Nested case control

Vignette Author

2019-03-11

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-tim (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.

They get their diagnosis, but we don't observe anything else about them before five years after (so data is truncated?). Then we're interested in the time from the index date (first diagnosis + 5 years) until they get diagnosed with dementia. We can potentially sample an observation every time we have a star in Figure 1.

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"

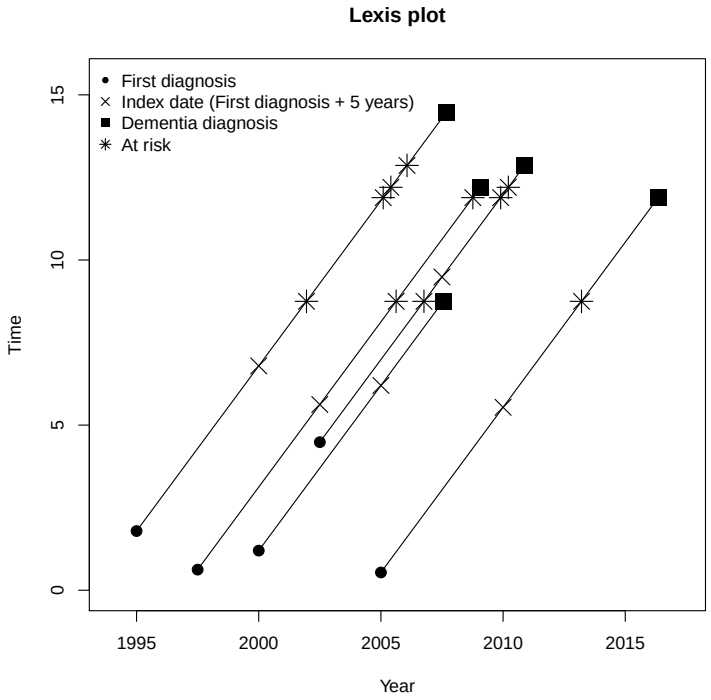

Figure 1: Lexis plot that I should figure (pun intended) out how to use.

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 it is expensive to measure one covariate so we want to use nested case control, but we have another 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 $s(i)$.

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 lose 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.

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. What if we have other covariates and/or the true parameter is different from 0?

Let's say we have a simple setup: two binary variables - one for treatment and another for sex. Let's say 10 % of patients are treated in the study. We let the censoring times equal the 1 % quantile of the failure times so that we have exactly 99 % censoring (so we have type 2 censoring). We use 5 controls per case for the nested case control design. We simulate 5000 observations and repeat the simulation 1000 times.

The true model used for the simulations is a Cox-Weibull model of the form:

$$\lambda(t) = Y(t)\alpha_0(t)\exp(\beta_1 X_{\text{treat}} + \beta_2 X_{\text{sex}})$$

with $\beta_{\text{treat}} = \beta_{\text{sex}} = 0.2$, scale parameter equal to 1/100, and shape parameter equal to 2. Precision is here defined as standard error of parameter estimate, so relative precision is the standard error of the parameter estimate using nested case control devided by the standard error of the parameter estimate using the Cox model.

How does the relative precision depend on the sample size? The simulation has been run with sample sizes of 5000, 25000, and 100000 observations leading to median relative precisions of 1.11, 1.11, and 1.11 respectively. So the relative precision of nested case control compared to the Cox model does not seem to depend on sample size even in the case of multiple covariates (should be possible to show theoretically?)..

How does the relative precision depend on m, the number of sampled controls per case? The results of the simulation are summarized in Figure 2.

It matters a lot when we go from 1 control to, say, 3 controls, but the added precision of having more controls seems to decrease very rapidly. The median relative precision is 1.54 when we have 1 control, and 1.11 when we have 5 controls. The relative precision is 1.06 with 10 controls and 1.03 with 20 controls so most of the reduction in relative precision has already happened at 5 controls.

How does the relative precision depend on the true parameter value? The simulation has been run for parameter values of -0.4, -0.2, 0, 0.2, 0.7 and 1.4 corresponding to hazard rates of 0.67, 0.82, 1, 1.22, 2.01, and 4.06. The median relative precisions were 1.06, 1.08, 1.10, 1.11, 1.16, and 1.25 respectively so it seems like the relative precision is increasing in the value of the true parameter.

This little simulation study is only one very simple scenario so we should be careful not to conclude too confidently based on it, but to make a long story short: it seems like the effectiveness of the nested case control design is not too imprecise - it lets us go from a sample size of 5000 to 300 with a standard error, which is only 11 % higher.

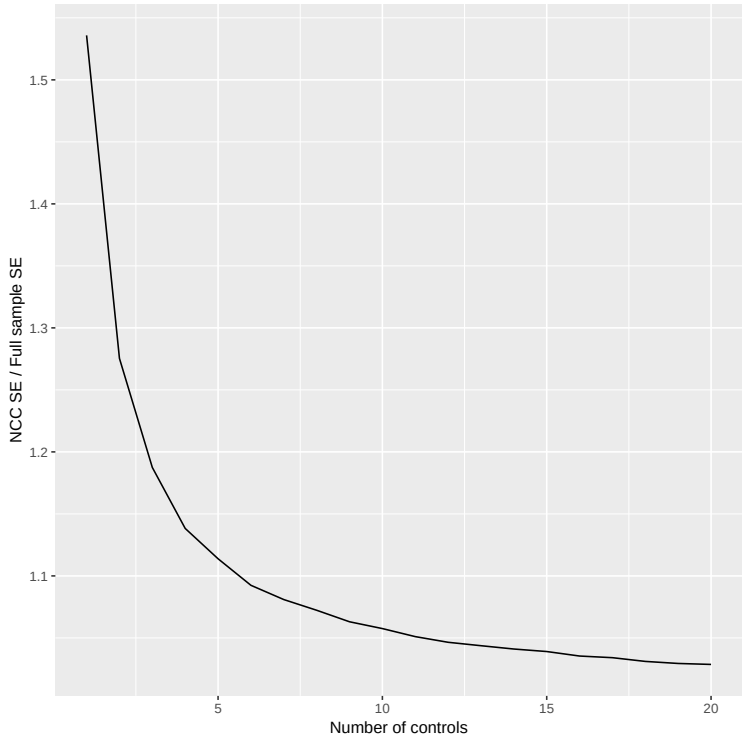

Figure 2: Median relative standard error of NCC parameter estimate to full cohort parameter estimate.