Survival Analysis for Cohorts with Missing Covariate Information

by Hormuzd A. Katki and Steven D. Mark

NestedCohort fits Kaplan-Meier and Cox Models to estimate standardized survival and attributable risk for studies where covariates of interest are observed on only a sample of the cohort. Missingness can be either by happenstance or by design (for example, the case-cohort and case-control within cohort designs).

Introduction

Most large cohort studies have observations with missing values for one or more exposures of interest. Exposure covariates that are missing by chance (missing by happenstance) present challenges in estimation well-known to statisticians. Perhaps less known is that most large cohort studies now include analyses of studies which deliberately sample only a subset of all subjects for the measurement of some exposures. These "missingness by design" studies are used when an exposure of interest is expensive or difficult to measure. Examples of different sampling schemes that are used in missing by design studies are the case-cohort, nested case-control, and case-control studies nested within cohorts in general (Mark and Katki (2001); Mark (2003); Mark and Katki (2006)). Missingness by design can yield important cost savings with little sacrifice of statistical efficiency (Mark (2003); Mark and Katki (2006)). Although for missingness-by-happenstance, the causes of missingness are not controlled by the investigator, the validity of any analysis of data with missing values depends on the relationship between the observed data and the missing data. Except under the strongest assumption that missing values occur completely at random (MCAR), standard estimators that work for data without missing values are biased when used to analyze data with missing values.

Mark (2003); Mark and Katki (2006) propose a class of weighted survival estimators that accounts for either type of missingness. The estimating equations in this class weight the contribution from completely observed subjects by the inverse probability of being completely observed (see below), and subtract an 'offset' to gain efficiency (see above references). The probabilities of being completely observed are estimated from a logistic regression. The predictors for this logistic regression are some (possibly improper) subset of the covariates for which there are no missing values; the outcome is an indicator variable denoting whether each observation has measurements for all covariates. The predictors may include the outcome variables (time-to-event), exposure variables that are measured on all subjects, and any other variables measured on the entire cohort. We refer to variables that are neither the outcome, nor in the set of exposures of interest (e.g. any variable used in the estimation of the Cox model), as auxiliary variables.

The weighted estimators we propose are unbiased when the missing mechanism is missing-at-random (MAR) and the logistic regression is correctly specified. For missing-by-design, MAR is satisfied and the correct logistic model is known. If there is any missing-by-happenstance, MAR is unverifiable. Given MAR is true, a logistic model saturated in the completely-observed covariates will always be correctly specified. In practice, given that the outcome is continuous (time-to-event), fitting saturated models is not feasible. However, fitting as rich a model as is reasonably possible not only bolsters the user's case that the model is correctly specified, but also improves efficiency (Mark (2003); Mark and Katki (2006)). Also, auxiliary variables can produce impressive efficiency gains and hence should be included as predictors even when not required for correct model specification (Mark (2003); Mark and Katki (2006)).

Our R package **NestedCohort** implements much of the methodology of Mark (2003); Mark and Katki (2006). The major exception is that it does not currently implement the finely-matched nested case-control design as presented in appendix D of Mark (2003); frequency-matching, or no matching, in a case-control design are implemented. In particular, **NestedCohort**

1. estimates not just relative risks, but also absolute and attributable risks. **NestedCohort** can estimate both non-parametric (Kaplan-Meier) and semi-parametric (Cox model) survival curves for each level of the exposures also attributable risks that are standardized for confounders.

2. allows cases to have missing exposures. Standard nested case-control and case-cohort software can produce biased estimates if cases are missing exposures.

3. produces unbiased estimates when the sampling is stratified on any completely observed variable, including failure time.

4. extracts efficiency out of auxiliary variables available on all cohort members.

5. uses weights *estimated* from a correctly-specified sampling model to greatly increase the efficency of the risk estimates compared to using the 'true' weights (Mark (2003); Mark and Katki (2006)).

6. estimates relative, absolute, and attributable risks for vectors of exposures. For relative risks, any covariate can be continuous or categorical.

NestedCohort has three functions that we demonstrate in this article.

1. `nested.km`: Estimates the Kaplan-Meier survival curve for each level of categorical exposures.

2. `nested.coxph`: Fits the Cox model to estimate relative risks. All covariates and exposures can be continuous or categorical.

3. `nested.stdsurv`: Fits the Cox model to estimate standardized survival probabilities, and Population Attributable Risk (PAR). All covariates and exposures must be categorical.

Example study nested in a cohort

In Mark and Katki (2006), we use our weighted estimators to analyze data on the association of *H.Pylori* with gastric cancer and provide simulations that demonstrate the increases in efficiency due to using estimated weights and auxiliary variables. In this document, we present a second example. Abnet et al. (2005) observe esophageal cancer survival outcomes and relevant confounders on the entire cohort. We are interested in the effect of concentrations of various metals, especially zinc, on esophageal cancer. However, measuring metal concentrations consumes precious esophageal biopsy tissue and requires a costly measurement technique. Thus we measured concentrations of zinc (as well as iron, nickel, copper, calcium, and sulphur) on a sample of the cohort. This sample oversampled the cases and those with advanced baseline histologies (i.e. those most likely to become cases) since these are the most informative subjects. Due to cost and availability constraints, less than 30% of the cohort was sampled. For this example, **NestedCohort** will provide adjusted hazard ratios, standardized survival probabilities, and PAR for the effect of zinc on esophageal cancer.

Specifying the sampling model

Abnet et al. (2005) used a two-phase sampling design to estimate the association of zinc concentration with the development of esophageal cancer. Sampling probabilities were determined by case-control status and severity of baseline esophageal histology. The sampling frequencies are given in the table below:

Baseline Histology	Case	Control	Total
Normal	14 / 22	17 / 221	31 / 243
Esophagitis	19 / 26	22 / 82	41 / 108
Mild Dysplasia	12 / 17	19 / 35	31 / 52
Moderate Dysplasia	3 / 7	4 / 6	7 / 13
Severe Dysplasia	5 / 6	3 / 4	8 / 10
Carcinoma In Situ	2 / 2	0 / 0	2 / 2
Unknown	1 / 1	2 / 2	3 / 3
Total	56 / 81	67 / 350	123 / 431

The column "baseline histology" contains, in order of severity, classification of pre-cancerous lesions. For each cell, the number to the right of the slash is the total cohort members in that cell, the left is the number we sampled to have zinc observed (i.e. in the top left cell, we measured zinc on 14 of the 22 members who became cases and had normal histology at baseline). Note that for each histology, we sampled roughly 1:1 cases to controls (frequency matching), and we oversampled the more severe histologies (who are more informative since they are more likely to become cases). Thirty percent of the cases could not be sampled due to sample availability constraints.

Since the sampling depended on case/control status (variable `ec01`) crossed with the seven baseline histologies (variable `basehist`), this sampling scheme will be accounted for by each function with the statement '`samplingmod="ec01*basehist"`'. This allows each of the 14 sampling strata its own sampling fraction, thus reproducing the sampling frequencies in the table.

NestedCohort requires that each observation have nonzero sampling probability. For this table, each of the 13 non-empty strata must have have someone sampled in it. Also, the estimators require that there are no missing values in any variable in the sampling model. However, if there is missingness, for convenience, **NestedCohort** will remove from the cohort any observations that have missingness in the sampling variables and will print a warning to the user. There should not be too many such observations.

Kaplan-Meier curves

To make non-parametric (Kaplan-Meier) survival curves by quartile of zinc level, use `nested.km`. These Kaplan-Meier curves have the usual interpretation: they do not standardize for other variables, and do not account for competing risks.

To use this, provide both a legal formula as per the `survfit` function and also a sampling model to calculate stratum-specific sampling fractions. Note that the '`survfitformula`' and '`samplingmod`' require their arguments to be inside double quotes. The

'data' argument is required: the user must provide the data frame within which all variables reside in. This outputs the Kaplan-Meier curves into a survfit object, so all the methods that are already there to manipulate survfit objects can be used[1].

To examine survival from cancer within each quartile of zinc, allowing different sampling probabilities for each of the 14 strata above, use nested.km, which prints out a table of risk differences versus the level named in 'exposureofinterest'; in this case, it's towards "Q4" which labels the 4th quartile of zinc concentration:

```
> library(NestedCohort)
> mod <- nested.km(survfitformula =
+          "Surv(futime01,ec01==1)~znquartiles",
+          samplingmod = "ec01*basehist",
+          exposureofinterest = "Q4", data = zinc)

Risk Differences vs. znquartiles=Q4 by time 5893
        Risk Difference  StdErr       95% CI
Q4 - Q1         0.28175 0.10416  0.07760 0.4859
Q4 - Q2         0.05551 0.07566 -0.09278 0.2038
Q4 - Q3         0.10681 0.08074 -0.05143 0.2651

> summary(mod)
[...]

308 observations deleted due to missing
                znquartiles=Q1
 time n.risk n.event survival std.err    95% CI
  163  125.5    1.37    0.989  0.0108 0.925 0.998
 1003  120.4    1.57    0.976  0.0169 0.906 0.994
 1036  118.8    1.00    0.968  0.0191 0.899 0.990
[...]

                znquartiles=Q2
 time n.risk n.event survival std.err    95% CI
 1038  116.9    1.57    0.987  0.0133 0.909 0.998
 1064  115.3    4.51    0.949  0.0260 0.864 0.981
 1070  110.8    2.33    0.929  0.0324 0.830 0.971
[...]
```

summary gives the lifetable. Although summary prints how many observations were 'deleted' because of missing exposures, the 'deleted' observations still contribute to the final estimates via estimation of the sampling probabilities. Note that the lifetable contains the weighted numbers of those at risk and who had the developed cancer.

The option 'outputsamplingmod' returns the sampling model that the sampling probabilities were calculated from. Examine this model if warned that it didn't converge. If 'outputsamplingmod' is TRUE, then nested.km will output a list with 2 components, the survmod component being the Kaplan-Meier survfit object, and the other samplingmod component being the sampling model.

Plotting Kaplan-Meier curves

Make Kaplan-Meier plots with the plot function for survfit objects. All plot options for survfit objects can be used.

```
> plot(mod,ymin=.6,xlab="time",ylab="survival",
+      main="Survival by Quartile of Zinc",
+      legend.text=c("Q1","Q2","Q3","Q4"),
+      lty=1:4,legend.pos=c(2000,.7))
```

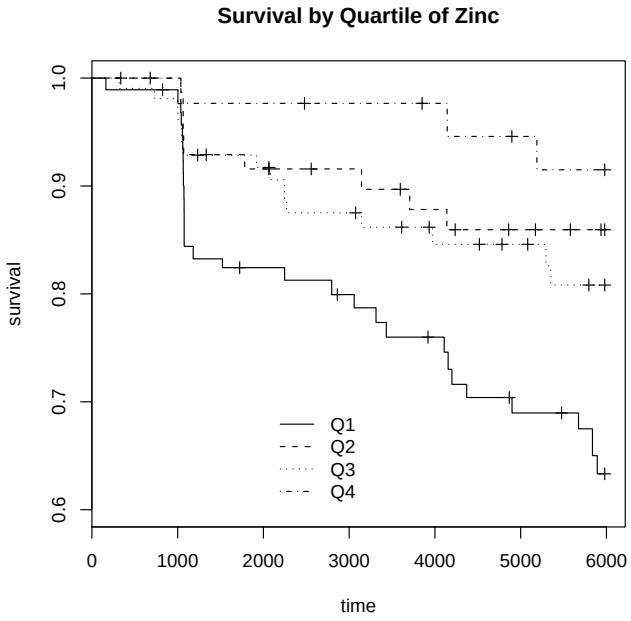

Figure 1: Kaplan-Meier plots by nested.km().

nested.km has some restrictions:

1. All variables are in a dataframe denoted by the 'data' argument.
2. No variable in the dataframe can be named o.b.s.e.r.v.e.d. or p.i.h.a.t.
3. 'survfitformula' must be a valid formula for survfit objects: All variables must be factors.
4. It does not support staggered entry into the cohort. The survival estimates will be correct, but their standard errors will be wrong.

Cox models: relative risks

To fit the Cox model, use nested.coxph. This function relies on coxph that is already in the survival package, and imitates its syntax as much as possible. In this example, we are interested in estimating the effect of zinc (as zncent, a continuous variable standardized to 0 median and where a 1 unit change is an

[1] nested.km uses the weights option in survfit to estimate the survival curve. However, the standard errors reported by survfit are usually quite different from, and usually much smaller than, the correct ones as reported by nested.km.

increase of 1 quartile in zinc) on esophageal cancer, while controlling for sex, age (as agepill, a continuous variable), smoking, drinking (both ever/never), baseline histology, and family history (yes/no). We use the same sampling model ec01*basehist as before. The output is edited for space:

```
> coxmod <- nested.coxph(coxformula =
+   "Surv(futime01,ec01==1)~sex+agepill+basehist+
     anyhist+zncent",
+   samplingmod = "ec01*basehist", data = zinc)

> summary(coxmod)
[...]
                          exp(coef) lower/upper.95
sexMale                        0.83    0.38   1.79
agepill                        1.04    0.99   1.10
basehistEsophagitis            2.97    1.41   6.26
basehistMild Dysplasia         4.88    2.19  10.88
basehistModerate Dysplasia     6.95    2.63  18.38
basehistSevere Dysplasia      11.05    3.37  36.19
basehistNOS                    3.03    0.29  30.93
basehistCIS                   34.43   10.33 114.69
anyhistFamily History          1.32    0.61   2.83
zncent                         0.73    0.57   0.93

[...]
Wald test = 97.5  on 10 df,   p=2.22e-16
```

This is the exact same coxph output, except that the R^2, overall likelihood ratio and overall score tests are not computed. The overall Wald test is correctly computed.

nested.coxph has the following restrictions

1. All variables are in the dataframe in the 'data' argument.
2. No variable in the dataframe can be named o.b.s.e.r.v.e.d. or p.i.h.a.t.
3. You must use Breslow tie-breaking.
4. No 'cluster' statements are allowed.

However, nested.coxph does allow staggered entry into the cohort, stratification of the baselize hazard via 'strata', and use of 'offset' arguments to coxph (see help for coxph for more information).

Standardized survival and attributable risk

nested.stdsurv first estimates hazard ratios exactly like nested.coxph, and then also estimates survival probabilities for each exposure level as well as Population Attributable Risk (PAR) for a given exposure level, standardizing both to the marginal confounder distribution in the cohort. For example, the standardized survival associated with exposure Q and confounder J is

$$S_{std}(t|Q) = \int S(t|J,Q)dF(J).$$

In contrast, the crude observed survival is

$$S_{crude}(t|Q) = \int S(t|J,Q)dF(J|Q).$$

The crude S is the observed survival, so the effects of confounders remain. The standardized S is estimated by using the observed J distribution as the standard, so J is independent of Q. For more on direct standardization, see Breslow and Day (1987)

To standardize, the formula for a Cox model must be split in two pieces: the argument 'exposures' denotes the part of the formula for the exposures of interest, and 'confounders' which denotes the part of the formula for the confounders. All variables in either part of the formula must be factors. In either part, do not use '*' to specify interaction, use interaction.

In the zinc example, the exposures are 'exposures="znquartiles"', a factor variable denoting which quartile of zinc each measurement is in. The confounders are 'confounders="sex+agestr+basehist+anyhist"', these are the same confounders in the hazard ratio example, except that we must categorize age as the factor agestr. 'timeofinterest' denotes the time at which survival probabilities and PAR are to be calculated at, the default is at the last event time. 'exposureofinterest' is the name of the exposure level to which the population is to be set at for computing PAR; 'exposureofinterest="Q4"' denotes that we want PAR if we could move the entire population's zinc levels into the fourth quartile of the current zinc levels. 'plot' plots the standardized survivals with 95% confidence bounds at 'timeofinterest' and returns the data used to make the plot. The output is edited for space.

```
> mod <- nested.stdsurv(outcome =
+  "Surv(futime01,ec01==1)",
+  exposures = "znquartiles",
+  confounders = "sex+agestr+basehist+anyhist",
+  samplingmod = "ec01*basehist",
+  exposureofinterest = "Q4", plot = T, main =
+  "Time to Esophageal Cancer
                          by Quartiles of Zinc",
+  data = zinc)

Std Survival for znquartiles by time 5893
      Survival  StdErr 95% CI Left 95% CI Right
Q1      0.5054 0.06936      0.3634       0.6312
Q2      0.7298 0.07768      0.5429       0.8501
Q3      0.6743 0.07402      0.5065       0.7959
Q4      0.9025 0.05262      0.7316       0.9669
Crude   0.7783 0.02283      0.7296       0.8194

Std Risk Differences vs.
                  znquartiles = Q4 by time 5893
        Risk Difference  StdErr       95% CI
Q4 - Q1          0.3972 0.09008  0.22060 0.5737
Q4 - Q2          0.1727 0.09603 -0.01557 0.3609
Q4 - Q3          0.2282 0.08940  0.05294 0.4034
```

```
Q4 - Crude        0.1242 0.05405  0.01823 0.2301

PAR if everyone had znquartiles = Q4
  Estimate StdErr 95% CI Left 95% CI Right
PAR 0.5602 0.2347     -0.2519       0.8455
```

The first table shows the survival for each quartile of zinc, standardized for all the confounders, as well as the 'crude' survival, which is the observed survival in the population (so is not standardized). The next table shows the standardized survival differences vs. the exposure of interest. The last table shows the PAR, and the CI for PAR is based on the $\log(1 - \text{PAR})$ transformation (this is often very different from, and superior to, the naive CI without transformation). `summary(mod)` yields the same hazard ratio output as if the model had been run under `nested.coxph`.

The plot is in figure 2. This plots survival curves; to plot cumulative incidence (1-survival), use '`cuminc=TRUE`'. The 95% CI bars are plotted at `timeofinterest`. All `plot` options are usable: e.g. '`main`' to title the plot.

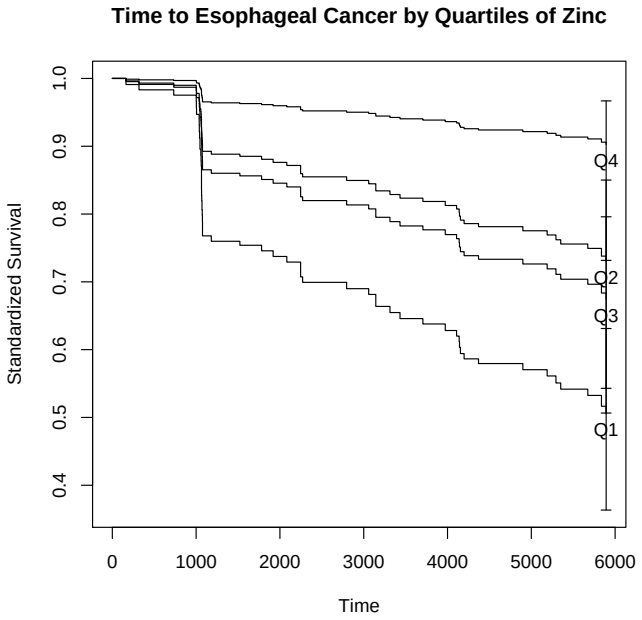

Figure 2: Survival curves for each zinc quantile, standardized for confounders

`nested.stdsurv` has some restrictions:

1. All variables are in the dataframe in the '`data`' argument.

2. No variable in the dataframe can be named `o.b.s.e.r.v.e.d.` or `p.i.h.a.t.`

3. The variables in the '`exposures`' and '`confounders`' must be factors, even if they are binary. In these formulas, never use '*' to mean interaction, use `interaction`.

4. It does not support staggered entry into the cohort.

5. It does not support the baseline hazard to be stratified. '`cluster`' and '`offset`' arguments are not supported either.

6. It only allows Breslow tie-breaking.

Including auxiliary variables

In this analysis, we used the smallest correctly-specified logistic model to predict sampling probabilities. To illustrate the use of an auxiliary variable, let's pretend we have a categorical surrogate named `znauxiliary`, a cheaply-available but non-ideal measure of zinc concentration, observed on the full cohort. The user could sample based on `znauxiliary` to try to improve efficiency. In this case, `znauxiliary` must be included as a sampling variable in the sampling model with `samplingmod="ec01*basehist*znauxiliary"`. Note that auxiliary variables must be observed on the entire cohort.

Even if sampling is not based on `znauxiliary`, it can still be included in the sampling model as above. This is because, even though `znauxiliary` was not explicitly sampled on, if `znauxiliary` has something to do with zinc, and zinc has something to do with either `ec01` or `basehist`, then one is implicitly sampling on `znauxiliary`. The simulations in (Mark and Katki (2006)) show the efficiency gain from including auxiliary variables in the sampling model. Including auxiliary variables will always reduce the standard errors of the risk estimates.

Multiple exposures

Multiple exposures (with missing values) are included in the risk regression just like any other variable. For example, if we want to estimate the esophageal cancer risk from zinc and calcium jointly, the Cox model would include `cacent` as a covariate. Cutting calcium into quartiles into the variable `caquartiles`, include it as an exposure with `nested.stdsurv` with '`exposures="znquartiles+caquartiles"`'.

Full cohort analysis

NestedCohort can be used if all covariates are observed on the full cohort. You can estimate the standardized survival and attributable risks by setting '`samplingModel="1"`', to force equal weights for all cohort members. `nested.km` will work exactly as `survfit` does. The Cox model standard errors will be those obtained from `coxph` with '`robust=TRUE`'.

Bibliography

Abnet, C. C., Lai, B., Qiao, Y.-L., Vogt, S., Luo, X.-M., Taylor, P. R., Dong, Z.-W., Mark, S. D., and Dawsey, S. M. (2005). Zinc concentration in esophageal biopsies measured by x-ray fluorescence and cancer risk. *Journal of the National Cancer Institute*, 97(4):301–306.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. Volume II: The Design and Analysis of Cohort Studies.* IARC Scientific Publications, Lyon.

Mark, S. D. (2003). Nonparametric and semiparametric survival estimators,and their implementation, in two-stage (nested) cohort studies. *Proceedings of the Joint Statistical Meetings*, 2675–2691.

Mark, S. D. and Katki, H. A. (2001). Influence Function Based Variance Estimation and Missing Data Issues in Case-Cohort Studies. *Lifetime Data Analysis*, 7:331–344.

Mark, S. D. and Katki, H. A. (2006). Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (sampled) cohort studies with missing case data. *Journal of the American Statistical Association*, 101(474):460–471.

Hormuzd A. Katki
Division of Cancer Epidemiology and Genetics
National Cancer Institute, NIH, DHHS, USA
katkih@mail.nih.gov

Steven D. Mark
Department of Preventive Medicine and Biometrics
University of Colorado Health Sciences Center
Steven.Mark@UCHSC.edu