ can be estimated either through a closed-form estimator or through bootstrapping the estimating equations.
Huang’s (2002) Approach

- Estimate $V$ by $V_n$ the same as in the induced smoothing approach first.
- Let $V_n = C_n^T C_n$ be the Cholesky decomposition of $V_n$. Then we solve estimating equations

$$n^{-1} U_n(\beta_k) = n^{-1/2} c_j, \quad j = 1, \ldots, p,$$

where $c_j$ is the $j$th column of $C_n$.
- Let $q_{nj}$ be the solution, $j = 1, \ldots, p$, and let $Q_n$ be the matrix whose $j$ the column is $q_{nj} - \hat{\beta}_n$.
- Then $Q_n^T Q_n$ is an estimate of $\Sigma$. 
Zeng and Lin’s (2008) Approach

- Matrix $V$ is estimated as in the induced smoothing approach.
- The slope matrix $A$ is estimated by resampling.
- Let $Z_b$, $b = 1, \ldots, B$, be $B$ realizations of a $p$-dimensional standard normal random vector.
- Let $U_{nj}$ be the $j$th component of $U_n$.
- The $j$th row of $A$, $j = 1, \ldots, p$ is estimated by the least squares estimate of the regression coefficients when regressing $n^{-1/2} U_{nj} (\hat{\beta}_n + n^{-1/2} Z_b)$ on $n^{-1/2} Z_b$, $i = 1, \ldots, n$. 
Simulation Design

- Model:
  \[ T_i = 2 + X_{1i} + X_{2i} + X_{3i} + \epsilon_i. \]
- \( X_{1i} \) follows Bernoulli with rate 0.5, \( X_{2i} \) and \( X_{3i} \) are uncorrelated standard normal.
- The error term, \( \epsilon_i \), either followed standard normal, standard logistic or standard Gumbel.
- Full cohort sample size: 1500 and 3000
- Censoring percentage: 97%.
- Average case cohort size: 150 and 300.
## Simulation Result 1: Logistic Marginal Error Distribution

- **Cohort size = 1500**

| $\beta$ | JN   | IS   | JN   | IS   | JNMB | ISMB | ISB  | HB   | ZLB  | JNMB | ISMB | ISB  | HB   | ZLB  | CP(%) |
|---------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------|
| $\beta_1$ | 1.029 | 1.034 | 0.565 | 0.567 | 0.523 | 0.525 | 0.533 | 0.524 | 0.530 | 0.934 | 0.935 | 0.936 | 0.934 | 0.941 |
| $\beta_2$ | 1.036 | 1.044 | 0.316 | 0.316 | 0.283 | 0.284 | 0.290 | 0.295 | 0.288 | 0.920 | 0.927 | 0.922 | 0.930 | 0.926 |
| $\beta_3$ | 1.036 | 1.043 | 0.320 | 0.320 | 0.281 | 0.281 | 0.288 | 0.314 | 0.285 | 0.929 | 0.936 | 0.933 | 0.945 | 0.936 |

- **Case cohort size = 300:**

| $\beta$ | JN   | IS   | JN   | IS   | JNMB | ISMB | ISB  | HB   | ZLB  | JNMB | ISMB | ISB  | HB   | ZLB  | CP(%) |
|---------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------|
| $\beta_1$ | 1.034 | 1.036 | 0.512 | 0.513 | 0.481 | 0.481 | 0.488 | 0.483 | 0.487 | 0.937 | 0.939 | 0.937 | 0.931 | 0.938 |
| $\beta_2$ | 1.025 | 1.029 | 0.279 | 0.279 | 0.253 | 0.252 | 0.261 | 0.264 | 0.260 | 0.922 | 0.927 | 0.925 | 0.925 | 0.925 |
| $\beta_3$ | 1.039 | 1.042 | 0.279 | 0.279 | 0.256 | 0.256 | 0.264 | 0.280 | 0.264 | 0.922 | 0.926 | 0.927 | 0.941 | 0.932 |

- **JN:** Jin et al. (2003)’s method, **IS:** induced smoothing method, **JNMB:** naive bootstrap approach with JN, **ISMB:** naive bootstrap approach with IS, **ISB:** A with IS, V with bootstrap, **HB:** Huang (2002), V with bootstrap, **ZLB:** A with Zeng and Lin (2008), V with bootstrap
Simulation Result 2: Logistic Marginal Error Distribution

- Cohort size = 3000

| \( \beta \) | EST | Empirical SE | Estimated SE | CP(%) |
|---|---|---|---|---|
| JN | IS | JN | IS | JNMB | ISMB | ISB | HB | ZLB | JNMB | ISMB | ISB | HB | ZLB |
| \( \beta_1 \) | 1.025 | 1.030 | 0.474 | 0.477 | 0.475 | 0.476 | 0.479 | 0.468 | 0.478 | 0.958 | 0.955 | 0.952 | 0.939 | 0.955 |
| \( \beta_2 \) | 1.043 | 1.050 | 0.271 | 0.272 | 0.255 | 0.255 | 0.258 | 0.261 | 0.255 | 0.932 | 0.934 | 0.936 | 0.933 | 0.936 |
| \( \beta_3 \) | 1.053 | 1.060 | 0.255 | 0.255 | 0.257 | 0.256 | 0.259 | 0.285 | 0.256 | 0.945 | 0.948 | 0.947 | 0.963 | 0.946 |

- Case cohort size = 150:

- Case cohort size = 300:

| \( \beta \) | EST | Empirical SE | Estimated SE | CP(%) |
|---|---|---|---|---|
| JN | IS | JN | IS | JNMB | ISMB | ISB | HB | ZLB | JNMB | ISMB | ISB | HB | ZLB |
| \( \beta_1 \) | 1.001 | 1.003 | 0.374 | 0.375 | 0.372 | 0.373 | 0.375 | 0.368 | 0.374 | 0.952 | 0.956 | 0.948 | 0.945 | 0.951 |
| \( \beta_2 \) | 1.024 | 1.028 | 0.212 | 0.212 | 0.204 | 0.204 | 0.207 | 0.207 | 0.205 | 0.934 | 0.940 | 0.942 | 0.938 | 0.941 |
| \( \beta_3 \) | 1.016 | 1.020 | 0.212 | 0.212 | 0.204 | 0.204 | 0.206 | 0.222 | 0.205 | 0.940 | 0.938 | 0.938 | 0.954 | 0.938 |
Timing Result in seconds with censoring percentage 97%

| dist | JNMB | ISMB | ISB | HB  | ZLB  |
|------|------|------|-----|-----|------|
| N    | 114.205 | 82.415 | 1.670 | 1.752 | 3.092 |
| L    | 112.427 | 65.485 | 1.444 | 1.519 | 2.713 |
| G    | 137.899 | 98.884 | 1.992 | 2.089 | 3.676 |

Full cohort size = 1500
Case cohort size = 150

| dist | JNMB | ISMB | ISB | HB  | ZLB  |
|------|------|------|-----|-----|------|
| N    | 407.666 | 161.214 | 3.350 | 3.493 | 6.102 |
| L    | 388.146 | 132.569 | 2.883 | 3.029 | 5.311 |
| G    | 445.477 | 188.789 | 3.884 | 4.034 | 7.072 |

Case cohort size = 300

| dist | JNMB | ISMB | ISB | HB  | ZLB  |
|------|------|------|-----|-----|------|
| N    | 187.968 | 150.941 | 3.273 | 3.437 | 5.976 |
| L    | 145.954 | 120.800 | 2.755 | 2.894 | 5.138 |
| G    | 266.283 | 194.313 | 4.138 | 4.336 | 7.523 |

Full cohort size = 3000
Case cohort size = 150

| dist | JNMB | ISMB | ISB | HB  | ZLB  |
|------|------|------|-----|-----|------|
| N    | 619.144 | 300.366 | 6.242 | 6.553 | 11.527 |
| L    | 538.123 | 236.015 | 5.342 | 5.609 | 9.992 |
| G    | 818.494 | 369.259 | 7.558 | 7.920 | 13.926 |

Case cohort size = 300
Application to the National Wilm’s Tumor Study Group (NWTSG) Cohort Studies

- D’Angio et al (1989), Green et al (1998).
- The dataset is available in the Survival package in R.
- Wilm’s Tumour is a rare kidney cancer in young children.
- Interest: assessing the relationship between the tumour histology and tumour relapse.
- Assume central histology measurement is available only for case-cohort sample
- Covariates: Stages of Wilms’ tumours (I - IV) and age
- $n = 4,028$, # of cases $= 571$ (86% censoring rate), average subcohort size $= 663$
### Data Analysis Results

| Effects | EST IS | SE ISMR | SE ISB | SE HB | SE ZLB |
|---------|--------|---------|--------|-------|--------|
| (time)  | 3727.774 | 22.478  | 24.264 | 36.306 |        |
| histol  | −3.415 | 0.409 | 0.458 | 0.425 | 0.458 |
| age     | −0.188 | 0.074 | 0.080 | 0.243 | 0.080 |
| stage2  | −1.303 | 0.589 | 0.620 | 0.541 | 0.624 |
| stage3  | −1.412 | 0.580 | 0.616 | 0.603 | 0.617 |
| stage4  | −2.120 | 0.664 | 0.710 | 0.762 | 0.712 |

**Case-Cohort Analysis:**

| (time) | (145938.4) | (1261.0) | (1304.3) | (879.5) |
|--------|-------------|----------|-----------|--------|
| histol | −2.861      | 0.137    | 0.146     | 0.144  |
| age    | −0.156      | 0.026    | 0.029     | 0.028  |
| stage2 | −1.231      | 0.231    | 0.223     | 0.233  |
| stage3 | −1.347      | 0.216    | 0.224     | 0.223  |
| stage4 | −1.966      | 0.248    | 0.242     | 0.244  |