Nested case-control and case-cohort studies

Ørnulf Borgen
Department of Mathematics
University of Oslo
NORBIS course
University of Oslo
4-8 December 2017

Outline:
- Radiation and breast cancer data
- Nested case control studies
- Case-cohort studies
- Stratified case-cohort studies


Radiation and breast cancer
(e.g. Hrubec et al., Cancer Research, 1989)

- We will for illustration use data from a cohort of 1720 women discharged from two tuberculosis sanatoria in Massachusetts 1930-1956
- Radiation doses to the breasts due to fluoroscopic examinations have been estimated for all women
- The women have been followed until end of 1980 by which time it was 75 breast cancer cases
- We want to a study the effect of radiation exposure on breast cancer risk
- We have data for the full cohort, but will use it to illustrate nested case-control and case-cohort studies

Cohort data and model

Observe events (e.g. occurrences of a disease) in a cohort of \( n \) individuals

\[ T_1 < T_2 < \cdots < T_d \]

The individual having an event at time \( T_j \) (the case) is denoted \( i_j \)

(arrows are censored observations)
Counting process for individual $i$

$$N_i(t) = \sum I\{T_j \leq t, i_j = i\}$$

has intensity process $\lambda_i(t)$ of the Cox regression form

$$\lambda_i(t) = Y_i(t) \alpha(t) \exp \left( \beta^T x_i \right)$$

at risk risk indicator

baseline hazard

hazard ratio

(relative risk)

Partial likelihood

$$L(\beta) = \prod_{i \in R} \exp \left( \beta^T x_i \right) \sum_{r \in R} \exp \left( \beta^T x_i \right)$$

We need covariate information for everyone at risk

**Cohort sampling**

- Cohort studies need information on covariates for all individuals at risk
- Expensive to collect and check covariate information for all individuals in large cohorts
- In biomarker studies it will also imply a waste of valuable biological material
- Using a cohort sampling design one only needs to collect covariate information for the cases and a sample of controls
- Two types of cohort sampling designs:
  - Matched designs: nested case-control
  - Unmatched designs: case-cohort

**Classical nested case-control design**

Select at random $m-1$ controls among those at risk when a case occurs (case excluded), i.e. match on study time

Illustration for two controls per case ($m = 3$)

We will derive a partial likelihood

Sampled risk set $\tilde{R}_j$ at time $T_j$ consists of the case $i_j$ and its sampled controls

Introduce the counting processes

$$N_{i,r}(t) = \sum I\{t_j \leq t, i_j = i, \tilde{R}_j = r\}$$

Their intensity processes are given by

$$\lambda_{i,r}(t) = \lambda_i(t) \cdot \pi(r | t, i) = Y_i(t) \cdot \exp \left( \beta^T x_i \right) \alpha(t) \pi(r | t, i)$$

where

$$\pi(r | t, i) = \binom{n(t) - 1}{m - 1}$$

is the probability of selecting the set $r$ as the sampled risk set if individual $i$ fails at time $t$
Introduce the aggregated processes

\[ N_i(t) = \sum_{l \in r} N_{ix}(t) \quad \lambda_i(t) = \sum_{l \in r} \lambda_{ix}(t) \]

The probability that individual \( i \) fails given that a failure occurs at \( t \) and given that the sampled risk set is \( r \):

\[ \pi(i \mid t, r) = \frac{\lambda_{ix}(t)}{\lambda_i(t)} = \frac{Y_i(t) \alpha_{i0}(t) \exp(\beta^T x_i)}{\sum_{l \in r} Y_{lx}(t) \alpha_{0x}(t) \exp(\beta^T x_l) \pi(l \mid r, t)} \]

The partial likelihood is a product of such factors over all failure times and sampled risk sets:

\[ L_{\text{acc}}(\beta) = \prod_{r_j} \frac{\lambda_{i_j \tilde{r}_j}(T_j)}{\lambda_{i_j}(T_j)} = \prod_{r_j} \frac{\exp(\beta^T x_{i_j})}{\sum_{l \in r_j} \exp(\beta^T x_l)} \]

The partial likelihood looks like the full cohort partial likelihood, but the sums are only over the sampled risk sets.

---

**Radiation and breast cancer example**

Select nested case-control data with 2 controls per case (m=3)

Consider model with \( \log_2(\text{dose}+1) \) as only covariate:

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\beta} )</th>
<th>( se(\hat{\beta}) )</th>
<th>( Z )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>0.491</td>
<td>0.162</td>
<td>3.04</td>
<td>0.002</td>
</tr>
<tr>
<td>Nested case-control</td>
<td>0.540</td>
<td>0.237</td>
<td>2.28</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Empirical relative efficiency for nested case-control relative to full cohort:

\[
\text{relative efficiency} = \frac{\text{Var(cohort)}}{\text{Var(nested case-control)}} = \frac{0.162^2}{0.237^2} = 0.47
\]

Asymptotic relative efficiency is \((m-1)/m\) for a model with one covariate with no effect

---

A note of stratification/matching

Sometimes we would like to adopt a stratified Cox model where the intensity process for an individual \( i \) in stratum \( c \) takes the form:

\[ \lambda_i(t) = Y_i(t) \alpha_{i0}(t) \exp(\beta^T x_i) \]

For the radiation and breast cancer example, one could e.g. stratify on sanatorium

For cohort data we then use a modification of the partial likelihood where the risk sets only contain those under observation in the same stratum as the case.
Similarly, in a nested case-control study, the controls are selected from the case’s stratum (matching)

Illustration for one control per case

The partial likelihood for nested case-control data remains unchanged

Prentice proposed to base estimation on the pseudo likelihood

\[ L_p(\beta) = \prod_{i} \frac{\exp(\beta'x_i)}{\sum_{j \in SC \cup \text{Cases}} Y_j(T_j) \exp(\beta'x_i)} \]

Note that a case outside the subcohort is only used at its event time.

The pseudo-likelihood does not possess likelihood properties.

Nevertheless, we may prove that \( \hat{\beta} \) is approximately multivariate normally distributed around the true value of \( \beta \).

But the covariance matrix may not be estimated by the observed information and likelihood ratio tests do not apply.

However, Wald tests may still be used.

Classical case-cohort design

Select by simple random sampling without replacement a subcohort \( SC \) consisting of a fraction \( m/n \) of the full cohort.

Illustration for \( n=8 \) and \( m=4 \)

A small efficiency gain may be obtained by including the cases at all times they are at risk and using inverse probability weighted (IPW) estimation.

We then maximize the IPW pseudo-likelihood

\[ L_w(\beta) = \prod_{i} \frac{\exp(\beta'x_i)}{\sum_{j \in SC \cup \text{Cases}} Y_j(T_j) \exp(\beta'x_i) w_j} \]

Here the weights are \( w_i = 1 \) for all cases and \( w_i = n^0/m^0 \) for subcohort members who do not experience the event, where \( n^0 \) and \( m^0 \) are the number of cohort and subcohort members, respectively, who do not experience the event.
Radiation and breast cancer data

Select subcohort with 150 women

Model with $\log_2(dose+1)$ as only covariate:

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}$</th>
<th>$se(\hat{\beta})$</th>
<th>$Z$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>0.491</td>
<td>0.162</td>
<td>3.04</td>
<td>0.002</td>
</tr>
<tr>
<td>Nested case-control</td>
<td>0.540</td>
<td>0.237</td>
<td>2.28</td>
<td>0.022</td>
</tr>
<tr>
<td>Prentice</td>
<td>0.519</td>
<td>0.215</td>
<td>2.41</td>
<td>0.016</td>
</tr>
<tr>
<td>IPW</td>
<td>0.524</td>
<td>0.211</td>
<td>2.48</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Nested case-control and case-cohort studies tend to have similar relative efficiencies when they are based on (about) the same number of individuals

Here nested case-control uses data for 210 women and case-cohort uses data for 218 women

Stratified case-cohort sampling

From classical sampling theory we know that we may often improve our estimates by using stratified sampling

Often some information, e.g. a surrogate for the exposure of main interest, may be available for all cohort members

In the radiation and breast cancer example the number of fluoroscopic examinations is a surrogate for exposure dose

For case-cohort sampling, we may stratify the cohort into $S$ sampling strata based on the surrogate information and select the subcohort by stratified sampling

Stratified nested case-control sampling (often denoted «counter-matching») is not considered in this course

We may here use the IPW pseudo-likelihood

$$L_w(\beta) = \prod_{T_i \in SC\cup Cases} \frac{\exp(\beta^T x_i)}{\sum_{T_j \in SC\cup Cases} Y_i(T_j) \exp(\beta^T x_j) w_i}$$

Now the weights are $w_i = 1$ for all cases and $w_i = n_i^0 / m_i^0$ for subcohort members from sampling stratum $s$ who do not experience the event, where $n_i^0$ and $m_i^0$ are the number of cohort and subcohort members from stratum $s$ who do not experience the event
**Radiation and breast cancer data**

Select subcohort by stratified sampling with 50 women in each of the sampling strata:
1: no fluoroscopic examinations (698 women)
2: 1-149 fluoroscopic examinations (765 women)
3: 150 fluoroscopic examinations or more (257 women)

Model with $\log_2(dose+1)$ as only covariate:

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}$</th>
<th>se($\hat{\beta}$)</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort:</td>
<td>0.491</td>
<td>0.162</td>
<td>3.04</td>
<td>0.002</td>
</tr>
<tr>
<td>Nested case-control:</td>
<td>0.540</td>
<td>0.237</td>
<td>2.28</td>
<td>0.022</td>
</tr>
<tr>
<td>Prentice:</td>
<td>0.519</td>
<td>0.215</td>
<td>2.41</td>
<td>0.016</td>
</tr>
<tr>
<td>IPW:</td>
<td>0.524</td>
<td>0.211</td>
<td>2.48</td>
<td>0.013</td>
</tr>
<tr>
<td>Stratified case-cohort:</td>
<td>0.506</td>
<td>0.184</td>
<td>2.75</td>
<td>0.006</td>
</tr>
</tbody>
</table>

**Points to consider when choosing a cohort sampling design**

- Statistical efficiency
- Ease of the statistical analysis
- Choice of time scale
- Planning the study workflow
- Multiple endpoints and reuse of controls
- Need for matching
- Stratified sampling

See section 17.5 in the handbook chapter for a discussion of these points