Nested case control

Jeppe EH Madsen

2019-05-03

Epidemiologic cohort studies are often used for assessing the variation in rates of morbidity and mortality due to factors present in the population. Because the outcome of interest can be very rare, cohort studies may require a lot of subjects to reliably be able to answer the question at hand. It can, however, be very expensive to collect covariate information for all subjects. One solution to this problem is nested case control where only a subset of the non-failures are used for the analysis. This works because the extra statistical power of the study gained by each additional non-failure is very small compared to that of the failures (cases) when there are many non-failures.

The case-cohort design works in the following way: For each observed failure-time (case), sample $m-1$ non-failures (controls) without replacement. Nested case-control is similar, but with the difference that the sampling is done, only among those at risk at the cases failure-time.

Visualization using Lexis diagram

The Lexis diagram plots the year of birth on the x-axis and the age on the y-axis. Someone who is born in 2000 will be represented by a diagonal line starting at (2000,0) and going through (2001,1), (2002,2) and so on. Let's say we have a group of patients as in Figure 1.

The stars represent times where the different patients could be sampled as controls (even though they all eventually become cases in the plot) and the squares are the actual failure times. This also means that the same person can get sampled more than once.

Nested case control for the Cox proportional hazards model

It turns out, somewhat surprisingly, that the estimator ($\hat{\beta}$) of the regression parameters (β) in the Cox model isn't influenced by the sampling and is thus the same as usually. The fact that we have fewer non-failures per failure in the data than in the real world does, however, mean that we can't use the usual Breslow estimator of the integrated baseline hazard. Let $\mathcal{R}_j$ be the set of all those at risk at failure-time t_j-, and $\tilde{\mathcal{R}}_j$ be the set of the case at t_j and the $m-1$ controls. Let $n(t_j)$ be the total number at risk at time t_j, and $Z_l(t)$ a vector of time-dependent covariates for observation l. Then the estimator for the integrated baseline hazard turns out to be

$$\hat{A}(t;\hat{\beta}) = \sum_{t_j<t} \frac{1}{\sum_{l\in\tilde{\mathcal{R}}_j} \exp\left(\hat{\beta}^T Z_l(t_j)\right) n(t_j)/m}.$$

So each observation is given a higher weight according to how many controls have been sampled and how big the original data set is. Note also that we just need the covariate value at time t_j even though it's time-dependent.

Connection to conditional logistic regression

Conditional logistic regression deals with the probability of events for observations in different strata given that we know how many events we observe in each stratum. This corresponds exactly to the situation we are in when we have a nested case control design, since each stratum corresponds to a failure time (so we know that we have exactly one failure time in each stratum). The β-parameters in this type of model correspond exactly to the β-parameters we get when we estimate the Cox model for the nested case control design. This also means that we should be very careful when interpreting the β-parameters - in "normal"

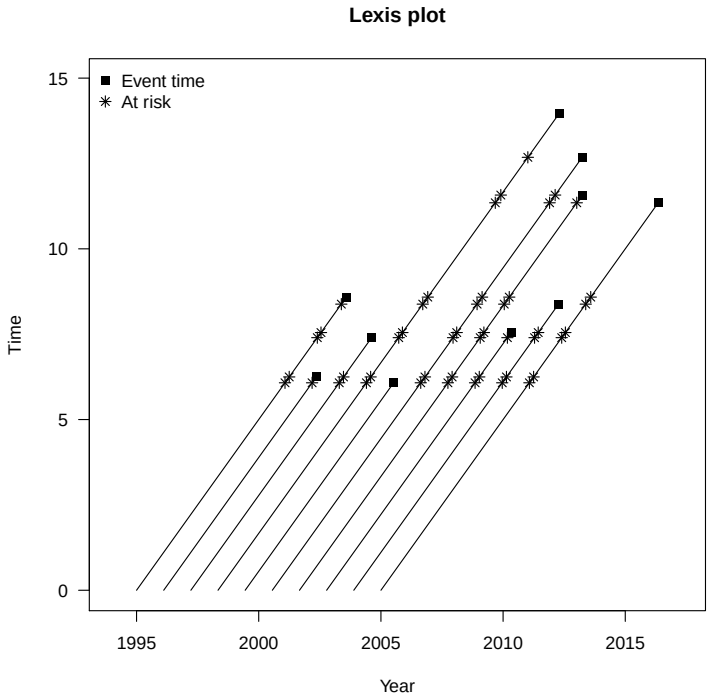

Figure 1: Lexis plot of hypothetical study.

logistic regression, it is easy to interpret the β-parameters since they equal the logarithm of the odds ratio between individuals with covariates $(Z_1, ..., Z_k + 1, ..., Z_K)$ and $(Z_1, ..., Z_k, ..., Z_K)$. We know, however, that this isn't the case in the Cox model - here they equal the logarithm of the *hazard rate* between individuals with covariates $(Z_1, ..., Z_k + 1, ..., Z_K)$ and $(Z_1, ..., Z_k, ..., Z_K)$. Hence, parameters should be interpreted as hazard rates rather than odds ratios when using conditional logistic regression.

Counter-matching / stratified sampling

A way to improve effeciency of nested case control is to use counter-matching. Let's say we have a study where we want to use nested case control, but we have a covariate which is very easy and cheap to collect (e.g. sex). Then split data into S strata (so two strata in the case of sex - or maybe more these days, but you get the idea). If we have a case in stratum $s(i)$ at time t, sample without replacement m_s among those at risk at time t in stratum $s \neq s(i)$, and $m_{s(i)} - 1$ from stratum $s(i)$. It could be smart, for instance, to only sample controls among women when we have a female case and likewise for men. Intuitively that makes it easier to determine the effect of a covariate on outcome because sex doesn't blur the picture.

How well does it work?

In case of only one covariate with the true parameter equal to 0, the variance of the estimator for the full cohort design to the variance of the nested case control design is $m/(m-1)$ independently of censoring and covariate distributions.

Advantages and disadvantages

There are both advantages and disadvantages to using a nested case control design. Some of the advantages are:

- We only need covariate information for a subset of observations.
- We only need covariate information for the observations at the times they are sampled for, even if covariates are time-dependent.
- If covariates are time-fixed, then we can ignore the matching between a given case and its controls, which leads to more efficient estimators.

Some disadvantages are:

- We loose statistical power when only using a subset of observations.
- If we have more than one kind of outcome, we need to sample controls for every type of outcome for nested case control, but not if we use the case-cohort design.
- We have to wait until cases occur to get covariate information when using nested case control, but controls can be chosen right away when using a case-cohort design.