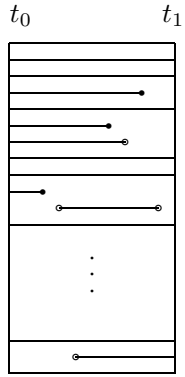
Epidemiology for biostatisticians

January 2007.

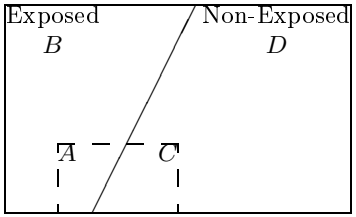
Cohort sampling.

Per Kragh Andersen

Study base: population followed from t_0 to t_1.



Population



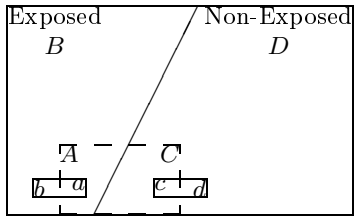
The population at t_0 consists of $A+B$ exposed and $C+D$ non-exposed individuals.

At t_1, A out of the exposed and C out of the non-exposed have developed the disease.

That is,

$$\text{Relative risk} = \frac{A/(A+B)}{C/(C+D)} \qquad \text{Odds ratio} = \frac{A/B}{C/D} = \frac{A \cdot D}{B \cdot C}$$

Population: cohort sample at t_0



Cohort sample:

Exposed:	$k_1(A+B)$	$=$	k_1A	$+$	k_1B
		$\sim$	a		$\sim b$
Non-exposed:	$k_2(C+D)$	$=$	k_2C	$+$	k_2D
		$\sim$	c		$\sim d$

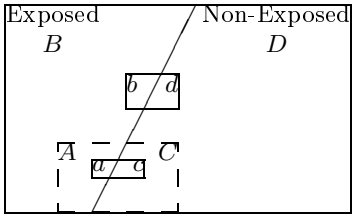
Then: $\dfrac{a}{a+b} \sim \dfrac{k_1A}{k_1A+k_1B} = \dfrac{A}{A+B}$

$\dfrac{c}{c+d} \sim \dfrac{k_2C}{k_2C+k_2D} = \dfrac{C}{C+D}$

$\Rightarrow$ *We can estimate relative risk*

AND odds ratio, since $\dfrac{a \cdot d}{b \cdot c} \sim \dfrac{k_1A \cdot k_2D}{k_1B \cdot k_2C} = \dfrac{A \cdot D}{B \cdot C}$

Population: case-control sample at t_1



Case-control sample: sample controls among disease-free at t_1

Diseased:	$k_3(A+C)$	$=$	k_3A	$+$	k_3C
(cases)		$\sim$	a		$\sim c$
Non-diseased:	$k_4(B+D)$	$=$	k_4B	$+$	k_4D
(controls)		$\sim$	b		$\sim d$

Then: $\dfrac{a}{a+b} \sim \dfrac{k_3A}{k_3A+k_4B}$, $\dfrac{c}{c+d} \sim \dfrac{k_3C}{k_3C+k_4D}$

$\Rightarrow$ *We canNOT estimate relative risk*

BUT odds ratio, since $\dfrac{a \cdot d}{b \cdot c} \sim \dfrac{k_3A \cdot k_4D}{k_4B \cdot k_3C} = \dfrac{A \cdot D}{B \cdot C}$

Exposure odds ratio.

	cases	controls
Exposed	a	b
Non-exposed	c	d

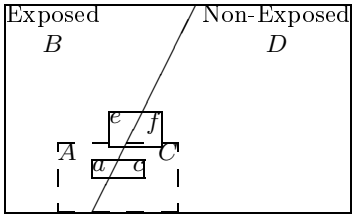
Odds for being exposed among cases $= a/c$, odds for being exposed among controls $= b/d$

$\Rightarrow$ exposure odds ratio $= \frac{a/c}{b/d} = \frac{ad}{bc}$, i.e. the exposure OR estimates the disease OR.

We can even do logistic regression (NB: intercept!).

When disease is "rare": $OR \approx RR \approx RR$, however epidemiologists don't like OR and they don't like the rare disease assumption.

Population: case-cohort (or -referent) sample



Alternative design: case-cohort, i.e. sample cases at t_1 and take a random sample at t_0.

$$\begin{array}{lrcccl} \textit{Cases:} & k_3(A+C) & = & k_3A & + & k_3C \quad \text{(as} \\ & & & \sim a & & \sim c \quad \text{before)} \end{array}$$

Sample from the whole population at t_0:

$$\begin{array}{rcccl} k(A+B+C+D) & = & k(A+B) & + & k(C+D) \\ & & \sim e & & \sim f \end{array} \quad \text{Then}$$

$$\frac{a/e}{c/f} \approx \frac{k_3 \cdot A/k \cdot (A+B)}{k_3 \cdot C/k \cdot (C+D)} = \frac{A/(A+B)}{C/(C+D)} = \textit{relative risk}$$

$\Rightarrow$ We *can estimate relative risk*

(using an "OR-type formula")

The problem is that the statistical analysis of the *RR*-estimate *gets complicated*. In "the usual 2 by 2 table":

	cases	"non-cases"
Exposed	a	e
Non-exposed	c	f

the columns are *not independent* since the "non-cases" (here: sample from the study base) may contain diseased individuals.

However, *SE* formulas exist and regression analysis is possible using software for logistic regression (Schouten et al., *Stat. in Med.*, 1993).

Incidence sampling of controls.

Sample *in* the interval from t_0 to t_1.

Rate: $\frac{\text{cases}}{\text{pyrs}}$

Rate ratio $\frac{A/Y_1}{C/Y_0}$, 1$\sim$ exposed, 0$\sim$ non-exposed.

In a case-control study we observe cases: $a = k_3A, c = k_3C$.

If controls are sampled proportionally to their pyrs-contribution, $b \sim rY_1, d \sim rY_0$ then the rate ratio can be estimated from the case-control data:

$$\frac{a/b}{c/d} \approx \frac{k_3A/rY_1}{k_3C/rY_0} = \frac{A/Y_1}{C/Y_0}.$$

Inference?

If *SE* is available, then stratified analysis can be carried out; regression?

Problems.

There are problems with:

- inference in case-cohort design
- inference for incidence sampling
- censoring
- delayed entry

These problem can be handled satisfactorily using survival analysis methods for the cohort.

Cox regression models for intensity of type l.

$$\lambda_{li}(t) = \lambda_{l0}(t)\exp(\beta_l^{\mathsf{T}} Z_i(t))$$

β_l estimated from Cox partial likelihood:

$$L(\beta_l) = \prod_{i=1}^{n}\prod_{t}\left(\frac{\exp(\beta_l^{\mathsf{T}} Z_i(t))}{\sum_{j\in R_l(t)}\exp(\beta_l^{\mathsf{T}} Z_j(t))}\right)^{\mathrm{d}N_{li}(t)}$$

$\Lambda_{l0}(t) = \int_0^t \lambda_{l0}(u)\mathrm{d}u$ estimated by the Nelson-Aalen type estimator

$$\widehat{\Lambda_{l0}}(t) = \int_0^t \Big(\sum_{j\in R_l(u)} \exp(\widehat{\beta_l}^{\mathsf{T}} Z_j(u))\Big)^{-1}\mathrm{d}N_{l\cdot}(u)$$

Large-sample properties derived using martingale methods (see e.g., Andersen, Borgan, Gill and Keiding, 1993, Theorems VII.2.1-3)

The Danish National Birth Cohort Study.

A cohort of 100000 pregnant woman and their children was established. (No. 100000 recruited Sept. 2002.)

- 4 Computer Assisted Telephone Interviews: 12 and 30 weeks of gestation, and at 6 and 18 months
- 3 blood samples: 6-8 and 26 weeks of gestation, and chord blood at birth

Thereby obtain "exposure register" to match Danish disease registers (cancer-, hospital discharge-) and investigate Barker's "programming hypotheses". (J. Olsen et al., 2001, *Scand. J. Pub. Health.*). Two short-term studies:

- Fever in early pregnancy and risk of fetal death
- Occupational exposure and risk of childhood leukemia

Fever in early pregnancy and risk of fetal death.

In animal studies: Hyperthermia may induce fetal death.

Here: study effect of fever in early (human) pregnancy on risk of fetal death

Data: 24,041 pregnant women recruited October 1997 to April 1999 to The Danish National Birth Cohort Study and interviewed (CATI).

Information on:

- fever incidents
- reproductive history
- smoking
- alcohol
- occupation
- ...

Fever in early pregnancy and risk of fetal death.

Outcome data from National Discharge Registry: 1168 fetal deaths

Andersen, Vastrup, Wohlfahrt, Andersen, Olsen and Melbye, *Lancet*, 2002:

Cox regression model with

- Time variable = gestational days (i.e., time since last menstrual period)
- Time of entry = time of consent
- Fever variables time-dependent, obtained in first interview
- Sub-analysis for women interviewed "prospectively" (here, time of entry = time of interview)

Results.

Exposure	Fetal deaths	Fetus-weeks	*RR* (95% c.i.)
No fever	986	545292	1
Fever wk. 1-16	147	103191	0.95 (0.80-1.13)
1. trim.	76	7064	0.92 (0.71-1.16)
2. trim.	54	55222	0.95 (0.71-1.27)
3. trim	17	40905	1.16 (0.69-1.97)

No effects of: time of fever, max. temp., no. of days with fever

All adjusted for: maternal age, parity, previous fetal deaths, occupation (in daycare), smoking, alcohol, coffee.

The Danish Adoption Register.

Register with information on 14427 children adopted away to unrelated parents between 1924 and 1947. Information on:

- ADoptee
- Adoptive Mother, Adoptive Father
- Biological Mother, Biological Father

That is: name, date of birth, address of adoptive parents, date of transfer, date of formal adoption, biological and adoptive siblings.

Aim: study relation between (early) cause-specific mortality among

- ADoptee and Biological parents
- ADoptee and Adoptive parents

and thereby evaluate genetic and environmental effects.

"Old" study.

1003 AD's born 1924-26 followed until 1982:

Sørensen, Nielsen, Andersen, Teasdale *NEJM* (1988).

Status 1982	AD	BF	BM	AF	AM
Alive in DK	765	114	367	64	163
Emigrated	75	32	27	4	8
Disappeared	1	4	2	1	0
Not followed	0	146	26	39	7
Dead	119	664	538	852	782
Total	960	960	960	960	960

"Old" study.

Cox regression model with lifetime of AD as outcome and information on lifetimes of parents coded as explanatory variables: Estimated hazard ratios (95% c.i.) for "at least 1 parent dead (from relevant cause) before age 70".

Cause	B/A	RR	c.i.
All	B	1.85	1.17-2.92
All	A	0.80	0.55-1.16
Natural	B	1.49	0.92-2.39
Natural	A	0.96	0.65-1.41
Infection	B	5.00	1.73-14.4
Infection	A	1.00	0.34-2.97
Vascular	B	1.92	0.78-4.73
Vascular	A	1.50	0.65-3.46
Cancer	B	0.87	0.26-2.88
Cancer	A	1.49	0.56-3.97

Later analyses: "frailty" models.

Data requirements in Cox model.

For all event times T_{li} we need $Z_j(T_{li})$ for all individuals, j, at risk for a type l event at T_{li} (i.e. $j \in R_l(T_{li})$).

- Childhood leukemia example: possible model $\lambda_i(age) = \lambda_0(age)\exp(\beta Z_i)$, where $Z_i = 1$ if is mother was exposed to a given chemical; need blood samples for 100000 women
- Adoption example, whole data set: possible model (cause l) $\lambda_{l,AD}(age) = \lambda_{l0}(age)\exp(\beta Z_{AD})$, where $Z_{AD} = 1$ if one of AD's adoptive parents died from cause l before age a_0; need to trace all adoptive parents; information before 1968 not computerized

$\Rightarrow$ SAMPLING of the cohort!

Two types of sampling design.

- (1): Nested case-control sampling: at *each* type l failure time T_{li}, select a simple random sample $\widetilde{R_l}(T_{li})$ of size m with $i \in \widetilde{R_l}(T_{li})$ and estimate β_l from the (partial) likelihood:

$$L_{NCC}(\beta_l) = \prod_{i=1}^{n}\prod_{t}\Big(\frac{\exp(\beta_l^{\mathsf{T}} Z_i(t))}{\sum_{j\in\widetilde{R_l}(t)}\exp(\beta_l^{\mathsf{T}} Z_j(t))}\Big)^{\mathrm{d}N_{li}(t)}$$

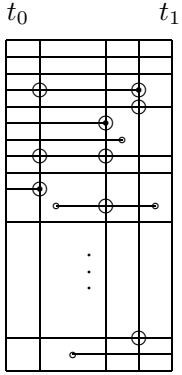
- (2): Case-cohort sampling: at time 0 select a random sample $\mathcal{S}$ (the *"sub-cohort"*) of size M and estimate β_l from the ("pseudo") likelihood:

$$L_{CC}(\beta_l) = \prod_{i=1}^{n}\prod_{t}\Big(\frac{\exp(\beta_l^{\mathsf{T}} Z_i(t))}{\sum_{j\in\mathcal{S}_l(t)}\exp(\beta_l^{\mathsf{T}} Z_j(t))}\Big)^{\mathrm{d}N_{li}(t)}$$

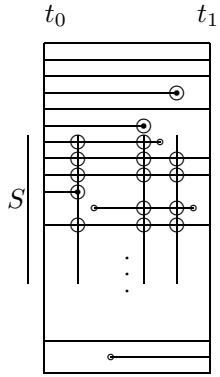
  where $\mathcal{S}_l(t) = (\mathcal{S}\cup\{i\})\cap R_l(t)$

Must be able to obtain covariate information for sampled persons.

Nested case-control study



Case cohort study



Nested case-control study

Estimation of rate ratio θ:

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Case-control set}}\theta}\right)$$

Compare matched case-control study.

Case cohort study.

"Pseudo-likelihood"

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Comparison group}}\theta}\right)$$

The comparison group is the case plus what is left of S at the present failure time.

Computations

Because of the similarity with the Cox partial likelihood, standard software may be used for parts of the analyses:

- `SAS PROC PHREG`, but wrong SE's in case-cohort study. Add-on macros exist. Correct results for NCC study
- `STATA, EPICURE`

Notes on designs.

- Nested case-control sampling:
  - other sampling methods than simple random may require different weighting of the terms
  - a new sample is selected at each failure time
  - only covariates for the "cases" and for the sampled "controls" are needed
- Case-cohort sampling:
  - the same sub-cohort is used at each failure time
  - in particular, the same sub-cohort is used for all event types
  - only covariates for the "cases" and for the sub-cohort are needed

Score equations for β, full cohort.

(Andersen, Borgan, Gill and Keiding, 1993):

$$U_{FC}(\beta) = \sum_{l=1}^{k} \int_0^\infty \sum_{i=1}^{n} (Z_i(t) - E_l(\beta_l, t)) \mathrm{d}N_{li}(t)$$

where

$$E_l(\beta_l, t) = \frac{\sum_{i \in R_l(t)} \exp(\beta_l^\mathsf{T} Z_i(t)) Z_i(t)}{\sum_{i \in R_l(t)} \exp(\beta_l^\mathsf{T} Z_i(t))}.$$

For $\beta = \beta_0$ (true value):

$$U_{FC}(\beta_0) = \sum_{l=1}^{k} \int_0^\infty \sum_{i=1}^{n} (Z_i(t) - E_l(\beta_{l0}, t)) \mathrm{d}M_{li}(t)$$

is a *martingale*.

Score equations for β, nested case-control study.

(Borgan, Goldstein and Langholz, *Ann. Statist.*, 1995):

$$U_{NCC}(\beta) = \sum_{l=1}^{k} \int_0^\infty \sum_{r \in \mathcal{P}} \sum_{i \in r} (Z_i(t) - E_{lr}(\beta_l, t)) \mathrm{d}N_{li,r}(t)$$

where $\mathcal{P}$ is the power set of $\{1, \ldots, n\}$ and

$$E_{lr}(\beta_l, t) = \frac{\sum_{i \in r \cap R_l(t)} \exp(\beta_l^\mathsf{T} Z_i(t)) Z_i(t)}{\sum_{i \in r \cap R_l(t)} \exp(\beta_l^\mathsf{T} Z_i(t))}.$$

For $\beta = \beta_0$ (true value):

$$U_{NCC}(\beta_0) = \sum_{l=1}^{k} \int_0^\infty \sum_{r \in \mathcal{P}} \sum_{i \in r} (Z_i(t) - E_l(\beta_{l0}, t)) \mathrm{d}M_{li,r}(t)$$

is a *martingale*.

Score equations for β, case-cohort study.

(Self and Prentice, *Ann. Statist*, 1988; Sørensen and Andersen, *Biometrika*, 2000):

$$U_{CC}(\beta) = \sum_{l=1}^{k} \int_0^\infty \sum_{i=1}^{n} (Z_i(t) - E_l^{\mathcal{S}}(\beta_l, t)) \mathrm{d}N_{li}(t)$$

where

$$E_l^{\mathcal{S}}(\beta_l, t) = \frac{\sum_{i \in \mathcal{S}_l(t)} \exp(\beta_l^\mathsf{T} Z_i(t)) Z_i(t)}{\sum_{i \in \mathcal{S}_l(t)} \exp(\beta_l^\mathsf{T} Z_i(t))}.$$

For $\beta = \beta_0$ (true value):

$$U_{CC}(\beta_0) \approx U_{FC}(\beta_0) + \sum_{i=1}^{n} (1 - \frac{n}{M} V_i) \sum_{l=1}^{k} X_{li}(\beta_{l0})$$

($V_i = I(i \in \mathcal{S})$) is a *martingale plus* a term which creates a correlation between score contributions.

Asymptotic results for β-estimates.

<u>Full cohort:</u>

$$\sqrt{n}(\widehat{\beta} - \beta_0) \sim \mathcal{N}(0, \Sigma^{-1})$$

(Σ^{-1} block diagonal if no β_l components are assumed identical.)

<u>Nested case-control:</u>

$$\sqrt{n}(\widetilde{\beta} - \beta_0) \sim \mathcal{N}(0, \widetilde{\Sigma}^{-1})$$

($\widetilde{\Sigma}^{-1}$ block diagonal if no β_l components are assumed identical.)

<u>Case-cohort:</u>

$$\sqrt{n}(\widehat{\beta_{\mathcal{S}}} - \beta_0) \sim \mathcal{N}(0, \Sigma_{\mathcal{S}}^{-1} + \frac{1-\pi}{\pi} \Sigma_{\mathcal{S}}^{-1} \Delta \Sigma_{\mathcal{S}}^{-1})$$

($\Sigma_{\mathcal{S}}^{-1}$ block diagonal if no β_l components are assumed identical but $\mathcal{S}$ creates a correlation between different β_l-estimates, $\pi = \lim M/n$.)

In all 3 cases: Σ estimated consistently by $-\frac{1}{n}$(obs. inf.).

Estimation of baseline hazards.

- FC:

$$\widehat{\Lambda_{l0}}(t) = \int_0^t (\sum_{j \in R_l(u)} \exp(\widehat{\beta_l}^{\mathsf{T}} Z_j(u)))^{-1} \mathrm{d}N_{l\cdot}(u)$$

- NCC:

$$\widetilde{\Lambda_{l0}}(t) = \int_0^t (\frac{Y_l(u)}{m} \sum_{j \in \widetilde{R_l}(u)} \exp(\widetilde{\beta_l}^{\mathsf{T}} Z_j(u)))^{-1} \times \mathrm{d}N_{l\cdot}(u)$$

- CC:

$$\widehat{\Lambda_{l0,\mathcal{S}}}(t) = \int_0^t (\frac{Y_l(u)}{M} \sum_{j \in \mathcal{S}_l(u)} \exp(\widehat{\beta_{l,\mathcal{S}}}^{\mathsf{T}} Z_j(u)))^{-1} \times \mathrm{d}N_{l\cdot}(u)$$

Asymptotic results available.

Other nested case-control sampling designs.

Matching.

Example: Lung cancer incidence, smoking possible confounder.

Many smoking cases, perhaps relatively few smoking controls ⇒ random sampling of $m-1$ controls will give few controls per smoking case and more controls per non-smoking case.

Matching on smoking may be efficient.

- Availability of data?
- Inability to estimate effect of smoking

$$\theta_{case} = \exp(\beta_1 \cdot \text{exposure}_{case} + \beta_2 \cdot \text{smoke}_{case})$$

$$\theta_{control} = \exp(\beta_1 \cdot \text{exposure}_{control} + \beta_2 \cdot \text{smoke}_{control})$$

where exposure is 0 or 1 and and where the value of smoke is the same for case and controls, i.e. $\exp(\beta_2)$ cancels out in log partial likelihood:

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Case-control set}} \theta}\right).$$

Example of matched, nested c-c study.

Ylitalo, Sørensen, Josefson, Magnusson, Andersen, Pontén, Adami, Gyllensten, Melbye, *Lancet*, 2000.

- 146889 women screened between 1969 and 1995 in Uppsala county cervix cancer screening program: (732887 smears taken)
- 478 cases of cervix cancer in situ (CIS) identified through the Swedish cancer register
- 5 (potential) controls selected per case from the calendar time risk set, matched on time of entry into cohort (= time of first smear) and on age. NO matching on number of smears.
- 1 of the 5 controls randomly selected for inclusion. If the selected control had only one smear then a second control was selected. (→ 608 controls.)

- Exposure, HPV-16 viral load, ascertained from the 2081/1754 available smears.

Why do a nested case-control study?

- To avoid making cytological analyses of *many* smears.

Why match

- on age? Standard, age is a confounder.
- on time of first smear? To make "exposure quality" similar for cases and controls.

Results: Dose-response effect of viral load on risk of CIS.

In this study (and in many other nested c-c studies): possible to estimate absolute risk.

Counter-matching.

To do the matched study, the confounder must be known for every one.

Suppose instead that exposure is known for every one but the confounder may be costly to obtain. Then:

- Matching on exposure is possible, but disastrous!
- Information on exposure may be used when selecting controls

E.g. in a given risk set: $N_1 = 10$ exposed, $N_0 = 100$ non-exposed. Simple random sampling then leads to uneven (and inefficient) exposure distribution in sampled case-control set. Instead, let the case-control set consist of $m = 5 + 1 = n_0 + n_1 = 3 + 3$ non-exposed/exposed individuals, i.e. if case is exposed then sample 2 exposed + 3 non-exposed controls and if case is non-exposed then sample 3 exposed + 2 non-exposed controls.

The confounder is ascertained for the sampled case-control set.

In the log-likelihood: Members of the case-control sets must be *weighted differently*:

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Case-control set}} w \cdot \theta}\right).$$

Here: $w = \frac{N_1}{n_1} = 10/3$ for exposed

$w = \frac{N_0}{n_0} = 100/3$ for non-exposed

"Counter-matching": $m - 1 = 1$, case and control must have different exposure status.

Counter-matching on surrogate exposure is also possible.

Analysis: computer program must be able to deal with different weights: `"OFFSET"` in `SAS PROC PHREG`.

"New" adoption case-cohort study.

All AD's (12301) followed until 1993, also siblings and half-siblings (both biologic and adoptive).

It is VERY time consuming to find all those individuals in non-computerised records prior to 1968.

Therefore, *case-cohort study*:

- all 1403 dead AD's traced (including entire "family")
- random sub-cohort of 1683 chosen and traced (1480 new)
- analyses similar to the "old" study performed on the case cohort sample

Cox regression model with lifetime of AD as outcome and information on lifetimes of parents coded as explanatory variables: Estimated hazard ratios (95% c.i.) for "at least 1 parent dead (from relevant cause) before age 70". (Petersen, Andersen & Sørensen, *Gen. Epi.*, 2005.)

Cause	B/A	*RR*	c.i.
All	B	1.27	1.08-1.50
All	A	0.92	0.80-1.07
Natural	B	1.24	1.01-1.52
Natural	A	0.88	0.74-1.05
Infection	B	1.35	0.80-2.27
Infection	A	0.97	0.62-1.51
Vascular	B	1.51	1.05-2.17
Vascular	A	0.84	0.57-1.23
Cancer	B	1.03	0.72-1.49
Cancer	A	1.07	0.77-1.48

References.

- Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag.
- Borgan, Ø., Goldstein, L., Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Stat.* **23**, 1749-1778.
- Self, S.G., Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Stat.* **16**, 64-81.