Nested case-control studies: Should one break the matching?

Ørnulf Borgan

Department of Mathematics

University of Oslo

Based on ongoing work with Ruth Keogh

Outline:

- Cohort sampling designs and "classical" inference for one endpoint
  - Matched: nested case-control
  - Unmatched: case-cohort
- Breaking the matching in nested case-control studies
- Two endpoints
- Simulations I
- Using the full cohort
  - ML-estimation
  - Multiple imputation
- Simulations II
- Concluding remarks

Cohort data and model (one endpoint)

Observe events (e.g. deaths from or onset of a disease) in a cohort C of n individuals

For each individual i we have two vectors of covariates:

- Covariates of main interest: $\mathbf{x}_i = (x_{i1}, \ldots x_{ip})^T$
- Confounding variables: $\mathbf{z}_i = (z_{i1}, \ldots z_{iq})^T$

The hazard for the time of an event for the i-th individual is assumed to follow a Cox regression model:

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{\beta}^T \mathbf{x}_i + \boldsymbol{\gamma}^T \mathbf{z}_i\right)$$

Note that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ describes the effects of the covariates of main interest

Due to censoring, we do not observe the event times

For each individual $i \in C$ we only observe (T_i, D_i) where

- T_i is the minimum of the event time and a censoring time
- $D_i = 1$ if T_i equals the event time; $D_i = 0$ otherwise

Introduce:

- the risk set $R(t) = \{i \mid T_i \geq t\}$ and $R_i = R(T_i)$
- the set of all cases $E = \{i \mid D_i = 1\}$

The regression coefficients are estimated by maximizing Cox's partial likelihood

$$L_{\text{co}}(\boldsymbol{\beta}) = \prod_{i \in E} \frac{\exp\left(\boldsymbol{\beta}^T \mathbf{x}_i + \boldsymbol{\gamma}^T \mathbf{z}_i\right)}{\sum_{j \in R_i} \exp\left(\boldsymbol{\beta}^T \mathbf{x}_j + \boldsymbol{\gamma}^T \mathbf{z}_j\right)}$$

Cohort sampling

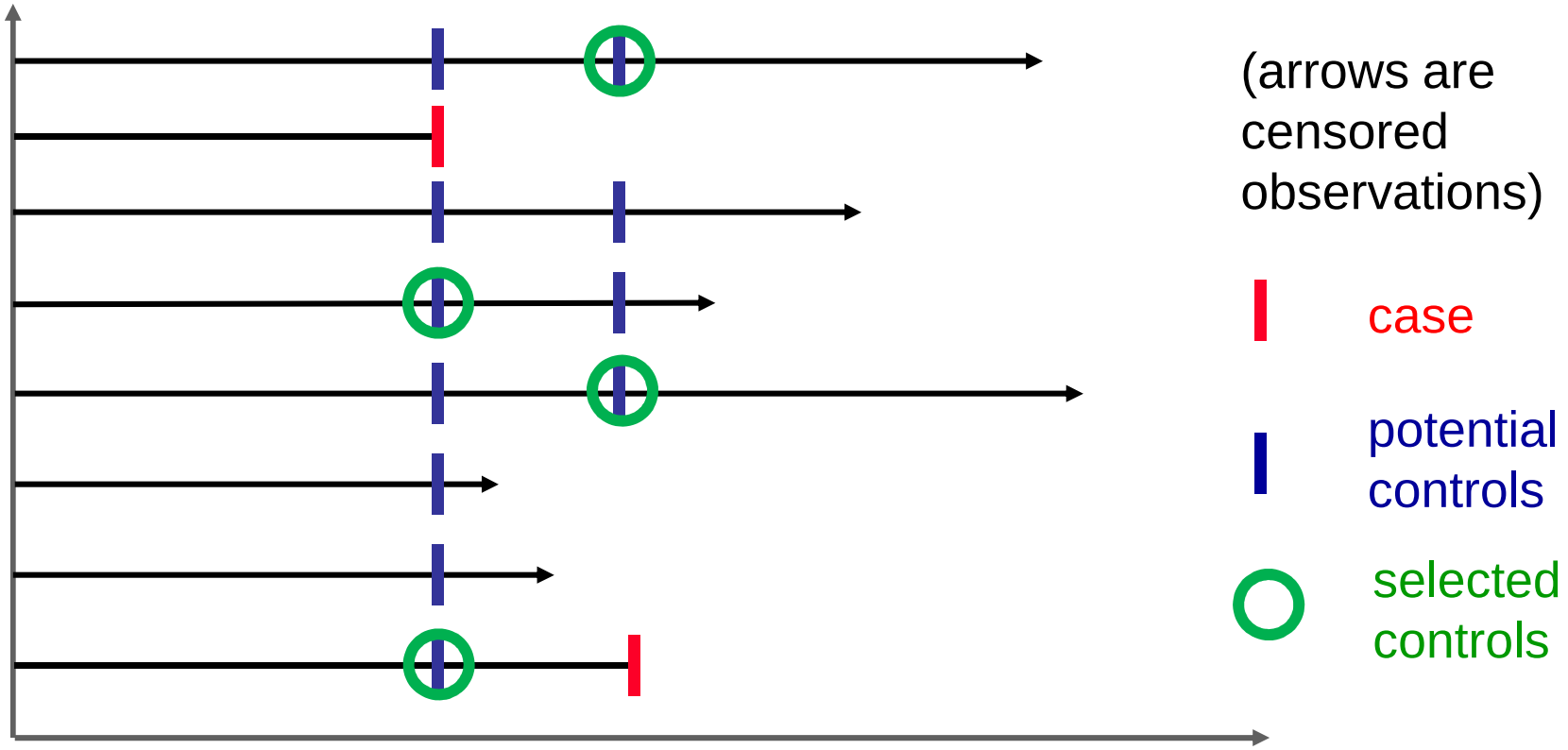
- Cohort studies need information on covariates for all individuals at risk
- Expensive to collect and check covariate information for all individuals in large cohorts
- In biomarker studies it will also imply a waste of valuable biological material
- Using a cohort sampling design one only needs to 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

Nested case-control studies

Select m –1 controls among those at risk when a case occurs, i.e. match on study time (and on confounding variables)

We will assume that controls are selected by simple random sampling (Thomas 1977), but note that stratified sampling and other sampling schemes also apply (Borgan et al 1995)

Illustration for m = 3 (with no additional matching):



We will assume that the confounding variables $\mathbf{z}_i$ can only take a finite number of different values $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k)}$ and are observed in the full cohort

Then the cohort may be divided in k strata according the values of the $\mathbf{z}_i$

If an individual i in stratum s experiences an event at time T_i one selects at random $m-1$ controls from the remaining individuals at risk in stratum s

The set consisting of the case i and the $m-1$ controls is called a sampled risk set and denoted $\tilde{R}_i$

The covariates of main interest $\mathbf{x}_i$ are ascertained for the individuals in the sampled risk sets, but are not needed for the remaining individuals in the cohort

For matched nested case-control data one cannot estimate the effects of the confounding variables $\mathbf{z}_i$

But the regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ for the covariates of main interest may be estimated by maximizing the partial likelihood

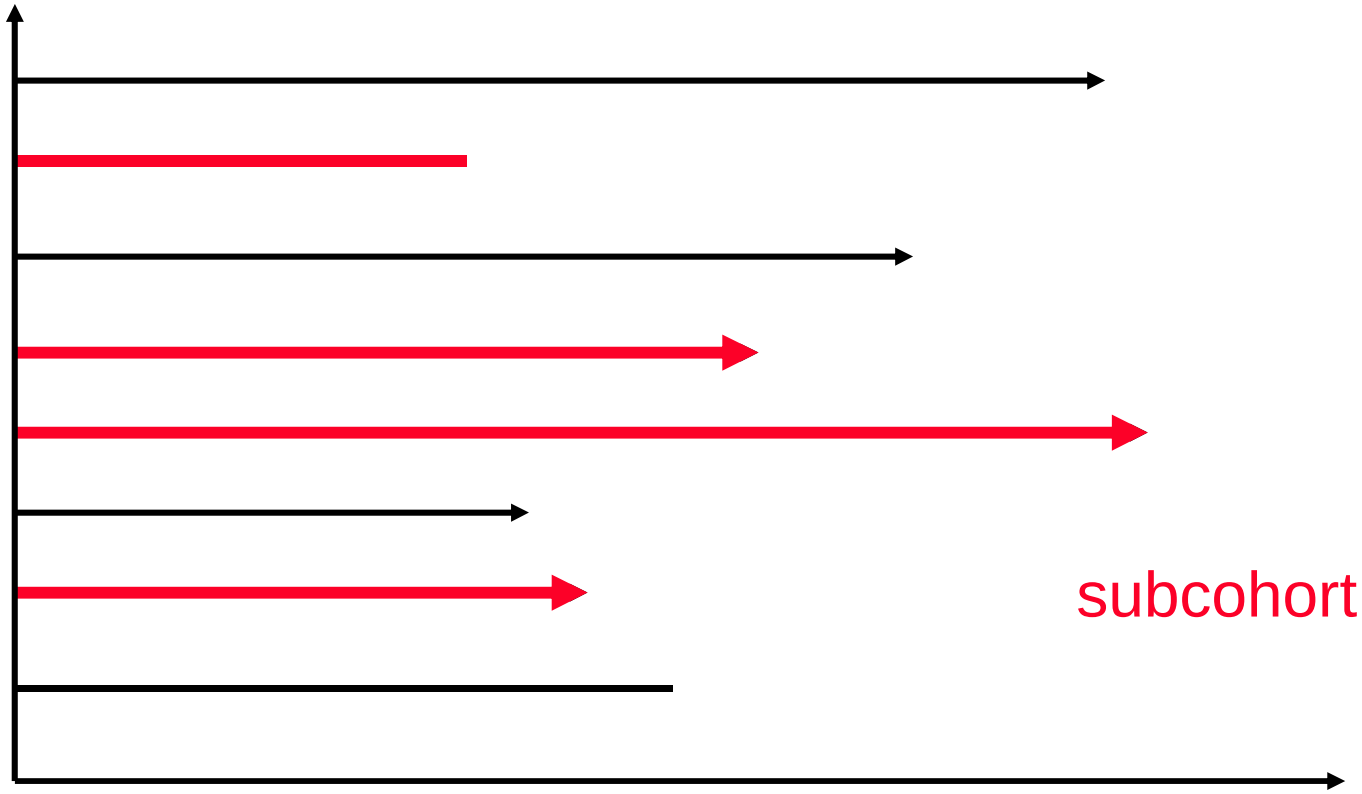
$$L_{\text{ncc}}(\boldsymbol{\beta}) = \prod_{i \in E} \frac{\exp\left(\boldsymbol{\beta}^T \mathbf{x}_i\right)}{\sum_{j \in \tilde{R}_i} \exp\left(\boldsymbol{\beta}^T \mathbf{x}_j\right)}$$

The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ enjoys "the usual properties" of ML-estimators (Borgan et al. 1995)

Case-cohort studies

Select a subcohort S of size m_s from the full cohort by simple random sampling (Prentice 1986) or stratified sampling (Borgan et al 2000)

Illustration for n=8 and m_s=4



The covariates of main interest $\mathbf{x}_i$ are ascertained for all cases and for the individuals in the subcohort, but are not needed for the remaining individuals in the cohort

The regression coefficients may be estimated by maximizing a pseudo likelihood of the form

$$L_{\text{ipw}}(\boldsymbol{\beta}) = \prod_{i \in E} \frac{\exp\left(\boldsymbol{\beta}^T \mathbf{x}_i + \boldsymbol{\gamma}^T \mathbf{z}_i\right) w_i}{\sum_{j \in S_i} \exp\left(\boldsymbol{\beta}^T \mathbf{x}_j + \boldsymbol{\gamma}^T \mathbf{z}_j\right) w_j}$$

The choice of S_i and the weights w_j depend of the sampling design and the method of estimation

For simple random sampling, Prentice original suggestion was to use $S_i = (S \cap R_i) \cup \{i\}$ and all $w_j = 1$

An alternative for simple random sampling is to use inverse probability weighting (IPW):

$S_i = (S \cup \{\text{Cases}\}) \cap R_i$ and $w_j = 1/p_j$ where p_j is the probability that j is included in the case-cohort sample

For cases $p_j = 1$

For non-cases $p_j = \dfrac{\#\text{non-cases in subcohort}}{\#\text{non-cases in cohort}}$

IPW may also be used for stratified case-cohort designs, but then the p_j for non-cases will differ between strata

Note that the maximum pseudo likelihood estimators do not enjoy "the usual properties" of ML-estimators so variance estimation needs special attention

Breaking the matching in nested case-control studies

By breaking the matching, nested case-control data may be treated as case-cohort data with a non-standard sampling schemes for the controls

Estimation may be based on an IPW pseudo likelihood with weights $w_j = 1/p_j$ where p_j is the probability that j is included in the nested case-control sample

Since the matching is broken, it is crucial that the confounding variables $\mathbf{z}_i$ are included in the pseudo likelihood

The inclusion probabilities are $p_j = 1$ for cases

For controls (who do not later become a case) there are more options

One possibility is to use the Kaplan-Meier type weights (Samuelsen 1997)

$$p_j = 1 - \prod_{\substack{T_i \text{ where } j \text{ is} \\ \text{possible control}}} \left(1 - \frac{m-1}{\text{number of possible controls at } T_i}\right)$$

Another option is to estimate the p_j's using logistic regression for the non-cases (Samuelsen et al 2007):

$$p_j = \frac{\exp\left(\xi_0 + \xi_1 T_j + \boldsymbol{\xi}_2^T \mathbf{z}_j\right)}{1 + \exp\left(\xi_0 + \xi_1 T_j + \boldsymbol{\xi}_2^T \mathbf{z}_j\right)}$$

The robust sandwich estimator seems to work fine for variance estimation (in most situations)

Two endpoints

Consider the situation with two endpoints (e.g. two diseases) and assume that the occurrence of one endpoint precludes the occurrence of the other

The situation may be described by a competing risks model with hazard for the e-th endpoint for the i-th individual given by

$$h_{ei}(t) = h_{e0}(t)\exp\left(\boldsymbol{\beta}_e^T \mathbf{x}_i + \boldsymbol{\gamma}_e^T \mathbf{z}_i\right) \qquad e = 1, 2$$

For nested case-control data, controls are matched to the cases, so one needs a separate set of controls for each endpoint

If one breaks the matching, one may pool the controls and use all controls for both endpoints

Question:

Should one break the matching and analyze the nested case-control data as case-cohort data?

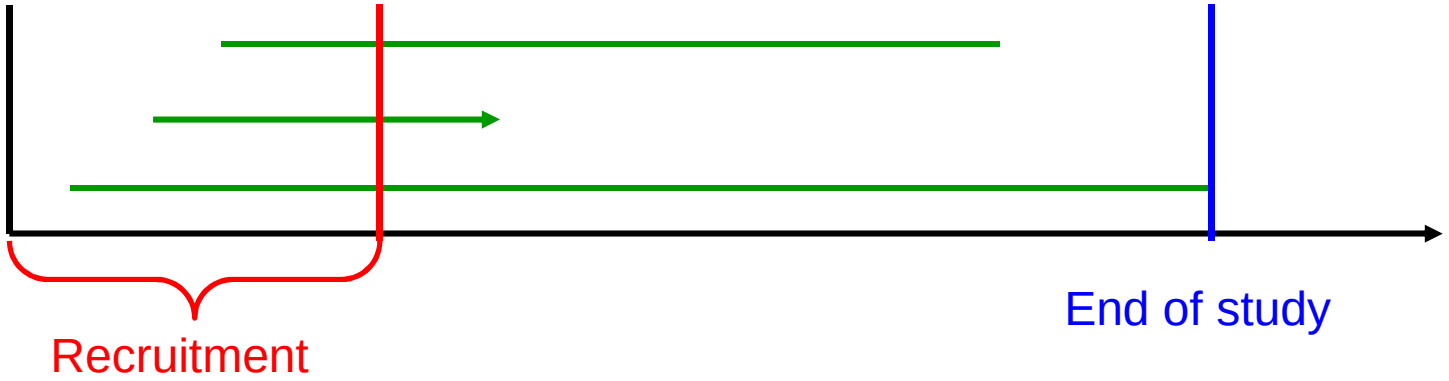
The answer depends on the situation at hand

We will use simulations to illustrate:

- problems that may occur if one breaks the matching
- benefits that one may obtain by breaking the matching

Basic simulation model (one endpoint)

Recruit cohort over 5 years, follow-up for maximum 15 years



- Recruitment is uniform over (0,5) years
- Age at entry is uniform over (50,70) years
- Three study centres of equal size
- Two covariates of main interest (normal with sd=1)
  - one "expensive" (observed for cases and controls)
  - one "cheap" (observed for all)

Events are generated from a Cox model with hazard depending on age, centre and the two main covariates

The regression coefficients are 0.50 for the main covariates

Simulate 1000 cohorts each of size 10 000 with 2.5% events and select one control per case matched on centre, age at entry, and time of recruitment

Results for the "expensive" covariate:

	Mean estimate	Mean sd	Empirical sd	Coverage
Cohort	0.502	0.064	0.063	95.5%
Nested case-control	0.505	0.107	0.109	94.9%
IPW estimation	0.520	0.102	0.107	92.8%

Results for the "cheap" covariate are similar

The two methods of estimation for nested case-control data have comparable performance

Simulating a batch effect

Assume that the "expensive" covariate is a biomarker, and that a case and its controls are analyzed in the same batch

Events are generated using the true value of the "expensive" covariate, but in the data used for estimation we add a normal measurement error with sd=0.50

Results for the "expensive" covariate:

	Mean estimate	Mean sd	Empirical sd	Coverage
Nested case-control	0.504	0.107	0.107	95.6%
IPW estimation	0.432	0.089	0.094	85.4%

Breaking the matching gives biased estimates for IPW estimation

Simulating two end-points

Simulate 1000 cohorts each of size 10 000 with 2.5% events of the first cause and 1% events of the second cause, and select one matched control per case

Results for the "expensive" covariate:

	Mean estimate	Mean sd	Empirical sd	Coverage
Cohort	0.503	0.064	0.063	94.7%
Nested case-control	0.511	0.108	0.109	95.5%
IPW estimation	0.515	0.091	0.093	94.0%
Cohort	0.506	0.101	0.099	95.1%
Nested case-control	0.534	0.175	0.175	95.8%
IPW estimation	0.518	0.123	0.121	95.3%

1st type

2nd type

An efficiency gain is obtained, especially for the rare end-point, by using IPW estimation with the controls for both endpoints combined

Using the full cohort

Often some of the covariates of main interest are available for the full cohort C while "expensive" covariates are only observed for the case-control sample $\tilde{C}$

Write $\mathbf{x}_i = (\mathbf{x}_i^{(\mathrm{cc})}, \mathbf{x}_i^{(\mathrm{all})})$ where $\mathbf{x}_i^{(\mathrm{cc})}$ is only observed for $i \in \tilde{C}$ while $\mathbf{x}_i^{(\mathrm{all})}$ is observed for all $i \in C$

For nested case-control (and case-cohort) sampling, information on $\mathbf{x}_i^{(\mathrm{cc})}$ is missing by design for $i \in C \setminus \tilde{C}$

For one endpoint the full likelihood takes the form
(conditional on the covariates available for everyone)

$$\prod_{i \in \tilde{C}} P(T_i, D_i, \mathbf{x}_i^{(\mathrm{cc})} \mid \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i) \times \prod_{i \in C \setminus \tilde{C}} P(T_i, D_i \mid \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i)$$

(Saarela et al 2008)

The full likelihood may be written:

$$\prod_{i \in \tilde{C}} P(T_i, D_i \mid \mathbf{x}_i^{(\mathrm{cc})}, \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i)\, dP(\mathbf{x}_i^{(\mathrm{cc})} \mid \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i)$$

$$\times \prod_{i \in C \setminus \tilde{C}} \int_{\mathbf{x}} P(T_i, D_i \mid \mathbf{x}_i^{(\mathrm{cc})} = \mathbf{x},\ \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i)\, dP(\mathbf{x} \mid \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i)$$

For full likelihood inference we assume that

- censoring does not depend on $\mathbf{x}_i^{(\mathrm{cc})}$

and we need models for

- the censored survival times given covariates (Cox)
- the conditional distribution of $\mathbf{x}_i^{(\mathrm{cc})}$ given $\mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i$

It may be quite demanding to maximize the full likelihood due to the second factor, in particular for large cohorts

Multiple imputation offers an alternative to maximum likelihood which is easier to implement (Keogh & White 2013)

Consider the situation where $x_i^{(\mathrm{cc})}$ is scalar and assume that

$$x_i^{(\mathrm{cc})} \mid \mathbf{x}_i^{(\mathrm{all})}, \mathbf{z}_i \sim N(\psi_0 + \boldsymbol{\psi}_1^T \mathbf{x}_i^{(\mathrm{all})} + \boldsymbol{\psi}_2^T \mathbf{z}_i, \sigma^2)$$

We may then for $i \in C \setminus \tilde{C}$ use the imputation model

$$x_i^{(\mathrm{cc})} = \theta_0 + \theta_1 D_i + \theta_2 H_0(T_i) + \boldsymbol{\theta}_3^T \mathbf{x}_i^{(\mathrm{all})} + \boldsymbol{\theta}_4^T \mathbf{z}_i + \boldsymbol{\theta}_5^T \mathbf{x}_i^{(\mathrm{all})} H_0(T_i) + \boldsymbol{\theta}_6^T \mathbf{z}_i H_0(T_i) + \varepsilon_i$$

where in practice we may replace the cumulative baseline hazard $H_0(T_i)$ by the Nelson-Aalen estimator (White & Royston 2009)

Basic simulation model

Results for the "expensive" covariate:

	Mean	Mean sd	Empirical sd	Coverage
Cohort	0.502	0.064	0.063	95.5%
Nested case-control	0.505	0.107	0.109	94.9%
IPW estimation	0.520	0.102	0.107	92.8%
Multiple imputation	0.490	0.095	0.094	95.1%

Results for the "cheap" covariate:

	Mean	Mean sd	Empirical sd	Coverage
Cohort	0.499	0.064	0.062	95.5%
Nested case-control	0.508	0.108	0.108	95.0%
IPW estimation	0.522	0.102	0.109	92.4%
Multiple imputation	0.500	0.069	0.066	95.9%

Multiple imputation gives an improvement, especially for the covariate available for the full cohort

Simulating a batch effect

Results for the "expensive" covariate (measured with error)

	Mean	Mean sd	Empirical sd	Coverage
Nested case-control	0.504	0.107	0.107	95.6%
IPW estimation	0.432	0.089	0.094	85.4%
Multiple imputation	0.487	0.094	0.091	94.4%

Multiple imputation seems to work fine when there is a batch effect

Simulating two end-points

Results for the "expensive" covariate:

	Mean estimate	Mean sd	Empirical sd	Coverage
Cohort	0.503	0.064	0.063	94.7%
Nested case-control	0.511	0.108	0.109	95.5%
IPW estimation	0.515	0.091	0.093	94.0%
Multiple imputation	0.498	0.086	0.084	94.8%
Cohort	0.506	0.101	0.099	95.1%
Nested case-control	0.534	0.175	0.175	95.8%
IPW estimation	0.518	0.123	0.121	95.3%
Multiple imputation	0.500	0.116	0.111	95.9%

1st type

2nd type

Multiple imputation works fine when there are two endpoints (both endpoints are used in the imputation)

For the "cheap" covariate multiple imputation gives results close to the full cohort

Multiple imputation is promising, but it assumes that:

- censoring does not depend on $x_i^{(cc)}$
- $x_i^{(cc)} \mid \mathbf{x}_i^{(all)}, \mathbf{z}_i \sim N(\psi_0 + \boldsymbol{\psi}_1^T \mathbf{x}_i^{(all)} + \boldsymbol{\psi}_2^T \mathbf{z}_i, \sigma^2)$

Both assumptions are fulfilled for the results shown above

What happens if the assumptions are violated?

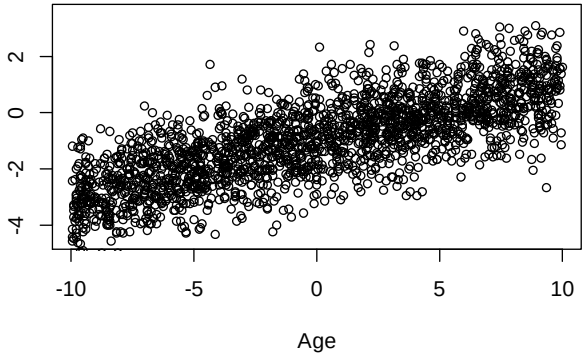
The first assumption is violated if censoring by a competing cause depends on $x_i^{(cc)}$

The second assumption is violated if the conditional distrubution of $x_i^{(cc)}$ given $\mathbf{x}_i^{(all)}, \mathbf{z}_i$ is wrongly specified

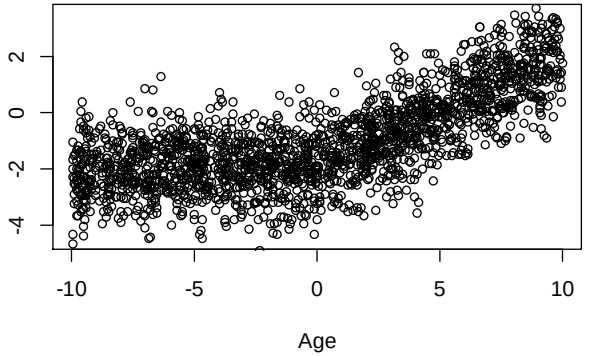
We will look at violation of the second assumption

Misspecified relation between $x_i^{(cc)}$ and age

Assumed relation:



True relation:



Results for the "expensive" covariate:

	Mean	Mean sd	Empirical sd	Coverage
Cohort	0.501	0.060	0.060	94.9%
Nested case-control	0.509	0.108	0.107	95.1%
IPW estimation	0.450	0.093	0.092	90.8%
Multiple imputation	0.435	0.087	0.083	87.2%

Multiple imputation gives biased estimate when the relation between $x_i^{(cc)}$ and age is wrongly specified

Also IPW estimation gives a bias here

Concluding remarks

- Nested case-control (and case-cohort) studies are useful when one or more covariates are "expensive"

Partial likelihood vs IPW estimation

- For one endpoint, the statistical efficiency of partial likelihood and IPW estimation is about the same, and partial likelihood estimation is the "default method" to use
- When there are more than one endpoint, IPW estimation makes it possible to "reuse controls". Thereby one may save biological material and/or obtain an efficiency gain (in particular for rare endpoints)
- When "close matching" is needed one should not break the matching and use IPW estimation

Using the full cohort

- Methods using the full cohort give more efficient estimates, especially for covariates available for the full cohort (and when there are few controls)
- But more modelling assumptions are needed (with the risk of model misspecification)

ML-estimation

- It is demanding to maximize the full likelihood in large cohorts
- User friendly software needs to be developed if ML-estimation is going to be used

Multiple imputation

- Multiple imputation is promising
- But further studies are needed to investigate its performance under model misspecification (which we are currently doing)
- Multiple imputation does not easily handle models with interaction between $\mathbf{x}_i^{(cc)}$ and $\mathbf{x}_i^{(all)}$

Limitations of our study

- Have only considered right-censoring. But all the methods are easily extended to handle left-truncation as well
- Have only considered fixed covariates. But partial likelihood estimation apply to time-varying covariates, and IPW estimation may also be used if the full paths of the time-varying covariates of cases and controls may be obtained
- ML-estimation and multiple imputation cannot handle time-varying covariates