cchs: An R Package for Stratified Case-Cohort Studies

by Edmund Jones

Abstract The **cchs** package contains a function, also called `cchs`, for analyzing data from a stratified case-cohort study, as used in epidemiology. For data from this type of study, `cchs` calculates Estimator III of Borgan et al. (2000), which is a score-unbiased estimator for the regression coefficients in the Cox proportional hazards model. From the user's point of view, the function is similar to `coxph` (in the **survival** package) and other widely used model-fitting functions. Convenient software has not previously been available for Estimator III since it is complicated to calculate. SAS and S-Plus code-fragments for the calculation have been published, but `cchs` is easier to use and more efficient in terms of time and memory, and can cope with much larger datasets. It also avoids several minor approximations and simplifications.

Introduction

One common type of study in epidemiology is the cohort study, in which participants are recruited and followed over time. Participants who experience a certain type of event (e.g., a heart attack or the diagnosis of a disease) are called cases, and participants who do not are called non-cases or controls. A related type of study is the case-cohort study, which is nested within a cohort study. In a case-cohort study, some of the covariates are only measured for the cases and the subcohort—the subcohort is a randomly selected subset of the cohort. These covariates are not measured for the other participants. A case-cohort study can often estimate effect-sizes almost as accurately as the corresponding cohort study, but at much lower cost since measuring all the covariates on all the participants is expensive.

The case-cohort dataset contains only the cases and the subcohort, and is usually analyzed with a Cox proportional hazards regression model (Cox, 1972). A common way to fit the Cox model is to use the estimator of Prentice (1986), who proposed the case-cohort design. More advanced kinds of analysis sometimes use information on the full cohort, such as the values of those covariates that were measured for all the participants.

In a stratified case-cohort study, the selection of the subcohort is stratified. The cohort is divided into strata and each stratum is assigned a sampling fraction, which is the number of participants in that stratum who will be in the subcohort divided by the total number of participants in the stratum. The purpose of the stratification is usually to improve the efficiency of the estimator. To fit a Cox model to data from a stratified case-cohort study, one possible estimator is the time-fixed version of Estimator III of Borgan et al. (2000). This estimator is optimal in the sense that it is score-unbiased; this criterion is explained in the section "Comparison with other estimators". In this article, "Estimator III" means the time-fixed version of Borgan et al.'s Estimator III, unless otherwise stated.

The subject of this article is the **cchs** package (version 0.4.0; Jones, 2017), which has made Estimator III available in a convenient form. The package consists of a single function, also called `cchs`, which calculates the estimator, and an example dataset, called `cchsData`.

Previous software for calculating Estimator III includes SAS code (SAS Institute Inc., 1999) that appears in Langholz and Jiao (2007a) and S-Plus code (Tibco Software Inc., 2007) in the appendix of Cologne et al. (2012). S-Plus is no longer available for purchase but is very similar to R; Cologne et al. (2012)'s code works in R if one function, `match.data.frame`, is imported from S-Plus.

The previously published SAS and S-Plus code-fragments both analyze specific datasets with specific models. They could be rewritten to work more generally, but even then with some datasets they would give slightly inaccurate results, and with others they would use far too much computational time or memory to be usable, as explained later in the article. Langholz and Jiao wrote more general SAS code (available at Langholz and Jiao, 2007b), but this still has the problems with inaccurate results and computational resources and is less easy to use than `cchs`. Cologne et al. (2012) state that Estimator III can be calculated by an application called Epicure, which is now sold by Risk Sciences International. The estimator was also calculated by Borgan et al. themselves, but their code is not publicly available.

The `cch` function in the **survival** package (Therneau, 2017; Therneau and Grambsch, 2000) can calculate several estimators for case-cohort data, including the estimator of Prentice (1986) and Estimators I and II of Borgan et al. (2000). The functionality of `cchs` could be added to `cch` as a single additional option. But Estimator III is complicated to calculate and `cchs` has numerous features that `cch` lacks. Other software for stratified case-cohort studies includes the **survey** and **NestedCohort** packages, which use inverse probability weights (Lumley, 2004, 2017a; Mark and Katki, 2006; Katki and Mark, 2008), and the code for Estimator II in Samuelsen et al. (2007).

The next five sections discuss the model, the estimator, and how to use `cchs`. The function is easy to use for anyone who is accustomed to the syntax of model-fitting functions in R such as `lm` or `coxph`. The subsequent sections discuss how `cchs` works internally compared to the previously published SAS and S-Plus code. The previously published code makes several approximations and is unable to cope with large datasets, but `cchs` avoids these drawbacks by manipulating the data in an economical way, leading to huge savings in computational time and memory.

Parts of `cchs` are based on code from `coxph`, `cch`, and the appendix of Cologne et al. (2012), and the mathematical formulas and some of the computational methods are from Therneau and Li (1999), Borgan et al. (2000), Langholz and Jiao (2007a), and Cologne et al. (2012), but a major part of the computational methods is original, as discussed in the section "How cchs works".

The idea of creating `cchs` came from work on EPIC-InterAct (InterAct Consortium, 2011) and EPIC-CVD (Danesh et al., 2007), two stratified case-cohort studies that are nested within the cohort from EPIC, the European Prospective Investigation of Cancer and Nutrition (Bingham and Riboli, 2004). EPIC-InterAct is a study of incident type 2 diabetes with data from 29 centers in eight European countries, where the subcohort was stratified by center or country. EPIC-CVD is a study of cardiovascular disease that uses most of the same centers as EPIC-InterAct and the same subcohort in those centers. In Jones et al. (2015), data from EPIC-InterAct was used to compare five different models for data from stratified case-cohort studies, and a pre-release version of `cchs` was used to fit two of those models.

The model and the estimator

The Cox model for survival-time data is commonly defined by this equation:

$$h_i(t) = h_0(t) \exp(\beta^\top z_i),$$

where

h_i is the hazard function for participant i,

t is the time to the event,

h_0 is the baseline hazard function,

β is the regression coefficients (log hazard ratios), and

z_i is the covariates for participant i.

If all the event-times are distinct, then the partial likelihood is

$$\prod_j \frac{\exp(\beta^\top z_{i_j})}{\sum_k Y_k(t_j) \exp(\beta^\top z_k)},$$

where

t_j is the jth event-time,

i_j is the participant whose event happened at time t_j,

$$Y_k(t_j) = \begin{cases} 1 & \text{if } t_j \in (t_{0k}, t_{1k}] \\ 0 & \text{otherwise} \end{cases},$$

t_{0k} is the entry-time for participant k (often zero for all participants), and

t_{1k} is the exit-time for participant k (i.e. the time at which they had the event or were censored).

The product is over all the cases and the sum is over all the participants. The time-interval $(t_{0k}, t_{1k}]$ is participant k's "at-risk time" and $Y_k(t_j)$ is an indicator variable for whether participant k is at risk at time t_j.

For case-cohort data, z_i is only known for the cases and subcohort members, so the partial likelihood cannot be evaluated and it is impossible to estimate the regression coefficients using the method of maximum likelihood. Instead, the coefficients are estimated by maximizing pseudo-likelihoods, which are approximations of the partial likelihood. The pseudo-likelihood for Estimator III is

$$\prod_j \frac{\alpha_{s(i_j)}^{-1} \exp(\beta^\top z_{i_j})}{\sum_{k \in R(t_j)} Y_k(t_j) \alpha_{s(k)}^{-1} \exp(\beta^\top z_k)},$$

where

$s(i)$ is the stratum that contains participant i,

α_s is the sampling fraction for stratum s (the proportion of stratum s that appears in the subcohort),

$$R(t_j) = \begin{cases} C & \text{if } i_j \in C \\ C \cup \{i_j\} \setminus \{J_{s(i_j)}\} & \text{if } i_j \notin C' \end{cases}$$

C is the subcohort, and

J_s is a participant randomly selected from $C \cap s$ (where s is a stratum).

Borgan et al. (2000) called this a "swapper" method because of the way $R(t_j)$ is defined for $i_j \notin C$. As shown above, $J_{s(i_j)}$ is removed from C and i_j is added. If $J_{s(i_j)}$ is not at risk at time t_j, then its removal from $R(t_j)$ makes no difference to the pseudo-likelihood since $Y_{J_{s(i_j)}}(t_j) = 0$. It is recommended to use an asymptotic covariance estimator rather than a robust one (Jiao, 2001); see the section "The calculation of the covariance matrix".

Langholz and Jiao (2007a) discuss two situations in which a case-cohort study might be stratified, which they call "exposure stratified" and "confounder stratified." In the exposure stratified situation, the strata are defined by levels of one or more covariates, and the purpose of the stratification is to increase the efficiency of the estimator. In the confounder stratified situation, the strata are defined by a covariate that is believed to confound the relation between the exposure of interest and the event-time (e.g., the center in a multi-center study). For an exposure stratified study, it is usual to fit a single unstratified Cox model to the data, and this is is the situation for which Estimator III was designed. For a confounder stratified study, it is more natural to fit a stratified Cox model—in which each stratum s has its own baseline hazard $h_{0s}(t)$ and $h_i(t) = h_{0s(i)}(t)\exp(\beta^\top z_i)$—and for this, the estimator of Prentice (1986) should be used; but it is still possible to fit an unstratified model using Estimator III.

Comparison with other estimators

Borgan et al. (2000) described three estimators for the Cox model with data from a stratified case-cohort study. These have been investigated in several simulation studies: Borgan et al. used cohorts of size 1000 with subcohorts of size 100, and found that there was little to choose between Estimators II and III, but both performed better than Estimator I; Cologne et al. (2012) made similar findings but reported that Estimator III did slightly better than Estimator II when the sampling fraction was very small; and Samuelsen et al. (2007) found that Estimator II performed well.

Of the three estimators, only Estimator III is score-unbiased. This is one of the main desirable criteria for estimators for case-cohort data. It means that the expectation of the pseudo-score (the derivative of the pseudo-likelihood) is zero when the coefficients take their true values (Godambe, 1960, 1976; Lindsay, 1982; Severini, 2011; Cologne et al., 2012). On the other hand, Estimator II has smaller asymptotic variance (this was hinted at in Borgan et al., 2000, stated in Samuelsen et al., 2007, and explained most clearly in Samuelsen et al., 2006).

Recall that "Estimator III" refers to the time-fixed version of that estimator, as calculated by `cchs`. Borgan et al. also described time-dependent versions of the three estimators, which have smaller asymptotic variance (Kulich and Lin, 2004). The time-dependent version of Estimator III is score-unbiased (Borgan et al., 2000), but it is even more complicated to calculate than the time-fixed version.

Estimator II can be improved using covariate data on the whole cohort, if that is available (Breslow et al., 2009a,b; Yan et al., 2017). The estimator of Prentice (1986) can also be adapted to work with stratified case-cohort studies. Its pseudo-likelihood is

$$\prod_j \frac{\exp(\beta^\top z_{i_j})}{\sum_{k\in C\cup\{i_j\}} Y_k(t_j)\exp(\beta^\top z_k)}.$$

This estimator is only score-unbiased if for $i_j \notin C$ the C in the above expression is replaced by $(C \cap s(i_j))$. This is less efficient than Estimator III since the denominator includes fewer participants.

For stratified case-cohort data it is also possible to use general methods based on inverse probability weights. These methods are suitable for a wide range of complex sampling schemes (Mark and Katki, 2006; Lumley, 2017b), but no claim is made that the estimators are score-unbiased. This approach has been used by Gray (2009), Liu et al. (2012), and Payne et al. (2016). In other research, Kulich and Lin (2004) presented a framework that covered many case-cohort estimators and the possibility of stratified subcohort-selection, and Zeng and Lin (2007) and Kim et al. (2013) described wide classes of sampling schemes and regression models for survival analysis, of which the stratified case-cohort design and the Cox model are special cases.

How to use cchs

From the user's point of view, cchs is reasonably similar to coxph. The most important arguments are:

formula, which specifies the model (the left-hand side must be a "Surv" object),

data, a data frame or environment that contains the variables in the formula,

inSubcohort, an indicator for whether each participant is in the subcohort,

stratum, the stratum variable,

samplingFractions, the sampling fraction for each stratum,

cohortStratumSizes, the size of each stratum in the full cohort, and

precision, discussed below.

If the data is supplied as a data frame, it is assumed to be in the usual form where each row corresponds to one observation (i.e., one participant in the case-cohort study) and each column corresponds to a variable. The inSubcohort and stratum arguments can either be columns or elements of data or separate objects that are named in the call to cchs.

One of samplingFractions or cohortStratumSizes must be specified, but not both. If the latter is specified, then cchs uses it to calculate the former—for each stratum, it divides the size of the subcohort by the size of the cohort, which is given by cohortStratumSizes. If the sampling fractions are not stored as a column of data, then they can be passed to cchs as a named vector. For example, if the stratum variable takes the values "London" and "Paris" and the sampling fractions are 0.1 and 0.2 respectively, then this vector should be c(London = 0.1, Paris = 0.2).

The precision argument must be set if there are any tied event-times. Estimator III requires all event-times to be distinct, so cchs changes tied event-times by small amounts, as suggested by Langholz and Jiao (2007a), and it uses precision to work out suitable values. If the times are all integers, then precision should be set to 1. In datasets from EPIC-InterAct and EPIC-CVD, the times are recorded to the nearest day but stored as numbers of years, so precision should be set to 1 / 365.25.

Additional arguments are discussed in "Obscure aspects of the calculation".

An example analysis

cchsData is an artificial stratified case-cohort dataset that was created from the nwtco dataset in the **survival** package. The original nwtco data comes from two clinical trials run by the National Wilms Tumor Study Group in the United States (D'Angio et al., 1989; Green et al., 1998). Different artificial case-cohort datasets based on nwtco have been used by Kulich and Lin (2004), Breslow et al. (2009a), and Breslow et al. (2009b).

In cchsData, the event is relapse of Wilms tumor and the subcohort-selection is stratified by the result of a histological examination. The most important variables are time (the time to the event or censoring), isCase (the event-indicator), inSubcohort (the indicator for being in the subcohort), localHistol (the stratum variable), and sampFrac (the stratum-specific sampling fractions). The times are stored as numbers of days, and some of them are tied, so the precision argument of cchs should be set to 1 (or 1/2, 1/3, etc.).

The sampling fractions in cchsData are 5% and 20%. The subcohort is about half as big as the set of cases. In case-cohort studies in general, the sampling fraction is typically in the area of 5% but sometimes larger (Sharp et al., 2014), and the subcohort is usually larger than the set of cases (Juraschek et al., 2013; Lamb et al., 2013; Jones et al., 2015) but occasionally smaller (Huxley et al., 2013).

The following command uses Estimator III to fit a Cox model to cchsData. The covariates are age at diagnosis (ageAtDiagnosis) and stage of cancer at diagnosis (stage):

```
> cchs(Surv(time, isCase) ~ ageAtDiagnosis + stage, data = cchsData,
+ inSubcohort = inSubcohort, stratum = localHistol,
+ samplingFractions = sampFrac, precision = 1)
```

The cchs function returns an object of S3 class "cchs" that contains many of the same elements as a "coxph" object and several additional elements. When a "cchs" object is printed, information about the dataset and the changes to the tied event-times appears before the coefficients. The output from the above command is:

```
Call: cchs(formula = Surv(time, isCase) ~ stage + ageAtDiagnosis,
    data = cchsData, inSubcohort = inSubcohort, stratum = localHistol,
    samplingFractions = sampFrac, precision = 1)
Number of observations/rows: 785, of which
        subcohort non-cases: 214
        subcohort cases:     48
        non-subcohort cases: 523
179 of 571 discretized event-times were changed by up to 0.01 to deal
 with ties.
Number of strata for subcohort-selection: 2
Coefficients (HR = hazard ratio, CI = 95% confidence interval, SE = standard
error):
                     HR   CIlower  CIupper           p     logHR    SElogHR
ageAtDiagnosis 1.101685 1.029489 1.178944 0.005104173 0.09684087 0.03458127
stageII        1.397695 0.915168 2.134635 0.121219767 0.33482413 0.21606100
stageIII       1.794611 1.156081 2.785816 0.009150320 0.58478850 0.22436755
stageIV        2.367953 1.382691 4.055282 0.001686931 0.86202589 0.27449190
```

This output can be interpreted like output from any other function that fits the Cox model. For example, the hazard is estimated to be 79% higher for patients who were at stage III when diagnosed compared to patients who were at stage I, adjusting for age at diagnosis. The level for the confidence intervals in the output is determined by the confidenceLevel argument to cchs, whose default is 0.95.

If the output of cchs is stored as an object called result, then the coefficients are available as result$coefficients and the covariance matrix is available as result$var. The table at the bottom of the output is stored as result$coeffsTable for convenience. For example, this can be used to get the confidence intervals for the hazard ratio for ageAtDiagnosis:

```
> result <- cchs(Surv(time, isCase) ~ ageAtDiagnosis + stage, data = cchsData,
+ inSubcohort = inSubcohort, stratum = localHistol,
+ samplingFractions = sampFrac, precision = 1)
> result$coeffsTable["ageAtDiagnosis", c("CIlower", "CIupper")]
```

Other issues for the user

Calling cchs twice with the same arguments will not necessarily give exactly the same results, which is unusual for estimators that maximize a likelihood or pseudo-likelihood. This happens for two reasons: first, Estimator III involves randomness in the choice of $J_{s(i_j)}$; and second, if there are tied event-times, these are changed in a random way. To get exactly the same results, set the random seed (using set.seed) just before the call to cchs.

Another approach would be to call cchs multiple times, using the replicate command, then combine the results, perhaps by using Rubin's rules (Rubin, 1987; Schafer, 1999), which are normally used for multiple imputation. With this approach, for each term in the model the coefficient estimate is the mean of the separate estimates, and the variance is the mean of the separate variances plus a term to account for the between-replication variability. It might be worth expressly comparing the within-replication variability (the separate variance estimates) with the between-replication variability (the variability due to the randomness of the estimator). Figure 1 shows the point-estimates and confidence intervals that appear in the previous section, and the quartiles and ranges of the point-estimates produced by 1000 calls to cchs with the same arguments. Here, the between-replication variability is small, and very small compared to the within-replication variability.

In coxph, model formulas can contain certain special terms. Of these, cchs allows offset terms, but it does not allow cluster (for identifying correlated groups of rows), strata (for fitting a stratified Cox model), or tt (for time-varying covariates). If the data are from a stratified case-cohort study and it is desired to fit a stratified Cox model where the strata in the baseline hazard are the same as the strata for the subcohort-selection, then Prentice's estimator should be used (Langholz and Jiao, 2007a).

All error and warning messages produced by cchs are designed to be meaningful. This is achieved by checking the arguments and raising an error if any of them have illegal values, and by adding extra messages to any errors or warnings raised by model.frame or coxph when they are called internally by cchs. The addition of the extra messages results in the call stack, which can be viewed by typing traceback(), being longer and more complicated than usual. Therefore if traceback() is going to be used to diagnose problems, it is advisable to switch the extra messages off by setting annotateErrors = FALSE.

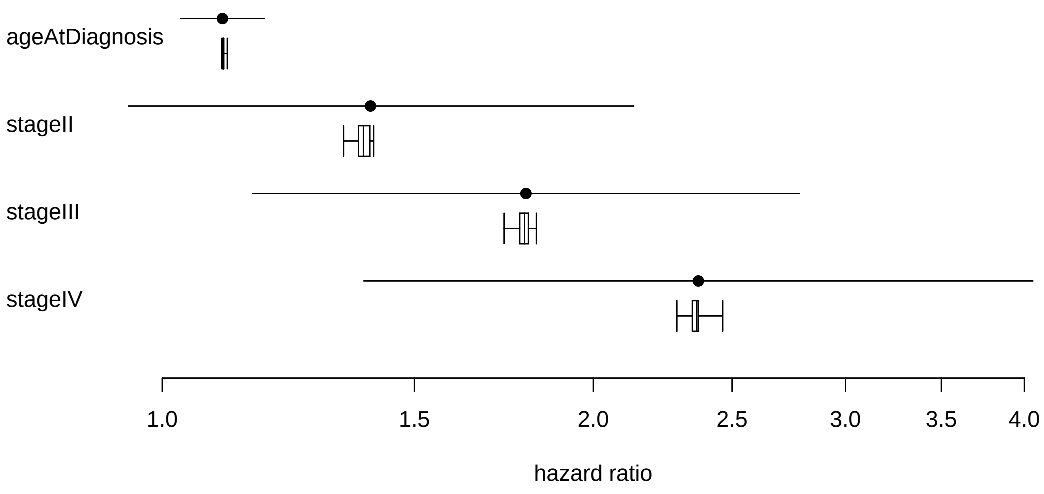

Figure 1: The dots and horizontal lines show the point-estimates and confidence intervals from the output in "An example analysis"; this is within-replication variability. The boxes and whiskers show the hazard ratios from 1000 replications of cchs, using the dataset and model in the same section; this is between-replication variability. The boxes show the quartiles and the whiskers show the full ranges.

How cchs works

The cchs function works differently from the previously published SAS and S-Plus code in several ways, and these novel aspects of its implementation affect the amount of computational time and memory that it uses and enable it to work with much bigger datasets. They also affect the numbers in the output. These issues are discussed here for cautious or curious users who want to know how cchs works or compare it with other software for Estimator III.

Functions that calculate estimators for the Cox model with case-cohort data invariably work by manipulating the data or the model-formula in various ways—for example, duplicating rows and changing entry-times, or appending weights to the rows and including the weight in the model-formula—and then passing the data and model-formula to a second model-fitting function (in R, coxph) whose original purpose is to fit the Cox model with the standard estimator. The manipulation is done in such a way that when the second function attempts to maximize the partial likelihood, what it actually maximizes is the pseudo-likelihood for the case-cohort estimator. The second function returns the estimates of the regression coefficients and their covariance matrix, and the top-level function usually then makes adjustments to the covariance matrix.

This method works because the standard partial likelihood and the various pseudo-likelihoods all have similar forms. They have the same number of multiplicative terms, one for each case, and the same values in the numerators (apart from multiplicative factors, which are easy to deal with). The only differences are the sets of participants who appear in the denominators and, for some pseudo-likelihoods, multiplicative factors in the denominators. These sets are dealt with by duplicating or excluding rows and changing entry- and exit-times, and any multiplicative factors are dealt with by using an offset term in the model-formula.

For Estimator III, the manipulations are rather complicated because of the swapping. The cchs function manipulates the data as follows:

1. Define a small number ϵ, which is smaller than any of the differences between consecutive event-times.
2. For each $i_j \notin C$, change the entry-time to $t_j - \epsilon$, where t_j is the exit-time.
3. For each stratum s:

   3.1 Choose a random participant J_s from $C \cap s$. Denote the entry-time for J_s by t_{0s} and the exit-time by t_{1s}.

   3.2 Let $U_s = \{t_j : i_j \in s, i_j \notin C, t_{0s} < t_j \leq t_{1s}\}$. Sort U_s into increasing order as $u_{s(1)}, \ldots, u_{s(|U_s|)}$.

   3.3 Replace J_s's row by $|U_s| + 1$ new rows with variables as follows.

   Row 1: entry-time t_{0s}, exit-time $u_{s(1)} - \epsilon$, event-indicator zero.
   Row 2: entry-time $u_{s(1)}$, exit-time $u_{s(2)} - \epsilon$, event-indicator zero.

...

Row $|U_s|$: entry-time $u_{s(|U_s|-1)}$, exit-time $u_{s(|U_s|)} - \epsilon$, event-indicator zero.

Row $|U_s| + 1$: entry-time $u_{s(|U_s|)}$, exit-time t_{1s}, event-indicator as in the original row.

These steps are best understood by referring to the earlier definition of $R(t_j)$ and bearing in mind that coxph takes a row's at-risk time to be the half-open interval, (entry-time, exit-time]. The purpose of step 1 is to choose ϵ to be sufficiently small that steps 2 and 3 do not have any unintended effects on other terms in the pseudo-likelihood. Step 2 has the effect of removing the non-subcohort cases from $R(t_j)$, except for i_j itself, which is not removed. Step 3 makes the further changes to $R(t_j)$ that are needed if $i_j \notin C$. Note that it deals with one J_s, not one i_j, at a time. Step 3.1 chooses J_s. Step 3.2 defines U_s to be the event-times that lie in the range of J_s's at-risk time and correspond to non-subcohort cases in s. Step 3.3 excises a short stretch of at-risk time from J_s at each time in U_s by splitting J_s's row into multiple rows with separate stretches of at-risk time.

The effect of step 3 is to remove $J_{s(i_j)}$ from $R(t_j)$ when $i_j \notin C$, as required by the definition of $R(t_j)$. For a non-subcohort case in s whose event-time is not in U_s, it makes no difference whether $J_{s(i_j)}$ is in $R(t_j)$ or not, because $Y_{J_{s(i_j)}}(t_j)$ is zero; so for such cases this manipulation is not necessary.

One other change to the data is needed. An extra column is created that contains $-\log \alpha_{s(k)}$ for participant k, and this is included as an offset term in the model-formula that cchs passes to coxph. This has the effect of inserting the $\alpha_{s(k)}^{-1}$ factors that are needed in the denominators.

The calculation of the covariance matrix

The cchs function uses an asymptotic covariance estimator, as recommended by Jiao (2001). Internally, after it manipulates the data and calls coxph, the final step is to adjust the covariance matrix. This is done by using the matrix of "dfbeta" residuals, D, produced by `residuals.coxph`; d_{ij} is the change in the estimate of regression coefficient j when row i is dropped from the data (or an approximation of that change). The covariance matrix is adjusted by adding $\sum_s m_s(1-\alpha_s)\,\mathrm{Cov}\,D_{C\cap s}$, where $m_s = |C \cap s|$, $D_{C\cap s}$ is the rows of D that correspond to $C \cap s$, and $\mathrm{Cov}\,D_{C\cap s}$ is the empirical covariance matrix $D_{C\cap s}^{\top} D_{C\cap s}/(m_s - 1)$.

This adjustment corresponds to Equation (17) in Borgan et al. (2000). The use of D is based on the formula for the adjustment to the covariance matrix with unstratified case-cohort data, which appears on page 102 of Therneau and Li (1999) and as Equation (5) in Langholz and Jiao (2007a). There is a small discrepancy between the formulas in these two publications. To explain this, let D_C be the rows of D that correspond to C and let $\alpha = m/n$ be the sampling fraction, where m is the size of the subcohort and n is the size of the cohort. According to Therneau and Li, the quantity to add to the covariance matrix is $(1-\alpha)D_C^{\top} D_C$, but according to Langholz and Jiao it is $\frac{m(n-m)}{n}\,\mathrm{Cov}\,D_C = (1-\alpha)\frac{m}{m-1} D_C^{\top} D_C$. These expressions differ by a factor of $m/(m-1)$, which is unimportant if m is large, but makes a difference if m is small. The cchs function does the calculation in the same way as Langholz and Jiao and Cologne et al. (2012).

How the SAS code and S-Plus code work

The previously published SAS and S-Plus code-fragments for Estimator III manipulate the data in a completely different way from cchs:

1. Define a small number ϵ, which is smaller than any of the differences between consecutive event-times.
2. For each $i_j \notin C$, change the entry-time to $t_j - \epsilon$.
3. Replace the dataset with a new dataset that is constructed as follows. For each i_j:

   3.1 Make one copy of all the rows.

   3.2 Drop those of the new rows that are not at risk at t_j.

   3.3 For all rows, set the entry-time to $t_j - \epsilon$ and the exit-time to t_j, and set the event-indicator to 1 for i_j and 0 for the other rows.

4. For each $i_j \notin C$, choose a random participant $J_{s(i_j)}$ from $C \cap s(i_j)$ and remove the row that contains $J_{s(i_j)}$ and has exit-time t_j.

Step 3 radically changes the dataset and greatly increases its size. The result of the manipulations is that when the data is passed to coxph in S-Plus or `proc phreg` in SAS, each row corresponds to exactly one additive term in the numerator or denominator of one multiplicative term of the pseudo-likelihood. The $\alpha_{s(k)}^{-1}$ terms and the adjustment to the covariance matrix are dealt with in the same way as in cchs.

The advantage of this method is that steps 3 and 4 are relatively simple—the "swapping" in step 4 just consists of choosing and removing a single row each time. The disadvantages are that if the original dataset is medium or large in size then this method uses a gigantic amount of memory and requires coxph to deal with a gigantic dataset. For step 3.1, the S-Plus code creates a data frame that contains one row for every combination of a case and a participant. In EPIC-CVD, a typical dataset that was supplied to an investigator contained about $31,000$ participants, of whom $14,000$ were cases, and this took up about 4.4MB as a binary file. So the data frame created by step 3 would have $31,000 \times 14,000 \approx 434$ million rows and take up about 62GB. This is too big for a desktop computer to store in memory, and even after some rows are removed in step 4, the data frame would be far too big for coxph to process in a reasonable amount of time.

In contrast, when cchs is used on the same dataset, the manipulated dataset has fewer than $31,000 + 14,000 = 45,000$ rows. This $45,000$ was calculated by supposing that for every case an extra row is created, whereas in reality some cases are in the subcohort and it is likely that some "swapper" rows are not at risk at the relevant event-times, and for these no extra rows are created, so the true number of rows in the manipulated dataset is less than $45,000$.

The manipulation of the data by the SAS and S-Plus code can be regarded as maximal, in the sense that each row ultimately corresponds to a single additive term in a single multiplicative term of the pseudo-likelihood. The manipulation by cchs can be regarded as minimal, since it makes only the manipulations that are strictly necessary in order for coxph to calculate the correct pseudo-likelihood, and it leaves the data with the smallest possible number of rows.

Obscure aspects of the calculation

Three obscure issues arise in calculating Estimator III. Under normal circumstances users of cchs can ignore these, but for development or testing it may be necessary to understand these issues or, as described below, set the relevant logical arguments to different values.

The first issue has to do with the dropNeverAtRiskRows argument, which determines whether rows should be dropped if their at-risk time does not contain any of the event-times. The dropped rows make no difference to the regression coefficients output by cchs, but they do affect the covariance-estimates and confidence intervals, because of the approximation that residuals.coxph uses to calculates the dfbeta residuals. The SAS code and S-Plus code drop these rows (in step 3.2), but Borgan et al. (2000) and Langholz and Jiao (2007a) make no mention of this issue. In cchs, dropNeverAtRiskRows is TRUE by default but can be set to FALSE if desired.

The second issue has to do with the dropSubcohEventsDfbeta argument. When the adjustment to the covariance matrix is calculated, the correct way is to use all the rows of the matrix of dfbeta residuals that correspond to subcohort members. The SAS code and S-Plus code both omit the rows that correspond to subcohort cases at their event-times, presumably because this makes the code simpler and is unlikely to change the output greatly. By default, cchs calculates the adjustment to the covariance matrix in the strictly correct way. But if dropSubcohEventsDfbeta is TRUE, then it calculates the adjustment in the same way as the SAS and S-Plus code, as follows. For each $i_j \in C$, it splits that row of the data at $t_j - \epsilon$ and replaces it with two rows; the first row has entry-time t_{0i_j}, exit-time $t_j - \epsilon$, and event-indicator 0; and the second row has entry-time $t_j - \epsilon$, exit-time t_j, and event-indicator 1. When calculating the adjustment to the covariance matrix, cchs excludes the row of the dfbeta residuals that corresponds to the second of these two rows.

The third issue is that the "swapper" $J_{s(i_j)}$ is supposed to be selected once for each stratum (see Section 3 of Borgan et al., 2000), but the SAS and S-Plus both select it once for each case. The cchs function selects the swapper correctly. It has no logical argument to specify how the swapper should be selected, since it manipulates the data in a completely different way from the SAS and S-Plus code.

Discussion

In case-cohort studies, stratified selection of the subcohort seems to be common (Sharp et al., 2014), so there are likely to be plenty of investigations where cchs may be useful. The way in which cchs manipulates the data makes it faster and more computationally efficient than the previously published SAS and S-Plus code. The three obscure issues described in the previous section should not greatly affect most analyses, but will make a difference with small datasets.

Acknowledgments

The author would like to thank Michael Sweeting for his helpful comments. This work was funded by the UK Medical Research Council (G0800270), the British Heart Foundation (SP/09/002), the UK National Institute for Health Research (Cambridge Biomedical Research Centre), the European Research Council (268834), and the European Commission Framework Programme 7 (HEALTH-F2-2012-279233).

Bibliography

S. Bingham and E. Riboli. Diet and cancer—the European Prospective Investigation into Cancer and Nutrition. *Nature Reviews Cancer*, 4(3):206–215, 2004. URL https://doi.org/10.1038/nrc1298. [p485]

Ø. Borgan, B. Langholz, S. O. Samuelsen, L. Goldstein, and J. Pogoda. Exposure stratified case–cohort designs. *Lifetime Data Analysis*, 6(1):39–58, 2000. URL https://doi.org/10.1023/A:1009661900674. [p484, 485, 486, 490, 491]

N. E. Breslow, T. Lumley, C. M. Ballantyne, L. E. Chambless, and M. Kulich. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49, 2009a. URL https://doi.org/10.1007/s12561-009-9001-6. [p486, 487]

N. E. Breslow, T. Lumley, C. M. Ballantyne, L. E. Chambless, and M. Kulich. Using the whole cohort in the analysis of case–cohort data. *American Journal of Epidemiology*, 169(11):1398–1405, 2009b. URL https://doi.org/10.1093/aje/kwp055. [p486, 487]

J. Cologne, D. L. Preston, K. Imai, M. Misumi, K. Yoshida, T. Hayashi, and K. Nakachi. Conventional case–cohort design and analysis for studies of interaction. *International Journal of Epidemiology*, 41(4): 1174–1186, 2012. URL https://doi.org/10.1093/ije/dys102. [p484, 485, 486, 490]

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34(2):187–220, 1972. URL https://www.jstor.org/stable/2985181. [p484]

J. Danesh, R. Saracci, G. Berglund, E. Feskens, K. Overvad, S. Panico, S. Thompson, A. Fournier, F. Clavel-Chapelon, M. Canonico, R. Kaaks, J. Linseisen, H. Boeing, T. Pischon, C. Weikert, A. Olsen, A. Tjønneland, S. P. Johnsen, M. K. Jensen, J. R. Quirós, C. A. G. Svatetz, M.-J. S. Pérez, N. Larrañaga, C. N. Sanchez, C. M. Iribas, S. Bingham, K.-T. Khaw, N. Wareham, T. Key, A. Roddam, A. Trichopoulou, V. Benetou, D. Trichopoulos, G. Masala, S. Sieri, R. Tumino, C. Sacerdote, A. Mattiello, W. M. M. Verschuren, H. B. B. de Mesquita, D. E. Grobbee, Y. T. van der Schouw, O. Melander, G. Hallmans, P. Wennberg, E. Lund, M. Kumle, G. Skeie, P. Ferrari, N. Slimani, T. Norat, and E. Riboli. EPIC-Heart: The cardiovascular component of a prospective study of nutritional, lifestyle and biological factors in 520,000 middle-aged participants from 10 European countries. *European Journal of Epidemiology*, 22(2):129–141, 2007. URL https://doi.org/10.1007/s10654-006-9096-8. [p485]

G. J. D'Angio, N. Breslow, J. B. Beckwith, A. Evans, E. Baum, A. Delorimier, D. Fernbach, E. Hrabovsky, B. Jones, P. Kelalis, H. B. Othersen, M. Tefft, and P. R. M. Thomas. Treatment of Wilms' tumor: Results of the third National Wilms' Tumor Study. *Cancer*, 64(2):349–360, 1989. URL https://doi.org/10.1002/1097-0142(19890715)64:2<349::AID-CNCR2820640202>3.0.CO;2-Q. [p487]

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960. URL https://doi.org/10.1214/aoms/1177705693. [p486]

V. P. Godambe. Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63(2):277–284, 1976. URL https://doi.org/10.1093/biomet/63.2.277. [p486]

R. J. Gray. Weighted analyses for cohort sampling designs. *Lifetime Data Analysis*, 15(1):24–40, 2009. URL https://doi.org/10.1007/s10985-008-9095-z. [p486]

D. M. Green, N. E. Breslow, J. B. Beckwith, J. Z. Finklestein, P. E. Grundym, P. R. Thomas, T. Kim, S. J. Shochat, G. M. Haase, M. L. Ritchey, P. P. Kelalis, and G. J. D'Angio. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the National Wilms' Tumor Study Group. *Journal of Clinical Oncology*, 16(1):237–256, 1998. URL https://doi.org/10.1200/JCO.1998.16.1.237. [p487]

R. R. Huxley, F. L. Lopez, R. F. MacLehose, J. H. Eckfeldt, D. Couper, C. Leiendecker-Foster, R. C. Hoogeveen, L. Y. Chen, E. Z. Soliman, S. K. Agarwal, and A. Alonso. Novel association between plasma matrix metalloproteinase-9 and risk of incident atrial fibrillation in a case-cohort study: The Atherosclerosis Risk in Communities study. *PLOS One*, 8(3):1–8, 2013. URL https://doi.org/10.1371/journal.pone.0059052. [p487]

InterAct Consortium. Design and cohort description of the interact project: An examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC study. *Diabetologia*, 54(9):2272–2282, 2011. URL https://doi.org/10.1007/s00125-011-2182-9. [p485]

J. Jiao. *Comparison of Variance Estimators in Case–Cohort Studies*. PhD thesis, University of Southern California, 2001. URL http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll16/id/242761. [p486, 490]

E. Jones. *cchs: Cox Model for Case–Cohort Data with Stratified Subcohort-Selection*, 2017. URL https://cran.r-project.org/package=cchs. R package version 0.4.0. [p484]

E. Jones, M. J. Sweeting, S. J. Sharp, and S. G. Thompson. A method making fewer assumptions gave the most reliable estimates of exposure–outcome associations in stratified case–cohort studies. *Journal of Clinical Epidemiology*, 68(12):1397–1405, 2015. URL https://doi.org/10.1016/j.jclinepi.2015.04.007. [p485, 487]

S. P. Juraschek, G. P. S. Shantha, A. Y. Chu, E. R. Miller III, E. Guallar, R. C. Hoogeveen, C. M. Ballantyne, F. L. Brancati, M. I. Schmidt, J. S. Pankow, and J. H. Young. Lactate and risk of incident diabetes in a case-cohort of the Atherosclerosis Risk in Communities (ARIC) study. *PLOS One*, 8(1):1–7, 2013. URL https://doi.org/10.1371/journal.pone.0055113. [p487]

H. A. Katki and S. D. Mark. Survival analysis for cohorts with missing covariate information. *R News*, 8(1):14–19, 2008. URL https://www.r-project.org/doc/Rnews/Rnews_2008-1.pdf#section*.31. [p484]

J. P. Kim, W. Lu, T. Sit, and Z. Ying. A unified approach to semiparametric transformation models under general biased sampling schemes. *Journal of the American Statistical Association*, 108(501): 217–227, 2013. URL https://doi.org/10.1080/01621459.2012.746073. [p486]

M. Kulich and D. Lin. Improving the efficiency of relative-risk estimation in case–cohort studies. *Journal of the American Statistical Association*, 99(467):832–844, 2004. URL https://doi.org/10.1198/016214504000000584. [p486, 487]

M. M. Lamb, M. D. Simpson, J. Seifert, F. W. Scott, M. Rewers, and J. M. Norris. The association between IgG4 antibodies to dietary factors, islet autoimmunity and type 1 diabetes: The diabetes autoimmunity study in the young. *PLOS One*, 8(2):1–7, 2013. URL https://doi.org/10.1371/journal.pone.0057936. [p487]

B. Langholz and J. Jiao. Computational methods for case–cohort studies. *Computational Statistics & Data Analysis*, 51(8):3737–3748, 2007a. URL https://doi.org/10.1016/j.csda.2006.12.028. [p484, 485, 486, 487, 488, 490, 491]

B. Langholz and J. Jiao. Case–cohort computation. http://web.archive.org/web/20100720003853/http://hydra.usc.edu/timefactors/Examples/Case-cohort%20computational%20methods/Case-cohort%20computation.html, 2007b. [p484]

B. Lindsay. Conditional score functions: Some optimality results. *Biometrika*, 69(3):503–512, 1982. URL https://doi.org/10.1093/biomet/69.3.503. [p486]

D. Liu, T. Cai, and Y. Zheng. Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics*, 68(4):1219–1227, 2012. URL https://doi.org/10.1111/j.1541-0420.2012.01787.x. [p486]

T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(8):1–19, 2004. URL https://doi.org/10.18637/jss.v009.i08. [p484]

T. Lumley. *survey: Analysis of Complex Survey Samples*, 2017a. URL https://cran.r-project.org/package=survey. R package version 3.32-1. [p484]

T. Lumley. *Two-Phase Designs in Epidemiology*, 2017b. URL https://cran.r-project.org/web/packages/survey/vignettes/epi.pdf. [p486]

S. D. Mark and H. A. Katki. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. *Journal of the American Statistical Association*, 101(474):460–471, 2006. URL https://doi.org/10.1198/016214505000000952. [p484, 486]

R. Payne, M. Neykov, M. K. Jensen, and T. Cai. Kernel machine testing for risk prediction with stratified case cohort studies. *Biometrics*, 72(2):372–381, 2016. URL https://doi.org/10.1111/biom.12452. [p486]

R. L. Prentice. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11, 1986. URL https://doi.org/10.1093/biomet/73.1.1. [p484, 486]

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. URL https://doi.org/10.1002/9780470316696. [p488]

S. O. Samuelsen, H. Ånestad, and A. Skrondal. Stratified case–cohort analysis of general cohort sampling designs. Technical report, Department of Mathematics, University of Oslo, 2006. URL https://www.duo.uio.no/handle/10852/10351. [p486]

S. O. Samuelsen, H. Ånestad, and A. Skrondal. Stratified case–cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*, 34(1):103–119, 2007. URL https://doi.org/10.1111/j.1467-9469.2006.00552.x. [p484, 486]

SAS Institute Inc. *SAS Software, Version 8*. Cary, NC, 1999. URL http://www.sas.com/. [p484]

J. L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15, 1999. URL https://doi.org/10.1177/096228029900800102. [p488]

T. A. Severini. Frequency properties of inferences based on an integrated likelihood function. *Statistica Sinica*, 21(1):433–447, 2011. URL https://www.jstor.org/stable/24309279. [p486]

S. J. Sharp, M. Poulaliou, S. G. Thompson, I. R. White, and A. M. Wood. A review of published analyses of case–cohort studies and recommendations for future reporting. *PLOS One*, 9(6):e101176, 2014. URL https://doi.org/10.1371/journal.pone.0101176. [p487, 491]

T. M. Therneau. *survival: A Package for Survival Analysis in S*, 2017. URL https://cran.r-project.org/package=survival. R package version 2.41-3. [p484]

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000. [p484]

T. M. Therneau and H. Li. Computing the Cox model for case–cohort designs. *Lifetime Data Analysis*, 5 (2):99–112, 1999. URL https://doi.org/10.1023/A:1009691327335. [p485, 490]

Tibco Software Inc. *S-Plus Software, Version 8*. Palo Alto, CA, 2007. URL http://www.tibco.com/. [p484]

Y. Yan, H. Zhou, and J. Cai. Improving efficiency of parameter estimation in case-cohort studies with multivariate failure time data. *Biometrics*, 73(3):1042–1052, 2017. URL https://doi.org/10.1111/biom.12657. [p486]

D. Zeng and D. Y. Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society B*, 69(4):507–564, 2007. URL https://doi.org/10.1111/j.1369-7412.2007.00606.x. [p486]

Edmund Jones
Cardiovascular Epidemiology Unit
Department of Public Health and Primary Care
University of Cambridge
United Kingdom
edmundjones79@gmail.com