A Pseudolikelihood Approach to Analysis of Nested Case-Control Studies

Yubing Yao, Yiding Zhang

February 17, 2017

Nested Case Control Study-Introduction

1. Nested Case Control Study is a study design, constructed of case-control within cohort study.
2. Typically all the cases in the cohort study will be selected. Then for each case, a specified number of controls (usually 1-5) are selected among those haven't developed disease by the occurrence time of the case. At each time controls are selected independently.
3. Nested case-control design can be matched, not matched or counter-matched.
4. or many research questions, the nested case-control design potentially offers impressive reductions in costs and efforts of data collection and analysis compared with the full cohort approach, with relatively minor loss in statistical efficiency.

Nested Case Control Study-Example

1. In one cohort study-Nurses' Health Study, 91,523 women did not have cancer at baseline at the start of the study and were followed for 14 years, 2,341 women had developed breast cancer by 1993.
2. If we are interested in the association between gene expression and breast cancer incidence, it would be very expensive and possibly wasteful of precious blood specimen to assay all 89,000 women without breast cancer.
3. In this situation, one may choose to assay all of the cases, and also, for each case, select a certain number of women to assay from the risk set of participants who have not yet failed (i.e. those who have not developed breast cancer before the particular case in question has developed breast cancer).

Advantages and disadvantages of Nested Case Control Study

1. **Advantages**:
    - Efficient-not all subjects of cohort require expensive diagonastic testing or biological assay
    - Flexible-allows testing hypotheses not anticipated when the cohort was drawn
    - Reduce selection bias-cases and noncases sampled from same population
    - Reduced information bias-risk factor exposure can be assessed with investigator blind to case status
2. **Disadvantage**: Reduces power (from parent cohort) because of reduced sample size by: $1/(c+1)$, where c = number of controls per case

Another type of case control study within cohort : case-cohort design

1. A case-cohort study is similar to a nested case-control study in that the cases and non-cases are within a large cohort study; cases and non-cases are identified at the event time, after baseline.

2. In a case-cohort study, the cohort members were assessed for risk factors at any time prior to event time. Non-cases are randomly selected from the cohort, forming a subcohort. No matching is performed.

Cox Model and Partial Likelihood

1. For survival outcome(partially censoring time to event outcomes) Cox proprosed a proportional hazard model:

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i(t))$$

   where regression parameters-$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)', \lambda_0(t)$ is baseline hazard function.

2. An intuitive understanding of partial likelihood defined by Cox in 1975 is that the estimation of parameters of interest-$\boldsymbol{\beta}$ in the proportional hazard model didn't involve the baseline hazard function-$\lambda_0(t)$, which is relatively a constant term with respect to $\boldsymbol{\beta}$, thus we can reduce full likelihood function based on proportional hazard model to partial likelihood removing those terms only related to $\lambda_0(t)$.

Cox Model and Partial Likelihood in cohort studies

1. **Data setting in cohort study with time to event outcome**:
   - Assume n individuals, each individual i enters the study at age b_i and is followed up until age c_i and the disease is developed at age T_i and if $b_i < T_i \leq c_i$, the time event at T_i is record; if $T_i > c_i$, then the T_i is right censored at c_i.
   - p possibly time-dependent coariates-$\boldsymbol{Z}_i(t) = (Z_{1i}(t), \ldots, Z_{pi}(t))'$.
2. The Cox partial likelihood can be expressed as:

$$L_c(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i(X_i)\}}{\sum\limits_{j \in \mathcal{R}_i} \exp\{\boldsymbol{\beta}'\boldsymbol{Z}_j(X_i)\}} \right]^{D_i} = \prod_{i=1}^{n} \left[\frac{\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i(X_i)\}}{S^{(0)}(\boldsymbol{\beta}, X_i)} \right]^{D_i}$$

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)', X_i = \min(T_i, c_i), D_i = \boldsymbol{I}(X_i = T_i), \mathcal{R}_i = \{j : X_j \geq X_i > b_j\}$ and

$$S^{(0)}(\boldsymbol{\beta}, t) = \sum_{j : b_j < t \leq X_j} \exp\{\boldsymbol{\beta}'\boldsymbol{Z}_j(t)\}$$

Cox Model and Partial Likelihood in nested case-control studies

1. **Data setting in nested case control study with time to event outcome**:
   - For each case from cohort study(X_i for $D_i = 1$), define $Y_i = \#\mathcal{R}_i$, the number of individuals at risk for subject i.
   - one sample m a set $\widetilde{\mathcal{R}}_{i0} = \{j_{i1}, j_{i2}, \dots, j_{im}\}$ where $m < Y_i$ from $\mathcal{R}_i \backslash \{i\}$.
2. Then the partial likelihood of Cox's model in nested case control study with setting above is:

$$\tilde{L}_c(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta' \boldsymbol{Z}_i(X_i)\}}{\sum\limits_{j \in \widetilde{\mathcal{R}}_i} \exp\{\beta' \boldsymbol{Z}_j(X_i)\}} \right]^{D_i}$$

where $\widetilde{\mathcal{R}}_i = \widetilde{\mathcal{R}}_{i0} \cup \{i\}$

Pseudolikelihood in nested case control studies

1. For the study of time-to-event outcomes, case-cohort and nested case-control designs, and their stratified version, exposure-stratified case-cohort and counter-matching designs, have been widely applied

2. The Cox model (proportional hazard model), widely used for modeling time to event outcomes, is not appropriate for nested case control studies, where controls matching with cases are biased samples from cohort studies and control sets at different event times could be dependent.

Inclusion probabilities in a nested case-control study

1. The data from cohort study-$\mathfrak{F} = \{(b_i, X_i, D_i); i = 1, 2, \ldots, n\}$
2. Conditional on the data $\mathfrak{F}$ from the conhort, the proability that some individual j is ever selected as a control in the nested case-control study is given by:

$$p_{0j} = 1 - \prod_{b_j < X_i < X_j} \left(1 - \frac{m}{Y_i - 1} D_i\right)$$

3. Define p_j conditional on $\mathfrak{F}$, the probability being included as a case or control in nested case-control study.

$$p_j = \begin{cases} 1 & if \quad D_j = 1, \\ p_{0j} & if \quad D_j = 0, \end{cases}$$

Denote V_{j0} the indicator that individual j is ever selected as a control, $V_j = \max(D_j, V_{j0})$ the indicator that individual j is ever either a case or a control.

Estimation based on pseudolikelihood in nested case control studies

A pseudolikelihood approach is proposed incoporating the inclusion probability of the controls into likelihood function in nested case-control studies. Thus for the proportional hazard model the regression parameters-β can be esimated by maximizing the semiparametric pseudolikelihood:

$$\tilde{L}_p(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta' \boldsymbol{Z}_i(X_i)\}}{\sum_{j \in \widetilde{\mathcal{R}}_i} \widetilde{S}^{(0)}(\beta, X_i)} \right]^{D_i}$$

where $\widetilde{S}^{(0)}(\beta, X_i) = \sum_{j \in \mathcal{R}_i} \frac{V_j}{p_j} \exp\{\beta' \boldsymbol{Z}_j(X_i)\} = \sum_{j \in \widetilde{\mathcal{R}}_i} \frac{1}{p_j} \exp\{\beta' \boldsymbol{Z}_j(X_i)\}$

Simulation Study

Model: Cox model with Weibull baseline hazard.

$$\lambda_i(t) = \theta t^{\theta-1} \exp(\beta Z_i + \gamma),$$

where

$$\begin{aligned} Z_i &\sim \text{Uniform}[0,1], \\ \beta &= 1, \\ \gamma &= \text{intercept} \\ \theta &= 2(\text{shape parameter}) \end{aligned}$$

Simulation Study

Simulation scenarios:

- Cohort size $n = 1000$.
- Nested case control rate $m = 1$ or 3.
- Censoring $C_i \perp Z_i$ and $C_i \not\perp Z_i$ ($C_i \propto Z_i$); 125 expected case subjects (87.5%).
- Estimators: Parametric pseudolikelihood, Semiparametric pseudolikelihood, Semiparametric partial likelihood.
- Repeat 500 times.

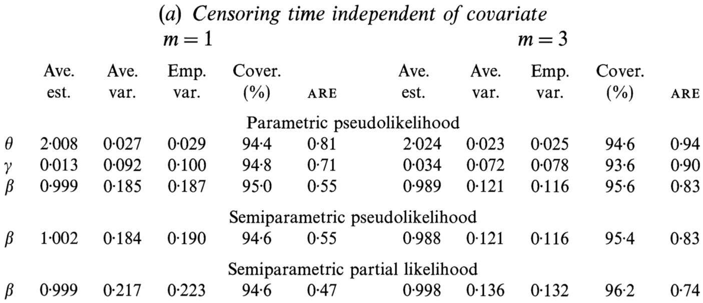
Simulation Results

Estimates that we care about:

- Average of parameter estimates (AVE.*est*): $\bar{\hat{\theta}} = \frac{1}{500}\sum_{i=1}^{500}\hat{\theta}_i$
- Average variance (AVE.*var*): $\bar{var}(\hat{\theta}) = \frac{1}{500}\sum_{i=1}^{500} var_i(\hat{\theta}_i)$
- Empirical variances (Emp.*var*) $Empvar = \frac{1}{500}\sum_{i=1}^{500}(\hat{\theta}_i - \bar{\hat{\theta}})^2$
- Coverage probabilities of $(1-\alpha)\%$ confidence intervals: $p = \frac{\text{\# of true } \theta \text{ within 95\% CIs of } \hat{\theta}s}{500}$
- Asymptotic relative efficiency (ARE): $ARE = \frac{var_{cohort}(\hat{\theta})}{var_{simulation}(\hat{\theta})}$.

Simulation Results

1.Censoring $C_i \perp Z_i$:



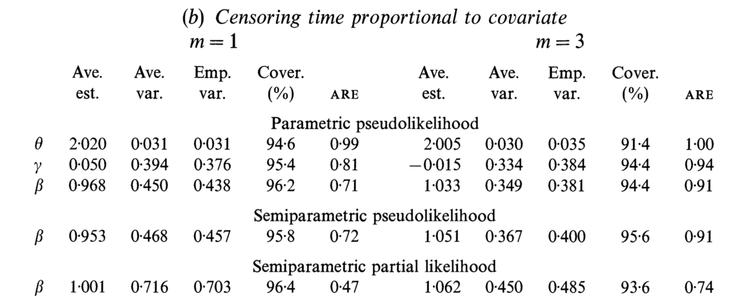
(a) *Censoring time independent of covariate*

	$m=1$					$m=3$				
	Ave. est.	Ave. var.	Emp. var.	Cover. (%)	ARE	Ave. est.	Ave. var.	Emp. var.	Cover. (%)	ARE
	Parametric pseudolikelihood									
θ	2·008	0·027	0·029	94·4	0·81	2·024	0·023	0·025	94·6	0·94
γ	0·013	0·092	0·100	94·8	0·71	0·034	0·072	0·078	93·6	0·90
β	0·999	0·185	0·187	95·0	0·55	0·989	0·121	0·116	95·6	0·83
	Semiparametric pseudolikelihood									
β	1·002	0·184	0·190	94·6	0·55	0·988	0·121	0·116	95·4	0·83
	Semiparametric partial likelihood									
β	0·999	0·217	0·223	94·6	0·47	0·998	0·136	0·132	96·2	0·74

- $m = 1$ and $Z_i \perp C_i$: All three methods have good and similar point estimate of β. Asymptotic relative efficiencies of β are similar.
- $m = 3$ and $Z_i \perp C_i$: Point estimate of β is more close to true value in semiparametric partial likelihood. ARE are higher in semiparametric pseudolikelihood method and parametric pseudolikelihood method then simiparametric partial likelihood method.

Simulation Results

2.Censoring $C_i \not\perp Z_i$:



(*b*) *Censoring time proportional to covariate*

	$m=1$					$m=3$				
	Ave. est.	Ave. var.	Emp. var.	Cover. (%)	ARE	Ave. est.	Ave. var.	Emp. var.	Cover. (%)	ARE
	Parametric pseudolikelihood									
θ	2·020	0·031	0·031	94·6	0·99	2·005	0·030	0·035	91·4	1·00
γ	0·050	0·394	0·376	95·4	0·81	−0·015	0·334	0·384	94·4	0·94
β	0·968	0·450	0·438	96·2	0·71	1·033	0·349	0·381	94·4	0·91
	Semiparametric pseudolikelihood									
β	0·953	0·468	0·457	95·8	0·72	1·051	0·367	0·400	95·6	0·91
	Semiparametric partial likelihood									
β	1·001	0·716	0·703	96·4	0·47	1·062	0·450	0·485	93·6	0·74

- $m = 1$ and $Z_i \propto C_i$: Point estimate of β of semiparametric partial likelihood method is more close to the true value, but its ARE is much less then 1 as compared with semiparametric pseudolikelihood and parametric pseudolikelihood method (0.47 vs 0.72 and 0.71).
- $m = 3$ and $Z_i \propto C_i$: Point estimate of β of parametric pseudolikelihood method is more close to the true value. ARE of semiparametric partial likelihood is much less then 1 as compared with semiparametric pseudolikelihood and parametric

- All point estimates are close to true value ($\theta = 2$, $\gamma = 0$, $\beta = 1$).
- All average variance and empirical variance are close to each others, except for $m = 3$ and $Z_i \propto C_i$, the average variances are lower than the empirical variances and which also happened on the cohort.
- The coverage proportions agreed with 95% confidence interval except for the value of 91.4% on θ when $m = 3$ and $Z_i \propto C_i$.
- ARE of θ and γ are higher than β in all schemes.
- $m = 1$ vs $m = 3$: Point estimate of βs do not change a lot, but ARE of all methods in $m = 3$ scheme are higher than those in $m = 1$ scheme.
- $Z_i \perp C_i$ vs $Z_i \propto C_i$: Point estimate of βs do not change a lot, but ARE of all methods in $Z_i \propto C_i$ scheme are higher than those in $Z_i \perp C_i$ scheme.
- Semiparametric pseudolikelihood always has highest ARE of β in all scheme.

Discussion

- Semiparametric pseudolikelihood estimator is relatively efficient compared to the partial likelihood estimator.
- Langholz & Thomas (1990): Improvement seemed to be at best moderate and NCC could be considerably better under heavy right censoring or delayed entry.
- Weighting techniques may be inefficient (parametric model without covariate).
- Individuals can be reused. (traditional estimator's limitation)
- Have advantages in unmatched case-control studies.

Thank You!