Survival Analysis I (CHL5209H)

Olli Saarela

Dalla Lana School of Public Health
University of Toronto

olli.saarela@utoronto.ca

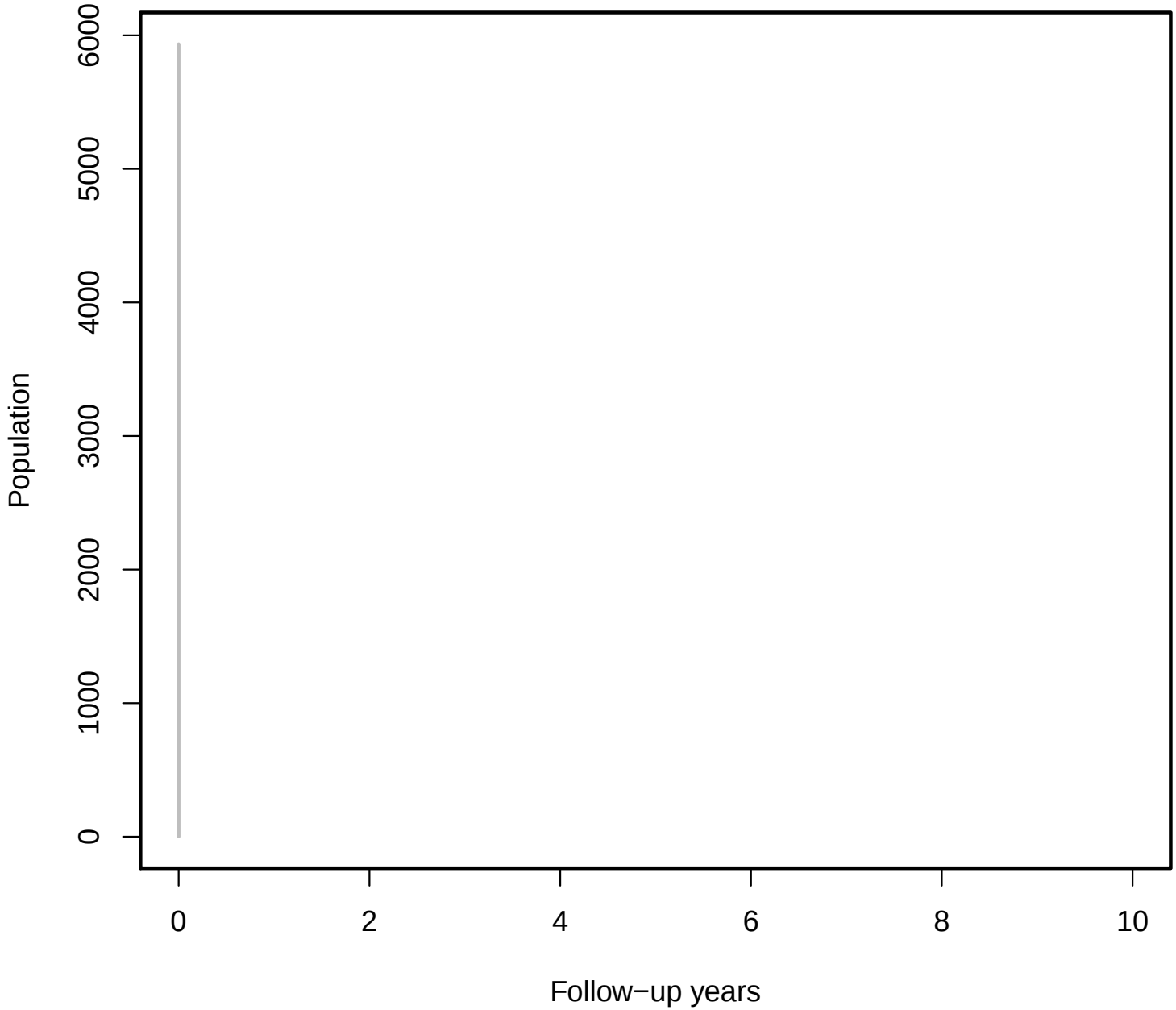
March 26, 2019

Time matching/risk set sampling/incidence density sampling/nested case-control design

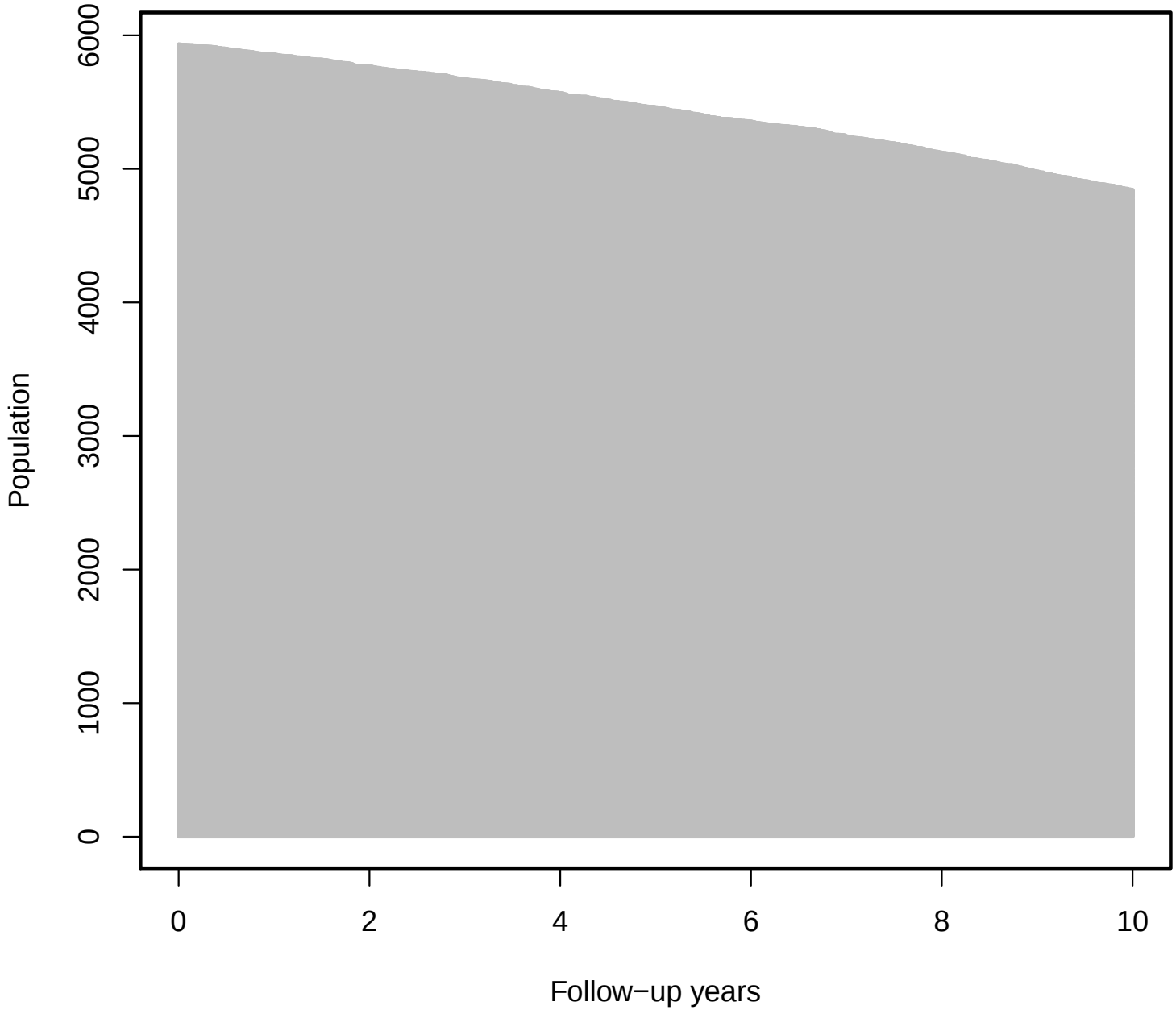
Motivation

- Suppose that a cohort (that is, a closed population) of size n has been recruited through participation in a health examination survey.
- Some baseline characteristics (e.g. blood pressure, cholesterol, BMI, smoking habits, prevalent health conditions and their treatment) are recorded on all cohort members. Denote these characteristics by X.
- We are interested in whether particular genetic/biomarkers (denoted by Z), determined from plasma/serum samples collected at the examination, are associated with health outcomes in the cohort.
- However, carrying out the measurements on all cohort member would be expensive, and no association study can be carried out at this point anyway. (Why?)
- Solution: store (freeze) the biological material to wait for later use.

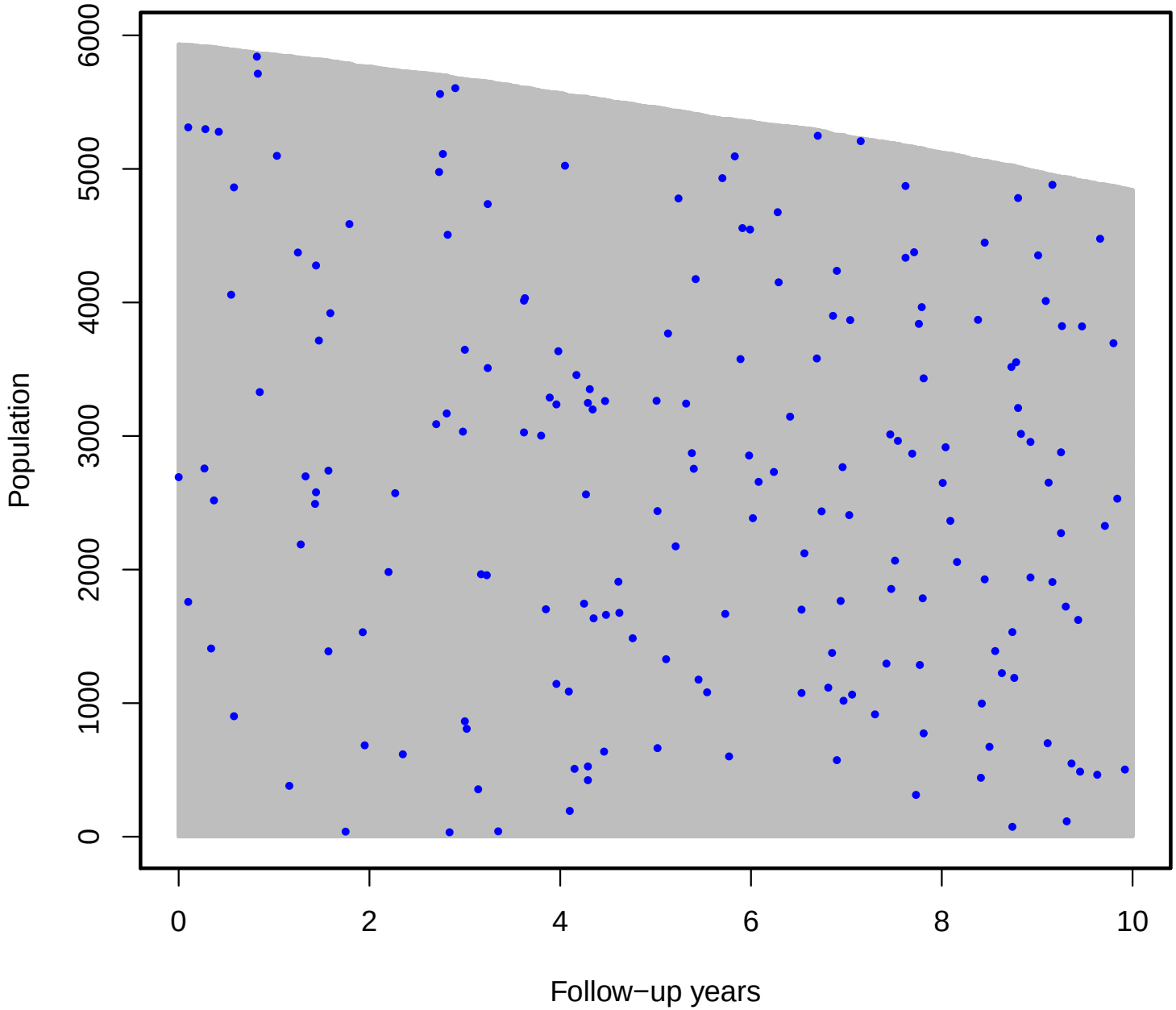
Before the follow-up



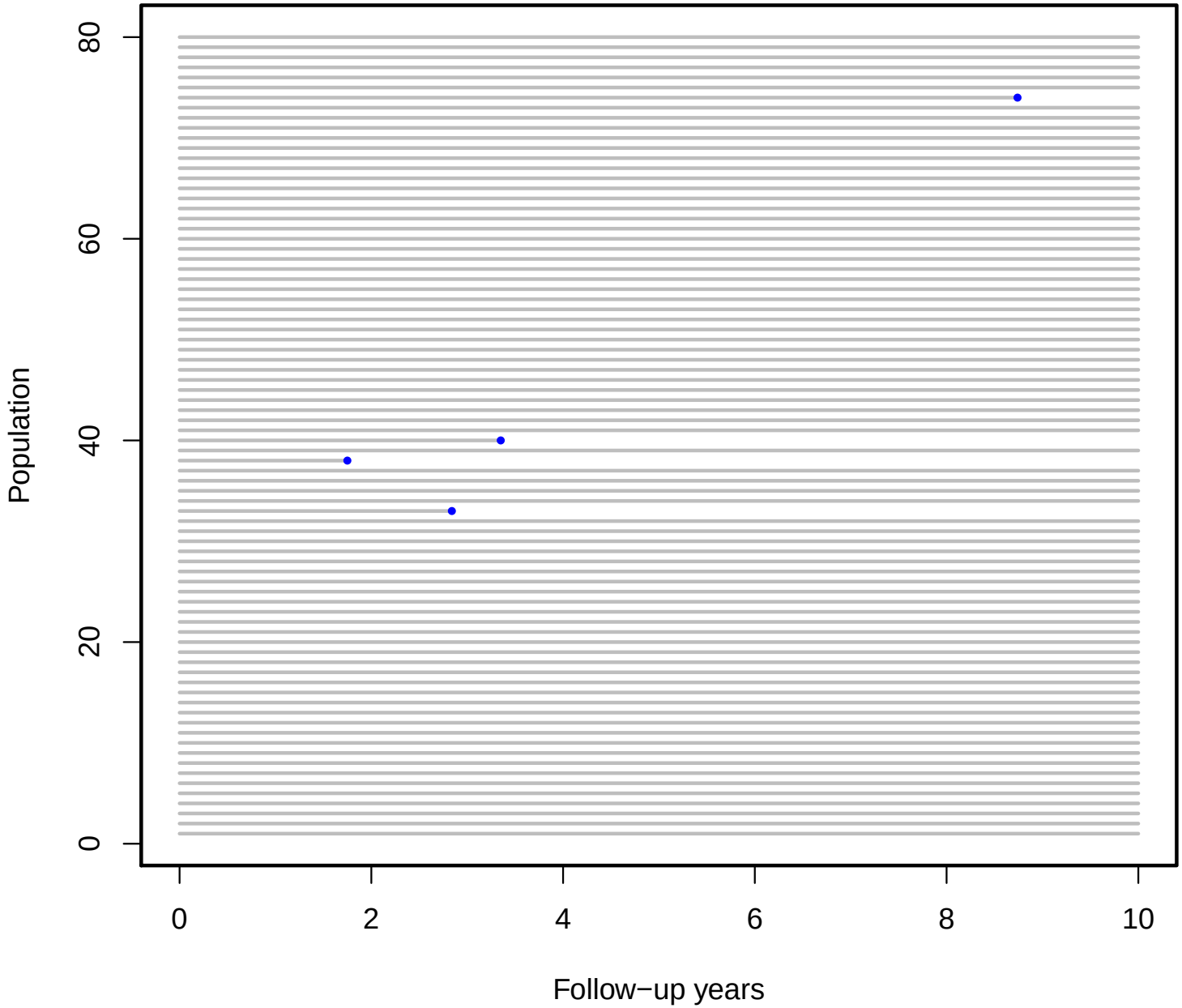
After 10 years of follow-up



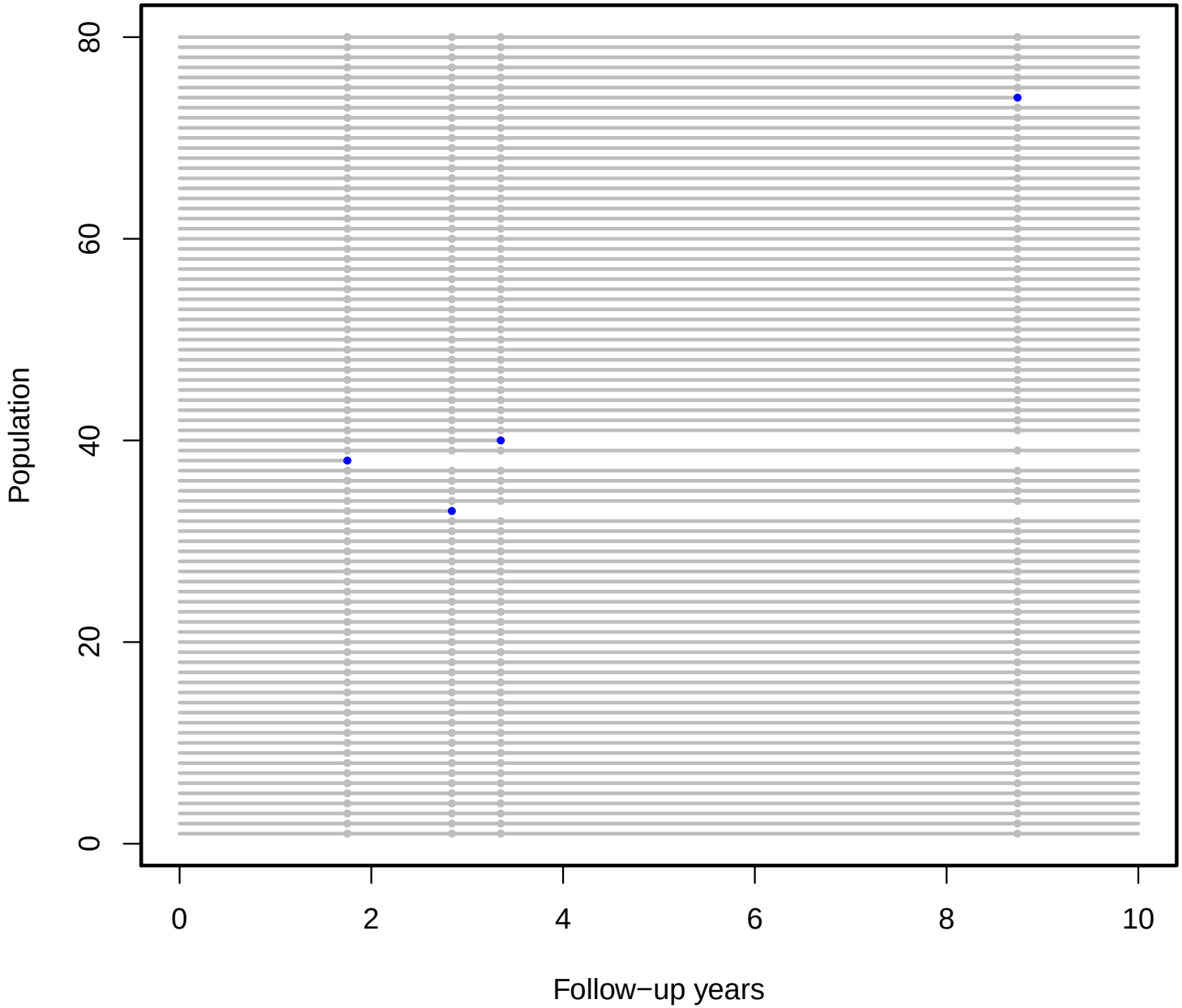
Case series (184 incident stroke events)



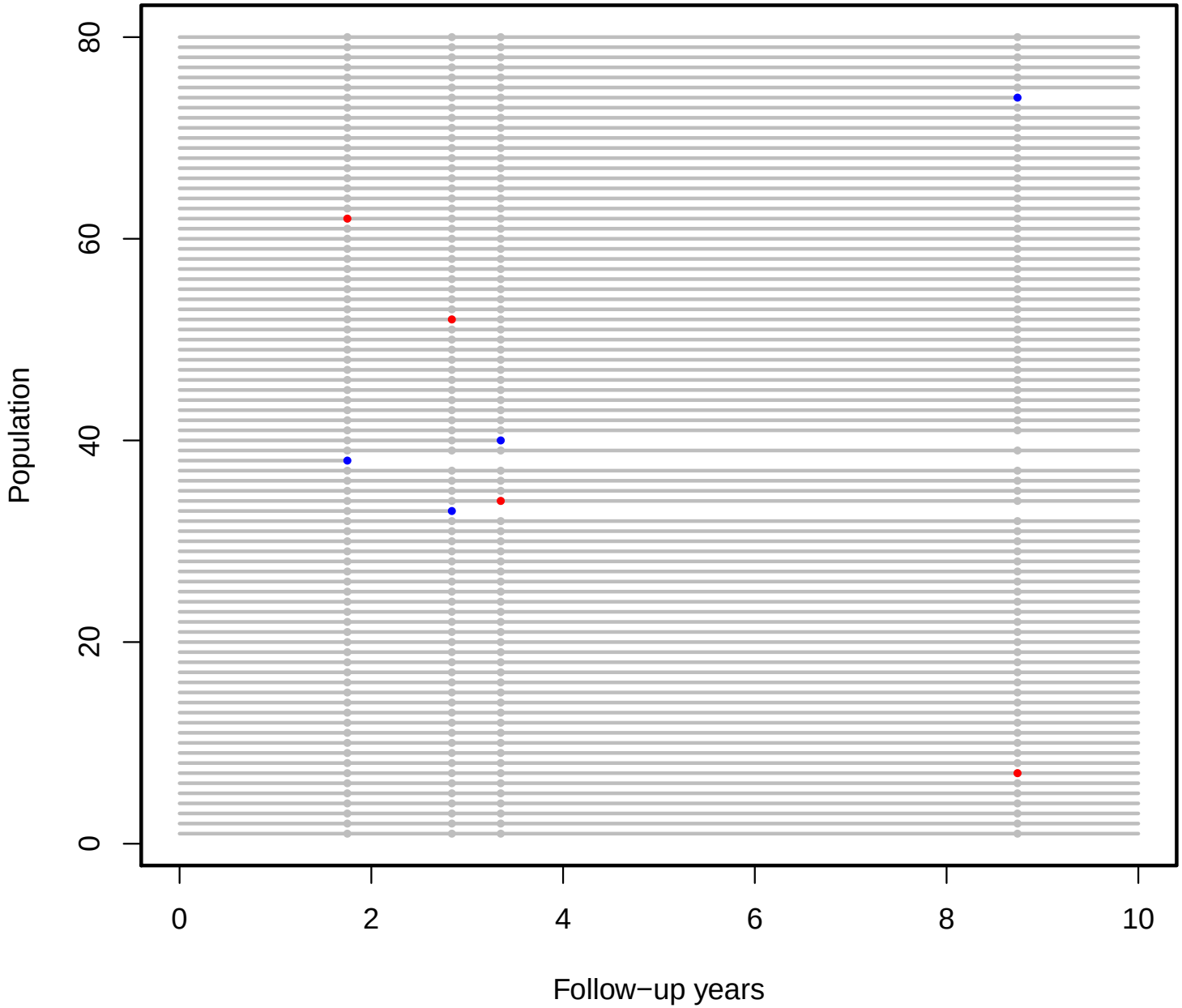
Zoom in to see more



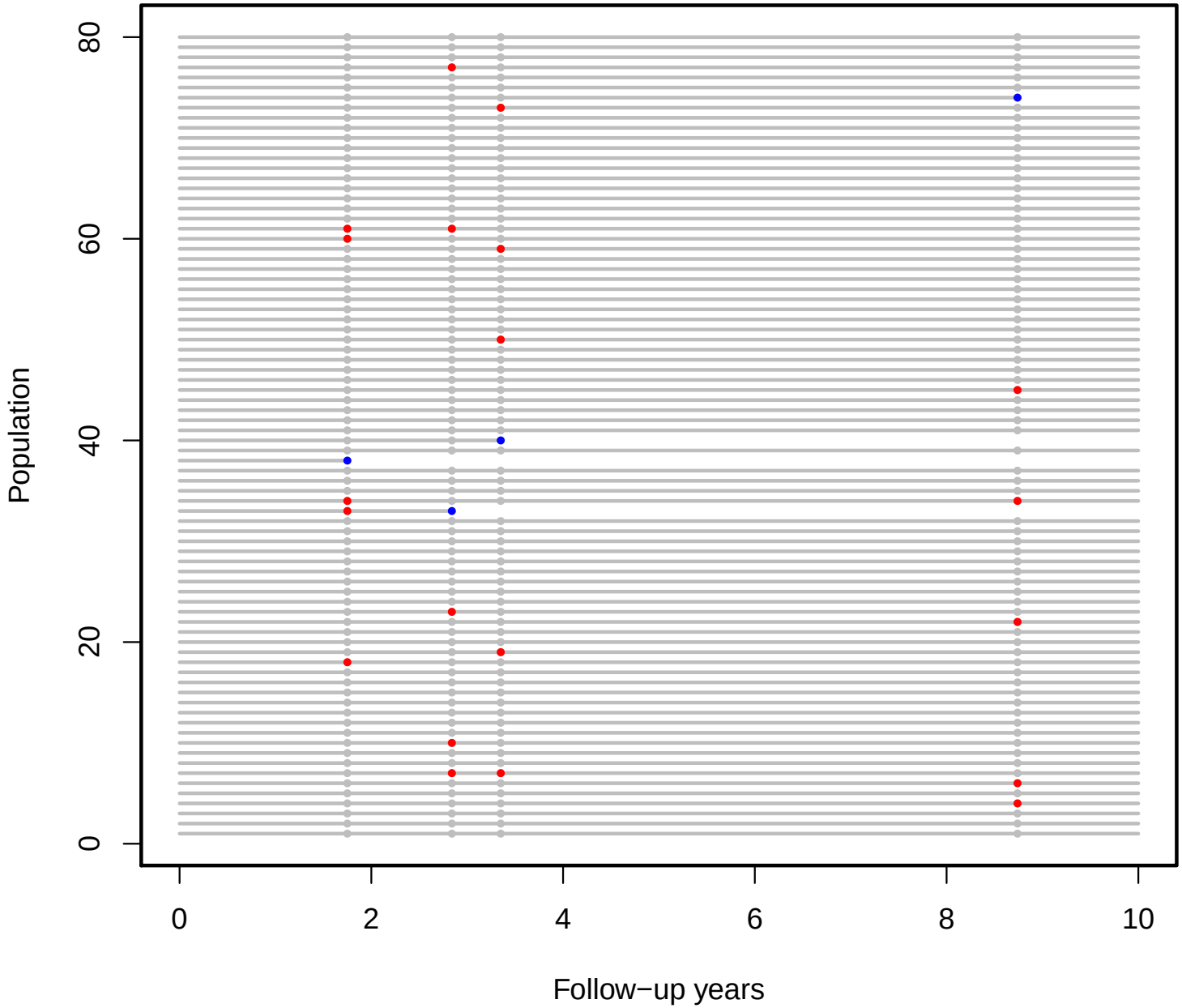
Identify risk sets



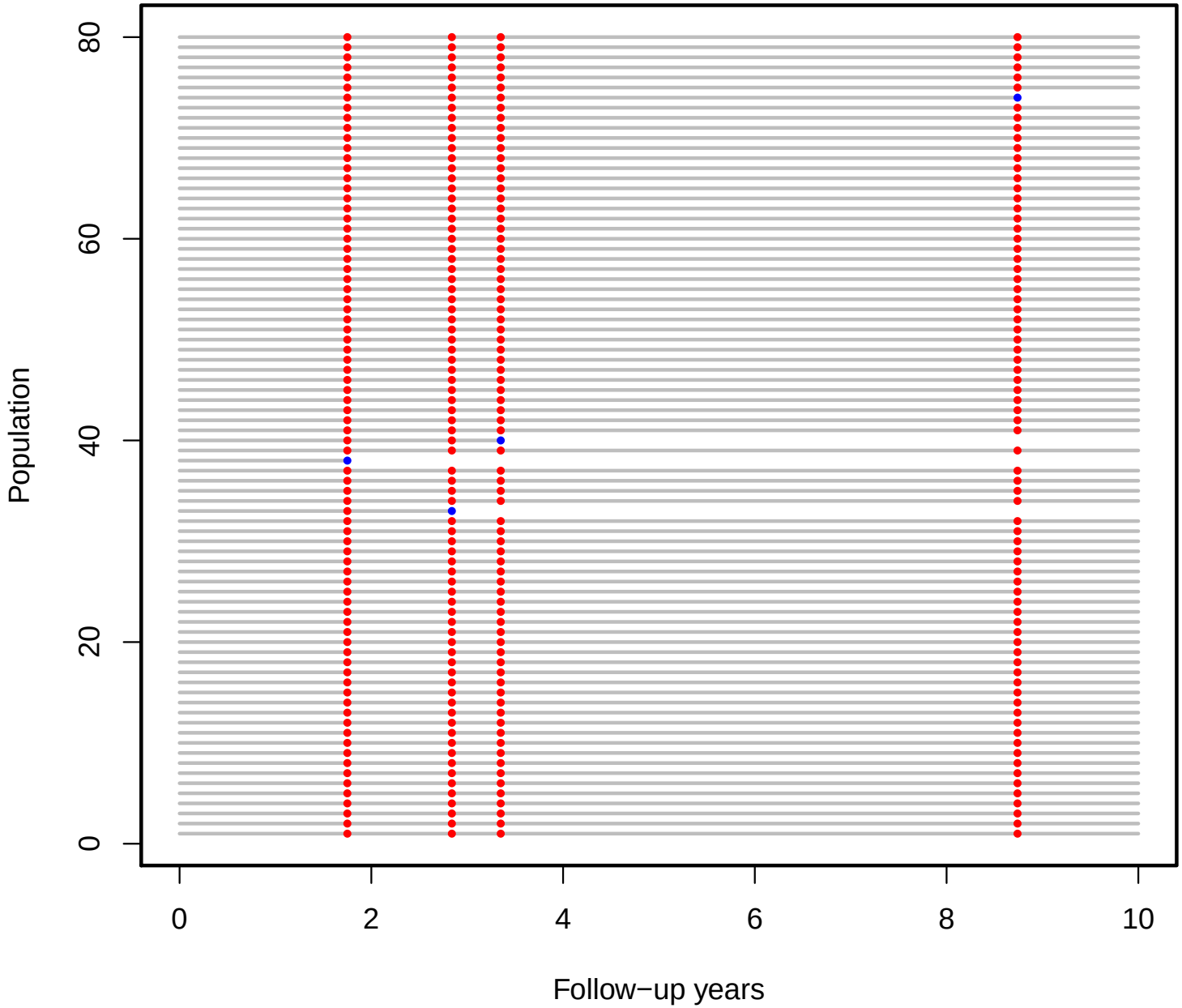
Sample one control per case



Sample 5 controls per case



Or, sample the whole riskset



Nested case-control study

- From each risk set, m controls are selected randomly, and independently from the previous or later sampled risk sets.
- Because of this, one individual can contribute a control to more than one case, and individuals with an outcome event can contribute a control before their event time.
- The case and the m controls are now matched by time.
- Selecting more than $m = 5$ controls per case no longer substantially improves efficiency. This is why cost savings are possible through nested case-control designs.
- The covariate measurements Z are collected on all individuals with an event of interest, and the pooled set of individuals contributing the sampled controls.
- To ensure comparability of the measurements, the cases and controls may need to to be further matched w.r.t. factors such as storage time, storage conditions, freeze-thaw cycles, and analytic batch (e.g. plate).

Conditional logistic regression

Matched case-control studies

- Before considering the analysis of nested case-control studies, we consider how matched case-control studies are generally analyzed.
- Use i to index sets of cases and controls, matched with respect to some characteristics (other than time).
- Ley $i1, i2, \ldots$ index individuals within the matched set.
- For every matched set we can specify the logistic regression model

$$P(D_{ij} = 1 \mid z_{ij}, x_{ij}) = \frac{\exp\{\alpha_i + \beta z_{ij} + \gamma' x_{ij}\}}{1 + \exp\{\alpha_i + \beta z_{ij} + \gamma' x_{ij}\}},$$

where z_i is the exposure of interest, and x_i are other non-matched characteristics that we want to include in the model.

Elimination of nuisance parameters

- In this model, there are as many intercept terms as there are matched sets.
- We would like to avoid estimating these nuisance parameters, as they are not of interest to us.
- The exposure effect was assumed the same across the matched sets.
- It turns out that certain conditioning argument helps.
- As an example, consider 1-1 matched design, where each matched set includes one case and one control.
- As the likelihood contribution, we use the conditional probability

$$P(D_{i1} = 1 \mid D_{i1} + D_{i2} = 1, z_{i1}, x_{i1}, z_{i2}, x_{i2}; \theta_i),$$

where $\theta_i = (\alpha_i, \beta, \gamma)$.

Elimination of nuisance parameters (2)

- Using the definition of conditional probability, and assuming the outcomes of the two individuals independent, we can write this as

$$
\begin{aligned}
&P(D_{i1} = 1 \mid D_{i1} + D_{i2} = 1, z_{i1}, x_{i1}, z_{i2}, x_{i2}; \theta_i) \\
&= \frac{P(D_{i1} = 1, D_{i2} = 0 \mid z_{i1}, x_{i1}, z_{i2}, x_{i2}; \theta_i)}{P(D_{i1} + D_{i2} = 1 \mid z_{i1}, x_{i1}, z_{i2}, x_{i2}; \theta_i)} \\
&= \frac{\begin{array}{r} P(D_{i1} = 1 \mid z_{i1}, x_{i1}; \theta_i) \\ \times P(D_{i2} = 0 \mid z_{i2}, x_{i2}; \theta_i) \end{array}}{\displaystyle\sum_{(d_1,d_2)\in\{(0,1),(1,0)\}} \begin{array}{r} P(D_{i1} = d_1 \mid z_{i1}, x_{i1}; \theta_i) \\ \times P(D_{i2} = d_2 \mid z_{i2}, x_{i2}; \theta_i) \end{array}}.
\end{aligned}
$$

Elimination of nuisance parameters (3)

- Substituting in the logistic regression model gives

$$
\begin{aligned}
&P(D_{i1}=1 \mid D_{i1}+D_{i2}=1, z_{i1}, x_{i1}, z_{i2}, x_{i2}; \theta_i) \\
&= \frac{\frac{\exp\{\alpha_i+\beta z_{i1}+\gamma' x_{i1}\}}{1+\exp\{\alpha_i+\beta z_{i1}+\gamma' x_{i1}\}} \times \frac{1}{1+\exp\{\alpha_i+\beta z_{i2}+\gamma' x_{i2}\}}}{\sum_{(d_1,d_2)\in\{(0,1),(1,0)\}} \frac{\exp\{d_1(\alpha_i+\beta z_{i1}+\gamma' x_{i1})\}}{1+\exp\{\alpha_i+\beta z_{i1}+\gamma' x_{i1}\}} \times \frac{\exp\{d_2(\alpha_i+\beta z_{i2}+\gamma' x_{i2})\}}{1+\exp\{\alpha_i+\beta z_{i2}+\gamma' x_{i2}\}}} \\
&= \frac{\exp\{\alpha_i+\beta z_{i1}+\gamma' x_{i1}\}}{\exp\{\alpha_i+\beta z_{i1}+\gamma' x_{i1}\}+\exp\{\alpha_i+\beta z_{i2}+\gamma' x_{i2}\}} \\
&= \frac{\exp\{\beta z_{i1}+\gamma' x_{i1}\}}{\exp\{\beta z_{i1}+\gamma' x_{i1}\}+\exp\{\beta z_{i2}+\gamma' x_{i2}\}}.
\end{aligned}
$$

- The intercept term α_i canceled out. Note also that if $z_{i1}=z_{i2}$, the likelihood contribution is uninformative of β.

Back to time matching

- The regression coefficient β in the previous likelihood expression was a log-odds ratio.
- In the time-matched setting, instead of logistic regression, we would specify a Cox model such as

$$\lambda_{ij}(t) = \lambda_{0i}(t) \exp\{\beta z_{ij} + \gamma' x_{ij}\},$$

where $\lambda_{0i}(t)$ is a baseline hazard function specific to the matched set i.

- Consider the 1-1 time matched design, where one control per case is selected randomly from the riskset at the event time of the case.
- Now the likelihood contribution is given by the conditional probability of individual 1 in risk set i experiencing an outcome event at time t_i, given that we know that one of the two individuals experienced an outcome event, that is,

$$P(\mathrm{d}N_{i1}(t_i) = 1 \mid \mathrm{d}N_{i1}(t_i) + \mathrm{d}N_{i2}(t_i) = 1, \mathcal{F}_{t_i^-}; \theta_i),$$

Back to time matching (2)

- Here the observed history $\mathcal{F}_{t_i^-}$ records the covariate values $\{z_{i1}, x_{i1}, z_{i2}, x_{i2}\}$ and that the two individuals are still at risk at this moment, that is, $\{Y_{i1}(t_i) = 1, Y_{i2}(t_i) = 1\}$.
- A similar calculation to before can be used to motivate that

$$P(\mathrm{d}N_{i1}(t_i) = 1 \mid \mathrm{d}N_{i1}(t_i) + \mathrm{d}N_{i2}(t_i) = 1, \mathcal{F}_{t_i^-}; \theta_i)$$
$$= \frac{\exp\{\beta z_{i1} + \gamma' x_{i1}\}}{\exp\{\beta z_{i1} + \gamma' x_{i1}\} + \exp\{\beta z_{i2} + \gamma' x_{i2}\}}.$$

- The baseline hazards canceled out of the expression.
- The regression coefficient β is a log-hazard ratio.
- The functional form of the likelihood contribution is the same as before.
- Product of such contributions over the risksets is no longer a conditional probability, but still gives a partial likelihood.

Conditional logistic regression

- Conditioning on the total number of cases in non-time matched and time-matched case-control designs results in the same functional form of the likelihood expression.
- However, since the underlying models are different (logistic vs. Cox), the parameter being estimated is different.
- But, since the likelihood expression has the same functional form, in both cases it can be maximized using the same software.

Conditional logistic regression

- Conditioning on the total number of cases in non-time matched and time-matched case-control designs results in the same functional form of the likelihood expression.
- However, since the underlying models are different (logistic vs. Cox), the parameter being estimated is different.
- But, since the likelihood expression has the same functional form, in both cases it can be maximized using the same software.
- On terminology: the terms time matching/risk set sampling/incidence density sampling/nested case-control design all mean the same particular kind of sampling mechanism for the controls.

Conditional logistic regression in R

- The conditional logistic likelihood can be maximized using the `clogit` function of the R survival package

```
clogit(formula, data, weights, subset, na.action,
       method=c("exact", "approximate",
                "efron", "breslow"), ...)
```

  which uses a model formula of the form

```
case.status~exposure+strata(matched.set)
```

- Here the strata variable identifies the time matched risksets.
- Ties in the data mean that more than one event occurred at the same time; the argument `method` specifies how the ties are handled.

Exact method for ties

- The conditioning approach generalizes if ties are present.
- For example, if two events occurred at time t_i and one control was sampled from the riskset, we would get

$$P\left(\mathrm{d}N_{i1}(t_i) = 1, \mathrm{d}N_{i2}(t_i) = 1 \mid \sum_{l=1}^{3} \mathrm{d}N_{il}(t_i) = 2, \mathcal{F}_{t_i^-}; \theta_i\right)$$
$$= \frac{\exp\{\beta z_{i1} + \gamma' x_{i1}\} \exp\{\beta z_{i2} + \gamma' x_{i2}\}}{\begin{array}{l}\exp\{\beta z_{i1} + \gamma' x_{i1}\} \exp\{\beta z_{i2} + \gamma' x_{i2}\} \\ \quad + \exp\{\beta z_{i1} + \gamma' x_{i1}\} \exp\{\beta z_{i3} + \gamma' x_{i3}\} \\ \quad\quad + \exp\{\beta z_{i2} + \gamma' x_{i2}\} \exp\{\beta z_{i3} + \gamma' x_{i3}\}\end{array}}.$$

- This is the so-called exact methods for handling ties; the others are approximations.