Stratified case-cohort analysis of general sampling designs

Southampton, November 2006

Sven Ove Samuelsen¹, Hallvard Ånestad² & Anders Skrondal³

¹University of Oslo, ²Ullevål University Hospital, ³London School of Economics

Email: osamuels@math.uio.no
Objectives

- Show how covariance matrices for an estimator proposed by Borgan, Langholz, Samuelsen, Goldstein & Pogoda (2000) *Exposure stratified case-cohort designs*, LIDA 6 may be calculated from standard software.

- Point out that the "local averaging estimator" of Chen, K (2001) *Generalized case-cohort sampling*, JRSSB 63 may be viewed as a post-stratified case-cohort estimator.

- Extend the framework of Chen (2001) to covariate dependent sampling designs such as stratified case-cohort and counter matching designs.

The estimators of Chen (2001) and extensions may thus be fitted as stratified case-cohort estimators.
Outline:

• Introduction
• Stratified case-cohort: Variance estimation
• Chen’s Generalized case-cohort and stratified case-cohort
• Extension of generalized case-cohort
• Conclusion
Background: Unmatched case-control

Sample proportion

- \( \pi_1 \approx 1 \) of individuals with disease (cases)
- \( \pi_0 \) (small) of individuals without disease (controls)
- Fit logistic regression model to case-control data = MLE

Alternatively: Maximize weighted log-likelihood

\[
\tilde{l}(\beta) = \sum_{i=1}^{n} \frac{V_i}{\pi(i)} [Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)]
\]

where \( Y_i \) indicator of disease, \( p_i \) prob. of disease, \( V_i \) indicator of being sampled and \( \pi(i) \) prob. that ind. \( i \) sampled.

Advantage of weighted likelihood: "Consistent" estimates when model is wrong (for wrong model)
Survival analysis and case-control

Traditional case-control designs ignore time. Need to extend to designs that takes length follow-up into account.

Two such designs:

- Nested case-control (Thomas, 1977) very similar to matched case-control design
- Case-cohort (Prentice, 1986) more like unmatched case-control
Case-cohort study (Prentice, 1986)

Covariate information collected on

- cases

- individuals in a randomly selected subcohort

\[ \text{Diagram:} \]

- Cases
- At risk
- Subcohort

\[ \text{Time on study:} \]

\[ t_1, t_2, t_3, t_4 \]
In counting process formulation:

- Cohort \( \{(N_i(t), Y_i(t), Z_i) : i = 1, 2, \ldots, n, 0 \leq t \leq \tau\} \)
  - \(N_i(t)\) indicator of event in \([0, t]\)
  - \(Y_i(t)\) indicator of being at risk at \(t\)
  - \(Z_i\) covariates
- Cases: \(N_i(\tau) = 1\)
- Subcohort: \(V_i^0\) indicator of being sampled from cohort with sampling fraction \(p\)
- Case-cohort: \(\{(N_i(t), Y_i(t), Z_i) : \max(V_i^0, N_i(\tau)) = 1\}\)

Usually: Want to fit a proportional hazards (Cox’) model.
Surrogate = rough measure of exposure available cohort.

• Divide the cohort into $L$ strata according to surrogate variable with $n_l^0$ individuals in stratum $l$
• Select subcohort with stratified sampling, $m_l^0$ individuals from stratum $l$
• Obtain covariates on cases and subcohort
• Modify strata after observing case status: $n_l$ non-cases in stratum $l$ of which $m_l$ in sub-cohort

Modified incl. prob:

$$p_i = \begin{cases} 
1 & \text{for cases} \\
\frac{m_l}{n_l} = \pi_l & \text{for non-cases in stratum } l
\end{cases}$$
Estimation

May fit proportional hazards model

\[ \lambda_i(t) = \exp(\beta' Z_i) \lambda_0(t) \]

by maximizing weighted log-partial likelihood

\[ \tilde{l}(\beta) = \sum_{i=1}^{n} \int_0^\tau \left[ \beta' Z_i - \log(\tilde{S}^{(0)}(\beta, t)) \right] dN_i(t) \]

where

\[ \tilde{S}^{(0)}(\beta, t) = \sum_{i=1}^{n} \frac{V_i}{p_i} Y_i(t) \exp(\beta' Z_i) \]

and \( V_i = \max(N_i(\tau), V_i^0) \) is the indicator of being in case-cohort sample.
Distributional results stratified case-cohort

Maximizer $\hat{\beta}$ approximately normal with covariance matrix that may be estimated by

$$\tilde{I}^{-1} + \sum_{l=1}^{L} m_l \frac{1 - \pi_l}{\pi_l^2} \tilde{I}^{-1} \tilde{\Delta}_l \tilde{I}^{-1}$$

where $\tilde{I}$ information matrix and $\tilde{\Delta}_l$ covariance matrix of

$$X_i = \int_{0}^{T} [Z_i - \frac{\tilde{S}^{(1)}(\tilde{\beta}, t)}{\tilde{S}^{(0)}(\tilde{\beta}, t)}] Y_i(t) \exp(\tilde{\beta}'Z_i) \frac{dN^\bullet(t)}{\tilde{S}^{(0)}(\tilde{\beta}, t)}$$

over stratum $l$ excluding the cases.

Note: For non-cases $-X_i$ are score-influence terms.
Calculation of covariance matrix: Observe that

\[
\frac{1}{\pi_i^2} \tilde{I}^{-1} \Delta_l \tilde{I}^{-1} = \frac{1}{m_l - 1} \sum (D_i - \bar{D}_l)(D_i - \bar{D}_l)^T
\]

where \( D_i \) are the DFBETAS, i.e.

\[
D_i = -\tilde{I}^{-1} X_i / p_i
\]

which can be computed in most software (Therneau & Li, 1999).

Example in R/Splus:

```r
strcox<-coxph(Surv(time,d) ~ z1+z2,weights=1/p)
dfb<-resid(strcox,type='dfbeta')

gma<-numeric(0)
for (str in 1:no.strata){
  indst<-(1:length(time))[stratum==str]
  gma<-gma+(1-m[str]/n[str])*m[str]*var(dfb[indst,])
}
adjvar<-stratcox$var+gma
```
Simulation

\[ T_i \sim \exp(\beta' Z_i) 2t \quad \text{(Weibull)} \]
\[ C_i \sim U[0, 0.5], Z_i \sim U[0, 1], \beta = 1 \]

Roughly 12.5% cases

Strata defined by \( Z_i < (>) 0.5 \)

13% sampling fraction both strata

Results 5000 repetitions:

<table>
<thead>
<tr>
<th></th>
<th>Average ( \hat{\beta} )</th>
<th>Ave. var-est</th>
<th>Empir. Var</th>
<th>Rob. Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 1000 )</td>
<td>1.023</td>
<td>0.198</td>
<td>0.210</td>
<td>0.250</td>
</tr>
<tr>
<td>( n = 10000 )</td>
<td>1.003</td>
<td>0.0192</td>
<td>0.0186</td>
<td>0.0244</td>
</tr>
</tbody>
</table>
Conservatism of robust variance estimator

The robust variance estimator works well for standard case-cohort data.

But in the simulations for stratified case-cohort it was clearly conservative.

The reason is that the robust variance estimator is asymptotically equivalent to using uncentered second order moment

\[ \frac{1}{m_l} \sum D_i D_i^\top \]

instead of centered second order moment

\[ \frac{1}{m_l} \sum (D_i - \bar{D}(l))(D_i - \bar{D}(l))^\top \]

"Generalized case-cohort designs", Chen (2001):

Class of designs:

- Finite set of sampling steps
- Simple random sample of controls and cases in each step
- Sampling may not depend on covariates or surrogate variables

Generalized case-cohort includes:

- Case-cohort
- Nested case-control: Control sampling from risk set
- "Traditional" case-control: Controls are sampled after follow-up from individuals that did not become cases.
Chen: local averaging

Same analysis method proposed for the different designs!

- Partition $0 = s_0 < s_1 < \ldots < s_L = \tau$
- By some control sampling scheme $m_l$ controls selected from $n_l$ censored in $(s_{l-1}, s_l]$
- Weight controls by inverse sampling fraction $\frac{n_l}{m_l} = \frac{1}{\pi_l}$
- Maximize weighted partial likelihood

(Chen also considers case sampling, here: all cases sampled)

But by

- defining strata intervals $I_l = (s_{l-1}, s_l]$ for censoring times
- conditioning on $m_l$ sampled from $n_l$ in $I_l$

this amounts to post-stratification on censoring interval.
**Post-stratified case-cohort**

Since "local averaging" equivalent to post-stratification on
- Case-status
- Time to censoring (or event)

model fitting and variance estimation by can be accomplished by methods of Borgan et al. (2000).

Actually, Borgan et al. redefined sampling fraction after observing the number of cases in the subcohort:
Moderate form of post-stratification.
## Simulations (unstratified) case-cohort

### Panel A: Censoring independent of the covariate

<table>
<thead>
<tr>
<th></th>
<th>Cohort (Cox)</th>
<th>Case-cohort with post-stratification on case-status only</th>
<th>5 intervals</th>
<th>10 intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean estimate</td>
<td>1.001</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.10</td>
<td>0.25</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>—</td>
<td>0.25</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.10</td>
<td>0.26</td>
<td>0.26</td>
<td>0.28</td>
</tr>
</tbody>
</table>

### Panel B: Censoring time proportional to the covariate

<table>
<thead>
<tr>
<th></th>
<th>Cohort (Cox)</th>
<th>Case-cohort with post-stratification on case-status only</th>
<th>5 intervals</th>
<th>10 intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean estimate</td>
<td>1.00</td>
<td>1.02</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.34</td>
<td>0.58</td>
<td>0.38</td>
<td>0.36</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>—</td>
<td>0.59</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.33</td>
<td>0.60</td>
<td>0.38</td>
<td>0.36</td>
</tr>
</tbody>
</table>
Nested case-control (NCC)

Design:

- At each event time $t_j$ sample $m$ controls from risk set.
- Covariates for case and controls at time $t_j$.
Estimators with nested case-control data

  - Conditional logistic regression
  - Stratified Cox-regression
- Compute inclusion probabilities $1 - \prod [1 - \frac{mY_i(t-dN\cdot(t))}{Y(t)-1}]$ (Samuelsen, 1997)
- Local averaging or post-stratification censoring interval (Chen, 2001)
- Inclusion probabilities by GAM: Smooth indicator of being sampled against censoring time
Probability of being sampled in NCC study
Simulations nested case-control, \( m = 1 \) control per case

Panel A: *Censoring time independent of the covariate*

<table>
<thead>
<tr>
<th>Inclusion probabilities</th>
<th>Traditional Thomas (1977)</th>
<th>Samuelsen (1997)</th>
<th>GAM</th>
<th>Post-stratified 10 intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean estimate</td>
<td>1.02</td>
<td>1.02</td>
<td>1.012</td>
<td>1.02</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.22</td>
<td>–</td>
<td>–</td>
<td>0.19</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>–</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.22</td>
<td>0.19</td>
<td>0.19</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Panel B: *Censoring time proportional to the covariate*

<table>
<thead>
<tr>
<th>Inclusion probabilities</th>
<th>Traditional Thomas (1977)</th>
<th>Samuelsen (1997)</th>
<th>GAM</th>
<th>Post-stratified 10 intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean estimate</td>
<td>1.03</td>
<td>0.98</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.71</td>
<td>–</td>
<td>–</td>
<td>0.35</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>–</td>
<td>0.53</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.73</td>
<td>0.48</td>
<td>0.35</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Chen’s Local averaging uses

that interval lengths \( s_j - s_{j-1} \to 0 \)
(i.e. \( s_j - s_{j-1} = o_p(n^{0.5}) \)).

For case-cohort this is not a necessary requirement:

- Will improve large sample efficiency
- But likely main efficiency gain achieved with a few strata
- For small sample efficiency strata should not be too small

For nested case-control:

- Estimates may be biased with long intervals - since inclusion probability depends on length of observation
- Strata not too large or small
Covariate (surrogate) dependent sampling

Chen (2001) unnecessary restricts attention to sampling designs that do not depend on covariates or surrogates of covariates.

It is straightforward to generalize to post-stratification both on

- Surrogate variables
- Censoring interval

Examples:

- Stratified case-cohort
- Counter-matching (Langholz & Borgan, 1995)
- Bernoulli sampling based on surrogate information (Kalbfleisch & Lawless, 1988, Robins et. al, 1995)
Simulations stratified case-cohort
(with censoring independent of covariate)

<table>
<thead>
<tr>
<th></th>
<th>Original stratification scheme</th>
<th>Post-stratified 5 intervals</th>
<th>Post-stratified 10 intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean estimate</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.20</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>0.25</td>
<td>0.26</td>
<td>0.28</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.21</td>
<td>0.16</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Efficiency improvement can be explained by inspecting score-influence term

\[
X_i = \int_0^\tau \left[ Z_i - \frac{\tilde{S}^{(1)}(\tilde{\beta}, t)}{\tilde{S}^{(0)}(\tilde{\beta}, t)} \right] Y_i(t) \exp(\tilde{\beta}' Z_i) \frac{dN^\bullet(t)}{\tilde{S}^{(0)}(\tilde{\beta}, t)}
\]

which depends on stratum - through covariate - and also on followup-time.
Counter-matching (Langholz & Borgan, 1995)

= stratified nested case-control \((n_k(t) \text{ at risk in stratum } k)\)
  
  - Case at time \(t_j\) in stratum \(l\)
  - Sample \(m_{l'}\) controls at risk in stratum \(l' \neq l\)
  - Sample \(m_l - 1\) controls at risk in stratum \(l\)

Data fitted with
  
  - Stratified Cox-regr. with offsets \(\log(n_k(t)/m_k)\) (L & B)
  - Strata-wise incl. prob. similar to Samuelsen (1997)
  - Local averaging inclusion probabilities
  - Smoothed inclusion probabilities (by GAM)
## Simulation counter-matching

<table>
<thead>
<tr>
<th></th>
<th>Traditional counter-matching</th>
<th>Inclusion probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Similar to Samuelsen (1997)</td>
</tr>
<tr>
<td>Mean estimate</td>
<td>1.01</td>
<td>1.03</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.12</td>
<td>—</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>—</td>
<td>0.20</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.12</td>
<td>0.16</td>
</tr>
</tbody>
</table>

In contrast with stratified case-cohort post-stratification on time does not seem to improve on the traditional method of Langholz & Borgan.

The generality of this (and other) simulation results should be investigated.
**Bernoulli sampling**

Individuals are drawn independently with given inclusion probability $p_i$ depending on case-status and surrogate variables.

Fit Cox-model with weighted partial likelihoods where weights are inverse inclusion probabilities given as:

- Specified sampling probabilities (Kalbfleisch & Lawless, 1988)
- Observed sampling fractions $\frac{m_l}{n_l}$ (Breslow & Wellner, 2006-7, SJS)
- Observed sampling fractions also post-stratified on censoring interval
## Simulation Bernoulli sampling

<table>
<thead>
<tr>
<th></th>
<th>Original sampling fraction</th>
<th>Corrected sampling fraction</th>
<th>Post-stratified scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean estimate</td>
<td>1.04</td>
<td>1.04</td>
<td>1.03</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.25</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>Mean robust variance</td>
<td>0.25</td>
<td>0.25</td>
<td>0.27</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.28</td>
<td>0.21</td>
<td>0.17</td>
</tr>
</tbody>
</table>

- Results very similar to stratified case-cohort
- With specified sampling fractions robust variance and estimated variance similar
Conclusions

- Variance estimation easy for stratified case-cohort data
- Methods of Chen (2001) can be fitted as stratified case-cohort
- Generalization of Chen’s approach to surrogate dependent sampling possible
- Efficiency gains are often possible, but was not found for counter-matching

- Inverse probability weighting generally not fully efficient, but other methods (e.g. Kulich & Lin, 2004, Scheike & Juul, 2004) more difficult to implement.
Case-cohort vs. traditional nested case-control

- Nested case-control more efficient than traditional case-cohort - with much censoring
- IWP can make nested case-control more efficient
- Much easier to fit models outside proportional hazards with weighting techniques
- Competing risks: With trad NCC controls sampled to one type of case can not be used as control for other types, not a problem with weighting
- Missing data for case or control: Bigger problem for trad. NCC than with weighting

So there are good reasons for using weighting techniques with NCC data.
Case-cohort vs. traditional nested case-control, II

However, NCC is usually matched on more factors than time, such as

- Sex
- Calendar time
- Community

This makes estimation of inclusion probabilities harder

- Multidimensional non-parametric estimates
- Sparse strata
- May need model for inclusion probabilities
For more details see:


Statistical Research Report 1, Department of Mathematics, University of Oslo

http://www.math.uio.no/eprint/