Survival Analysis for Case-Cohort Studies

Petr Klášterecký

Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
e-mail: petr.klasterecky@matfyz.cz

Abstract: This paper gives an overview of some data analysis methods commonly used in survival analysis and their modifications for case-cohort studies. Most attention is paid to the Cox proportional hazards model for studying relative risk in failure-time studies.

Key Words: survival analysis, case-cohort studies, Cox proportional hazards model, partial likelihood

Introduction

In epidemiological and clinical research the so called cohort studies are routinely conducted to assess the effects of various covariates on a failure time. After being randomly selected, the cohort of subjects is typically followed for a given time period, the covariates are measured according to the study design and the failure or censoring time is recorded for each individual. The analysis is then performed using classical techniques such as the Cox regression, accelerated failure time models etc. The main drawback of this type of study is its cost when following a large number of subjects, since the covariates of interest may include e.g. expensive medical laboratory tests. If, however, the disease prevalence is very low in the population, it would be necessary to conduct a large cohort study to ensure the enrolling of a sufficient number of cases (i.e subjects experiencing the event of interest) so that the analysis would even be meaningful.

Case-cohort studies were introduced by R.L. Prentice [Prentice] in 1986 in order to reduce the cost by, roughly speaking, observing much fewer subjects. Rather than following the whole cohort, a random sample called subcohort of the individuals is selected and followed. In addition to that any subject enters the study when experiencing the event of interest. This way the cases are “sampled” with probability one and their covariate values are recorded at (and potentially also after) their failure times while only a relatively small number of the remaining individuals are followed from the beginning of the study. The setup introduced in [Prentice] also indicates that there is no “intended” information available for the cases up to their failure time unless the case is by chance sampled to the subcohort. The case-cohort study data are then often analysed using modifications of the full-data partial likelihood.

Survival analysis overview

Let us first introduce some standard notation as used e.g. in [Fleming and Harrington]: suppose there are \( n \) subjects in the study and denote for each \( i = 1, \ldots, n \) by \( T_i \) a (non-negative) random failure time, \( C_i \) a (positive) censoring time of the \( i \)th individual, \( X_i = \min(T_i, C_i) \) the censored failure time, \( \delta_i = I_{[X_i = T_i]} \) and \( Z_i(t) \) possibly time-dependent \( p \)-dimensional vector of covariates. Note that we are only able to observe the triplets \( (X_i, \delta_i, Z_i(t)) \), \( i = 1, \ldots, n \). Let us further introduce \( N(t) = I_{[T_{(1)} \leq t, \Delta = 1]} \) (a stochastic counting process indicating if the event of interest has been observed) and the “at risk” process \( Y(t) = I_{[X > t]} \). Note that \( Y(t) = 0 \) implies that a potential failure at time \( t \) cannot be observed. The important inference questions in this setting most often concern the relative risk estimation and the conditional distribution of failure time, given covariates.
The Cox proportional hazards model

The concepts of proportional hazards models and partial likelihood were first presented by D. R. Cox in [Cox 1972] and [Cox 1975]. Since then, these results were reviewed in many books and we shall only point out the basic ideas here.

Assume first for simplicity that no covariates depend on time and let \( S(t|Z) = P[T > t|Z] \) be the conditional survival function. The conditional hazard function is then defined by

\[
\lambda(t|Z) = \lim_{h \downarrow 0} \frac{1}{h} P[t \leq T < t+h|T \geq t, Z]
\]

Now suppose that, for any two covariate values \( Z_1 \) and \( Z_2 \), the associated failure rates have a fixed ratio over time. Then \( \lambda(t|Z_1) = k(Z_1, Z_2)\lambda(t|Z_2) \) where \( k \) is a nonnegative real valued function which does not depend on time - the hazard rates are proportional. If we denote \( \lambda_0(t) = \lambda(t|Z = 0) \) (usually called the baseline hazard rate), then

\[
\lambda(t|Z) = \lambda_0(t)k(Z, 0) = \lambda_0(t)g(Z).
\]

The function \( g(Z) \) must not be negative (we are modelling a hazard rate) with \( g(0) = 1 \) and its dependence on \( Z \) should not be too complicated in order to keep the model practically useful. The dependence on \( Z \) is thus most often modelled through a linear term \( \beta^T Z = \beta_1 Z_1 + \cdots + \beta_p Z_p \) while \( g \) is usually taken to be \( g(x) = e^x \). The model (1) then becomes

\[
\lambda(t|Z) = \lambda_0(t) \exp(\beta^T Z),
\]

which is the most common form of the proportional hazards model. For \( \beta_j, \ j = 1, \ldots, p \) now \( e^{\beta_j} \) shows directly the relative risk of failure for an individual with covariate vector \( Z_1, \ldots, Z_{j-1}, Z_j + 1, Z_{j+1}, \ldots, Z_p \) as compared to an individual with covariates \( Z_1, \ldots, Z_{j-1}, Z_j, Z_{j+1}, \ldots, Z_p \).

The conditional survival function \( S(t|Z) \) is now modelled as

\[
S(t|Z) = \{S_0(t)\}^{\exp(\beta^T Z)}
\]

because of the relationship \( S(t|Z) = \exp(- \int_0^t \lambda(u|Z)du) \). Note that time-dependent covariates can be introduced into the proportional hazards model simply through a function

\[
\lambda(t|Z(t)) = \lim_{h \downarrow 0} \frac{1}{h} P[t \leq T < t+h|T \geq t, Z(t)].
\]

However, the function \( \lambda(t) \) defined in (3) can no longer be viewed as a traditional hazard function and cannot be used to compute a conditional survival function \( S(t|Z(t)) \).

Partial likelihood

The vector of parameters, \( \beta \), has to be estimated from the observed data in the proportional hazards model (1) or (2), respectively. The method of maximum likelihood is a very popular method of estimating parameters in classical statistics such as generalized linear models. If, however, the unknown baseline hazard function (rate) \( \lambda_0(t) \) is present, the full likelihood approach would become very cumbersome. Therefore the so called partial likelihood is used and will be presented in this section.

Suppose \( X \) is a random vector with a density \( f_X(x, \theta) \), where \( \theta \) is a vector parameter \( (\phi, \beta) \) with \( \phi \) being a nuisance parameter. The full likelihood approach would require joint maximization over the whole parameter vector \( \theta \) which may not be feasible in high dimensions. The idea behind the partial likelihood approach is that the joint likelihood may often be written as a product of two parts; one of them depends on \( \beta \) only (does not involve \( \phi \)) and only this part is used for inference.
about $\beta$. The latter one may depend on both $\beta$ and $\phi$ - this part of the information about $\beta$ is thus discarded. When this idea is applied to the failure time variables with covariates it follows that if there are no ties in the observed failure times, then the partial likelihood is given by

$$ L \prod_{i=1}^{L} \frac{\exp(\beta^T Z_i(t))}{\sum_{i' \in R_i} \exp(\beta^T Z_i(t))}, $$

where $L$ denotes the total number of observed failure times, $R_i$ is the set of individuals who remain in risk (the risk set) and $Z_i(t)$ is the vector of covariates for the individual who fails as the $i$th one. One can see that together with the covariates the only information preserved and used from the original data are in fact the ranks of the original failure times. All other information (including for example the actual failure/censoring times) is discarded. A more general formula allowing for ties in the observed failure times or stratification can be found in [Fleming and Harrington], pp. 143-144. Using the counting process notation, the general (ties and time-dependent covariates allowing) formula for the partial likelihood is given by

$$ (4) \prod_{i=1}^{n} \prod_{s \geq 0} \left\{ \frac{Y_i(s) \exp(\beta^T Z_i(s))}{\sum_{j=1}^{n} Y_j(s) \exp(\beta^T Z_j(s))} \right\} \Delta N_i(s). $$

**Statistical inference based on partial likelihood**

Parameter estimates $\hat{\beta}$ can be obtained by setting the partial derivatives of (4) with respect to $\beta$ equal to 0 and solving the given system of equations. In order to keep the formulas as comprehensive as possible let us now introduce the following notation:

$$ S^{(0)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) \exp(\beta^T Z_i(t)) $$

$$ S^{(1)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Z_i(t) Y_i(t) \exp(\beta^T Z_i(t)) $$

The partial likelihood (4) may be also viewed as the value at $t = \infty$ of the process

$$ L(\beta, t) = \prod_{i=1}^{n} \prod_{0 \leq u \leq t} \left\{ \frac{Y_i(u) \exp(\beta^T Z_i(u))}{\sum_{j=1}^{n} Y_j(u) \exp(\beta^T Z_j(u))} \right\} \Delta N_i(u) $$

implying that the score vector $U(\beta) = \frac{\partial}{\partial \beta} \log L(\beta, \infty)$ is the value at $t = \infty$ of the process

$$ U(\beta, t) = \sum_{i=1}^{n} \int_{0}^{t} \{ Z_i(u) - E(\beta, u) \} dN_i(u), $$

where

$$ E(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} $$

can be thought of, roughly speaking, as the expected vector of covariates for the items still remaining at the risk set at time $t$. This way one can view the score vector as a kind of comparison of the observed and expected values of covariates. Using martingale methods one can prove under certain regularity conditions consistency of the partial maximum likelihood estimator $\hat{\beta}$ and the weak convergence of the normalized score process to a Gaussian stochastic process, so that there are some asymptotic results available similar to the ones from the classical maximum likelihood estimation.
Modifications for case-cohort studies

The usual methods based on the partial likelihood require the knowledge of covariates for all subjects which of course is not available under the case-cohort design. Remind that the only information that can be used here are the complete covariate histories for the controls (and cases, depending on further assumptions), covariate values for the cases at failure times and maybe some aditional values such as age, gender or some database entries for the whole cohort. It is thus clear that some modifications are needed in order to adapt the methods described above to the case-cohort problem.

Only the cases contribute directly to the summation in (5), whereas the controls’ influence is hidden in the at-risk covariate averages \( E(\beta, t) \). It is, however, this value of \( E \) that causes trouble because under the case-cohort design it contains unobserved data and thus cannot be calculated. Instead, the so called pseudoscores, which rely on the available data only, are used:

\[
U_C(\beta, t) = \sum_{i=1}^{n} \int_0^t \{ Z_i(u) - E_C(\beta, u) \} \, dN_i(u),
\]

where the case-cohort at risk average is given by \( E_C(\beta, t) = S_C^{(1)}(\beta, t) / S_C^{(0)}(\beta, t) \) with

\[
S_C^{(0)}(\beta, t) = n^{-1} \sum_{i=1}^{n} g_i(t) Y_i(t) \exp(\beta' Z_i(t))
\]

\[
S_C^{(1)}(\beta, t) = n^{-1} \sum_{i=1}^{n} g_i(t) Z_i(t) Y_i(t) \exp(\beta' Z_i(t)).
\]

The parameter vector \( \beta \) is then estimated by solving the equation \( U_C(\beta, \infty) = 0 \), the estimators defined in this way were shown to be consistent, asymptotically normal and their asymptotic variance can be estimated directly from the variance estimator of the score function \( U(\beta) \) (see [Prentice] for special cases and [Kulich and Lin] for a general treatment). The asymptotic distribution theory is further studied in [Self and Prentice] using a combination of martingale and finite population convergence results. Among others, the sufficiency conditions for validity of the asymptotic results for the full-cohort and the case-cohort estimators are compared there. A robust approach to the variance estimation is presented in [Barlow] - the variance estimator suggested there also falls into the class of so-called “sandwich estimators”. Still another approach to the estimation of the variance of \( \hat{\beta} \) which utilizes bootstrapping can be found in [Wacholder et al.]. All these results should be compared with extra caution, since the basic setup usually differs slightly from paper to paper. For example, the estimator studied in [Self and Prentice] differs slightly from the maximum pseudolikelihood estimator defined in [Prentice] because it is not based on the same pseudolikelihood function, both estimators are however shown to be quite generally asymptotically equivalent.

The weights \( g_i(t) \) eliminate the subjects with incomplete data by setting \( g_i(t) = 0 \) for such individuals; various proposals for the choice of \( g_i(t) \) have been published leading to different parameter estimators. If we denote by \( \alpha \) the sampling probability for each subject, by \( \xi_i = 1 \) if the subject \( i \) was sampled and \( \xi_i = 0 \) otherwise, then the weights usually involve a factor \( \xi_i/\alpha \) for two reasons. It eliminates subjects that were not selected to the subcohort and at the same time it allows a natural interpretation that each of the selected subjects represents a given proportion of individuals in the original cohort. If the true sampling probability \( \alpha \) is unknown, then its (possibly time-varying) estimator \( \hat{\alpha} \) can be used; the simplest estimator for \( \alpha \) is naturally the empirical proportion of sampled subjects. According to [Robins...
et al.], the resulting estimator can be more efficient when the estimator \( \hat{\alpha} \) is used even if the true value of \( \alpha \) is known.

In general there are two ways of treating the cases, resulting in two different classes of estimators which are being referred to as the N- and D-estimators, respectively. N-estimators treat the cases as if they were sampled with probability one at their failure time, but no covariates are recorded for them before this failure time unless they happen to fall into the subcohort. The subcohort is treated as a sample from all study objects regardless of the failure status and a natural view of the whole procedure is prospective. Thus, the general weights for an N-estimator are given by

\[ \varrho_i(t) = \xi_i / \alpha(t) \text{ for } t \leq T_i; \quad \varrho_i(T_i) = 1, \]

where the sampling probability of the individuals, \( \alpha(t) \), can be replaced by its estimator \( \hat{\alpha}(t) \). With this approach, the cases are weighted by 1 while the controls by the inverse of their (estimated) sampling probabilities.

Unlike N-estimators, the D-estimators work with the original cohort stratified according to failure status, cases forming a separate stratum “sampled” with probability one. Given the failure status, the sampling probabilities \( \alpha(t) \) only apply to the set of controls and a D-estimator is defined by weights

\[ \varrho_i(t) = \delta_i + (1 - \delta_i)\xi_i / \alpha(t), \]

where the controls’ sampling probabilities \( \alpha(t) \) can again be estimated from the data. This approach can be naturally viewed as a retrospective one, mainly due to stratification on the failure status.

Note that stratification on some other discrete variables can be induced with no principal complexity. Such stratification will just likely affect the sampling probabilities \( \alpha_k(t), k = 1, \ldots, K \), which may now be different in each of the \( K \) strata. As a consequence, the weights may also differ and more caution is thus needed in notation and computations.

Concluding remarks

In this paper a brief overview of some survival analysis methods used in case-cohort studies was presented. The original Prentice’s estimator belongs to the class of N-estimators and its basic asymptotic properties (such as asymptotic normality) were derived in [Prentice]. So far, not much attention was paid to certain topics, e.g. the effective use of the values of covariates, which are available for subjects not selected to the subcohort. A generalization addressing (among others) this issue is a class of doubly-weighted estimators introduced and studied in [Kulich and Lin].

An interesting open problem is also how to merge several subcohorts independently sampled from the same cohort. It is common that multiple subcohorts are drawn from a large study population to investigate the effects of different exposures on the same censored outcome. It would be useful to combine the covariate data sampled in this way in order to perform a single analysis of the joint effect of all such exposures. In this problem, only the cases have complete covariate information. A careful use of the available partial covariate information for the rest of the cohort will be necessary to solve this problem.

References


