PhD course 2023

Epidemiological methods in medical research

Matched case-control and cohort studies. Cohort sampling

Clayton & Hills, Ch. 18.3-4, 19, 23.3, 29, 33

9 March 2023

Per Kragh Andersen

Today's topics

- Frequency matched case controls studies
- Individually matched case control studies
- Matched cohort studies
- Cohort sampling
- Case cohort study with binary data

Logistic regression of BCG study

Table 23.2. Cases of leprosy and controls by age and BCG scar

	Leprosy cases		Healthy population		Odds ratio
BCG	−	+	−	+	estimate
Age 0–4	1	1	7593	11719	0.65
Age 5–9	11	14	7143	10184	0.89
Age 10–14	28	22	5611	7561	0.58
Age 15–19	16	28	2208	8117	0.48
Age 20–24	20	19	2438	5588	0.41
Age 25–29	36	11	4356	1625	0.82
Age 30–34	47	6	5245	1234	0.54
Total	159	101	34594	46028	0.48

ω = odds that a person (in the study) is a case

estimated by case/control ratios:

Table 23.3. Case/control ratio $(\times 10^3)$ by age and BCG scar

	BCG scar		
Age	Absent	Present	OR
0-4	0.13	0.08	0.65
5-9	1.54	1.37	0.89
10-14	4.99	2.91	0.58
15-19	7.25	3.45	0.48
20-24	8.20	3.40	0.41
25-29	8.26	6.77	0.82
30-34	8.96	4.86	0.58

$$\log(\widehat{OR}) = -0.547 \quad (SD = 0.141) \Rightarrow \widehat{OR} = 0.579 \quad (0.439, 0.763).$$

Fewer controls per case

Table 23.6. A simulated study with 1000 controls

		Cases		Controls	
	BCG	−	+	−	+
Age	0–4	1	1	101	137
	5–9	11	14	91	115
	10–14	28	22	82	101
	15–19	16	28	28	87
	20–24	20	19	25	69
	25–29	36	11	63	21
	30–34	47	6	56	24

Here, the number of controls per case varies considerably with age.

$\log(\widehat{OR}) = -0.548 \quad (SD = 0.161) \Rightarrow \widehat{OR} = 0.578 \quad (0.422, 0.792).$

Efficiency of case-control study

In the BCG study there were several controls per case.

If we have a case-control study with m controls per case then the ratio between the SD for the odds ratio and the SD for a corresponding rate ratio from a cohort study based on the same cases is

$$\sqrt{1+\frac{1}{m}}.$$

For different values of m this is:

m	1	2	3	4	5	10
Ratio	1.41	1.22	1.15	1.12	1.10	1.05

The ratio does not change much for $m \geq 4-5$.

Exercise

Calculate the number (m) of controls per case for the 'big' and the 'small' study and investigate whether the expression $\sqrt{1 + 1/m}$ is compatible with the ratio between the observed SD's (0.141 and 0.161, respectively).

Solution

The values of m are, respectively,

$$m_{\text{big}} = 80622/260 = 310.1 \text{ and } m_{\text{small}} = 1000/260 = 3.9.$$

The ratios are

$$\sqrt{1 + 1/m_{\text{small}}}/\sqrt{1 + 1/m_{\text{big}}} = 1.12 \text{ and } 0.161/0.141 = 1.14,$$

quite close!

Age matching (group matching)

Fewer controls may be used more efficiently by *matching*.

Table 23.6. A simulated four-to-one group-matched study

		Cases		Controls	
	BCG	−	+	−	+
Age	0–4	1	1	3	5
	5–9	11	14	48	52
	10–14	28	22	67	133
	15–19	16	28	46	130
	20–24	20	19	50	106
	25–29	36	11	126	62
	30–34	47	6	174	38

$\log(\widehat{OR}) = -0.572 \quad (SD = 0.155) \Rightarrow \widehat{OR} = 0.564 \quad (0.417, 0.764),$

i.e. the SD is (slightly) smaller than without matching.

Intercept ('Corner')?

We know that if π = risk of failure in the population

then $\omega = K \times \dfrac{\pi}{1-\pi}$

where

$$K = \frac{\text{Prob(a "failure" is included as case)}}{\text{Prob(a "survivor" is included as control)}}$$

Thereby,

$$\log \frac{\pi}{1-\pi} = \text{Corner} + \text{Age} + \text{BCG}$$

$$\Longrightarrow \log(\omega) = \log(K) + \text{Corner} + \text{Age} + \text{BCG}$$

same odds ratios when K does not depend on Age and BCG.

But the estimated Corner parameter cannot be interpreted.

Case/control ratios

In the matched study, the factor K does depend on Age!

	Absent	Present
0-4	0.33	0.20
5-9	0.23	0.27
10-14	0.42	0.17
15-19	0.35	0.22
20-24	0.40	0.18
25-29	0.29	0.18
30-34	0.27	0.16

Matched and unmatched analysis

	Unmatched		Age-matched	
Parameter	Estimate	SD	Estimate	SD
Corner	-8.880	0.7093	-1.0670	0.800
Age(1)	2.624	0.7340	-0.0421	0.827
Age(2)	3.583	0.7203	0.0119	0.812
Age(3)	3.824	0.7228	0.0713	0.814
Age(4)	3.900	0.7244	0.0244	0.816
Age(5)	4.156	0.7224	-0.1628	0.814
Age(6)	4.158	0.7213	-0.2380	0.813
BCG(1)	-0.547	0.1409	-0.5721	0.155

In the matched study we should correct for Age though the Age estimates in the model:

$$\log(\omega) = \text{Corner} + \text{Age} + \text{BCG}$$

cannot be interpreted.

	Cases		Controls		Odds
Stratum	+	−	+	−	ratio
1	89	11	80	20	2.0
2	67	33	50	50	2.0
3	33	67	20	80	2.0
Total	189	111	150	150	1.7

Table 18.4. Bias due to ignoring matching

Individual matching in case-control studies

The BCG study provided an example of *frequency matching*, i.e. when selecting controls it was assured that cases and controls have the same distribution of the matching variable (age).

However, a given case did not have his or her 'own private' control(s).

That is achieved when using *individual matching*, that is when matching is based on a variable like neighborhood, familial relation or the like (typically variables with *many* possible values), i.e., not easily quantifiable in a regression model.

Matching on *time* ('risk-set sampling' of controls, 'nested case-control study', 'incidence density sampling') is also a type of individual matching (more later).

Analysis of matched pairs

In the simplest case with a single binary exposure and one control individually matched to each case, data can be summarized as a two-by-two table of *pairs*:

History of case	History of control: Positive	History of control: Negative
Positive	$a = 26$	$b = 15$
Negative	$c = 7$	$d = 37$

Table 19.1. Tonsillectomy history in 85 matched pairs of Hodgkin cases and matched controls. Matched controls are same-sex siblings within a given age range. Originally: 174 cases and 472 controls.

Analysis of matched pairs

Each pair gives rise to a *stratum* and the Mantel-Haenszel methods apply.

	Exp	Unexp	Exp	Unexp	Exp	Unexp	Exp	Unexp
Case	1	0	1	0	0	1	0	1
Control	1	0	0	1	1	0	0	1

Only *discordant strata*, i.e. those where case and control have different exposure status contribute to estimator and test.

Exercise 19.1 (p.187).

How many strata of each of the four types are there?

Analysis of matched pairs

The Mantel-Haenszel estimate for the odds ratio for exposure (tonsillectomy) is simply the ratio

$$b/c = 15/7 = 2.14$$

between the numbers of the two different types of *discordant pairs*.

The SD of the corresponding log(odds ratio) is

$$\sqrt{\frac{1}{b} + \frac{1}{c}},$$

here 0.4577, leading to a 95% confidence interval for OR from 0.874 5.266.

The Mantel-Haenszel test reduces to *McNemar's* test:

$$\frac{(b-c)^2}{b+c} = \frac{64}{22} = 2.91, \quad P = 0.09.$$

Ignoring matching

The incorrect analysis corresponding to ignoring matching amounts to setting up the following 2 by 2 table of case-control status versus exposure and estimate the odds ratio:

History	Positive	Negative
Case	41	44
Control	33	52

Exercise 19.4 (p.188).

Argue why the resulting 2 by 2 table has these entries and show that the resulting odds ratio now becomes 1.47 ($= \frac{41 \cdot 52}{44 \cdot 33}$).

Note that this is closer to 1 (here: smaller) than the M-H estimate (2.14).

Individual matching: logistic regression model

If there are n matched case-control sets $i = 1, ..., n$ (e.g., each consisting of 1 case and 1 matched control) then the model for subject j in set i is:

$$\log(\text{odds}_{ij}) = \text{CORNER}_i + \text{EXPOSURE}_{ij},$$

i.e., one CORNER parameter for each set (i) containing the effect of the matching variables.

It turns out that the standard likelihood method does not work for this logistic regression model because of the many (n) 'nuisance parameters' (CORNER_i) that cannot be estimated. Instead, so-called *conditional likelihood* ('conditional logistic regression') may be used.

Individual matching: model

In the simplest case (1 control per case, 1 binary exposure) this actually gives exactly the same as the M-H analysis but using conditional logistic regression it is simple to

- include more than 1 control per case
- adjust for confounders that vary *within* matched sets

Note that variables that are constant within matched sets cannot be included (their effect will be 'absorbed' into the CORNERs).

Some times, 'individual' matching is performed on factors like sex and age. This is not really 'individual', and the structure with pairs need not be maintained in the analysis (but sex and age should still be (correctly!) accounted for).

Code for conditional logistic regression

In SAS, PROC LOGISTIC may be used via a STRATA command:

```
PROC LOGISTIC DATA=HODGKIN;
CLASS TONSIL/PARAM=GLM;
MODEL CASE=TONSIL;
STRATA SET;
RUN;
```

PROC PHREG may also be used for analyzing individually matched case-control studies but the code is less transparent.

Similarly in R: clogistic function in Epi package or coxph (or clogit) functions in survival package.

Matching: pros and cons

- Classical method to adjust for confounding in case-control studies
- Inability to estimate effect of match variables
- May increase efficiency for exposure effect
- Matching variables must be accounted for in analysis
- Risk of 'over-matching'
- Frequency matching vs. individual matching

Matched cohort studies

In a cohort study, the *exposed group* is frequently 'obvious':

- persons with a certain diagnosis
- persons in a certain occupation
- persons taking a certain drug

But how to select an unexposed group for comparison?

1. at random?
2. matched to the exposed on factors like sex and age?
3. matched to the exposed on more individual factors (neighborhood, familal relationship)?
4. matched to the exposed on *propensity score*?

Matched cohort studies

Similar considerations as for matched case-control studies apply

1. *at random?*
2. *matched to the exposed on factors like sex and age?*

   In these cases, adjustment for confounders will typically be done by including them in a regression model together with exposure (note that these are estimable).
3. *matched to the exposed on more individual factors (neighborhood, familal relationship)?*

   Here, a *stratified* Cox model respecting the individual matching may be used:

   $$\lambda_{ij}(t) = \lambda_{0i}(t) \exp(\beta_1 x_{ij1} + ... + \beta_p x_{ijp})$$

   where each matched set ('stratum', i) has its own baseline hazard and individuals (j) within matched sets have covariates x_{ij}.

Propensity score

If Z is a binary exposure and X the confounders then the *propensity score* is the probability of being exposed:

$$e(X) = P(Z = 1 \mid X).$$

Propensity score was introduced in an important paper by Rosenbaum and Rubin (1983) and has been used increasingly in recent years - often for doing *causal inference*, i.e. trying to analyze observational data to obtain answers that would otherwise require a randomized study.

The propensity score has an important *balancing property*:

The confounders X included in $e(X)$ have the same distribution for exposed ($Z = 1$) and unexposed ($Z = 0$) subjects with the same value of $e(X)$ ('X and Z are independent given $e(X)$').

How to analyze propensity-matched cohort data?

Typically, the individual matching is not kept in the analysis.

Rather, for studies with a follow-up time that varies among individuals, a *marginal* Cox model is used:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta Z_i)$$

because, according to the balancing property of the propensity score, exposed ($Z = 1$) and unexposed ($Z = 0$) individuals have the same distribution of X.

Cohort sampling, example 1: The Danish Adoption Register

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

- Adoptee (AD)
- Adoptive Mother, Adoptive Father (AM, AF)
- Biological Mother, Biological Father (BM, BF)

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 relatives
- ADoptee and Adoptive relatives

and thereby evaluate genetic and environmental effects.

"Old" study: parents

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". Time=age.

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

Cohort sampling, example 2: HPV and cervix cancer in situ

Josefson, Magnusson, Ylitalo, Sørensen, Qwarforth-Tubbin, Andersen, Melbye, Adami, Gyllensten *Lancet*, 2000, **355**, 2189-93.

- 146889 women screened between 1969 and 1995 in Uppsala county cervix cancer screening program: (732887 smears taken)
- 478 cases of cervix cancer in situ (CIN) identified through the Swedish cancer register
- Exposure, HPV-16 viral load, ascertained from smears.

Possible Cox models for examples

- Adoption example, whole data set: $\lambda_i(t) = \lambda_0(t)\exp(\beta x_i)$, where $x_i = 1$, if one of the adoptive parents for adoptee i died before age a_0 (t =age)
- HPV/cervix cancer example: $\lambda_i(t) = \lambda_0(t)\exp(\beta x_i)$, where $x_i =$ subject i's HPV viral load at time of first screening

Estimation in Cox model: maximize Cox's partial log-likelihood:

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

where $\theta_i = \exp(\beta x_i)$.

At all event times t_i we need the covariates x_j for all individuals, j, at risk for an event at t_i.

Cohort sampling

In Example 1 we need to trace all adoptive parents. However, information before 1968 is not computerized. Similarly for biological parents

In Example 2 we need data from all first smears.

$\Rightarrow$ *sampling* of the cohort!

Two types of sampling design

- (1): Nested case-control sampling: at *each* event time t_i, select a (simple random) sample (of size m) containing i and estimate β from the partial log-likelihood:

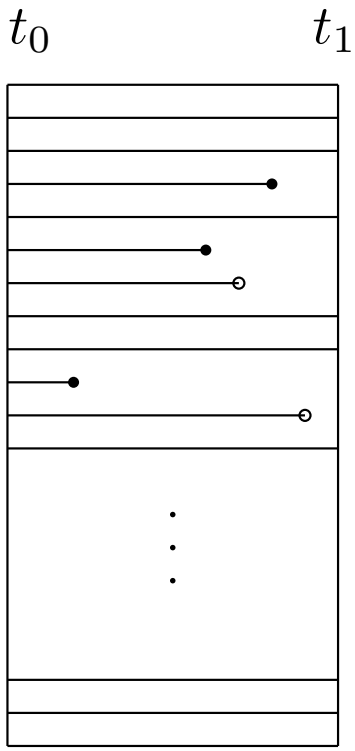
$$\sum_{\text{failures}} \log\left(\frac{\theta_{\text{(for case)}}}{\sum_{\text{Case-control set}}\theta}\right)$$

- (2): Case-cohort sampling: at time 0 select a random sample $\mathcal{S}$ (the *"sub-cohort"*) (with some sampling fraction q) and estimate β from the "pseudo" log-likelihood:

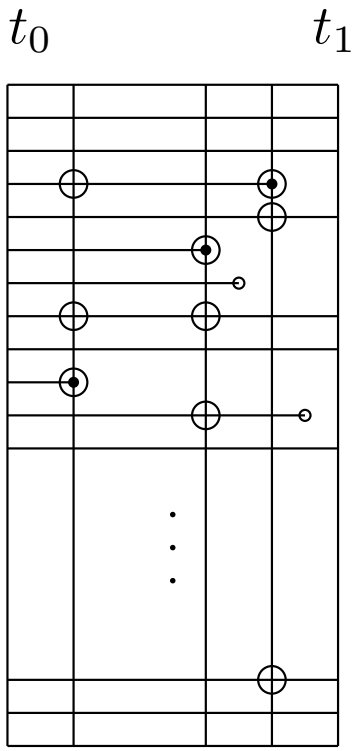
$$\sum_{\text{failures}} \log\left(\frac{\theta_{\text{(for case)}}}{\sum_{\text{Comparison group}}\theta}\right)$$

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

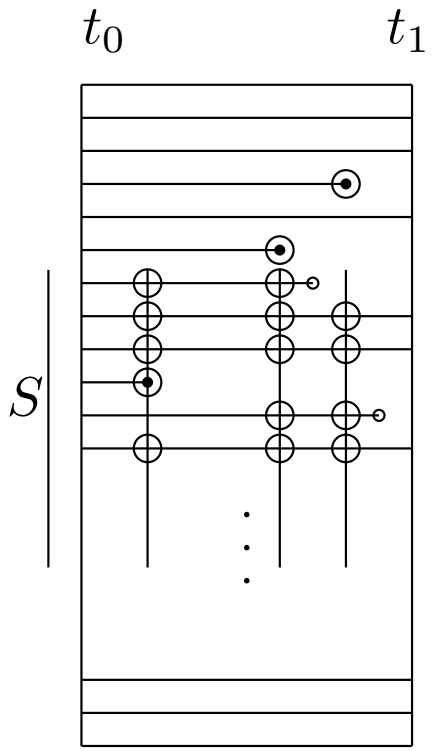
Cohort with incomplete follow-up



Nested case-control study



Case cohort study



Notes on designs

- Nested case-control sampling: other sampling methods than simple random may require different weighting of the terms
- Nested case-control sampling: a new sample is selected at each failure time $\Rightarrow$ if there are several case series then each series requires its own sets of controls
- Case-cohort sampling: the same sub-cohort is used at each failure time
- Case-cohort sampling: in particular, the same sub-cohort may be used for several case series
- Both designs: only covariates for the "cases" and for the sampled controls are needed
- Both designs: "relatively little" statistical precision is lost
- Both designs: covariate information of "similar quality" must be obtainable for all subjects no matter case/controls status
- Both designs: absolute risks may be estimated

Example: HPV and cervix cancer in situ

- 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. ($\rightarrow$ 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.
- Why match on time of first smear? To make "exposure quality" similar for cases and controls.

Results (using first smear)

Josefsson et al., *Lancet*, 2000, **355**, 2189-93.

Viral load	Cases/controls	$\exp(\beta)$
HPV 16 negative	354/578	1
Below 20 percentile	16/15	1.9 (0.8-4.2)
20-40 percentile	23/7	7.2 (2.7-19.1)
40-60 percentile	28/3	22.8 (5.5-95.0)
60-80 percentile	27/4	18.9 (5.5-64.9)
Above 80 percentile	30/1	59.0 (7.5-462.2)
Total	478/608	

Dose-response effect of viral load on rate of CIN.

In an accompanying paper (Ylitalo et al., *Lancet*, 2000, 2194-98), estimation of *absolute risk* of CIN was illustrated.

"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-computerized 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.) Time=age.

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

Individually matched vs. nested case-control

The likelihood (Cox's likelihood) for the nested case-control study is mathematically the same as the conditional likelihood for a $1:(m-1)$ individually matched case-control design.

This means that the same software packages can be used for both.

However, the results from a nested case-control study should be interpreted as *rate ratios* and those from conditional logistic regression as *odds ratios*.

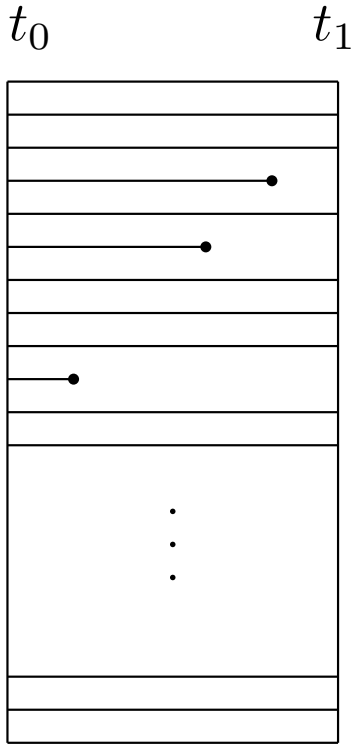
This is because *time* is explicitly involved in nested case-control studies (sampling from risk sets).

Computations

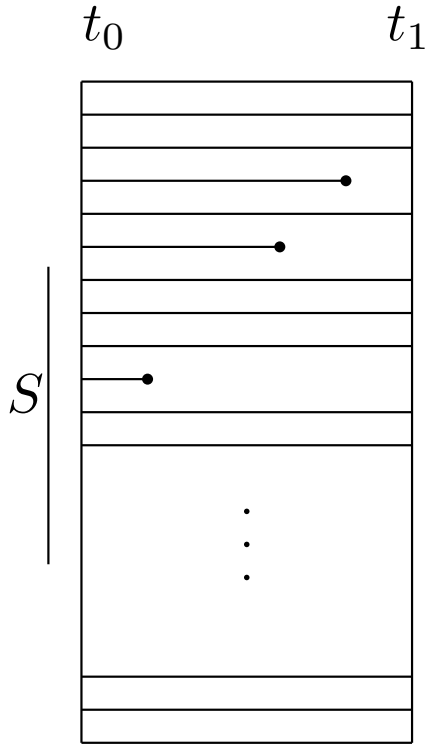
For the nested case-control study, standard software may be used. For the case-cohort study standard software may be used for parts of the analyses:

- Nested case-control study, `SAS: PROC PHREG, PROC LOGISTIC`, similarly in `R`: use `clogistic`, `coxph, clogit`
- Case-cohort study, `SAS`: construct weighted data set (e.g., Langholz and Jiao, (2007, *Comp. Stat. and Data Anal.*) and use `PHREG`; in `R: Epi` and `survival` packages may be adapted – also `cchs` package

Cohort with complete follow-up (no censoring)



Case cohort study for binary data (no censoring)



Analysis of binary case-cohort data

Include all cases ($Y_i = 1$) at t_1 and a random sample of the full cohort at t_0.

Let q be the corresponding sampling fraction.

This is the case-cohort design and the random sample is the sub-cohort.

Consider the *relative risk* model:

$$\log(P(Y_i = 1)) = c + d_1 x_{i1} + \cdots + d_k x_{ik},$$

i.e., the d-coefficients are log(relative risks).

These relative risk parameters may be estimated using *logistic regression* of an expanded data set, as follows.

Define $S_i = 1$ if subject i has been selected to the sub-cohort, $S_i = 0$ otherwise, and create a new expanded data set with:

- $D_i = 1$ if $Y_i = 1, S_i = 0$ (cases outside sub-cohort)
- $D_i = 0$ if $S_i = 1$ (all sub-cohort members)
- and add records with $D_i = 1$ if $Y_i = 1, S_i = 1$ (i.e., records for cases in sub-cohort are 'duplicated')

Then, in the expanded data set,

$$P(D_i = 1 \mid x_i) = \frac{P(Y_i = 1 \mid x_i)}{q + P(Y_i = 1 \mid x_i)}$$

$$= \frac{\exp(c - \log(q) + d_1 x_{i1} + \cdots + d_k x_{ik})}{1 + \exp(c - \log(q) + d_1 x_{i1} + \cdots + d_k x_{ik})}.$$

Use robust SD's (i.e., keep track of duplicated records).

Note that the intercept is 'contaminated' by the sampling fraction, q.