

# Semi-parametric Transformation Models for Case-cohort Study

Wenbin Lu

lu@stat.ncsu.edu

<http://www4.stat.ncsu.edu/~lu>

*NC State University*  
DEPARTMENT OF STATISTICS



# Outline

- **Semi-parametric models and estimations**

# Outline

- **Semi-parametric models and estimations**
- **Case-cohort study**

# Outline

- **Semi-parametric models and estimations**
- **Case-cohort study**
- **Weighted estimating equations (WEE) approaches for case-cohort study**

# Outline

- **Semi-parametric models and estimations**
- **Case-cohort study**
- **Weighted estimating equations (WEE) approaches for case-cohort study**
- **Numerical studies**

# Outline

- **Semi-parametric models and estimations**
- **Case-cohort study**
- **Weighted estimating equations (WEE) approaches for case-cohort study**
- **Numerical studies**
- **Remarks and future work**

# Semi-parametric survival models

- Cox proportional hazards model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta' Z)$$

# Semi-parametric survival models

- Cox proportional hazards model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta' Z)$$

- Proportional odds model

$$\frac{1 - S(t|Z)}{S(t|Z)} = \frac{1 - S_0(t)}{S_0(t)} \exp(\beta' Z)$$

# Semi-parametric survival models

- Cox proportional hazards model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta' Z)$$

- Proportional odds model

$$\frac{1 - S(t|Z)}{S(t|Z)} = \frac{1 - S_0(t)}{S_0(t)} \exp(\beta' Z)$$

- Linear transformation models

$$H(T) = -\beta' Z + \epsilon$$

$H$  - an unknown monotone increasing function

$\epsilon$  - error term

# Estimations for linear transformation models

- Cheng, Wei and Ying (1995) and Fine, Ying and Wei (1998)'s approach:

# Estimations for linear transformation models

- Cheng, Wei and Ying (1995) and Fine, Ying and Wei (1998)'s approach:
  - Main assumption: censoring time  $C$  is independent of covariates and  $T$

# Estimations for linear transformation models

- Cheng, Wei and Ying (1995) and Fine, Ying and Wei (1998)'s approach:
  - Main assumption: censoring time  $C$  is independent of covariates and  $T$
  - Unbiased estimating equation

$$\sum_{i \neq j, i, j=1}^N w_{ij}(\theta) \dot{\eta}_{ij}(\theta) \left[ \frac{\delta_j I\{\min(\tilde{T}_i, t_0) \geq \tilde{T}_j\}}{\hat{G}_n^2(\tilde{T}_j)} - \eta_{ij}(\theta) \right] = 0,$$

where  $\tilde{T}_i = \min(T_i, C_i)$ ,  $\delta_i = I(T_i \leq C_i)$ , and  $\theta = (\zeta, \beta)'$  with  $\zeta = H(t_0)$ , where  $t_0$  is a constant such that  $P(\tilde{T} > t_0) > 0$ .  $\hat{G}_n$  is the Kaplan-Meier estimator for the censoring distribution.  $\eta_{ij}(\theta)$  is a known function of  $\theta$ .

- Chen, Jin and Ying (2002)'s approach:

- Chen, Jin and Ying (2002)'s approach:
  - More general assumption:  $C$  is independent of  $T$  given covariates

- Chen, Jin and Ying (2002)'s approach:
  - More general assumption:  $C$  is independent of  $T$  given covariates
  - Martingale based estimating equations

$$\sum_{i=1}^N [dN_i(t) - Y_i(t)d\Lambda\{H(t) + \beta'Z_i\}] = 0, t \geq 0,$$

and

$$\sum_{i=1}^N \int_0^{\infty} Z_i [dN_i(t) - Y_i(t)d\Lambda\{H(t) + \beta'Z_i\}] = 0.$$

where  $N_i(t)$  and  $Y_i(t)$  are the usual counting and at-risk processes.

# Case-cohort study

- Introduced by Prentice (1986) for large epidemiological and other event history studies.

# Case-cohort study

- Introduced by Prentice (1986) for large epidemiological and other event history studies.
- Subcohort

# Case-cohort study

- Introduced by Prentice (1986) for large epidemiological and other event history studies.
- Subcohort
- Partial covariate information

# Case-cohort study

- Introduced by Prentice (1986) for large epidemiological and other event history studies.
- Subcohort
- Partial covariate information
- Advantage: saving the cost in many large medical studies

# Case-cohort study

- Introduced by Prentice (1986) for large epidemiological and other event history studies.
- Subcohort
- Partial covariate information
- Advantage: saving the cost in many large medical studies
- pseudo-likelihood approach for the PH model

# WEE approaches

- Kong, Cai & Sen's approach (2004):

# WEE approaches

- Kong, Cai & Sen's approach (2004):
  - Main assumption: censoring time  $C$  is independent of covariates and  $T$

# WEE approaches

- Kong, Cai & Sen's approach (2004):
  - Main assumption: censoring time  $C$  is independent of covariates and  $T$
  - Unbiased estimating equation

$$\sum_{i \neq j, i, j=1}^N \rho_{ij} w_{ij}(\theta) \dot{\eta}_{ij}(\theta) \left[ \frac{\delta_j I\{\min(\tilde{T}_i, t_0) \geq \tilde{T}_j\}}{\hat{G}_n^2(\tilde{T}_j)} - \eta_{ij}(\theta) \right] = 0,$$

where  $\rho_{ij} = \rho_i \rho_j$  with  $\rho_i = \delta_i + (1 - \delta_i)\xi_i/p$ . And  $\xi_i = 1/0$  denoting the subcohort indicator.  $p$  is the probability that a subject is selected into subcohort.

# WEE approaches (ctd)

- Our approach:

# WEE approaches (ctd)

- Our approach:
  - More general assumption:  $C$  is independent of  $T$  given covariates

# WEE approaches (ctd)

- Our approach:
  - More general assumption:  $C$  is independent of  $T$  given covariates
  - Martingale based estimating equations

$$\sum_{i=1}^N \rho_i [dN_i(t) - Y_i(t)d\Lambda\{H(t) + \beta' Z_i\}] = 0, t \geq 0,$$

and

$$\sum_{i=1}^N \int_0^{\infty} Z_i \rho_i [dN_i(t) - Y_i(t)d\Lambda\{H(t) + \beta' Z_i\}] = 0.$$

- Asymptotic properties of our estimates

Proposition. Under suitable regularity conditions, we have that

$$N^{\frac{1}{2}}(\hat{\beta} - \beta_0) \rightarrow N\{0, A^{-1}\Sigma(A^{-1})'\}$$

in distribution, as  $N \rightarrow \infty$ . Moreover,  $A$  and  $\Sigma$  can be consistently estimated by

$$\hat{A} = \frac{1}{N} \sum_{i=1}^N \int_0^{\tau} \rho_i\{Z_i - \bar{Z}(t)\} Z_i' \dot{\lambda}\{\hat{H}(t) + \hat{\beta}' Z_i\} Y_i(t) d\hat{H}(t),$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N \left[ \int_0^{\tau} \rho_i\{Z_i - \bar{Z}(t)\} d\hat{M}_i(t) \int_0^{\tau} \rho_i\{Z_i - \bar{Z}(s)\}' d\hat{M}_i(s) \right].$$

# Numerical study

- hazard function of error term  $\epsilon$

$$\lambda(t) = \exp(t) / \{1 + \gamma \exp(t)\}$$

where  $\gamma = 0, 1, 2$  (Dabrowska & Doksum, 1988).

# Numerical study

- hazard function of error term  $\epsilon$

$$\lambda(t) = \exp(t) / \{1 + \gamma \exp(t)\}$$

where  $\gamma = 0, 1, 2$  (Dabrowska & Doksum, 1988).

- $H(t)$  is chosen as  $\log(t)$  for  $\gamma = 0$ ,  $\log(e^t - 1)$  for  $\gamma = 1$ , and  $\log(0.5e^{2t} - 0.5)$  for  $\gamma = 2$ .

# Numerical study

- hazard function of error term  $\epsilon$

$$\lambda(t) = \exp(t) / \{1 + \gamma \exp(t)\}$$

where  $\gamma = 0, 1, 2$  (Dabrowska & Doksum, 1988).

- $H(t)$  is chosen as  $\log(t)$  for  $\gamma = 0$ ,  $\log(e^t - 1)$  for  $\gamma = 1$ , and  $\log(0.5e^{2t} - 0.5)$  for  $\gamma = 2$ .
- Covariate  $Z = (Z_1, Z_2)'$ , where  $Z_1$  follows  $U[0, 1]$  and  $Z_2$  follows  $\text{Ber}(0.5)$ .

# Numerical study

- hazard function of error term  $\epsilon$

$$\lambda(t) = \exp(t) / \{1 + \gamma \exp(t)\}$$

where  $\gamma = 0, 1, 2$  (Dabrowska & Doksum, 1988).

- $H(t)$  is chosen as  $\log(t)$  for  $\gamma = 0$ ,  $\log(e^t - 1)$  for  $\gamma = 1$ , and  $\log(0.5e^{2t} - 0.5)$  for  $\gamma = 2$ .
- Covariate  $Z = (Z_1, Z_2)'$ , where  $Z_1$  follows  $U[0, 1]$  and  $Z_2$  follows  $\text{Ber}(0.5)$ .
- $C$  follows  $U[0, c]$ .

# Numerical study

- hazard function of error term  $\epsilon$

$$\lambda(t) = \exp(t) / \{1 + \gamma \exp(t)\}$$

where  $\gamma = 0, 1, 2$  (Dabrowska & Doksum, 1988).

- $H(t)$  is chosen as  $\log(t)$  for  $\gamma = 0$ ,  $\log(e^t - 1)$  for  $\gamma = 1$ , and  $\log(0.5e^{2t} - 0.5)$  for  $\gamma = 2$ .
- Covariate  $Z = (Z_1, Z_2)'$ , where  $Z_1$  follows  $U[0, 1]$  and  $Z_2$  follows  $\text{Ber}(0.5)$ .
- $C$  follows  $U[0, c]$ .
- Full Cohort  $N = 1000$ .

# Numerical study

- hazard function of error term  $\epsilon$

$$\lambda(t) = \exp(t) / \{1 + \gamma \exp(t)\}$$

where  $\gamma = 0, 1, 2$  (Dabrowska & Doksum, 1988).

- $H(t)$  is chosen as  $\log(t)$  for  $\gamma = 0$ ,  $\log(e^t - 1)$  for  $\gamma = 1$ , and  $\log(0.5e^{2t} - 0.5)$  for  $\gamma = 2$ .
- Covariate  $Z = (Z_1, Z_2)'$ , where  $Z_1$  follows  $U[0, 1]$  and  $Z_2$  follows  $\text{Ber}(0.5)$ .
- $C$  follows  $U[0, c]$ .
- Full Cohort  $N = 1000$ .
- Two case-cohort designs considered.

- Covariate independent censoring ( $C \sim U[0, c]$ )

Study Design	$\beta_1 = 1$			$\beta_2 = -1$		
	CCI	CCII	FULL	CCI	CCII	FULL
(a) $\gamma = 1$						
Mean( $\hat{\beta}$ )	1.017	1.014	0.985	-1.046	-1.042	-1.014
SD( $\hat{\beta}$ )	0.538	0.447	0.362	0.338	0.286	0.236
Mean SE	0.591	0.483	0.382	0.336	0.282	0.235
CP	97.6	96.2	97.0	94.4	94.0	95.2
RE	0.45	0.66	1	0.49	0.68	1
RE*	0.34	0.47	0.65	0.47	0.60	0.73
(b) $\gamma = 2$						
Mean( $\hat{\beta}$ )	1.093	1.032	0.993	-1.047	-1.042	-1.011
SD( $\hat{\beta}$ )	0.620	0.512	0.391	0.362	0.309	0.246
Mean SE	0.642	0.522	0.416	0.359	0.300	0.247
CP	95.6	96.2	96.2	96.2	94.6	95.2
RE	0.40	0.58	1	0.46	0.63	1
RE*	0.35	0.48	0.65	0.46	0.59	0.73

- Covariate independent censoring (ctd)

$$\beta_1 = 1$$

$$\beta_2 = -1$$

$$(c) \gamma = 0$$

Mean( $\hat{\beta}$ )	1.014	0.999	0.975	-1.030	-1.028	-1.023
SD( $\hat{\beta}$ )	0.534	0.417	0.337	0.340	0.287	0.230
Mean SE	0.544	0.446	0.355	0.317	0.268	0.225
CP	93.8	96.2	94.6	93.4	93.8	95.0
RE	0.40	0.65	1	0.46	0.64	1
RE*	0.33	0.47	0.65	0.49	0.62	0.76

CCI - the first case-cohort design;

CCI - the second case-cohort design;

SD - sample standard deviation;

Mean SE - mean of estimated standard error;

CP - empirical coverage probability of 95% confidence interval;

RE - empirical relative efficiencies of our estimators.

RE\* - empirical relative efficiencies of Kong et al. (2004)'s estimators.

- Covariate dependent censoring ( $C$  follows Cox model)

Model	$\beta_c$	$\beta_1 = 1$				$\beta_2 = -1$			
		Mean( $\hat{\beta}$ )	SD( $\hat{\beta}$ )	SE	CP	Mean( $\hat{\beta}$ )	SD( $\hat{\beta}$ )	SE	CP
$\gamma = 1$	0	1.055	0.658	0.639	94.0	-1.055	0.398	0.366	93.2
	0.2	1.034	0.628	0.633	95.8	-1.051	0.368	0.364	95.8
	log 2	1.054	0.642	0.635	94.4	-1.032	0.395	0.363	92.0
$\gamma = 2$	0	1.022	0.670	0.695	95.2	-1.020	0.420	0.397	95.2
	0.2	1.061	0.645	0.698	97.8	-1.062	0.399	0.396	93.8
	log 2	1.028	0.707	0.693	95.0	-1.036	0.389	0.393	96.6
$\gamma = 0$	0	1.062	0.595	0.567	94.0	-1.060	0.357	0.335	93.4
	0.2	1.033	0.601	0.567	92.6	-1.054	0.357	0.335	93.0
	log 2	1.075	0.603	0.571	93.6	-1.067	0.343	0.333	94.6

SD - sample standard deviation;

SE - mean of estimated standard error;

CP - empirical coverage probability of 95% confidence interval.

# Remarks and future work

- Consider more general censoring assumption

# Remarks and future work

- Consider more general censoring assumption
- Easy to implement and allow a rigorous development of asymptotic normality with an explicit formula for the variance-covariance matrix

# Remarks and future work

- Consider more general censoring assumption
- Easy to implement and allow a rigorous development of asymptotic normality with an explicit formula for the variance-covariance matrix
- Subcohort sampling may depend on some available covariates for the full cohort

# Remarks and future work

- Consider more general censoring assumption
- Easy to implement and allow a rigorous development of asymptotic normality with an explicit formula for the variance-covariance matrix
- Subcohort sampling may depend on some available covariates for the full cohort
- More efficient estimation by adding some weight functions

# Remarks and future work

- Consider more general censoring assumption
- Easy to implement and allow a rigorous development of asymptotic normality with an explicit formula for the variance-covariance matrix
- Subcohort sampling may depend on some available covariates for the full cohort
- More efficient estimation by adding some weight functions
- Double robust ideas might also be used to improve efficiency