Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Combining cohorts in longitudinal surveys

by Iván A. Carrillo and Alan F. Karr

June 2013

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 149-182
Statistics Canada, Catalogue No. 12-001-X

Combining cohorts in longitudinal surveys

Iván A. Carrillo and Alan F. Karr[1]

Abstract

A question that commonly arises in longitudinal surveys is the issue of how to combine differing cohorts of the survey. In this paper we present a novel method for combining different cohorts, and using all available data, in a longitudinal survey to estimate parameters of a semiparametric model, which relates the response variable to a set of covariates. The procedure builds upon the Weighted Generalized Estimation Equation method for handling missing waves in longitudinal studies. Our method is set up under a joint-randomization framework for estimation of model parameters, which takes into account the superpopulation model as well as the survey design randomization. We also propose a design-based, and a joint-randomization, variance estimation method. To illustrate the methodology we apply it to the Survey of Doctorate Recipients, conducted by the U.S. National Science Foundation.

Key Words: Superpopulation parameters; Joint-randomization inference; Replication variance estimation; Rotating panel surveys; Multi-cohort longitudinal surveys; Weighted Generalized Estimating Equations.

1 Introduction

The Survey of Doctorate Recipients (SDR) is a National Science Foundation (NSF) longitudinal survey whose design incorporates features of both repeated panels and rotating panels. The purpose of the survey is to study U.S. doctorate recipients in science, engineering, and health fields. It is conducted approximately every two years. A detailed description of the SDR can be found at NSF (2012). In this paper we restrict our attention to the data collected from 1995 through 2008 (7 waves).

At any particular wave a new cohort is selected. The new cohort consists of a sample of recent graduates (from the previous two years) selected from the Doctorate Records File, which is a database constructed mainly from the Survey of Earned Doctorates (http://www.nsf.gov/statistics/srvydoctorates/). The selected individuals are kept in the sample, *i.e.*, interviewed every two years, until the age of 75, while living in the U.S. during the survey reference week, and

1. Iván A. Carrillo and Alan F. Karr, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, U.S.A. E-mail: ivan@niss.org and karr@niss.org.

while not institutionalized. However, *not* all the sampled graduates satisfying these characteristics are retained forever. Some individuals, rather than entire cohorts, are dropped from the sample in order to a) include the new graduates in the new cohorts and b) maintain a relatively constant sample size across waves. In Section 2.2 we describe how the selection of the individuals who are dropped is made.

Survey weights for cross-sectional analyses of the SDR are already available, but not for longitudinal analyses. Rather than requiring a *new* longitudinal weight for *all* the data, the method proposed in this paper is able to use the existing cross-sectional weights for longitudinal analyses without ignoring any data. We concentrate on estimation of parameters of statistical models of the effect of covariates on a response of interest, but the method can also be used for estimation of finite population quantities (Carrillo and Karr 2012). We focus on analysis of the SDR, but our method is applicable to any fixed-panel, fixed-panel-plus-'births', repeated-panel, rotating-panel, split-panel, or refreshment sample survey, as long as for each wave there is a cross-sectional weight to represent the population of interest at that wave. See Smith, Lynn and Elliot (2009), Hirano, Imbens, Ridder and Rubin (2001), and Nevo (2003) for definitions of all these types of longitudinal sample designs.

The SDR is a hybrid of repeated-panel and rotating panel designs. It is not purely a repeated-panel design because of the removal of some subjects at each wave. It is not purely a rotating-panel design because entire panels (or cohorts) are *not* removed, only individuals; additionally, the composition of the finite population of interest changes over time, unlike in a rotating panel survey.

Diggle, Heagerty, Liang and Zeger (2002) and Hedeker and Gibbons (2006) point out that, with longitudinal studies, contrary to a cross-sectional study, it is possible to separate age effect (actual change within subjects over time) and cohort effect (difference between units at the beginning of the study period).

Hedeker and Gibbons (2006) also suggest that since longitudinal studies allow for the measurement of time-varying explanatory variables (covariates), the statistical inferences about dynamic relationship between the outcome on interest (response) and these covariates are much stronger than those based on cross-sectional studies.

When we are interested in the marginal mean of a variable, possibly conditionally on some covariates, and not in measuring change, a longitudinal study is not necessary; a cross-sectional study suffices. However, even in this case, a longitudinal study tends to be more powerful, because each subject serves as his or her own control for any unmeasured characteristics (Diggle, *et al.* 2002).

Our approach differs from the existing alternatives in the literature, which have some limitations for analysis of such data, and in particular for application to the SDR. For example, Berger (2004a) and Berger (2004b) go into detail about the estimation of change using rotating samples, but they assume that the composition of the finite population does not change over time, which is not the case of the SDR. This assumption does not hold in many other large-scale surveys. Also, the methodology proposed by Berger is not easily generalizable to more than two waves. Similarly, Qualité and Tillé (2008) also assume the finite population is fixed over time. Hirano, *et al.* (2001) and Nevo (2003) present different methods of estimation assuming a fixed-panel plus refreshment for attrition design, but also assume the finite population composition is fixed over time.

A time series approach is utilized by McLaren and Steel (2000) and Steel and McLaren (2007) to estimate change and trend with survey data. Although their approach allows for the incorporation of within-subject association in the point estimates, they do not consider covariates in their models (beyond the implicit time covariates). Also, they only discuss the estimation of change for continuous variables.

Another alternative for analyzing longitudinal data is to fix the finite population of interest, except perhaps for deaths, which could be allowed. Studies of this kind are those where there are

data available only for a single cohort. For example, Vieira and Skinner (2008), Carrillo, Chen and Wu (2010), and Carrillo, Chen and Wu (2011) show some alternatives for modeling with single-cohort survey data. However, to use these kinds of analyses with multi-cohort surveys, one needs to ignore some (or many) available data, for example those data from subjects who are not common to all waves. An example of a weighting procedure of this type can be found in Ardilly and Lavallée (2007).

Finally, the approach of Larsen, Qing, Zhou and Foulkes (2011) is appealing, in principle, because it is the way survey practitioners generally proceed. An initial weight is adjusted, among other things for calibration to known totals, in this case totals by survey wave. Nonetheless, for rotating panels this method is still in its infancy; there are some things that are not completely clear how to carry out. For example, it is not clear what the initial weight should be: a constant weight?, the earliest available weight?, the average of the available weights for each case?, or the latest available weight? Also, in the case of dropouts, as there exist in the SDR, the authors do not clarify how to carry out a nonresponse adjustment with this method. Even more, it is not clear why a nonresponse adjustment for dropouts at, say, wave 4 should have any influence on the observations at wave 3, as this methodology permits since there is a single weight for each subject. Additionally, the authors mention that they estimated standard errors, but they do not indicate how to take into account all the features of the sampling design, such as changes over time in the stratification and weighting adjustment classes of the SDR. Our method, on the other hand, utilizes only cross-sectional weights and variance estimation methods, which have been studied thoroughly in the literature and are readily available for the SDR.

The rest of the paper is organized as follows. In the next section we give a description of the SDR design. After that, in Section 3, we propose a novel approach for longitudinal analysis of marginal mean models with multi-cohort surveys. Then we present the application of the methodology to the SDR. Finally we offer a few discussion points in Section 5.

2 The SDR design

2.1 Finite population

The SDR finite population of interest can be represented as in Table 2.1. At wave 1, *i.e.*, the first time of interest, there is a finite set, $U_{1(1)} = U_1$, of $N_{1(1)} = N_1$ Ph.D. holders, either recent or not, who satisfy the requirements of the SDR.

Table 2.1
SDR finite population

j:	1		2		3		…		$J-1$		J
	$U_{1(1)}$	$\supseteq$	$U_{2(1)}$	$\supseteq$	$U_{3(1)}$	$\supseteq$	…	$\supseteq$	$U_{J-1(1)}$	$\supseteq$	$U_{J(1)}$
	$N_{1(1)}$	$\geq$	$N_{2(1)}$	$\geq$	$N_{3(1)}$	$\geq$	…	$\geq$	$N_{J-1(1)}$	$\geq$	$N_{J(1)}$
			$U_{2(2)}$	$\supseteq$	$U_{3(2)}$	$\supseteq$	…	$\supseteq$	$U_{J-1(2)}$	$\supseteq$	$U_{J(2)}$
			$N_{2(2)}$	$\geq$	$N_{3(2)}$	$\geq$	…	$\geq$	$N_{J-1(2)}$	$\geq$	$N_{J(2)}$
						$\ddots$			$\vdots$		$\vdots$
									$U_{J-1(J-1)}$	$\supseteq$	$U_{J(J-1)}$
									$N_{J-1(J-1)}$	$\geq$	$N_{J(J-1)}$
											$U_{J(J)}$
											$N_{J(J)}$
	U_1		U_2		U_3		…		U_{J-1}		U_J
	N_1		N_2		N_3		…		N_{J-1}		N_J

At wave 2 only a subset of the subjects in $U_{1(1)}$ still satisfy the SDR requirements; we call this subset, of $N_{2(1)}$ subjects, $U_{2(1)}$. In addition, there is a set of new, recent Ph.D. recipients, who have obtained their degree since wave 1, and also satisfy the other requirements of the survey. This set of new graduates in scope is called $U_{2(2)}$ and is of size $N_{2(2)}$. Therefore, at wave 2, there is a total of $N_2 = N_{2(1)} + N_{2(2)}$ subjects in the population of interest $U_2 = U_{2(1)} \cup U_{2(2)}$.

At the next wave, wave 3, the same process occurs. Some people in $U_{2(1)}$ leave the population of interest and there are only $N_{3(1)}$ left in $U_{3(1)}$. The same thing happens with the set $U_{2(2)}$; only a subset $U_{3(2)}$ of $N_{3(2)}$ among them still satisfy the requirements of the SDR. Additionally, there are $N_{3(3)}$ recent graduates entering the population of interest; this set is called $U_{3(3)}$. In total, the finite population of interest at wave 3 is $U_3 = U_{3(1)} \cup U_{3(2)} \cup U_{3(3)}$, with $N_3 = N_{3(1)} + N_{3(2)} + N_{3(3)}$ subjects.

This procedure, of thinning of old cohorts and adding new cohorts, continues until the last wave of interest, wave J. We notice that the finite population of interest changes at every wave due to two main reasons. Firstly, some of the subjects in the old cohorts are no longer in scope at the current wave, and they are not part of the current target population. Secondly, the recent graduates are added to the target population in the current wave. We denote by $j = 1, 2, \ldots, J$ the wave of interest (outside the parenthesis) and by $j' = 1, 2, \ldots, J$ the cohort to which a subject belongs (inside the parenthesis), and therefore $U_{j(j')} = U_{\text{wave(cohort)}}$.

2.2 Sampling

The sampling design of the SDR has a similar structure to the finite population and is depicted in Table 2.2. At wave 1, a (complex) sample $s_{1(1)} = s_1$ of $n_{1(1)} = n_1$ subjects is selected from within the N_1 elements in U_1. Each element i in s_1 is interviewed and its data collected; also, there is a design weight $w_{i1} = 1 / \pi_{i1}$ associated with it, which is the inverse of its inclusion probability at wave 1.

Table 2.2
SDR Sample

j :	**1**		**2**		**3**		**…**		**$J-1$**		**J**
	$s_{1(1)}$	$\supseteq$	$s_{2(1)}$	$\supseteq$	$s_{3(1)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(1)}$	$\supseteq$	$s_{J(1)}$
	$n_{1(1)}$	$\geq$	$n_{2(1)}$	$\geq$	$n_{3(1)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(1)}$	$\geq$	$n_{J(1)}$
			$s_{2(2)}$	$\supseteq$	$s_{3(2)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(2)}$	$\supseteq$	$s_{J(2)}$
			$n_{2(2)}$	$\geq$	$n_{3(2)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(2)}$	$\geq$	$n_{J(2)}$
					$s_{3(3)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(3)}$	$\supseteq$	$s_{J(3)}$
					$n_{3(3)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(3)}$	$\geq$	$n_{J(3)}$
							$\ddots$		$\vdots$		$\vdots$
									$s_{J-1(J-1)}$	$\supseteq$	$s_{J(J-1)}$
									$n_{J-1(J-1)}$	$\geq$	$n_{J(J-1)}$
											$s_{J(J)}$
											$n_{J(J)}$
	s_1		s_2		s_3		$\cdots$		s_{J-1}		s_J
	n_1		n_2		n_3		$\cdots$		n_{J-1}		n_J

At the second wave, the elements in $s_{1(1)}$ who are not in scope anymore are simply dropped from the frame (though their observations at wave 1 are kept), and a subsample $s_{2(1)}$, of size $n_{2(1)}$, of those still in scope is selected. Not all the members in $s_{1(1)}$ who are still in scope at wave 2 are retained in the sample; this is in order to be able to make up room for the sample of the new Ph.D. recipients and still maintain more or less the same sample size as in wave 1. A sample $s_{2(2)}$ of size $n_{2(2)}$ is selected from $U_{2(2)}$; people in $s_{2(2)}$ form the second cohort. The total sample at wave 2 is $s_2 = s_{2(1)} \cup s_{2(2)}$, which is of size $n_2 = n_{2(1)} + n_{2(2)}$, which is approximately equal to n_1. All the people in s_2 are interviewed at wave 2. The design weights at wave 2, $w_{i2} = 1 / \pi_{i2}$, are such that the sample s_2 represents the population of interest at wave 2, namely U_2.

The same procedure is repeated at each wave, till the last one (J), where a subsample of the remaining subjects from each of the previous $J - 1$ cohorts is selected, and a new sample (the new cohort) $s_{J(J)}$ of recent graduates is selected from $U_{J(J)}$. At the last wave, all people in $s_J = \bigcup_{j'=1}^{J} s_{J(j')}$ are interviewed and a design weight $w_{iJ} = 1 / \pi_{iJ}$ is created for each person interviewed, so that s_J represents the finite population U_J.

With respect to how the selection of the individuals that are dropped is made, for example in 2008, according to NSF (2012), the subsample $s_{08} \setminus s_{08(08)}$ was selected by stratifying s_{06} "into 150 strata based on three variables: demographic group, degree field, and sex." They go on to explain that:

- the past practice of selecting the sample with probability proportional to size continued, where the measure of size was the base weight associated with the previous survey cycle. For each stratum, the sampling algorithm started by identifying and removing self-representing cases through an iterative procedure. Next, the non-self-representing cases within each stratum were sorted by citizenship, disability status, degree field, and year of doctoral degree award. Finally, the balance of the sample (*i.e.*, the total allocation minus the number of self-representing cases) was selected from each stratum systematically with probability proportional to size.

It is worth mentioning that up to 1989 the cohort (or more specifically the graduation year) was part of the stratifying variables (and weight-adjustment cells), but beginning in 1991 it has not been; it was replaced by the disability status. For more details about the subsampling procedure, including the description of the sample allocation, see NSF (2012) or Cox, Grigorian, Wang and Harter (2010).

From the preceding description, it is clear that the design of the SDR is not a rotating panel design. Beside the fact that the composition of the finite population of interest is changing over time, a rotating panel design would select, at time j, a new cohort from U_j, and not from $U_j \setminus U_{j-1}$ as the SDR does.

Another peculiarity of the SDR is that, at each wave j, a frame of the recent graduates $U_{j(j)}$ exists, from which the new cohort $s_{j(j)}$ can be selected straightforwardly. However, in other applications, the cost of building such a frame, *i.e.*, a frame of new members, may be excessive (particularly as it cumulates over waves), and the new cohort may need to be selected from U_j (as opposed to from $U_{j(j)}$). The method proposed in this paper can also be applied in such cases, as long as for the total sample at wave j, s_j, a cross-sectional weight can be created to represent U_j. We further discuss this topic in Section 3.2.

Notice that in the notation $s_{j(j')}$, the quantity j represents the wave to which the sample refers, and j' denotes the sample's cohort, *i.e.*, the wave at which the sample was first selected. The notation for the weights is w_{ij}, where the first subscript identifies the subject, and the second refers to the wave of interest, regardless of when the subject was first selected.

3 Methodology

3.1 Motivation

Assume that (in a non-survey context) interest lies in the $p \times 1$ vector parameter $\boldsymbol{\beta}$ in the following model:

$$\xi : \begin{cases} E[Y_{ij} \mid X_{ij}] = \mu_{ij} = g^{-1}(X'_{ij}\boldsymbol{\beta}), & j = 1, 2, \ldots, J, i = 1, 2, \ldots \\ \text{Var}[Y_{ij} \mid X_{ij}] = \phi\nu(\mu_{ij}), & j = 1, 2, \ldots, J, i = 1, 2, \ldots \\ \text{Cov}[Y_i \mid X_i] = \Sigma_i, & i = 1, 2, \ldots \\ Y_k \perp Y_l \mid X_k, X_l, & k \neq l = 1, 2, \ldots; \end{cases} \tag{3.1}$$

where Y_{ij} is the response variable for subject i at wave j, X_{ij} is a $p \times 1$ vector of covariates, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$, $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$ is a $p \times J$ matrix; $g(\cdot)$ is a monotonic one-to-one differentiable "link function"; $\nu(\cdot)$ is the "variance function" with known form; and $\phi > 0$ is the "dispersion parameter." Since, in general, the $J \times J$ covariance matrix Σ_i is hard to specify, we model it as $\text{Cov}[Y_i \mid X_i] = V_i = A_i^{1/2}\mathbf{R}(\alpha)A_i^{1/2}$, a "working" covariance matrix; where $A_i = \text{diag}[\phi\nu(\mu_{i1}), \phi\nu(\mu_{i2}), \ldots, \phi\nu(\mu_{iJ})]$ and $\mathbf{R}(\alpha)$ is a "working" correlation matrix, both of dimension $J \times J$, and α is a vector that fully characterizes $\mathbf{R}(\alpha)$ (see Liang and Zeger 1986).

To estimate $\boldsymbol{\beta}$ we select a (single-cohort) sample of n elements from model ξ and we (intend to) measure each of them at J occasions. If all the elements in the sample respond at every single occasion j, the task can be completed with the usual generalized estimating equation (GEE) methodology of Liang and Zeger (1986). However, in any study it is rarely the case that all subjects do respond at all waves. It is more common to have some elements in the sample who drop out of the study.

Under this situation, and assuming that the missing responses can be regarded as missing at random or MAR (see Rubin 1976), in particular that the dropout at a given wave does not depend on the current (unobserved) value, Robins, Rotnitzky and Zhao (1995) proposed to estimate $\boldsymbol{\beta}$ by solving the estimating equations: $\sum_{i=1}^{n}(\partial\boldsymbol{\mu}'_i / \partial\boldsymbol{\beta})V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iJ})'$, $\hat{\Delta}_i = \text{diag}[R_{i1}\hat{q}_{i1}^{-1}, R_{i2}\hat{q}_{i2}^{-1}, \ldots, R_{iJ}\hat{q}_{iJ}^{-1}]$, R_{ij} is the response indicator for subject i at wave j, and $\hat{q}_{ij}$ is an estimate of the probability that subject i is observed through wave j.

For survey applications, one would use the estimating equation $\sum_{i \in s}[w_i(\partial\boldsymbol{\mu}'_i / \partial\boldsymbol{\beta})V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)] = \mathbf{0}$, where w_i is the survey weight for subject i. Another way of writing this

equation is $\sum_{i \in s} (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1} \hat{\Delta}_{wi} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, with $\hat{\Delta}_{wi} = \text{diag}[w_i R_{i1} \hat{q}_{i1}^{-1}, w_i R_{i2} \hat{q}_{i2}^{-1}, \ldots, w_i R_{iJ} \hat{q}_{iJ}^{-1}]$.

We notice that the diagonal elements of $\hat{\Delta}_{wi}$ are simply wave-specific nonresponse-adjusted survey weights whenever the subject is observed, and are equal to zero whenever the subject is missing. This feature in and of itself suggests a solution to the multi-cohort problem, which will be presented in the next section.

3.2 A novel approach to combining cohorts in longitudinal surveys

Based on the discussion in the previous section, if we have a fixed-panel, fixed-panel-plus-'births', repeated-panel, rotating-panel, split-panel, or refreshment sample survey, we propose to estimate the superpopulation parameter $\boldsymbol{\beta}$ in model ξ by the solution to the estimating equations:

$$\Psi_s(\boldsymbol{\beta}) = \sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}; \tag{3.2}$$

where the sum is over the sample s, *i.e.*, over all the elements selected (for the first time) in any of the samples $s_{1(1)}, s_{2(2)}, \ldots, s_{J(J)}$. The diagonal matrix W_i is $W_i = \text{diag}[I_i(U_1) w_{i1}, I_i(U_2) w_{i2}, \ldots, I_i(U_J) w_{iJ}]$, with w_{ij} being the (nonresponse-adjusted) cross-sectional weight for subject i at wave j (as long as subject i is part of sample s_j) and $I_i(U_j)$ is the indicator of whether subject i belongs to finite population U_j or not. In Section 3.2.1 we argue why this is a reasonable estimation procedure, and in Section 3.2.2 we discuss the missing value issue.

The cross-sectional weights w_{ij}, in W_i, are such that the sample s_j represents U_j, when used in conjunction with said weights. This means that, for each observation i in sample s_j, there has to be a survey weight w_{ij}, which could be regarded as the number of units that such observation represents in U_j. However, remember that the sample s_j is composed of different sets of subjects, or different subsamples (the different cohorts), and the integration of these subsamples into a single cross-sectional weight variable w_{ij} may not be a straightforward task.

For the SDR, the construction of the cross-sectional weight for wave j is not too complicated as the different cohorts are selected independently, from non-overlapping populations. The base weight in that case is easy to compute, and all that remains is the adjustment for things like attrition and calibration to known totals in the population U_j.

On the other hand, in other situations, for example, when a frame of *new* members does not exist, the new cohort may need to be selected from the overall population at the given wave, or from a frame containing new members *plus* some old members, or from multiple frames. In such cases, the building of the cross-sectional weights may not be as straightforward, and the theory of multiple frames may need to be used. We refer the reader to the works of Lohr (2007) and Rao and Wu (2010), and references therein, for cases like that.

Expression (3.2) is a generalization of equation (2.25) in Vieira (2009). The latter is applicable only when all the subjects have the same number of observations or any missing responses can be regarded as missing completely at random or MCAR (see Rubin 1976). As discussed in Robins, *et al.* (1995), using such an equation when the missing responses are not MCAR produces inconsistent estimators; therefore, with a rotation scheme like that of the SDR, where not all subjects are dropped (or kept) with the same probabilities, its usage would not be appropriate. The adequacy of equation (3.2) in that case and when there are missing responses is addressed in sections 3.2.1 and 3.2.2, respectively. If all subjects have cross-sectional weights that do not vary over time (or have a single longitudinal weight) equation (3.2) reduces to equation (2.25) in Vieira (2009).

3.2.1 Unbiasedness

The unbiasedness property of the estimating function is important because, as Song (2007, Section 5.4) argues, it is the most crucial assumption in order to obtain a consistent estimator.

Let us define β_N, the so-called "census estimator," to be the solution to the following finite population estimating equation:

$$\Psi_U(\boldsymbol{\beta}_N) = \sum_{i \in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_N} V_i^{-1} \mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)) = \mathbf{0}, \tag{3.3}$$

where the sum is over U, *i.e.*, over all the elements who became members of the target population in any of $U_{1(1)}, U_{2(2)}, \ldots, U_{J(J)}$, and $\mathrm{I}_i(\mathrm{U}) = \mathrm{diag}[I_i(U_1), I_i(U_2), \ldots I_i(U_J)]$. In order to show design-unbiasedness of the estimating function $\Psi_s(\boldsymbol{\beta})$, we need to show that its design expectation is $\Psi_U(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}$.

The sampling design characteristics of a longitudinal survey can be thought of as those of a multiphase sample, as can be seen in Särndal, Swensson and Wretman (1992, Section 9.9). We therefore use the methodology of multiphase sampling for the derivations. We assume, without loss of generality, that there are only three waves; the derivations with just three waves show the patterns for general J, with respect to unbiasedness and variance.

As we mentioned earlier, we assume that w_{ij} is the cross-sectional weight for subject i at wave j, if that subject belongs to s_j, and zero otherwise. From the theory of multiphase sampling we have that for $i \in s_{1(1)}$, $w_{i1} = \pi_{i1}^{-1}$, $w_{i2} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}$, and $w_{i3} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}\pi_{i3|s_{2(1)}}^{-1}$; for $i \in s_{2(2)}$, $w_{i2} = \pi_{i2}^{-1}$ and $w_{i3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}$; and for $i \in s_{3(3)}$, $w_{i3} = \pi_{i3}^{-1}$; where π_{ij} is the inclusion probability of subject i in sample $s_{j(j)}$ and $\pi_{ij|s_{j-1(j')}}$ is the conditional inclusion probability of subject i in sample $s_{j(j')}$ given $s_{j-1(j')}$.

Using $E_p(\cdot)$ to denote the expectation with respect to the sampling design, we have:

$$E_p\left[\sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right] = E_p\left[\sum_{j=1}^{3} \sum_{i \in s_{j(j)}} B_i W_i \boldsymbol{e}_i\right]; \tag{3.4}$$

where $B_i = (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1}$ and $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$. For example, for $\sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i$ we obtain:

$$\begin{aligned} E_p\left[\sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i\right] &= E\left\{E\left[\sum_{i \in U_{2(2)}} B_i D_i \boldsymbol{e}_i \mid s_{2(2)}\right]\right\} = E\left\{\sum_{i \in U_{2(2)}} B_i D_i^{*} \boldsymbol{e}_i\right\} \\ &= \sum_{i \in U_{2(2)}} B_i D_i^{**} \boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i \in U_{2(2)}} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i, \end{aligned} \tag{3.5}$$

where $D_i = \text{diag}[0, I_i(U_2)w_{i2}I_i(s_{2(2)}), I_i(U_3)w_{i3}I_i(s_{3(2)})I_i(s_{2(2)})]$, $D_i^* = \text{diag}[0, (I_i(U_2)w_{i2}I_i(s_{2(2)})), (I_i(U_3)\pi_{i3|s_{2(2)}}I_i(s_{2(2)})) / (\pi_{i2}\pi_{i3|s_{2(2)}})]$, and $D_i^{**} = \text{diag}[0, (I_i(U_2)\pi_{i2}) / \pi_{i2}, \ (I_i(U_3)\pi_{i2}) / \pi_{i2}]$; similarly we can show that $E_p\left[\sum_{i\in s_{1(1)}} B_i W_i \boldsymbol{e}_i\right] = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$ and $E_p\left[\sum_{i\in s_{3(3)}} B_i W_i \boldsymbol{e}_i\right] = \sum_{i\in U_{3(3)}} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$. From these expressions and equation (3.4) we conclude that $E_p[\Psi_s(\boldsymbol{\beta})] = \Psi_U(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}$, which means that the estimating function $\Psi_s(\boldsymbol{\beta})$ is design-unbiased for the finite population estimating function.

Furthermore, as the target of inference is the superpopulation parameter, we need to guarantee that the model for μ_{ij} is such that $E_\xi(Y_{ij} - \mu_{ij}) = 0$ is satisfied, where $E_\xi(\cdot)$ represents the expectation with respect to model ξ. For if this is the case, we have:

$$E_{\xi p}[\Psi_s(\boldsymbol{\beta})] \overset{\text{def}}{=} E_\xi E_p[\Psi_s(\boldsymbol{\beta})] = E_\xi[\Psi_U(\boldsymbol{\beta})] = \sum_{i\in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \mathrm{I}_i(\mathrm{U}) E_\xi(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0};$$

so that the estimating function $\Psi_s(\boldsymbol{\beta})$ is model-design unbiased. The requirement $E_\xi(Y_{ij} - \mu_{ij}) = 0$ means that the mean model needs to be correctly specified; consequently, one needs to pay attention to residual diagnostics for the particular model being fitted.

3.2.2 A note on nonresponse

In the SDR, as in any other (longitudinal) survey, there is nonresponse. Some sampled individuals choose not to participate at all, whereas some subjects participate in some waves but not in others. The SDR remedies this situation by making a nonresponse adjustment to the cross-sectional survey weights.

Assume that the nonresponse adjustment at wave j is a multiplication by the inverse of the estimated wave j response probability $\hat{\pi}_{rij}$. For example, the nonresponse-adjusted weight for a person who *did* respond at wave 3 (and was first selected at wave 2), *i.e.*, for $i \in r_{3(2)}$, would be $w_{ri3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}\hat{\pi}_{ri3}^{-1}$.

We need to redefine the estimating equation, to include only the respondents, as $\Psi_r(\boldsymbol{\beta}) = \sum_{i\in r}(\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1} W_{ri}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, where the sum is over the respondent set r, *i.e.*,

over all the elements who belonged for the first time in any of the respondent sets $r_{1(1)}, r_{2(2)}, \ldots, r_{J(J)}$, and the matrix W_{ri} is $W_{ri} = \text{diag}[I_i(U_1)w_{ri1}, I_i(U_2)w_{ri2}, \ldots, I_i(U_J)w_{riJ}]$. Also, denote by $r_{j(j')}$ the set of cohort j' respondents at wave j. Obviously, $w_{rij} = 0$ if $i \notin r_j = \bigcup_{j'=1}^{j} r_{j(j')}$.

If additionally, the response mechanism (R) can be assumed to be MAR, we then have, for example for $\sum_{i \in r_{2(2)}} B_i W_{ri} \boldsymbol{e}_i$:

$$E_R\left\{\sum_{i\in r_{2(2)}} B_i W_{ri}\boldsymbol{e}_i\right\} = E_R\left\{\sum_{i\in s_{2(2)}} B_i D_i \boldsymbol{e}_i\right\} = \sum_{i\in s_{2(2)}} B_i D_i^{*}\boldsymbol{e}_i = \sum_{i\in s_{2(2)}} B_i D_i^{**}\boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i, \quad (3.6)$$

where $D_i = \text{diag}[0, I_i(U_2)w_{ri2}I_i(r_{2(2)}), I_i(U_3)w_{ri3}I_i(r_{3(2)})]$, $D_i^{*} = \text{diag}[0, (I_i(U_2)\pi_{ri2}) / (\pi_{i2} \times \hat{\pi}_{ri2}), (I_i(U_3)\pi_{ri3}) / (\pi_{i2}\pi_{i3|s_{2(2)}}\hat{\pi}_{ri3})]$, and $D_i^{**} = \text{diag}[0, I_i(U_2)w_{i2}, I_i(U_3)w_{i3}]$. The third equality in (3.6) requires that the nonresponse model used for $\hat{\pi}_{rij}$ satisfies $E_R[I_i(r_{j(j')})] \overset{\text{def}}{=} \pi_{rij} = \hat{\pi}_{rij}$. This means that in the model for $\hat{\pi}_{rij}$ we have to include as much available information, thought to influence the nonresponse propensity, as possible, in order for this assumption (*i.e.*, the MAR assumption) to be tenable. For example, if the nonresponse is thought to be independent across waves, one should include, in the model for $\hat{\pi}_{rij}$, as many variables from the corresponding wave as possible. If, on the other hand, it is reasonable to assume that the response propensity at a given wave depends on previous responses (and possibly response history), then those responses should be included in the response model, and so on.

The design as well as the model-design unbiasedness follow immediately from (3.6) together with the previous section. Hereafter we therefore ignore the issue of nonresponse for notational simplicity.

3.3 Variance and variance estimation

We now develop a (Taylor Series) linearization for the variance of the proposed estimator. The basic technique is due to Binder (1983). For simplicity in the derivations and notation we divide through by N; we redefine

$$\Psi_s(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \text{ and } \Psi_U(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

where $N = \sum_{j=1}^{J} N_j$. Let $\hat{\boldsymbol{\beta}}$ be our estimator, which satisfies $\Psi_s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, and let $\boldsymbol{\beta}_N$ be the "census estimator," which satisfies $\Psi_U(\boldsymbol{\beta}_N) = \mathbf{0}$. Assume $\boldsymbol{\beta}_N - \boldsymbol{\beta} = O_P(1/\sqrt{N_m})$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_P(1/\sqrt{n_m})$, with $N_m = \min\{N_1, N_2, \ldots, N_J\}$ and $n_m = \min\{n_1, n_2, \ldots, n_J\}$. We can write the total error of $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + (\boldsymbol{\beta}_N - \boldsymbol{\beta})$ = Sampling Error + Model Error. After some straightforward calculations, the total variance, or more precisely the total MSE, can be decomposed as:

$$V_{\text{Tot}} = E_{\xi p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' = V_{\text{Sam}} + 2 \otimes C_{\text{Sam-Mod}} + o(1/n_m), \tag{3.7}$$

where $2 \otimes A = A + A'$ for any matrix A, $V_{\text{Sam}} = E_\xi V_p$ is the "sampling variance" component, $2 \otimes C_{\text{Sam-Mod}}$ is the cross "sampling-model variance" component, $V_p = E_p[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)']$, $C_{\text{Sam-Mod}} = E_p C_\xi$, and $C_\xi = E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})'$. Furthermore, by Taylor series expansions we can obtain the following approximations: $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = [H(\boldsymbol{\beta}_N)]^{-1}\Psi_s(\boldsymbol{\beta}_N) + o_P(1/\sqrt{n_m})$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = [\hat{H}(\boldsymbol{\beta})]^{-1}\Psi_s(\boldsymbol{\beta}) + o_P(1/\sqrt{n_m})$, and $\boldsymbol{\beta}_N - \boldsymbol{\beta} = [H(\boldsymbol{\beta})]^{-1}\Psi_U(\boldsymbol{\beta}) + o_P(1/\sqrt{N_m})$, where we define $H(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}\mathrm{I}_i(\mathrm{U})(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})$ and $\hat{H}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}W_i(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})$.

We then get, for V_p and C_ξ in (3.7),

$$V_p = [H(\boldsymbol{\beta}_N)]^{-1}\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} + o_P(1/n_m), \tag{3.8}$$

$$\begin{aligned} C_\xi &= [\hat{H}(\boldsymbol{\beta})]^{-1}E_\xi[\Psi_s(\boldsymbol{\beta})\Psi_U'(\boldsymbol{\beta})][H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m) \\ &= N^{-1}[\hat{H}(\boldsymbol{\beta})]^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m), \end{aligned} \tag{3.9}$$

where $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = E_p[\Psi_s(\boldsymbol{\beta}_N)\Psi_s'(\boldsymbol{\beta}_N)]$ and $\hat{H}_{\Sigma V}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}[(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}W_i\Sigma_i \times V_i^{-1}(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})]$; the derivation of (3.9) can be found in the Appendix.

In conclusion, so far we have found that:

$$
\begin{aligned}
V_{\text{Tot}} &= E_\xi V_p + 2 \otimes E_p C_\xi + o(1/n_m) \\
&= E_\xi \left\{ [H(\boldsymbol{\beta}_N)]^{-1} \text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} \right\} \\
&\quad + 2 \otimes N^{-1} E_p \left\{ [\hat{H}(\boldsymbol{\beta})]^{-1} \hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} \right\} + o(1/n_m).
\end{aligned} \tag{3.10}
$$

In (3.10) all the terms can be estimated by "plugging in" the estimate $\hat{\boldsymbol{\beta}}$, except for the term $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$; this is the subject of the next section.

If the sampling fraction is small, *i.e.*, $n \ll N$, the first term in expression (3.10) is a good approximation for the total variance; *i.e.*, the expression for V_{Tot} is simply $E_\xi V_p$ (and lower order terms). If, on the other hand, the sampling fraction is large, both terms in (3.10) are required.

3.3.1 Design variance of the estimating function

In order to derive an expression for $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, we assume $J = 3$, as before. The methodology is that of two-phase sampling (more precisely, multiphase sampling), as discussed in chapter 9 of Särndal, *et al.* (1992). After some derivations (see Appendix), and defining $B_i = (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N} V_i^{-1}$, $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$, $\boldsymbol{e}_{i(1\cdots3)} = \boldsymbol{e}_i$, $\boldsymbol{e}_{i(2\cdots3)} = (0, e_{i2}, e_{i3})'$, and $\boldsymbol{e}_{i(3\cdots3)} = (0, 0, e_{i3})'$, we obtain:

$$
\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = \sum_{j=1}^{3} D_{(j)} = \sum_{j=1}^{3}\sum_{k=j}^{3} D_{(j)k}, \tag{3.11}
$$

where $D_{(j)} \overset{\text{def}}{=} N^{-2}\text{Var}_p\left(\sum_{i \in s_{j(j)}} B_i W_i \boldsymbol{e}_i \right) = \sum_{k=j}^{3} D_{(j)k}$, for $j = 1, 2, 3$,

$$
N^2 D_{(j)j} \overset{\text{def}}{=} \text{Var}\left[\sum_{i \in s_{j(j)}} w_{ij} B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(j\cdots3)} \right], \text{ for } j = 1, 2, 3,
$$

$$
N^2 D_{(j-1)j} \overset{\text{def}}{=} E\left\{ \text{Var}\left[\sum_{i \in s_{j(j-1)}} w_{ij} B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(j\cdots3)} \mid s_{j-1(j-1)} \right] \right\}, \text{ for } j = 2, 3,
$$

$$
N^2 D_{(1)3} \overset{\text{def}}{=} E\left\{ E\left[\text{Var}\left(\sum_{i \in s_{3(1)}} w_{i3} B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(3\cdots3)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\},
$$

and in the Appendix we show that:

$$N^2 D_{(j)k} = \mathrm{Var}\left[\sum_{i\in s_{k(j)}} w_{ik} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(k\cdots 3)}\right] - \mathrm{Var}\left[\sum_{i\in s_{k-1(j)}} w_{i,k-1} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_{i(k\cdots 3)}\right],$$

for $j = 1, 2, 3,$ and $3 \geq k > j$. In general, we have proved the following

Property 3.1 The (design) variance of $\Psi_s(\boldsymbol{\beta}_N)$ can be decomposed as:

$$\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$$

$$= \frac{1}{N^2}\sum_{j'=1}^{J}\sum_{j=j'}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_{j(j')}} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1(j')}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\} \quad (3.12)$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_j} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\}, \quad (3.13)$$

where we let $w_{i,j-1} = 0$ whenever $j = j'$, $w_{i0} = 0,$ and to get (3.13) we have changed variables and used the independence among cohorts.

In (3.11), (3.12), and (3.13) we have assumed that the cohorts are design-independent. However, in some cases this assumption may not be tenable; an example of such a case is the multiple frame situation discussed in the first part of Section 3.2. Another instance in which it may not be appropriate to assume cohort independence is when weight adjustments cross cohorts, which is the case of the SDR; we discuss this issue in Section 5. Calculations for the case of three cohorts, in the Appendix, show that (3.13) holds for the variance terms even without independence. The Appendix also identifies conditions under which it is a good approximation for the covariance terms.

3.3.2 Estimation

The estimation of V_{Tot} in (3.10) can be achieved as follows. $H(\boldsymbol{\beta}_N)$, $\hat{H}(\boldsymbol{\beta})$, and $H(\boldsymbol{\beta})$ can be estimated by $\hat{H}(\hat{\boldsymbol{\beta}})$. $\hat{H}_{\Sigma V}(\boldsymbol{\beta})$ can be estimated by $\hat{H}_{\Sigma V}(\hat{\boldsymbol{\beta}})$, where $\Sigma_i = \mathrm{Cov}[Y_i \mid X_i]$ can be estimated by $\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i'$.

We use (3.13) in Property 3.1 to estimate $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$. As long as there is a method to estimate the variance of (cross-sectional) Horvitz-Thompson (H-T) estimators, expression (3.13)

can be used. If we define $Z_{ij} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}$, we notice that each of the terms involved in the computation of (3.13), terms like $\mathrm{Var}_p\left[\sum_{i\in s_j} w_{ij} Z_{ij}\right]$, is simply the variance of a wave-j H-T estimator. Obviously, the variance estimation method needs to account for the sampling design as well as for any nonresponse and calibration adjustments performed, but this does not present any additional complications beyond what is found in any cross-sectional problem, as everything is implemented cross-sectionally. The SDR uses replication to estimate variances of cross-sectional estimators, but any method of design variance estimation can be used.

We use the cross-sectional replicate weights that SDR provides, but we do not re-estimate the parameter of interest at each replicate. First, note that we require replication only for the estimation of the "meat" $(\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)])$ of the design variance $(E_\xi V_p)$. Secondly, although $\hat{\boldsymbol{\beta}}$ does appear in the expression for the H-T estimator whose variance needs to be calculated (and re-calculated at each replicate), the work of Roberts, Binder, Kovačević, Pantel and Phillips (2003), who apply the "estimating function bootstrap" (Hu and Kalbfleisch 2000) to survey data, show that in a setting like ours, it is not necessary to re-compute the estimator at each replicate, but that the full-sample estimator suffices. This simplification speeds up the computation of the replicate estimates.

As a way of illustration, say we currently are at wave j, *i.e.*, we are estimating the j^{th} term in (3.13). The r^{th} replicate of the first term is $\sum_{i\in s_j} w_{ij}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{ij}^{(r)}$ is the r^{th} replicate weight for subject i at wave j, and the r^{th} replicate of the second term is $\sum_{i\in s_{j-1}} w_{i,j-1}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{i,j-1}^{(r)}$ is the r^{th} replicate weight for subject i at wave $j-1$.

4 Application to the SDR

The dataset we use is the restricted SDR data, under a license agreement from NSF. The SDR collects information about employment situation, principal employer, principal job, past employment, recent education, demographics, and disability, among others that vary from wave

to wave. We use only information requested in all the waves of interest: 1995, 1997, 1999, 2001, 2003, 2006, and 2008.

To illustrate our methodology, we constructed a model for individuals' salaries over time. The response is the log of salary (in the principal job), with an identity link function, and several covariates; modeling log of salary (as opposed to salary) is a standard practice. There are both time-independent covariates (such as gender) and time-dependent ones (such as employment sector). We have four major classes of covariates. The *Degree variables* are: degree field, years since degree, and age at graduation. The *Job variables* are: job field or category, sector, postdoc indicator, adjunct faculty indicator, hours worked per week in the principal job, weeks per year in the principal job, how related is the job to the doctoral degree, part-time for different reasons, number of months since started in the principal job, the starting month in the principal job, whether the employer/type of job has changed since previous wave, and whether changed employer/type of job since previous wave because was laid off or job terminated. The *Person's demographics* are: gender, citizenship status, race/ethnicity, presence of children in family, marital status, and spouse's working status. Finally, the "*Environment*" *variables* are: years since 1995, state (of employment), and the consumer price index (of the region of employment). The full list of variables, interactions, and categories can be found in Carrillo and Karr (2011). For categorical variables, the reference category is the one with the largest count.

The dataset for our model consists of 59,346 subjects and 190,693 observations, distributed as: $n_{95} = 30{,}234$, $n_{97} = 30{,}652$, $n_{99} = 26{,}732$, $n_{01} = 26{,}778$, $n_{03} = 24{,}956$, $n_{06} = 25{,}910$, and $n_{08} = 25{,}431$. Those data correspond to non-missing salaries between \$5,000 and \$999,995, for people with consistent ages across the waves, and with non-missing value for the variable indicating whether the (postsecondary educational institution) employer was public or private. The average (cross-sectional) survey weight for each of those waves are: $\bar{w}_{95} = 15.37$, $\bar{w}_{97} = 16.28$, $\bar{w}_{99} = 19.96$, $\bar{w}_{01} = 20.74$, $\bar{w}_{03} = 22.71$, $\bar{w}_{06} = 22.93$, and $\bar{w}_{08} = 24.88$.

The survey weights that we use for each wave are the final adjusted weights. These weights are the original design weights adjusted for nonresponse and post-stratification. However, the theory that we developed in Section 3 assumes that the weights are the inverse of the selection probabilities; in other words, the original design weights. This is a mismatch whose effect we plan to investigate in the future. On the other hand, the calculations in the last part of the Appendix (which do not assume anything about the weights) suggest that the effect of this mismatch is small.

The covariates and interactions that we considered were selected because they were suggested either by exploratory analyses or by the subject matter experts at the NSF. Carrillo and Karr (2011) present the estimated β coefficients in the model $y_{ij} = \log(\text{SALARY}_{ij}) = X'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}$, where X_{ij} includes the intercept along with the other covariates. This $\boldsymbol{\beta}$ corresponds to the one in model ξ, in Formula (3.1), and whose properties are discussed in Section 3. The working covariance matrix is estimated to be $\hat{V}_i = \hat{\phi}\mathbf{R}(\hat{\alpha})$, with $\hat{\phi} = \hat{\sigma}^2 = \left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij}\hat{e}_{ij}^2\right)\Big/\left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij} - p\right) = 0.196$, where $\hat{e}_{ij} = y_{ij} - X'_{ij}\hat{\boldsymbol{\beta}}$ and $p = 208$ is the number of covariates in X_{ij}, w_{ij} is the cross-sectional weight for subject i at wave j as long as $i \in s_j$ and zero otherwise. The estimate $\hat{\alpha}$ contains the $21 = (7 \times 6) / 2$ estimated auto-correlations $\hat{\alpha}_{jj'} = \hat{\alpha}_{j'j} = \left(\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}}\hat{e}_{ij}\hat{e}_{ij'}\right)\Big/\left(\hat{\phi}\left[\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}} - p\right]\right)$, for $j \neq j' =$ 1995, 1997, 1999, 2001, 2003, 2006, 2008, and $\hat{\alpha}_{jj} = 1$ for all j. These estimated values form the auto-correlation matrix:

$$\mathbf{R}(\hat{\alpha}) = \begin{pmatrix} 1 & \hat{\alpha}_{95,97} & \hat{\alpha}_{95,99} & \hat{\alpha}_{95,01} & \hat{\alpha}_{95,03} & \hat{\alpha}_{95,06} & \hat{\alpha}_{95,08} \\ & 1 & \hat{\alpha}_{97,99} & \hat{\alpha}_{97,01} & \hat{\alpha}_{97,03} & \hat{\alpha}_{97,06} & \hat{\alpha}_{97,08} \\ & & 1 & \hat{\alpha}_{99,01} & \hat{\alpha}_{99,03} & \hat{\alpha}_{99,06} & \hat{\alpha}_{99,08} \\ & & & 1 & \hat{\alpha}_{01,03} & \hat{\alpha}_{01,06} & \hat{\alpha}_{01,08} \\ & & & & 1 & \hat{\alpha}_{03,06} & \hat{\alpha}_{03,08} \\ & \text{sym} & & & & 1 & \hat{\alpha}_{06,08} \\ & & & & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.38 & 0.36 & 0.32 & 0.30 & 0.28 & 0.27 \\ & 1 & 0.42 & 0.36 & 0.33 & 0.32 & 0.31 \\ & & 1 & 0.46 & 0.38 & 0.36 & 0.34 \\ & & & 1 & 0.47 & 0.40 & 0.38 \\ & & & & 1 & 0.49 & 0.44 \\ & \text{sym} & & & & 1 & 0.55 \\ & & & & & & 1 \end{pmatrix}.$$

We now give some conclusions about salaries in the Ph.D. workforce based on the estimated coefficients, which appear in Carrillo and Karr (2011). First of all, a sensible estimate of mean salary considers the intercept, the hours worked per week (whose average is 47), and years since degree (average of 15); so that an estimate of the overall average is $\exp(9.4 + 47 \times 0.038 - 47^2 \times 0.0003 + 15 \times 0.03 - 15^2 \times 0.0006) = \$52,067$, for a subject with all other continuous covariates equal to zero and in the reference of all categorical covariates.

All other things being constant, women's salaries are about 93.4% those of men, whereas race does not seem to have an effect on salaries. The gender $\times$ years since 1995 interaction is not significant; therefore this salary differential is not changing over time. Notice that with a single year's data, we would not be able to evaluate the effect of time. Even more important than that, using only the data from a single wave, say 2008, we would not be able to assess whether the effect of being female is changing over time.

Doctorate holders with a management job have the highest salaries, followed by those in health occupations; on the other hand, those with the lowest salaries are the ones employed in "other" occupations, followed by those in political science.

Among employment sectors, salaries are highest in for-profit industry (20% higher than for the reference category of tenured faculty in public 4-year institutions), followed in order by the federal government, self-employment, non-profit industry, all of which are higher than the reference category. The lowest salaries are those in two-year colleges and in two- and four-year institutions for which tenure is not applicable.

The highest single negative effect on salaries also occurs within the education sector. Those with positions as adjunct faculty members have salaries that are approximately 59% of the salaries of comparable doctorate holders. Not surprisingly, postdoctoral salaries are only about 74% of the salaries of comparable people in other types of positions.

Sector is also a contributing factor to the hard-to-interpret dependence of salary on the starting month for the current position: salaries are lower for starting months of August and September.

Additional analyses show that the monthly effect is present only in the education sector, where, as we have seen, salaries are lower than in industry or government, and in which starting months of August and September are common. Therefore, sector is part of the answer, but not the entire answer. Finer-grained divisions of the education sector, using Carnegie classifications, further reduce, but do not remove, the significance of monthly effects. The SDR does not seem to contain sufficient data to remove the monthly effects entirely, so we have retained the SDR definition of sector.

People with degrees in computing and information sciences have the highest salaries (around 20% higher than in the biological sciences), followed by those in electrical and computer engineering and in economics (approximately 16% higher). Doctorate holders in agricultural and food sciences, environmental life sciences, earth, atmospheric, and ocean sciences, and in "other" social sciences have the lowest salaries. The "other" social sciences are the social sciences excluding economics and political science.

Married people have the highest salaries, followed by those who are in married-like relationships, widowed, separated, divorced, and never married. The latter have salaries only around 89% as high as the married ones; one could argue that there is some association between never married and age. The presence of children older than two is associated with higher salaries, but the presence of children younger than two is not.

Doctorate holders with jobs only somewhat related to their degree field make around 93% of what people with closely related jobs (the reference category) do. If the job is not related to the doctoral degree as the result of a change in career or professional interests, they make around 82% of what people with closely related jobs do. On the other hand, those with jobs not related for other reasons make only about 76% of what the reference category does.

There is an increase of around 3% for every additional year since doctorate graduation, although there is a diminishing effect for higher number of years. We interpret this as the effect

of experience. There is a small penalty for receiving the doctorate later in life; for every additional year of age at graduation, the salary reduces by 1%.

We also found that the regional Consumer Price Index (CPI) is significant. The higher the CPI, the higher the salary. We could not use the CPI associated with the labor market of employment because the SDR data do not identify geography beyond the state. We included the state in the model as a proxy for cost of living; the state effect is highly significant and some state coefficients are among the highest overall. The highest salaries are in California, Washington D.C. and its suburbs, and New York City and its suburbs. On the other hand, the lowest salaries are in Puerto Rico, Vermont, Montana, Maine, Idaho, South Dakota, North Dakota, and in the Territories/Abroad.

Having a part-time job due to being retired or semi-retired is significant and in several significant interactions. Because of this, we do not think that the available data present the full picture about retirement, for example, for people who are (semi-)retired and yet have full-time jobs.

Finally, we analyzed residuals; Figures 4.1 and 4.2 show a Box and Whisker plot of standardized residuals by year and a spaghetti plot of standardized residuals, respectively.

Figure 4.1 shows that the model fits reasonably well for all the reference years as most of the standardized residuals lie between -2 and 2. Also, the distributions of residuals do not seem to greatly differ from year to year.

From Figure 4.2 we also conclude that the model fits reasonably well for most people, as most of the lines fluctuate between -2 and 2. Nonetheless, there are a few people for which the model seems to greatly over-predict in 2003 and some few people for whom that happens in 2006. We included several terms in the model to correct this issue but clearly none seemed to do so completely.

The last thing we tried was to produce exploratory classification trees for these residual blips. We found that, in the dataset available, the only thing related to them was the survey mode. The

blips in 2003 are disproportionately high for web responses, and the blips in 2006 are disproportionately high for CATI responses. We conclude that either there is a mode effect in these two years or those respondents have something different, in those years, that is not included in the available variables.

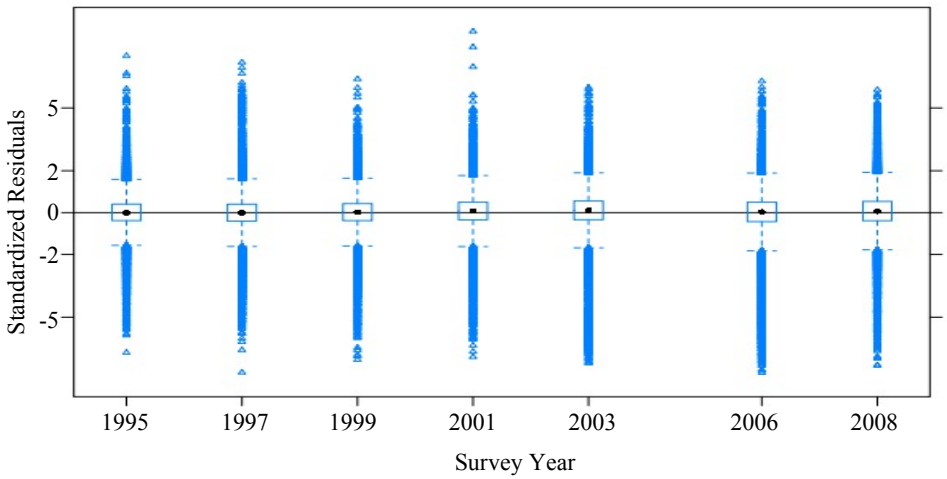

Figure 4.1 Box and Whisker plot of standardized residuals by year

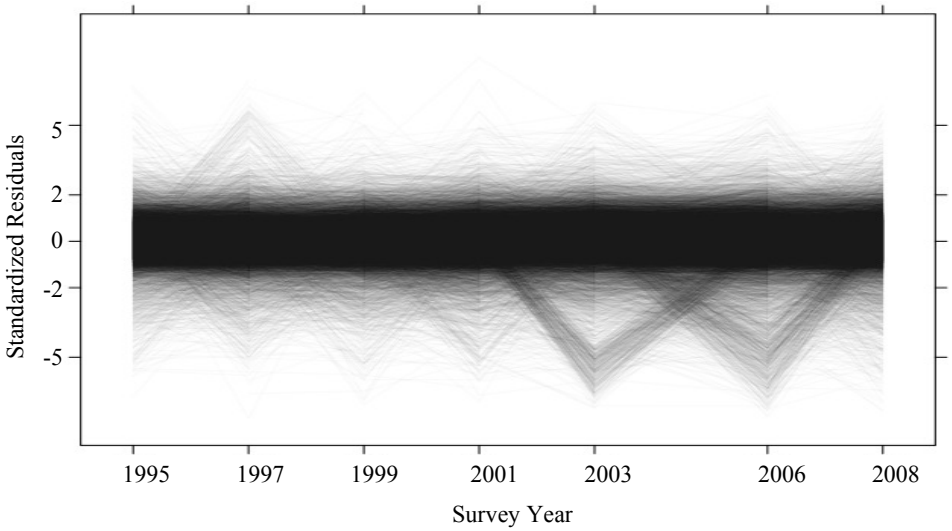

Figure 4.2 Spaghetti plot of standardized residuals

Finally, the plot of fitted values versus observed (which can be found in Carrillo and Karr 2011) also shows a similar story. For most observations the model performs well, apart from those few cases in 2003 and 2006 for whom there is large over-estimation.

5 Conclusions and future research

We have proposed a novel approach to combining different cohorts of a longitudinal survey. The major requirement of our method is that there is a cross-sectional survey weight for each wave, or that one can be built from available information. This weight should allow for statistical inference to the population of interest at the corresponding wave. In that case, our method should perform better than usual estimation procedures (where the auto-correlation is not incorporated) in many practical situations, in particular when there is a high auto-correlation among responses from the same subject.

In general, survey practitioners avoid as much as possible the use of multiple survey weights. However, in the case of rotating panels this is an appealing approach for at least two reasons. On the one hand, it allows for the use of all the available data in a clear and cohesive way in a single analysis procedure. On the other hand, we have shown how readily available cross-sectional survey weights can be directly used for longitudinal analysis, without the need to develop, store, and distribute an additional longitudinal weight or weights.

Our method is directly applicable to any kind of longitudinal survey as long as there are cross-sectional survey weights available (or these can be created) at each wave, and these weights represent the population of interest at the particular wave.

For the theory that we developed about the variance of the estimator proposed, we utilized the (cross-sectional) design weights w_{ij}, which are the inverse of the inclusion probabilities. Yet for the application in our model for salary in the SDR we used the final (cross-sectional) survey weights, which are not the original design weights, but adjusted (in the usual way) weights. This mismatch requires further exploration.

Similarly, in our derivations of the variance, we assumed that the cohorts were independent. However, the SDR does not totally satisfy this assumption for two reasons. Firstly, at any particular wave, the selection of the sample from the old cohorts is not performed independently across cohorts. In order to reduce the number of strata, since 1991 the NSF has collapsed strata

over year of degree receipt for the old cohorts. Additionally, the post-stratification adjustments made to the design weights do not condition over cohort either, and as a result, weights are shared across cohorts. This sampling selection scheme and weighting adjustment procedure violate the independence across cohorts. Some additional calculations (included in the Appendix) have shown that the independence among cohort is not such a crucial requirement for our variance estimation method to produce good approximations, as explained in Section 3.3.1. In future research we plan to evaluate in more detail the impact of this issue.

Acknowledgements

This research was supported by NSF grant SRS-1019244 to the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank Paul Biemer of RTI International, Stephen Cohen and Nirmala Kannankutty of the National Center for Science and Engineering Statistics at NSF, and Criselda Toto, formerly of NISS, for numerous insightful discussions during the research. We are also grateful to the Associate Editor and two referees for their useful suggestions.

Appendix - Proofs

- To develop an expression for C_ξ, we first simplify $\Psi_s(\beta)\Psi'_U(\beta)$. Let $\boldsymbol{F}_{i(k)} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots 3)}$ for $k = 1, 2, 3,$ then we have:

$$N^2\Psi_s(\beta)\Psi'_U(\beta) = \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in U} \boldsymbol{e}'_i \mathrm{I}_i(\mathrm{U}) B'_i = \left[\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right]\left[\sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\notin s} \boldsymbol{F}'_{i(1)}\right]$$

$$= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\notin s} \boldsymbol{F}'_{i(1)}$$

$$= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_i B'_i + \sum_{i\in s}\sum_{\substack{k\in s \\ k\neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k + \mathrm{A},$$

where $\mathrm{A} = \left(\sum_{i \in s} B_i W_i \boldsymbol{e}_i \right) \left(\sum_{i \notin s} \boldsymbol{F}'_{i(1)} \right)$, and let $\mathrm{B} = \sum_{i \in s} \sum_{\substack{k \in s \\ k \neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k$. The two sums in A are model-independent, $\boldsymbol{e}_i$ and $\boldsymbol{e}'_k$ (in B) are two model-independent terms, and A and B both have model-expectation zero; therefore, $E_\xi[\Psi_s(\boldsymbol{\beta})\, \Psi'_U(\boldsymbol{\beta})] = N^{-2} \sum_{i \in s} B_i W_i E_\xi[\boldsymbol{e}_i \boldsymbol{e}'_i]\, B'_i = N^{-2} \sum_{i \in s} B_i W_i \Sigma_i B'_i = N^{-1} \hat{H}_{\Sigma V}(\boldsymbol{\beta})$; equation (3.9) follows.

- We now develop the expression for $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, the design variance of the estimating function; we redefine $B_i = (\partial \boldsymbol{\mu}'_i / \partial \boldsymbol{\beta})\,|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}\, V_i^{-1}$ and $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$; then

$$\begin{aligned}
\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] &= \mathrm{Var}_p \left(\frac{1}{N} \sum_{i \in s} B_i W_i \boldsymbol{e}_i \right) \\
&= \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{1(1)}} B_i W_i \boldsymbol{e}_i + \sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i + \sum_{i \in s_{3(3)}} B_i W_i \boldsymbol{e}_i \right) \\
&= \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{1(1)}} B_i W_i \boldsymbol{e}_i \right) + \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i \right) \\
&\quad + \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{3(3)}} B_i W_i \boldsymbol{e}_i \right) = D_{(1)} + D_{(2)} + D_{(3)}, \qquad \text{(A.1)}
\end{aligned}$$

where, for line (A.1), we assume that the (three) cohorts are design-independent. Now, $N^2 D_{(1)} = \mathrm{Var}_p \left[\sum_{i \in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{I}_{i(1)}\}\, \boldsymbol{e}_i \right] = \mathrm{Var}_p \left[\sum_{i \in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{e}_i\}\, \boldsymbol{I}_{i(1)} \right]$, where $\mathrm{Diag}\{\boldsymbol{e}\}$ is, for a column vector $\boldsymbol{e}$, a diagonal matrix with diagonal entries being the elements of $\boldsymbol{e}$, and $\boldsymbol{I}_{i(1)} = \left(I_i(s_{1(1)}), I_i(s_{2(1)}) I_i(s_{1(1)}), I_i(s_{3(1)}) I_i(s_{2(1)}) I_i(s_{1(1)}) \right)'$. Similarly we can get $N^2 D_{(2)} = \mathrm{Var}_p \left[\sum_{i \in U_{2(2)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(2)} \right]$ and $N^2 D_{(3)} = \mathrm{Var}_p \left[\sum_{i \in U_{3(3)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{e}_i\}\, \boldsymbol{I}_{i(3)} \right]$, where $\boldsymbol{I}_{i(2)} = \left(0, I_i(s_{2(2)}), I_i(s_{3(2)}) I_i(s_{2(2)}) \right)'$, and $\boldsymbol{I}_{i(3)} = \left(0, 0, I_i(s_{3(3)}) \right)'$. Now, let us concentrate on $D_{(1)}$; letting $C_i = B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}$, we have:

$$
\begin{aligned}
N^2 D_{(1)} &= \mathrm{Var}_p\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)}\right) = \mathrm{Var}\left\{E\left[\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)}\right]\right\} \\
&\quad + E\left\{\mathrm{Var}\left[\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)}\right]\right\} \\
&= \mathrm{Var}\left\{E\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\} \\
&\quad + E\left\{\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right. \\
&\quad \left. + E\left[\mathrm{Var}\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\} \\
&= N^2 D_{(1)1} + N^2 D_{(1)2} + N^2 D_{(1)3}. \qquad \text{(A.2)}
\end{aligned}
$$

Let us do each of the terms in (A.2) in turn, beginning with $N^2 D_{1(1)}$, we have:

$$
E\left(\sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) = \sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}^{(1)}_{i(1)},
$$

where $\boldsymbol{I}^{(1)}_{i(1)} = (I_i(s_{1(1)}), I_i(s_{2(1)}) I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)}) I_i(s_{1(1)}))'$, then

$$
\begin{aligned}
&E\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right] \\
&\qquad = \sum_{i\in U_{1(1)}} C_i \boldsymbol{I}^{(2)}_{i(1)} = \sum_{i\in U_{1(1)}} B_i \mathrm{I}^{(1)}_i(\mathrm{U}) \mathrm{Diag}\{\boldsymbol{I}^{(2)}_{i(1)}\} \boldsymbol{e}_i \\
&\qquad = \sum_{i\in U_{1(1)}} F_i \left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}}\right]' \\
&\qquad = \sum_{i\in U_{1(1)}} F_i \mathbf{1}_3 \frac{I_i(s_{1(1)})}{\pi_{i1}} = \sum_{i\in U_{1(1)}} w_{i1(1)} \boldsymbol{F}_{i(1)} I_i(s_{1(1)}),
\end{aligned}
$$

where $\boldsymbol{I}^{(2)}_{i(1)} = (I_i(s_{1(1)}), \pi_{i2|s_{1(1)}} I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} \pi_{i2|s_{1(1)}} I_i(s_{1(1)}))'$, $\mathrm{I}^{(1)}_i(\mathrm{U}) = \mathrm{diag}[I_i(U_1)/\pi_{i1}$, $I_i(U_2) / (\pi_{i1}\pi_{i2|s_{1(1)}})$, $I_i(U_3) / (\pi_{i1}\pi_{i2|s_{1(1)}}\pi_{i3|s_{2(1)}})]$, $F_i = B_i \mathrm{I}_i(\mathrm{U}) \mathrm{Diag}\{\boldsymbol{e}_i\}$, and $\mathbf{1}_3 = (1, 1, 1)'$; this implies that $N^2 D_{(1)1} = \mathrm{Var}\left[\sum_{i\in U_{1(1)}} w_{i1} B_i \mathrm{I}_i(\mathrm{U})\, \boldsymbol{e}_i I_i(s_{1(1)})\right] = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i1} \boldsymbol{F}_{i(1)}\right]$.

For $N^2 D_{(1)2}$, we have:

$$
\begin{aligned}
&E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \\
&\quad = \sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(1)} = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i^{(1)}(\mathrm{U})\mathrm{Diag}\{\boldsymbol{I}_{i(1)}^{(3)}\}\boldsymbol{e}_i \\
&\quad = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}\right]' \\
&\quad = \sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]',
\end{aligned}
$$

where $\boldsymbol{I}_{i(1)}^{(3)} = (I_i(s_{1(1)}), I_i(s_{2(1)})I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)})I_i(s_{1(1)}))'$; then,

$$
\begin{aligned}
&\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(4)} \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[0, I_i(s_{2(1)}), I_i(s_{2(1)})]' \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} I_i(s_{2(1)})\mathbf{1}_{02} \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots 3)} \mid s_{1(1)}\right], \qquad \text{(A.3)}
\end{aligned}
$$

where $\boldsymbol{I}_{i(1)}^{(4)} = [\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]'$, $\mathbf{1}_{02} = (0, 1, 1)'$, and line (A.3) is because, conditional on $s_{1(1)}, \pi_{i2|s_{1(1)}}$ is constant and therefore the variance of that component is zero. This means that:

$$
\begin{aligned}
N^2 D_{(1)2} &= E\left\{\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\} \\
&= E\left\{\mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)} \mid s_{1(1)}\right]\right\} \\
&= \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)}\right] - \mathrm{Var}\left\{E\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)} \mid s_{1(1)}\right]\right\}
\end{aligned}
$$

$$= \text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)}\right] - \text{Var}\left\{E\left[\sum_{i\in s_{2(1)}} w_{i2|s_{1(1)}} w_{i1}\boldsymbol{F}_{i(2)} \mid s_{1(1)}\right]\right\}$$

$$= \text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)}\right] - \text{Var}\left\{\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i}\right\}.$$

We can, similarly, show that:

$$N^2 D_{(1)3} = E\left\{E\left[\text{Var}\left(\sum_{i\in U_{1(1)}} B_i W_i \text{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\}$$

$$= E\left\{E\left[\text{Var}\left(\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)}) I_i(s_{1(1)})\boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\}$$

$$= E\left\{\text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)})\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right]\right.$$

$$\left. -\text{Var}\left[E\left(\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)})\boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\}$$

$$= E\left\{\text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right] - \text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right]\right\}$$

$$= \text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)}\right] - \text{Var}\left[E\left(\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right)\right]$$

$$-\text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)}\right] + \text{Var}\left[E\left(\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right)\right]$$

$$= \text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)}\right] - \text{Var}\left[\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i(3)}\right]$$

$$-\text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)}\right] + \text{Var}\left[\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i(3)}\right].$$

With similar calculations, we obtain the corresponding expressions for $N^2 D_{(2)}$, $N^2 D_{(2)2}$, $N^2 D_{(2)3}$, and $N^2 D_{(3)} = N^2 D_{(3)3}$.

- Finally, we sketch the development of an expression for $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ without assuming independence among cohorts. First, notice that $\Psi_s(\boldsymbol{\beta}_N)$ can be written as:

$$\sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} w_{i1} & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix} + \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$- \sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$= \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i - \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_3} w_{i3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix};$$

letting $\boldsymbol{z}_i = B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i$, $\boldsymbol{z}_{i(2\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, e_{i2}, e_{i3}]'$, and $\boldsymbol{z}_{i(3\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, 0, e_{i3}]'$, $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ can be expanded as:

$$\mathrm{Var}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i \right] + \mathrm{Var}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)} \right]$$
$$+ \mathrm{Var}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)} \right] + \mathrm{Var}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$+ \mathrm{Var}_p \left[\sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right] - 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)} \right]$$
$$+ 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)} \right]$$
$$- 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right] + 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$- 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)} \right] + 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$- 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right] - 2\mathrm{Cov}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$+ 2\mathrm{Cov}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right] - 2\mathrm{Cov}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)}, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right]$$

$$
\begin{aligned}
&= \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_i\right] + \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 3)}\right] \\
&+ \mathrm{Var}_p\left[\sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 1)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 1)}, \sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right].
\end{aligned}
\tag{A.4}
$$

In this last expression, the first thing we notice is that *all* the diagonal elements in *all* the covariance terms are exactly equal to zero; this means that whether or not the cohorts are independent of one another, expression (3.13) is exact for the variance terms.

To analyze the importance of the covariance terms, we concentrate on the term in line (A.4); the conclusion for the other terms is the same; note that this term can be written as:

$$
2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} e_{i1}\\ e_{i2}\\ 0\end{pmatrix}, \sum_{i\in s_3} w_{i3}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0\\ 0\\ e_{i3}\end{pmatrix} - \sum_{i\in s_2} w_{i2}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0\\ 0\\ e_{i3}\end{pmatrix}\right];
$$

Property 3.1 states that if the *cohorts* are design-independent, all the covariance terms are exactly equal to zero. In addition to that, from this last expression we conclude, trivially, that if the *waves* are design-independent, all the covariance terms are equal to zero too. This formula for the term in line (A.4) also implies that if the individual weights do not vary greatly between consecutive waves, and there is a high overlap between consecutive waves, the covariance terms are not too large. Finally, if the overlap is small, it is reasonable to assume design-independence between the waves, and then the covariance terms can be safely approximated by zero.

References

Ardilly, P., and Lavallée, P. (2007). Weighting in rotating samples: The SILC survey in France. *Survey Methodology*, 33, 2, 131-137.

Berger, Y.G. (2004a). Variance estimation for change: An evaluation based upon the 2000 finnish labour force survey. Proceedings. European Conference on Quality and Methodology in Official Statistics.

Berger, Y.G. (2004b). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 4, 451-467.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Carrillo, I.A., Chen, J. and Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *The Canadian Journal of Statistics*, 38, 4, 540-554.

Carrillo, I.A., Chen, J. and Wu, C. (2011). A pseudo-GEE approach to analyzing longitudinal surveys under imputation for missing responses. *Journal of Official Statistics*, 27, 2, 255-277.

Carrillo, I.A., and Karr, A.F. (2011). Combining cohorts in longitudinal surveys. Technical Report 180, National Institute of Statistical Sciences, Research Triangle Park, NC. URL http://www.niss.org/sites/default/files/tr180.pdf.

Carrillo, I.A., and Karr, A.F. (2012). Estimating change with multi-cohort longitudinal surveys. In preparation.

Cox, B.G., Grigorian, K., Wang, R. and Harter, R. (2010). 2008 Survey of Doctorate Recipients Weighting Implementation Report, document prepared by the National Opinion Research Center (NORC) for the National Science Foundation (NSF).

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd Edition. Oxford University Press, New York.

Hedeker, D., and Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. New Jersey: John Wiley & Sons, Inc., Hoboken.

Hirano, K., Imbens, G.W., Ridder, G. and Rubin, D.B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69, 6, 1645-1659.

Hu, F., and Kalbfleisch, J.D. (2000). The estimating function bootstrap (Pkg: P449-495). *The Canadian Journal of Statistics*, 28, 3, 449-481.

Larsen, M.D., Qing, S., Zhou, B. and Foulkes, M.A. (2011). Calibration estimation and longitudinal survey weights: Application to the NSF Survey of Doctorate Recipients. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 1360-1374.

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lohr, S. (2007). Recent developments in multiple frame surveys. In *Joint Statistical Meeting of the American Statistical Association*, 3257-3264.

McLaren, C.H., and Steel, D.G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. *Survey Methodology*, 26, 2, 163-172.

National Science Foundation, National Center for Science and Engineering Statistics (2012). Survey of doctorate recipients. http://www.nsf.gov/statistics/srvydoctoratework/, accessed Feb. 09 2012.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21, 1, 43-52.

Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 2, 173-181.

Rao, J.N.K., and Wu, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 492, 1494-1503.

Roberts, G., Binder, D., Kovačević, M., Pantel, M. and Phillips, O. (2003). Using an estimating function bootstrap approach for obtaining variance estimates when modelling complex health survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, Halifax.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Smith, P., Lynn, P. and Elliot, D. (2009). Sample design for longitudinal surveys. In *Methodology of Longitudinal Surveys*, (Ed., P. Lynn). Wiley, Chichester, Chapter 2, 21-33.

Song, P.X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer Series in Statistics. New York: Springer.

Steel, D., and McLaren, C. (2007). Design and analysis of repeated surveys. Keynote lecture. International Conference on Quality Management of Official Statistics, Korea.

Vieira, M.D.T. (2009). Analysis of Longitudinal Survey Data: Allowing for the Complex Survey Design in Covariance Structure Models. VDM Verlag.

Vieira, M.D.T., and Skinner, C.J. (2008). Estimating models for panel survey data under complex sampling. *Journal of Official Statistics*, 24, 3, 343-364.