NISS

Combining Cohorts in Longitudinal Surveys

Iván A. Carrillo and Alan F. Karr,

Technical Report 180
December 2011

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC
www.niss.org

Combining Cohorts in Longitudinal Surveys

Iván A. Carrillo and Alan F. Karr [1]

December 6, 2011

Abstract

A question that commonly arises in longitudinal surveys is the issue of how to combine differing cohorts of the survey. The different cohorts can represent disjoint populations, a single population, or overlapping populations. In this paper we present, under a superpopulation approach, a course of action for combining different cohorts in a longitudinal survey with a repeated-panel/rotating-panel design; namely the Survey of Doctorate Recipients, conducted by the U.S. National Science Foundation. In this case the cohorts represent non-overlapping populations. The procedure builds upon the Weighted Generalized Estimation Equation method existing in the literature for handling missing waves in longitudinal studies. Although our method is set up under a joint-randomization framework, which takes into account the superpopulation model, our simulations show that the method also performs well for estimating well-defined finite population quantities, as well as superpopulation parameters. We also propose a design-based, and a joint-randomization, variance estimation method.

Key Words: Estimating change; Finite population parameters; Replication variance estimation; Rotating panel surveys; Superpopulation parameters; Weighted Generalized Estimating Equations.

[1] Iván A. Carrillo, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27713, USA. E-mail: ivan@niss.org; Alan Karr, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27713, USA

1 Introduction

The Survey of Doctorate Recipients (SDR) is a National Science Foundation 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 `http://nsf.gov/statistics/srvydoctoratework/`. 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. at survey reference week, and are 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.

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 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 neither a repeated-panel nor a rotating panel design, but has features of both. It is not a repeated-panel design because of the removal of some subjects at each wave. And it is not 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.

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, and 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, Imbens, Ridder, and Rubin (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 and do not mention other variables of interest, such as binary responses or counts.

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 rotating panel surveys, or in other words with multiple-cohort surveys, one needs to ignore some (many times a lot of) 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 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.

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 multiple-cohort surveys. Section 4 shows the superior performance of the proposed methodology, with respect to the usual estimation method, with a simple simulation exercise. Then we present the application of the methodology to the SDR. And finally we offer a few discussion points in Section 6.

2 The SDR Design

2.1 Finite Population

The SDR finite population of interest can be represented as in Figure 1. At wave 1, i.e. the first time of interest, there is a finite set, $U_{1(1)} = U_1$, of $N_{1(1)} = N_1$ PhD holders, either recent or not, who satisfy the requirements of the SDR (hold a doctoral degree in a science, engineering or health field, are non-institutionalized, live in the U.S., and are under the age of 76).

At wave 2, i.e. the second time of interest, only a subset of those 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 PhD recipients, who have obtained their degree since wave 1, and also satisfy

$$
\begin{array}{ccccccccccc}
j:\ 1 & & 2 & & 3 & & \cdots & & J-1 & & J \\
\hline
U_{1(1)} & \supseteq & U_{2(1)} & \supseteq & U_{3(1)} & \supseteq & \cdots & \supseteq & U_{J-1(1)} & \supseteq & U_{J(1)} \\
N_{1(1)} & \geq & N_{2(1)} & \geq & N_{3(1)} & \geq & \cdots & \geq & N_{J-1(1)} & \geq & N_{J(1)} \\
 & & U_{2(2)} & \supseteq & U_{3(2)} & \supseteq & \cdots & \supseteq & U_{J-1(2)} & \supseteq & U_{J(2)} \\
 & & N_{2(2)} & \geq & N_{3(2)} & \geq & \cdots & \geq & N_{J-1(2)} & \geq & N_{J(2)} \\
 & & & & U_{3(3)} & \supseteq & \cdots & \supseteq & U_{J-1(3)} & \supseteq & U_{J(3)} \\
 & & & & N_{3(3)} & \geq & \cdots & \geq & N_{J-1(3)} & \geq & N_{J(3)} \\
 & & & & & & \ddots & & \vdots & & \vdots \\
 & & & & & & & & U_{J-1(J-1)} & \supseteq & U_{J(J-1)} \\
 & & & & & & & & N_{J-1(J-1)} & \geq & N_{J(J-1)} \\
 & & & & & & & & & & U_{J(J)} \\
 & & & & & & & & & & N_{J(J)} \\
\hline
U_1 & & U_2 & & U_3 & & \cdots & & U_{J-1} & & U_J \\
N_1 & & N_2 & & N_3 & & \cdots & & N_{J-1} & & N_J
\end{array}
$$

Figure 1: SDR Finite Population

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 $U_2 = U_{2(1)} \cup U_{2(2)}$.

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 not in scope anymore at the current wave, and they do not make part of the current target population. And secondly, some people, namely the recent graduates, are added to the target population in the current wave. We denote by $j = 1,2,\dots,J$ the wave (outside the parenthesis) and by $j' = 1,2,\dots,J$ the cohort to which a subject belongs (inside the parenthesis).

2.2 Sampling

The sampling design of the SDR has a similar structure to the finite population and is depicted in Figure 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.

$$\begin{array}{cccccccccccc}
j: & 1 & & 2 & & 3 & & \cdots & & J-1 & & J \\ \hline
& s_{1(1)} & \supseteq & s_{2(1)} & \supseteq & s_{3(1)} & \supseteq & \cdots & \supseteq & s_{J-1(1)} & \supseteq & s_{J(1)} \\
& n_{1(1)} & \geq & n_{2(1)} & \geq & n_{3(1)} & \geq & \cdots & \geq & n_{J-1(1)} & \geq & n_{J(1)} \\
& & & s_{2(2)} & \supseteq & s_{3(2)} & \supseteq & \cdots & \supseteq & s_{J-1(2)} & \supseteq & s_{J(2)} \\
& & & n_{2(2)} & \geq & n_{3(2)} & \geq & \cdots & \geq & n_{J-1(2)} & \geq & n_{J(2)} \\
& & & & & s_{3(3)} & \supseteq & \cdots & \supseteq & s_{J-1(3)} & \supseteq & s_{J(3)} \\
& & & & & n_{3(3)} & \geq & \cdots & \geq & n_{J-1(3)} & \geq & n_{J(3)} \\
& & & & & & & \ddots & & \vdots & & \vdots \\
& & & & & & & & & s_{J-1(J-1)} & \supseteq & s_{J(J-1)} \\
& & & & & & & & & n_{J-1(J-1)} & \geq & n_{J(J-1)} \\
& & & & & & & & & & & s_{J(J)} \\
& & & & & & & & & & & n_{J(J)} \\ \hline
& s_1 & & s_2 & & s_3 & & \cdots & & s_{J-1} & & s_J \\
& n_1 & & n_2 & & n_3 & & \cdots & & n_{J-1} & & n_J
\end{array}$$

Figure 2: SDR Sample

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 sub-sample $s_{2(1)}$, 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 PhD 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 sub-sample 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.

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 particularity 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 exorbitant (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.

Just to reiterate, the notation we use is $s_{j(j')} = s_{\text{wave(cohort)}}$. This is, the quantity outside the parenthesis represents the wave to which the sample refers; and the quantity inside the parenthesis is the sample's cohort, i.e. the wave at which the sample was *first selected.* On the other hand, the notation for the weights is $w_{ij} = w_{\text{subject wave}}$; this is, 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 on the $p \times 1$ vector parameter $\boldsymbol{\beta}$ in the following generalized estimating equations (GEE) model:

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

where Y_{ij} is the response variable for subject i at wave j, X_{ij} is a $p \times 1$ vector of covariates associated with it, $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iJ})'$, $X_i = (X_{i1}, X_{i2}, \cdots, X_{iJ})$ 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 $\mathrm{Cov}[Y_i|X_i] = V_i = A_i^{1/2}\mathbf{R}(\alpha)A_i^{1/2}$, a "working" covariance matrix; where $A_i = \mathrm{diag}[\phi\nu(\mu_{ij})]$ 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 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 drop out of the study. In other words, it is common to have some subjects who respond at the beginning of the study and then do not respond after a certain wave; so that the latter responses for some subjects are not observed.

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 $\hat{\Delta}_i=\text{diag}[R_{i1}\hat{\pi}_{i1}^{-1}, R_{i2}\hat{\pi}_{i2}^{-1}, \ldots, R_{iJ}\hat{\pi}_{iJ}^{-1}]$, R_{ij} is the response indicator for subject i at wave j, and $\hat{\pi}_{ij}$ 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 $\hat{\Delta}_{wi}=\text{diag}[w_iR_{i1}\hat{\pi}_{i1}^{-1}, w_iR_{i2}\hat{\pi}_{i2}^{-1}, \ldots, w_iR_{iJ}\hat{\pi}_{iJ}^{-1}]$.

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 multiple-cohort problem. This is the subject of 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 super-population parameter $\boldsymbol{\beta}$, or the corresponding finite population quantity $\boldsymbol{\beta}_N$ (see "Unbiasedness" below), by the solution to the estimating equations:

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

where the sum is over the sample s, i.e. over all the elements selected (for the first time) in any of the samples $s_{1(1)}, s_{2(2)}, \ldots, s_{J(J)}$; and the diagonal matrix W_i is

$$W_i = \begin{bmatrix} I_i(U_1)w_{i1} & & & \mathrm{O} \\ & I_i(U_2)w_{i2} & & \\ & & \ddots & \\ \mathrm{O} & & & I_i(U_J)w_{iJ} \end{bmatrix},$$

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 the “Unbiasedness” section below we argue why this is a reasonable estimation procedure; and in “A Note on Nonresponse” we discuss the missing value issue.

The cross-sectional weights w_{ij}, in W_i, used in equation (1), 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 values that such observation represents in U_j. However, remember that the sample s_j is composed of different sets of subjects, or different sub-samples (the different cohorts), and the integration of these sub-samples 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 different, non-overlapping, populations. The base weight in that case is easy to compute; and all that remains, after that, 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.

3.2.1 Unbiasedness

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

If one is interested in a pure design-based analysis, with parameters of interest being finite-population quantities, in the present situation one could define the parameter of interest to be $\boldsymbol{\beta}_N$, the solution to the following finite population estimating equation:

$$\Psi_U(\boldsymbol{\beta}_N) = \sum_{i\in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_N} V_i^{-1} \mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)) = \mathbf{0}; \qquad (2)$$

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 other words, the target parameter $\boldsymbol{\beta}_N$ satisfies equation (2). 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, Sec. 9.9). We therefore use the methodology of multiphase sampling for the following derivations. Without loss of generality we assume that $J = 3$, i.e. there are only three waves; although it may seem too restrictive for real applications, the derivations with just three waves show the patterns for general J, in this section and with respect to the variance.

As we mentioned earlier, we assume that w_{ij} is the cross-sectional weight for subject i at wave j, as long as that subject belongs to s_j. 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:

$$E_p\Big[\sum_{i\in s}\underbrace{\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}}_{A_i}W_i\underbrace{(\boldsymbol{y}_i-\boldsymbol{\mu}_i)}_{\boldsymbol{e}_i}\Big] = E_p\Big[\sum_{j=1}^{3}\sum_{i\in s_{j(j)}} A_iW_i\boldsymbol{e}_i\Big]. \tag{3}$$

For example for $\sum_{i\in s_{2(2)}} A_iW_i\boldsymbol{e}_i$ we obtain:

$$\begin{aligned}
E_p\Big[\sum_{i\in s_{2(2)}} A_iW_i\boldsymbol{e}_i\Big] \;&=\; E\Big\{E\Big[\sum_{i\in U_{2(2)}} A_iD_i\boldsymbol{e}_i\Big|s_{2(2)}\Big]\Big\}\\
&= E\left\{\sum_{i\in U_{2(2)}} A_i\begin{bmatrix} 0 & & \mathrm{O}\\ & I_i(U_2)w_{i2}I_i(s_{2(2)}) & \\ \mathrm{O} & & \dfrac{I_i(U_3)\pi_{i3|s_{2(2)}}I_i(s_{2(2)})}{\pi_{i2}\pi_{i3|s_{2(2)}}}\end{bmatrix}\boldsymbol{e}_i\right\}\\
&= \sum_{i\in U_{2(2)}} A_i\begin{bmatrix} 0 & & \mathrm{O}\\ & \dfrac{I_i(U_2)\pi_{i2}}{\pi_{i2}} & \\ \mathrm{O} & & \dfrac{I_i(U_3)\pi_{i2}}{\pi_{i2}}\end{bmatrix}\boldsymbol{e}_i \;\overset{\text{def}}{=}\; \sum_{i\in U_{2(2)}} A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i,
\end{aligned}$$

where $D_i = \text{diag}[0,\ I_i(U_2)w_{i2}I_i(s_{2(2)}),\ I_i(U_3)w_{i3}I_i(s_{3(2)})I_i(s_{2(2)})]$; similarly we can show that $E_p[\sum_{i\in s_{1(1)}} A_iW_i\boldsymbol{e}_i] = \sum_{i\in U_{1(1)}} A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$ and $E_p[\sum_{i\in s_{3(3)}} A_iW_i\boldsymbol{e}_i] = \sum_{i\in U_{3(3)}} A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$. From these expressions and equation (3) 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.

If, on the other hand, the target of inference is the super-population parameter, we need to guarantee that the model for μ_{ij} is such that $E_\xi(Y_{ij}-\mu_{ij}) = 0$ is satisfied. For if this is the case, we have:

$$\begin{aligned}
E_\xi E_p[\Psi_s(\boldsymbol{\beta})] = E_\xi[\Psi_U(\boldsymbol{\beta})] \;&=\; E_\xi\Big[\sum_{i\in U}\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\Big]\\
&= \sum_{i\in U}\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})E_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i) \;=\; \mathbf{0};
\end{aligned}$$

so that the estimating function $\Psi_s(\boldsymbol{\beta})$ is model-design unbiased.

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 others participate at 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 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), 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, in the following way:

$$\Psi_r(\boldsymbol{\beta}) = \sum_{i \in r} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_{ri}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0};$$

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.

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

$$E_R\Big\{\sum_{i \in r_{2(2)}} A_i W_{ri} \boldsymbol{e}_i\Big\} = E_R\left\{\sum_{i \in s_{2(2)}} A_i \begin{bmatrix} 0 & & \text{O} \\ & I_i(U_2)w_{ri2}I_i(r_{2(2)}) & \\ \text{O} & & I_i(U_3)w_{ri3}I_i(r_{3(2)}) \end{bmatrix} \boldsymbol{e}_i\right\}$$

$$= \sum_{i \in s_{2(2)}} A_i \begin{bmatrix} 0 & & \text{O} \\ & \dfrac{I_i(U_2)\pi_{ri2}}{\pi_{i2}\hat{\pi}_{ri2}} & \\ \text{O} & & \dfrac{I_i(U_3)\pi_{ri3}}{\pi_{i2}\pi_{i3|s_{2(2)}}\hat{\pi}_{ri3}} \end{bmatrix} \boldsymbol{e}_i \qquad (4)$$

$$= \sum_{i \in s_{2(2)}} A_i \begin{bmatrix} 0 & & \mathrm{O} \\ & I_i(U_2)w_{i2} & \\ \mathrm{O} & & I_i(U_3)w_{i3} \end{bmatrix} \boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i \in s_{2(2)}} A_i W_i \boldsymbol{e}_i. \qquad (5)$$

Expressions (4) and (5) detail that the nonresponse model used for $\hat{\pi}_{rij}$ has to be such that $E_R[I_i(r_{j(j')})] = \pi_{rij} = \hat{\pi}_{rij}$. 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 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 (5) together with the previous section. Hereafter we therefore ignore the issue of nonresponse for notation 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

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

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 estimator of $\boldsymbol{\beta}$, as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + (\boldsymbol{\beta}_N - \boldsymbol{\beta}) = \text{Sampling Error} + \text{Model Error}.$$

After some straightforward calculations, the total variance, or more precisely the total MSE, can be decomposed as:

$$\begin{aligned} V_{\text{Tot}} &= E_{\xi p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \\ &= E_{\xi p}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)' + 2 \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})'] + o(1/n_m) \qquad (6) \\ &= E_\xi V_p + 2 \otimes E_p C_\xi + o(1/n_m) \ = \ V_{\text{Sam}} + 2 \otimes C_{\text{Sam-Mod}} + o(1/n_m), \end{aligned}$$

where $2 \otimes A = A + A'$, $V_p = E_p(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)'$, $C_\xi = E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})'$, $V_{\text{Sam}} = E_\xi V_p$ is the "sampling variance" component, $C_{\text{Sam-Mod}} = E_p C_\xi$, and $2 \otimes C_{\text{Sam-Mod}}$ is the cross "sampling-model variance" component. 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,

$$H(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i \in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \mathrm{I}_i(\mathrm{U}) \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \quad \text{and} \quad \hat{H}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}.$$

We then get, for V_p and C_ξ in (6),

$$\begin{aligned} V_p &= E_p(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)' \ = \ E_p\{[H(\boldsymbol{\beta}_N)]^{-1}\Psi_s(\boldsymbol{\beta}_N)\Psi_s'(\boldsymbol{\beta}_N)[H(\boldsymbol{\beta}_N)]^{-1}\} + o_p(1/n_m) \\ &= [H(\boldsymbol{\beta}_N)]^{-1} E_p[\Psi_s(\boldsymbol{\beta}_N)\Psi_s'(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} + o_p(1/n_m) \\ &= [H(\boldsymbol{\beta}_N)]^{-1} \mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} + o_p(1/n_m), \qquad (7) \end{aligned}$$

$$\begin{aligned} C_\xi &= E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})' \ = \ E_\xi\{[\hat{H}(\boldsymbol{\beta})]^{-1}\Psi_s(\boldsymbol{\beta})\Psi_U'(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1}\} + o_p(1/n_m) \\ &= [\hat{H}(\boldsymbol{\beta})]^{-1} E_\xi[\Psi_s(\boldsymbol{\beta})\Psi_U'(\boldsymbol{\beta})][H(\boldsymbol{\beta})]^{-1} + o_p(1/n_m) \\ &= \frac{1}{N}[\hat{H}(\boldsymbol{\beta})]^{-1} \hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} + o_p(1/n_m), \qquad (8) \end{aligned}$$

with

$$\hat{H}_{\Sigma V}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i \Sigma_i V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}};$$

the derivation of (8) can be found in Appendix A.

In conclusion, so far we have found that:

$$\boxed{\begin{aligned} V_{\text{Tot}} &= E_\xi V_p + 2 \otimes E_p C_\xi + o(1/n_m) \\ &= E_\xi \left\{ [H(\boldsymbol{\beta}_N)]^{-1} \text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} \right\} + \frac{2 \otimes E_p \left\{ [\hat{H}(\boldsymbol{\beta})]^{-1} \hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} \right\}}{N} \\ &\quad + o(1/n_m). \end{aligned}} \tag{9}$$

In (9) 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 interest lies on the census estimator, $\boldsymbol{\beta}_N$, only the first quantity in expression (9) is necessary; i.e. the expression for V_{Tot} is simply $E_\xi V_p$ (and lower order terms). Also, even if we are interested in the superpopulation quantity $\boldsymbol{\beta}$, but the sampling fraction is small, i.e. $n \ll N$, the first term is a good enough approximation for the total variance. If, on the other hand, the sampling fraction is large, and inference is about the superpopulation parameter, both terms in (9) 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, Swensson, and Wretman (1992). After some derivations (see Appendix A), and defining $A_i = (\partial \boldsymbol{\mu}_i'/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N} V_i^{-1}$, $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$, $\boldsymbol{e}_{i(1\cdots3)} = \boldsymbol{e}_i$, $\boldsymbol{e}_{i(2\cdots3)} = (0, e_{i2}, e_{i3})'$, and $\boldsymbol{e}_{i(3\cdots3)} = (0,0,e_{i3})'$, we obtain:

$$\begin{aligned} \text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] &= D_{(1)} + D_{(2)} + D_{(3)} \\ &= D_{(1)1} + D_{(1)2} + D_{(1)3} + D_{(2)2} + D_{(2)3} + D_{(3)3}, \end{aligned} \tag{10}$$

where $D_{(1)} \overset{\text{def}}{=} N^{-2}\text{Var}_p(\sum\limits_{i \in s_{1(1)}} A_i W_i \boldsymbol{e}_i) = D_{(1)1} + D_{(1)2} + D_{(1)3}$, $D_{(2)} \overset{\text{def}}{=} N^{-2}\text{Var}_p(\sum\limits_{i \in s_{2(2)}} A_i W_i \boldsymbol{e}_i) = D_{(2)2} + D_{(2)3}$, $D_{(3)} \overset{\text{def}}{=} N^{-2}\text{Var}_p(\sum\limits_{i \in s_{3(3)}} A_i W_i \boldsymbol{e}_i) = D_{(3)3}$,

$$N^2 D_{(1)1} \overset{\text{def}}{=} \text{Var}\big[\sum_{i \in s_{1(1)}} w_{i1} A_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(1\cdots3)}\big],$$

$$
\begin{aligned}
N^2 D_{(1)2} &\overset{\text{def}}{=} E\Big\{\text{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(2\cdots3)}\Big|\, s_{1(1)}\Big]\Big\}\\
&= \text{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(2\cdots3)}\Big] - \text{Var}\Big[\sum_{i\in s_{1(1)}} w_{i1}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(2\cdots3)}\Big],\\
N^2 D_{(1)3} &\overset{\text{def}}{=} E\Big\{E\Big[\text{Var}\Big(\sum_{i\in s_{3(1)}} w_{i3}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)} \mid s_{2(1)}, s_{1(1)}\Big)\Big| s_{1(1)}\Big]\Big\}\\
&= \text{Var}\Big[\sum_{i\in s_{3(1)}} w_{i3}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)}\Big] - \text{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)}\Big],\\
N^2 D_{(2)2} &\overset{\text{def}}{=} \text{Var}\Big\{\sum_{i\in s_{2(2)}} w_{i2}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(2\cdots3)}\Big\},\\
N^2 D_{(2)3} &\overset{\text{def}}{=} E\Big\{\text{Var}\Big[\sum_{i\in s_{3(2)}} w_{i3}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)} \mid s_{2(2)}\Big]\Big\}\\
&= \text{Var}\Big[\sum_{i\in s_{3(2)}} w_{i3}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)}\Big] - \text{Var}\Big[\sum_{i\in s_{2(2)}} w_{i2}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)}\Big],\\
N^2 D_{(3)} = N^2 D_{(3)3} &\overset{\text{def}}{=} \text{Var}\Big\{\sum_{i\in s_{3(3)}} w_{i3}A_i\text{I}_i(\text{U})\boldsymbol{e}_{i(3\cdots3)}\Big\}.
\end{aligned}
$$

In general, we have proved the following

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

$$
\boxed{
\begin{aligned}
&Var_p[\Psi_s(\boldsymbol{\beta}_N)]\\
&= \frac{1}{N^2}\sum_{j'=1}^{J}\sum_{j=j'}^{J}\Big\{Var_p\big[\sum_{i\in s_{j(j')}} w_{ij}A_iI_i(U)\mathbf{e}_{i(j\cdots J)}\big] - Var_p\big[\sum_{i\in s_{j-1(j')}} w_{i,j-1}A_iI_i(U)\mathbf{e}_{i(j\cdots J)}\big]\Big\}\\
&= \frac{1}{N^2}\sum_{j=1}^{J}\Big\{Var_p\big[\sum_{i\in s_j} w_{ij}A_iI_i(U)\mathbf{e}_{i(j\cdots J)}\big] - Var_p\big[\sum_{i\in s_{j-1}} w_{i,j-1}A_iI_i(U)\mathbf{e}_{i(j\cdots J)}\big]\Big\},
\end{aligned}
}
\qquad (11)
$$

where, we let $w_{i,j-1} = 0$ whenever $j = j'$, $w_{i0} = 0$, and to get the last line in (11) we have changed variables and used the independence among cohorts.

In (10) and (11) 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 6. Nonetheless, some calculations not assuming independence (not shown), for the case of few cohorts, show that, even in that case, the approximation given by the last line of expression (11) is a very good one for the variance terms (i.e. the terms in the diagonal), and still good for the covariance terms (albeit not as good as for the variance terms).

3.3.2 Estimation

The estimation of V_{Tot} in (9) can be done 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 $\Sigma_i = \text{Cov}[Y_i|X_i]$ can be estimated by $\hat{\boldsymbol{e}}_i\hat{\boldsymbol{e}}_i'$.

We use (11) in Property 3.1 to estimate $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$. As long as there is a method to estimate the variance of (cross-sectional) Horvitz-Thompson (H-T) estimators, expression (11) can be used. If we define $Z_{ij} = A_i \text{I}_i(\text{U})\boldsymbol{e}_{i(j\cdots J)}$, we notice that each of the terms involved in the computation of (11), terms like $\text{Var}_p\left[\sum_{i\in s_j} w_{ij}Z_{ij}\right]$, 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 non-response and calibration adjustments performed; but this does not present any additional complications beyond what is found in any cross-sectional problem as everything is done cross-sectionally. The SDR uses replication to estimate variances of cross-sectional estimators, but any of the available methods of design variance estimation can be used (see for example Wolter, 2007).

We use the cross-sectional replication 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" ($\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$) of the design variance ($E_\xi V_p$). And 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 (11). The r-th replicate of the first term is $\sum_{i\in s_j} w_{ij}^{(r)} A_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, 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 $\sum_{i\in s_{j-1}} w_{i,j-1}^{(r)} A_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{i,j-1}^{(r)}$ is the r-th replicate weight for subject i at wave $j-1$.

4 A Simple Simulation Example

To demonstrate the performance of the proposed estimator, compared to a "usual" estimating procedure, we carried out the following simulation. Assume that there is a finite population at time $j=0$ (the first wave) of $N_0 = N_{0(0)} = 20,000$ subjects. At time $j=1$ (the second wave) none of the original 20,000 subjects leave the population, but there are $N_{1(1)} = 6,000$ new individuals entering the finite population; for a total of $N_1 = 26,000$ subjects at the second wave. We generate the finite population form the model:

$$\xi : \begin{cases} Y_{ij} = \underbrace{3}_{\beta_0} + \underbrace{0.5}_{\beta_1} \times j + \varepsilon_{ij}, \quad j = 0,1, \quad i = 1,2,\ldots \\ \boldsymbol{\varepsilon}_i \sim \mathrm{MVN}\left(\binom{0}{0},\ \sigma^2\begin{pmatrix}1 & \alpha \\ \alpha & 1\end{pmatrix}\right), \quad i = 1,2,\ldots, \quad \sigma^2 = 25, \quad \alpha = 0.4. \end{cases}$$

At wave $j=0$ we select a simple random sample of size $n_0 = n_{0(0)} = 40$ subjects without replacement (SRS) from among the $N_0 = N_{0(0)} = 20,000$ original subjects, and interview them (at $j=0$). Of those, we keep (by SRS) $n_{1(0)} = 36$ for interview at the second wave ($j=1$), and drop the other four elements. Among the $N_{1(1)} = 6,000$ new subjects in the population, we select (by SRS) a sample of size $n_{1(1)} = 4$ (the second cohort) to be interviewed at $j=1$, and who replace the four we dropped from the original cohort.

The target parameter is the net change between the two times, $D_N = \bar{Y}_1 - \bar{Y}_0$. A real-world example of such a parameter is the difference in consumer price index (CPI) between two months one year apart. For example one may be interested in the change in CPI in the New York City labor market between August 2010 and August 2011. Clearly the population of interest (NYC residents) has changed between the two time points, and yet the difference in CPIs is a quantity of high interest to analysts.

We can estimate D_N by the "usual" estimator, which is the difference of the H-T estimators; i.e. $\hat{D}_N = \hat{\bar{Y}}_{1HT} - \hat{\bar{Y}}_{0HT}$; we refer to this estimator as the "independent" estimator, or "Ind", as this estimator does not take into account the auto-correlation.

Our proposal to estimate D_N, on the other hand, is to use the $\hat{\beta}_1$ from the $\hat{\boldsymbol{\beta}}$ that solves the estimating equations $\sum_{i\in s} X_i V_i^{-1} W_i(\boldsymbol{y}_i - X_i'\boldsymbol{\beta}) = \mathbf{0}$, with $W_i = \text{diag}[w_{i1}, w_{i2}]$, the diagonal matrix of cross-sectional survey weights for subject i, which come from the SRS characteristics. We refer to this estimator as "R", to signify that this estimator does take into account the auto-correlation by means of the working correlation matrix $\mathbf{R}(\alpha)$. It is important to point out that for the estimation procedure we do not use the true correlation matrix $\left(\begin{smallmatrix} 1 & 0.4 \\ 0.4 & 1 \end{smallmatrix}\right)$, as that would not be the case in real world applications; instead, we estimate it from the residuals (see for example Carrillo, Chen, and Wu, 2010 and Carrillo-García, 2008 for details).

A clear way to check that the "Ind" estimator $\hat{D}_N$ ignores the auto-correlation is that if we solve $\sum_{i\in s} X_i V_i^{-1} W_i(\boldsymbol{y}_i - X_i'\boldsymbol{\beta}) = \mathbf{0}$ with $\mathbf{R}(\alpha) = I_2$ (the size-2 identity), we recover $\hat{D}_N$ as second element of $\hat{\boldsymbol{\beta}}$. Additionally, note that the target parameter D_N *also* ignores the auto-correlation, as it is the second element of the solution to $\sum_{i\in U} X_i V_i^{-1} \mathrm{I}_i(U)(\boldsymbol{y}_i - X_i'\boldsymbol{\beta}) = \mathbf{0}$ when we use $\mathbf{R}(\alpha) = I_2$.

We carry out simulations for samples of (cross-sectional) sizes 40, 80, 120, 160, 200, 240, 280, 320, 360, 400, 440, 480, 520, and 560; keeping the same proportions as for the samples of size 40 explained before. For each sample size we select 1,000 samples and in each sample we calculate the two estimators.

We evaluate the two alternatives as estimators of the either the superpopulation parameter β_1 or the finite population quantity D_N along two traits, the (Monte Carlo) relative bias and the (Monte Carlo) mean square error (MSE). The relative bias of $\hat{\beta}_1$, as estimator of D_N, is defined as $1000^{-1}\sum_{l=1}^{1000}(\hat{\beta}_1^{(l)} - D_N)/D_N$, where $\hat{\beta}_1^{(l)}$ is the estimate of β_1 from the l-th simulation sample. The MSE of $\hat{\beta}_1$, as estimator of D_N, is defined as $1000^{-1}\sum_{l=1}^{1000}(\hat{\beta}_1^{(l)} - D_N)^2$. The definitions for estimator $\hat{D}_N$ and for target parameter β_1 are similar. Figures 3 and 4 show the simulation results.

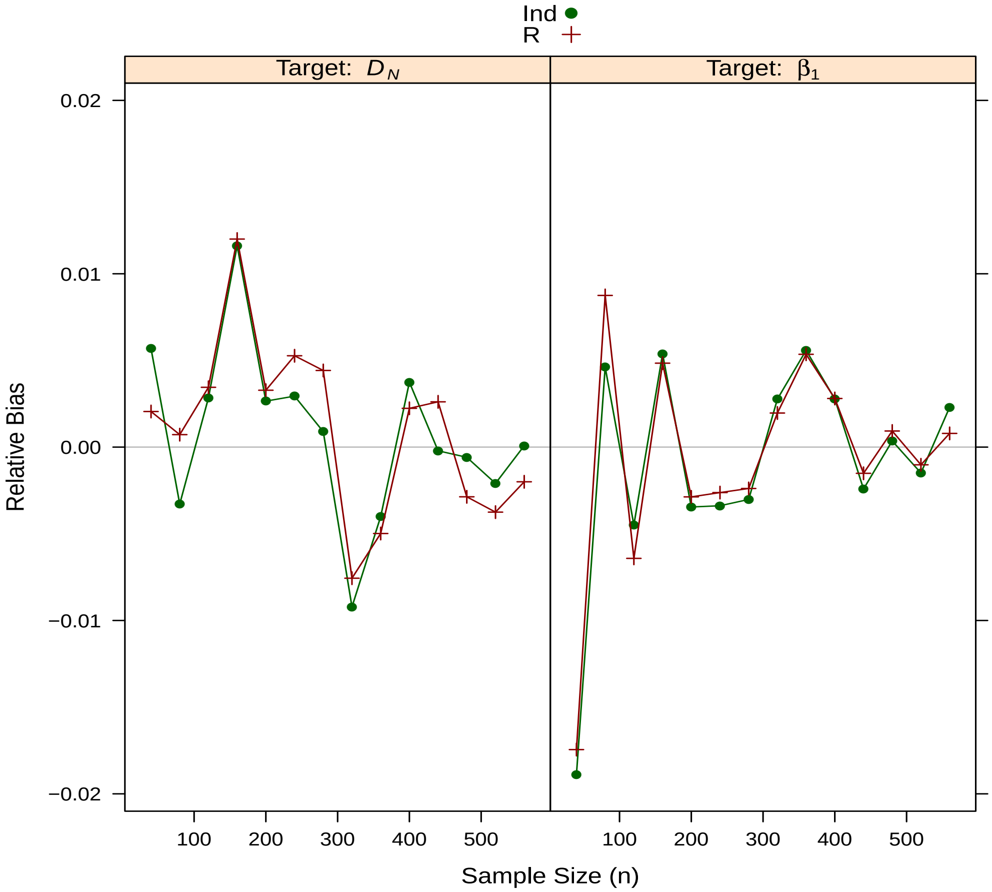

Figure 3: Relative Biases

We see that, with respect to the relative bias (Figure 3), neither estimator is consistently superior than the other. The relative bias of the estimator $\hat{\beta}_1$ is sometimes lower and some times higher than that of the estimator $\hat{D}_N$. The same conclusion holds true no matter what the target is, either β_1 or D_N.

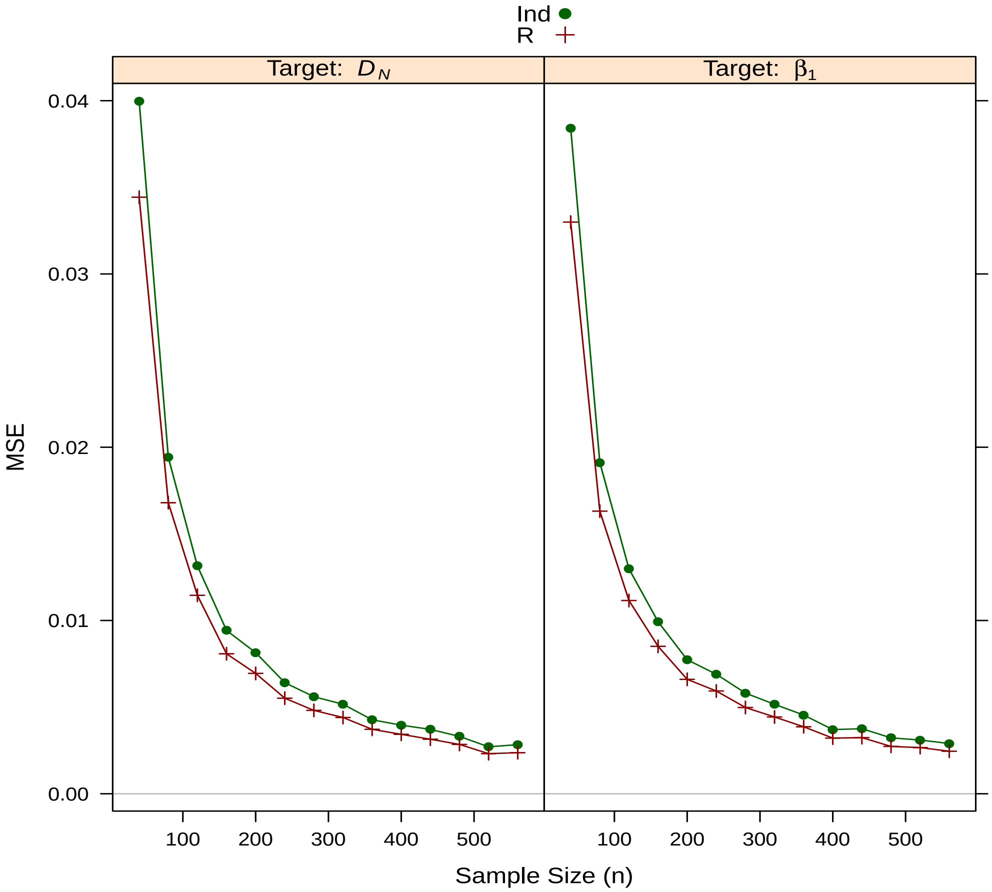

Figure 4: MSEs

Note that $\hat{D}_N$ is design-unbiased for D_N and model-unbiased for β_1; i.e., in theory, the biases of the "Ind" estimator are zero. The fact that this is not reflected in Figure 3 is purely due to Monte Carlo error. It is reassuring that the picture shows that the difference in (estimated) biases between the two methods is, in general, lower than the distance from the estimated bias of "Ind" from zero. We can then conclude that any bias of $\hat{\beta}_1$ can be safely ignored (for either parameter).

Furthermore, we observe that, in general, there does seem to be a decrease of bias (of either estimator) as the sample size increases. The fact that it is not a monotone decrease is also because of the Monte Carlo error.

As for the MSE (Figure 4), there is a different story. First of all, we do see a monotone decrease of MSE as the sample size increases. But more importantly, our proposed estimator performs *consistently better* than the "usual" estimator for all sample sizes. For *any* given sample size, the MSE of $\hat{\beta}_1$ is lower than that of $\hat{D}_N$. And this is true for both, the superpopulation parameter β_1 and the finite population parameter D_N.

It is not surprising that the estimator that takes into account the auto-correlation ($\hat{\beta}_1$) performs better than the one that does not ($\hat{D}_N$) when estimating the superpopulation parameter β_1. After all, the model (ξ) contains some auto-correlation between the observations from the same subject, and $\hat{D}_N$ completely ignores it, whereas $\hat{\beta}_1$ incorporates it.

On the other hand, somewhat surprisingly, our estimator $\hat{\beta}_1$, which *takes into account the auto-correlation*, is *also* the superior alternative for estimating the finite population quantity D_N, which *ignores the auto-correlation*. This result seems counter-intuitive; $\hat{D}_N$ even has *the same functional form* as D_N, whereas $\hat{\beta}_1$ does not!

We also did simulations with different values of the auto-correlation parameter α and different overlapping percentages between the two waves. We found the same kind of results in most cases. The only circumstance where $\hat{D}_N$ has (slightly) lower MSE than $\hat{\beta}_1$ is when α is zero or close to zero and the overlapping fraction in the sample is very low. In that case there may not be enough information to do a good job in estimating α and it would be better not to lose degrees of freedom and assume there is no auto-correlation. However, with the sort of auto-correlations frequently encountered in practice, this seems hardly the case. Auto-correlations of the magnitude of 0.4 are not uncommon in practice.

5 Application to the SDR

The dataset we use is the restricted SDR data, under a license agreement from NSF. The SDR collects several kinds of information from the selected doctoral recipients; the different cohorts are selected using as frame the SED. The information it collects every wave is about employment situation, the principal employer, the principal job, past employment, recent education, demographics, and disability. In addition to that, there is a part of the questionnaire that changes from wave to wave, concentrating on different topics; for example in 2001 questions about professional associations were included, in 2006 detailed information about postdoctoral appointments held was asked, and in 2008 the focus was on papers, books, inventions, and patents. We use only information collected in all the waves, 1995, 1997, 1999, 2001, 2003, 2006, and 2008.

The SDR sample sizes at each of those waves are the following: $n_{95} = 35,370$, $n_{97} = 35,667$, $n_{99} = 31,318$, $n_{01} = 31,366$, $n_{03} = 29,915$, $n_{06} = 30,817$, and $n_{08} = 29,974$. Each of those samples is composed of some people belonging to previous cohorts, who are being re-interviewed, and some new selected individuals. At each wave some subjects are dropped from the study because they have gone out of scope, some subjects are lost because of attrition, and some people are removed to make up space for the new cohort.

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 some time-independent covariates (like gender) and some time-dependent ones (like sector). We have four big classes of covariates. The **Degree variables**: degree field, years since degree, and age at graduation. The **Job variables**: 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**: gender, citizenship status, race/ethnicity, presence of children in family, marital status, and spouse's working status. And the **"Environment" variables**: years since 1995, state (of employment), and the consumer price index (of the region of employment). The full list of variables and categories can be found in Table 1 in the appendix.

Of the original 224,427 total observations and 64,975 subjects, we dropped several of them for various reasons. First of all, we kept only the 198,454 observations with non-missing salaries, corresponding to 60,637 individuals. Among these, there were 269 people with inconsistent ages across waves; we removed those people and were left with 197,418 observations. We also removed two people whom we considered to have "non-sensible" ages at doctorate graduation; this further got rid of three observations. After that, we removed the observations with a missing value for the variable indicating whether the (post-secondary education institution) employer was public or private; this leaves us with 59,855 individuals and 193,667 observations. Finally, we removed some "outlying" salaries; we removed any salary below $5,000 in order to make the histograms of log-salary as symmetric as possible, leaving 59,479 subjects and 191,195 observations; the very last step was to remove the few observations of salary at $999,996, which, from the exploratory graphs, seem to clearly be separated from the rest of the observations. We are then left with **59,474 subjects** and **191,079 observations**, distributed as: $\mathbf{n_{95} = 30,332}$, $\mathbf{n_{97} = 30,734}$, $\mathbf{n_{99} = 26,792}$, $\mathbf{n_{01} = 26,816}$, $\mathbf{n_{03} = 24,997}$, $\mathbf{n_{06} = 25,943}$, and $\mathbf{n_{08} = 25,465}$. The average (cross-sectional) survey weight for each of those waves are: $\mathbf{\bar{w}_{95} = 15.37}$, $\mathbf{\bar{w}_{97} = 16.29}$, $\mathbf{\bar{w}_{99} = 19.96}$, $\mathbf{\bar{w}_{01} = 20.74}$, $\mathbf{\bar{w}_{03} = 22.71}$, $\mathbf{\bar{w}_{06} = 22.94}$, and $\mathbf{\bar{w}_{08} = 24.88}$.

The covariates, and interactions, we considered were selected because they were suggested by either exploratory graphs, or exploratory classification trees, or by the subject matter experts at the NSF. We included, in previous versions of the model, more interactions than those found in the final model, in Table 2 (in Appendix B), but some interactions were dropped (sequentially) because they turned out to be insignificant. Nonetheless, we left in the model some variables or interactions that are not significant if we considered that the fact that they are not significant is important from the subject matter point of view.

Table 2, in the appendix, presents the estimated β coefficients in the following model:

$$y_{ij} = \log(\text{SALARY}_{ij}) = X_{ij}'\boldsymbol{\beta} + \varepsilon_{ij},$$

where X_{ij} includes an intercept along with the other variables and interactions in Table 2; and the working covariance matrix is estimated to be $\hat{V}_i = \hat{\phi}\mathbf{R}(\hat{\alpha})$, with

$$\hat{\phi} = \widehat{\sigma^2} = \frac{\sum_{i\in s}\sum_{j=95}^{08} w_{ij}\hat{e}_{ij}^2}{(\sum_{i\in s}\sum_{j=95}^{08} w_{ij}) - p} = 0.196,$$

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. And $\hat{\alpha}$ contains the $21 = (7\times 6)/2$ estimated auto-correlations $\hat{\alpha}_{jj'} = \hat{\alpha}_{j'j}$, with

$$\hat{\alpha}_{jj'} = \hat{\alpha}_{j'j} = \frac{\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}}\hat{e}_{ij}\hat{e}_{ij'}}{\hat{\phi}(\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}} - p)},$$

for $j \neq j' = 95, 97, 99, 01, 03, 06, 08$; 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 from the estimated coefficients in Table 2. First of all, we notice that the exponential of the intercept ($\exp(9.4) = \$12,144$) is not in the sensible mean of the observed salaries; for that we need to consider the hours worked per week (whose average is 47) and years since degree (average of 15); we then have that a more sensible estimate of the overall average is $\exp(9.4 + 47\times 0.04 - 47^2\times 0.0003 + 15\times 1.03 - 15^2\times 0.999) = \$51,952$. Obviously, the other continuous covariates (years since 1995, age at graduation, the region's CPI, and the number of months since started principal job) also intervene.

We can conclude that, 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 doctorate holders' salaries.

The fact that the gender×years since 1995 interaction is not significant implies that the penalty for being a woman is not changing over time. We see that 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.

With respect to the sector, the ones with highest salaries are the doctorates in for-profit industry (around 20% higher than that for a tenured person in public 4-year college), followed by the federal government. All the industry has higher salaries than colleges and universities; the lowest salaries are found in two-year colleges and people with non-applicable tenure in four-year colleges. The fact that tenured doctorates in private 4-year colleges have significantly lower salaries than the corresponding in public ones may be due to the fact that there is a big variety of such private institutions, whereas the public ones tend to be large.

People with doctorate degrees in computing and information sciences have the highest salaries (around 20% higher than in the biological sciences), followed by degree holders 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.

Married people have the highest salaries, followed by married-like, widowed, separated, divorced, and the never married. The last ones have salaries only around 89% as high as the married ones; one could argue that there probably is some association between never married and age. There does not seem to be a difference between people in families with no children and families with children <2. But the presence of children 2-5, 6-11, and 12+ *is* associated with higher salaries.

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

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. There is a small penalty for receiving the doctorate later in life; for every additional year of age at graduation, the current salary reduces to 99%.

The highest negative effect on salaries is having a position as adjunct faculty; they have salaries that are around 59% the salaries of other doctorate holders. Postdoctoral salaries are only about 74% of the average salary of comparable people in other types of positions.

We have 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 for the labor market of employment as we do not yet have a way to identify geography beyond the state. We decided to also included the state in the model (although we know that it cannot be a causal factor for salary) as a proxy for cost of living; we could use the CPI for that if we had the zip code of employment. Even so, the state effect is highly significant and some state coefficients are among the highest overall. The highest salaries tend to be in California, Washington D.C. and its suburbs, and New York City and its suburbs; we conclude this as the highest coefficients are for Washington D.C., California, New Jersey, New York, Delaware, Connecticut, Maryland, and Virginia. On the other hand, the lowest salaries seem to be in Puerto Rico, Vermont, Montana, Maine, Idaho, South Dakota, North Dakota, and the “others” (territories and abroad).

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

Finally we present some residual analysis. Figures 5, 6, and 7 show a Box and Whisker plot of standardized residuals by year, a spaghetti plot of standardized residuals, and a fitted vs. observed value plot, respectively.

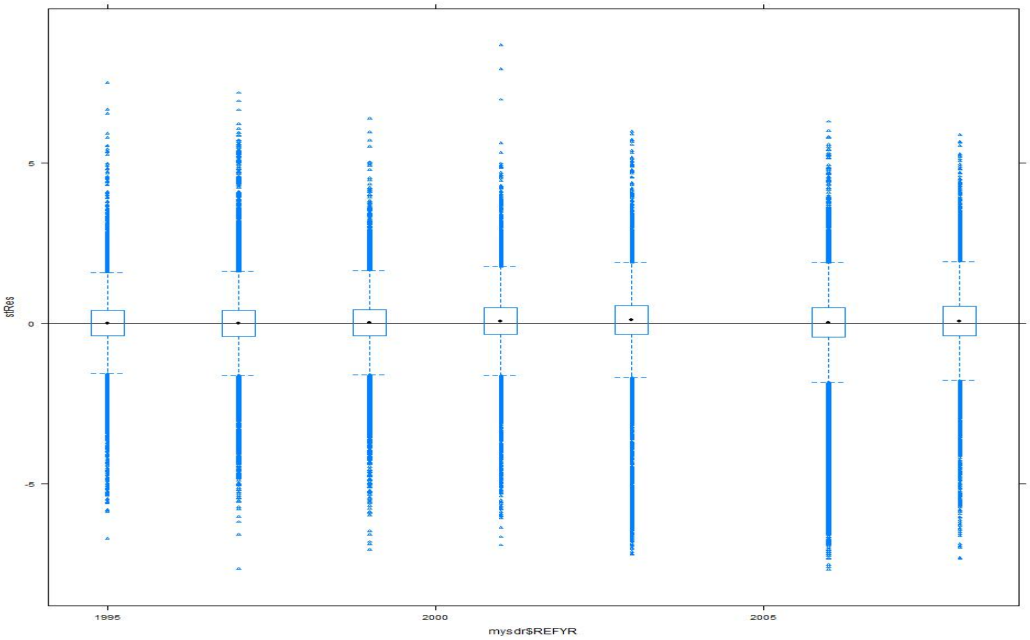

Figure 5: Box and Whisker Plot of Standardized Residuals by Year

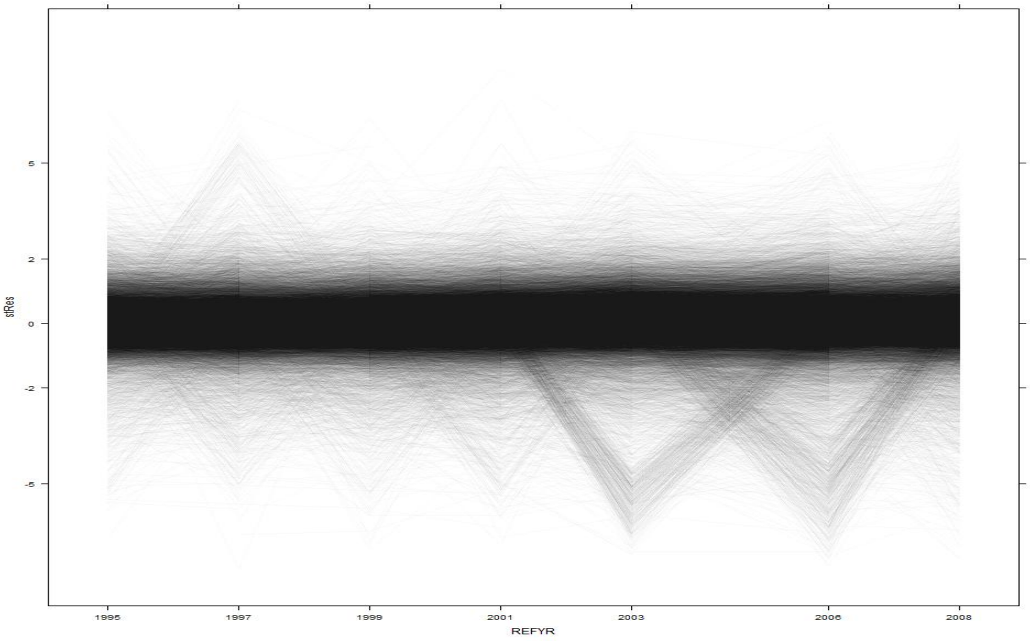

Figure 6: Spaghetti Plot of Standardized Residuals

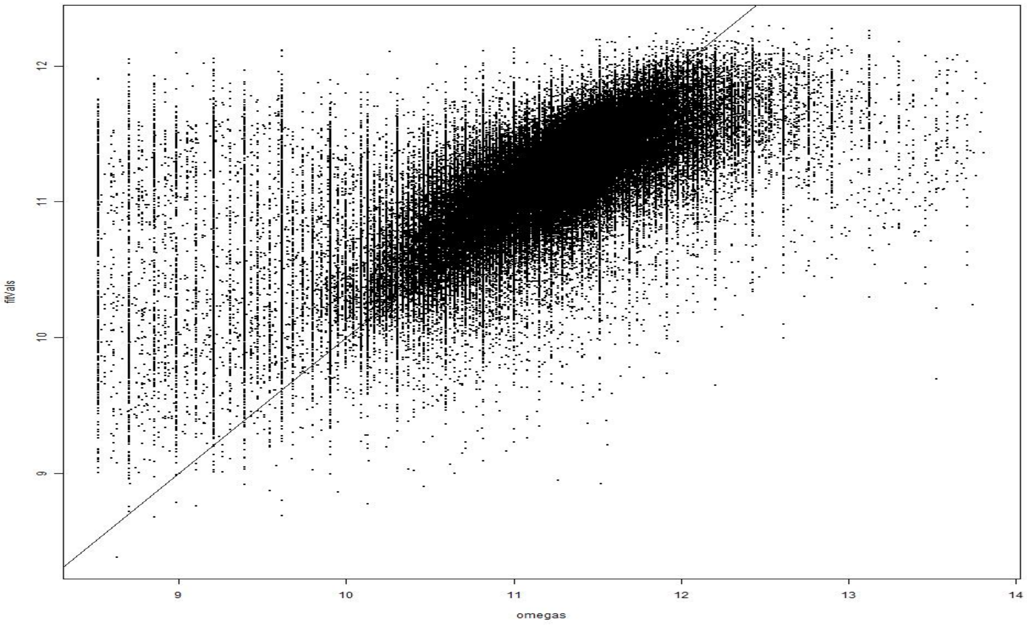

Figure 7: Fitted Values by Observed Values (log-salary)

Figure 5 shows that the model fits reasonably well for all the reference years as most of the standardized residuals are between the -2, 2 limits. Also, the distributions of residuals do not seem to greatly differ from year to year.

From Figure 6 we conclude that the model also fits reasonably well for most people, as most of the lines seem to fluctuate between the -2, 2 limits. 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; although some previously existing "blips" like these did go away.

The last thing we tried was to produce exploratory classification trees for these residual blips, and 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 respondents (in 2003); and the blips in 2006 are disproportionately high for CATI respondents (in 2006). 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. The last thing to notice is that there are not many cases in which the blip remains for more than one wave; for most of them, it goes down and then comes back to "normal" the next wave.

Finally, the plot of fitted values versus observed (Figure 7) also show 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, which are in the top left corner of Figure 7.

6 Discussion and Conclusions

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 represent 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.

Although the design of the SDR is strictly speaking neither a rotating-panel nor a repeated-panel design, 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 done 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 have shown that the independence among cohort is not such a crucial requirement for our variance estimation method to produce good approximations. In future research we plan to evaluate the impact of this issue.

Acknowledgments

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, all for numerous insightful discussions during the research.

Appendix A - Proofs

To develop an expression for C_ξ, we first simplify $\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta})$:

$$\begin{aligned}
N^2\Psi_s(\boldsymbol{\beta})\Psi'_U(\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)\sum_{i\in U}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'\mathrm{I}_i(\mathrm{U})V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\\
&=\Big[\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\Big]\Big[\sum_{i\in s}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'\mathrm{I}_i(\mathrm{U})V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}+\sum_{i\notin s}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'\mathrm{I}_i(\mathrm{U})V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\Big]\\
&=\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\sum_{i\in s}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'\mathrm{I}_i(\mathrm{U})V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\\
&\qquad+\underbrace{\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\sum_{i\notin s}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'\mathrm{I}_i(\mathrm{U})V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}}_{\text{A = Two model-independent summations}}\\
&=\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\\
&\qquad+\sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i\;\underbrace{(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_k-\boldsymbol{\mu}_k)'}_{\text{B = Model-independent terms}}\;\mathrm{I}_k(\mathrm{U})V_k^{-1}\frac{\partial\boldsymbol{\mu}_k}{\partial\boldsymbol{\beta}}\;+\;\mathrm{A}\\
&=\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\;+\;\mathrm{B}^*\;+\;\mathrm{A},
\end{aligned}$$

where A and B* both have model-expectation zero; therefore,

$$\begin{aligned}
E_\xi[\Psi_s(\boldsymbol{\beta})\Psi'_U(\boldsymbol{\beta})] &= E_\xi\Big[\frac{1}{N^2}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\Big]\\
&=\frac{1}{N^2}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_iE_\xi[(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)']V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\;=\;\frac{1}{N^2}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}V_i^{-1}W_i\Sigma_iV_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\\
&=\frac{1}{N}\hat{H}_{\Sigma V}(\boldsymbol{\beta}),
\end{aligned}$$

equation (8) follows.

We now develop the expression for $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, the design variance of the estimating function:

$$\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = \mathrm{Var}_p\Big\{\frac{1}{N}\sum_{i\in s}\underbrace{\frac{\partial\boldsymbol{\mu}'_i}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}V_i^{-1}}_{A_i}W_i[\underbrace{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_N)}_{\boldsymbol{e}_i}]\Big\}$$

$$
\begin{aligned}
&= \frac{1}{N^2}\mathrm{Var}_p\big(\sum_{i\in s_{1(1)}} A_i W_i \boldsymbol{e}_i + \sum_{i\in s_{2(2)}} A_i W_i \boldsymbol{e}_i + \sum_{i\in s_{3(3)}} A_i W_i \boldsymbol{e}_i\big) \\
&= \frac{1}{N^2}\mathrm{Var}_p\big(\sum_{i\in s_{1(1)}} A_i W_i \boldsymbol{e}_i\big) + \frac{1}{N^2}\mathrm{Var}_p\big(\sum_{i\in s_{2(2)}} A_i W_i \boldsymbol{e}_i\big) + \frac{1}{N^2}\mathrm{Var}_p\big(\sum_{i\in s_{3(3)}} A_i W_i \boldsymbol{e}_i\big) \qquad (12) \\
&= D_{(1)} + D_{(2)} + D_{(3)},
\end{aligned}
$$

where, for line (12), we assume that the (three) cohorts are design-independent. Now,

$$
\begin{aligned}
N^2 D_{(1)} &= \mathrm{Var}_p\big(\sum_{i\in s_{1(1)}} A_i W_i \boldsymbol{e}_i\big) \;=\; \mathrm{Var}_p\Big\{\sum_{i\in s_{1(1)}} A_i \begin{bmatrix} I_i(U_1)w_{i1} & & \mathrm{O} \\ & I_i(U_2)w_{i2} & \\ \mathrm{O} & & I_i(U_3)w_{i3} \end{bmatrix} \boldsymbol{e}_i\Big\} \\
&= \mathrm{Var}_p\Big\{\sum_{i\in U_{1(1)}} A_i \begin{bmatrix} I_i(U_1)w_{i1}I_i(s_{1(1)}) & & \mathrm{O} \\ & I_i(U_2)w_{i2}I_i(s_{2(1)})I_i(s_{1(1)}) & \\ \mathrm{O} & & I_i(U_3)w_{i3}I_i(s_{3(1)})I_i(s_{2(1)})I_i(s_{1(1)}) \end{bmatrix} \boldsymbol{e}_i\Big\} \\
&= \mathrm{Var}_p\Big\{\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \begin{bmatrix} I_i(s_{1(1)}) \\ I_i(s_{2(1)})I_i(s_{1(1)}) \\ I_i(s_{3(1)})I_i(s_{2(1)})I_i(s_{1(1)}) \end{bmatrix}\Big\} \;=\; \mathrm{Var}_p\Big[\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}\Big],
\end{aligned}
$$

where $\mathrm{Diag}\{\boldsymbol{e}_i\}$ is, for a column vector $\boldsymbol{e}_i$, a diagonal matrix with diagonal entries being the elements of $\boldsymbol{e}_i$, and $\boldsymbol{I}_{i(1)} = \big(I_i(s_{1(1)}),\ I_i(s_{2(1)})I_i(s_{1(1)}),\ I_i(s_{3(1)})I_i(s_{2(1)})I_i(s_{1(1)})\big)'$. Similarly we can get:

$$
N^2 D_{(2)} = \mathrm{Var}_p\Big[\sum_{i\in U_{2(2)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(2)}\Big] \quad \text{and} \quad N^2 D_{(3)} = \mathrm{Var}_p\Big[\sum_{i\in U_{3(3)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(3)}\Big],
$$

where $\boldsymbol{I}_{i(2)} = \big(0,\ I_i(s_{2(2)}),\ I_i(s_{3(2)})I_i(s_{2(2)})\big)'$, and $\boldsymbol{I}_{i(3)} = \big(0,\ 0,\ I_i(s_{3(3)})\big)'$. Now, let us concentrate on $D_{(1)}$:

$$
\begin{aligned}
N^2 D_{(1)} =& \mathrm{Var}_p\big[\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}\big] \\
=& \mathrm{Var}\Big\{E\big[\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{1(1)}\big]\Big\} \\
&+ E\Big\{\mathrm{Var}\big[\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{1(1)}\big]\Big\} \\
=& \mathrm{Var}\Big\{E\Big[E\big(\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]\Big\} \\
&+ E\Big\{\mathrm{Var}\Big[E\big(\sum_{i\in U_{1(1)}} A_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]
\end{aligned}
$$

$$+E\Big[\mathrm{Var}\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]\Big\}$$

$$= N^2D_{(1)1} + N^2D_{(1)2} + N^2D_{(1)3}\ . \tag{13}$$

Let us do each of the terms in (13) in turns, beginning with $N^2D_{1(1)}$, we have:

$$E\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) = \sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\begin{bmatrix} I_i(s_{1(1)}) \\ I_i(s_{2(1)})I_i(s_{1(1)}) \\ \pi_{i3|s_{2(1)}}I_i(s_{2(1)})I_i(s_{1(1)}) \end{bmatrix},$$

then,

$$\begin{aligned}
&E\Big[E\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big] \\
&= \sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\begin{bmatrix} I_i(s_{1(1)}) \\ \pi_{i2|s_{1(1)}}I_i(s_{1(1)}) \\ \pi_{i3|s_{2(1)}}\pi_{i2|s_{1(1)}}I_i(s_{1(1)}) \end{bmatrix} \\
&= \sum_{i\in U_{1(1)}} A_i\begin{bmatrix} \frac{I_i(U_1)}{\pi_{i1}} & & \mathrm{O} \\ & \frac{I_i(U_2)}{\pi_{i1}\pi_{i2|s_{1(1)}}} & \\ \mathrm{O} & & \frac{I_i(U_3)}{\pi_{i1}\pi_{i2|s_{1(1)}}\pi_{i3|s_{2(1)}}} \end{bmatrix}\begin{bmatrix} I_i(s_{1(1)}) & & \mathrm{O} \\ & \pi_{i2|s_{1(1)}}I_i(s_{1(1)}) & \\ \mathrm{O} & & \pi_{i3|s_{2(1)}}\pi_{i2|s_{1(1)}}I_i(s_{1(1)}) \end{bmatrix}\boldsymbol{e}_i \\
&= \sum_{i\in U_{1(1)}} A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\begin{bmatrix} I_i(s_{1(1)})/\pi_{i1} \\ I_i(s_{1(1)})/\pi_{i1} \\ I_i(s_{1(1)})/\pi_{i1} \end{bmatrix} = \sum_{i\in U_{1(1)}} A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\frac{I_i(s_{1(1)})}{\pi_{i1}} \\
&= \sum_{i\in U_{1(1)}} w_{i1(1)}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_iI_i(s_{1(1)}),
\end{aligned}$$

which implies that:

$$\begin{aligned}
N^2D_{(1)1} &= \mathrm{Var}\Big\{E\Big[E\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]\Big\} \\
&= \mathrm{Var}\big[\sum_{i\in U_{1(1)}} w_{i1}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_iI_i(s_{1(1)})\big] \ = \ \mathrm{Var}\big[\sum_{i\in s_{1(1)}} w_{i1}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(1\cdots 3)}\big].
\end{aligned}$$

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

$$\begin{aligned}
E\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \ &= \ \sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\begin{bmatrix} I_i(s_{1(1)}) \\ I_i(s_{2(1)})I_i(s_{1(1)}) \\ \pi_{i3|s_{2(1)}}I_i(s_{2(1)})I_i(s_{1(1)}) \end{bmatrix} \\
= \sum_{i\in U_{1(1)}} A_i\begin{bmatrix} \frac{I_i(U_1)}{\pi_{i1}} & & \mathrm{O} \\ & \frac{I_i(U_2)}{\pi_{i1}\pi_{i2|s_{1(1)}}} & \\ \mathrm{O} & & \frac{I_i(U_3)}{\pi_{i1}\pi_{i2|s_{1(1)}}\pi_{i3|s_{2(1)}}} \end{bmatrix}&\begin{bmatrix} I_i(s_{1(1)}) & & \mathrm{O} \\ & I_i(s_{2(1)})I_i(s_{1(1)}) & \\ \mathrm{O} & & \pi_{i3|s_{2(1)}}I_i(s_{2(1)})I_i(s_{1(1)}) \end{bmatrix}\boldsymbol{e}_i
\end{aligned}$$

$$= \sum_{i\in U_{1(1)}} A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} \begin{bmatrix} I_i(s_{1(1)})/\pi_{i1} \\ I_i(s_{2(1)})I_i(s_{1(1)})/\pi_{i1}\pi_{i2|s_{1(1)}} \\ I_i(s_{2(1)})I_i(s_{1(1)})/\pi_{i1}\pi_{i2|s_{1(1)}} \end{bmatrix} = \sum_{i\in s_{1(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} \begin{bmatrix} \pi_{i2|s_{1(1)}} \\ I_i(s_{2(1)}) \\ I_i(s_{2(1)}) \end{bmatrix},$$

then,

$$\begin{aligned}
&\mathrm{Var}\Big[E\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big] \\
&= \mathrm{Var}\Big[\sum_{i\in s_{1(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} \begin{bmatrix} \pi_{i2|s_{1(1)}} \\ I_i(s_{2(1)}) \\ I_i(s_{2(1)}) \end{bmatrix} \Bigm| s_{1(1)}\Big] \\
&= \mathrm{Var}\Big[\sum_{i\in s_{1(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} \begin{bmatrix} 0 \\ I_i(s_{2(1)}) \\ I_i(s_{2(1)}) \end{bmatrix} \Bigm| s_{1(1)}\Big] \qquad (14) \\
&= \mathrm{Var}\Big[\sum_{i\in s_{1(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}I_i(s_{2(1)}) \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \Bigm| s_{1(1)}\Big] \\
&= \mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)} \Bigm| s_{1(1)}\Big],
\end{aligned}$$

where line (14) 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:

$$\begin{aligned}
N^2D_{(1)2} &= E\Big\{\mathrm{Var}\Big[E\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]\Big\} \\
&= E\Big\{\mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)} \Bigm| s_{1(1)}\Big]\Big\} \\
&= \mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)}\Big] - \mathrm{Var}\Big\{E\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)} \Bigm| s_{1(1)}\Big]\Big\} \\
&= \mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)}\Big] - \mathrm{Var}\Big\{E\Big[\sum_{i\in s_{2(1)}} w_{i2|s_{1(1)}}w_{i1}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)} \Bigm| s_{1(1)}\Big]\Big\} \\
&= \mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)}\Big] - \mathrm{Var}\Big\{\sum_{i\in s_{1(1)}} w_{i1}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots3)}\Big\}.
\end{aligned}$$

We can, similarly, show that:

$$\begin{aligned}
N^2D_{(1)3} =& E\Big\{E\Big[\mathrm{Var}\big(\sum_{i\in U_{1(1)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]\Big\} \\
=& E\Big\{E\Big[\mathrm{Var}\big(\sum_{i\in s_{3(1)}} w_{i3}I_i(s_{2(1)})I_i(s_{1(1)})A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots3)} \mid s_{2(1)}, s_{1(1)}\big) \Bigm| s_{1(1)}\Big]\Big\}
\end{aligned}$$

$$
\begin{aligned}
=&E\Big\{\mathrm{Var}\Big[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{I}_i(s_{2(1)})A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ \Big|\ s_{1(1)}\Big]\\
&\quad-\mathrm{Var}\Big[E\big(\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{I}_i(s_{2(1)})A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ \big|\ s_{2(1)},s_{1(1)}\big)\ \Big|\ s_{1(1)}\Big]\Big\}\\
=&E\Big\{\mathrm{Var}\Big[\sum_{i\in s_{3(1)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ \Big|\ s_{1(1)}\Big]-\mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ \Big|\ s_{1(1)}\Big]\Big\}\\
=&\mathrm{Var}\Big[\sum_{i\in s_{3(1)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big]-\mathrm{Var}\Big[E\big(\sum_{i\in s_{3(1)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ |s_{1(1)}\big)\Big]\\
&\quad-\mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big]+\mathrm{Var}\Big[E\big(\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ |s_{1(1)}\big)\Big]\\
=&\mathrm{Var}\Big[\sum_{i\in s_{3(1)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big]-\mathrm{Var}\Big[\sum_{i\in s_{1(1)}} w_{i1}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big]\\
&\quad-\mathrm{Var}\Big[\sum_{i\in s_{2(1)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big]+\mathrm{Var}\Big[\sum_{i\in s_{1(1)}} w_{i1}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\big)\Big].
\end{aligned}
$$

Also, with similar calculations, we obtain:

$$
\begin{aligned}
N^2D_{(2)}=&\ \mathrm{Var}_p\Big[\sum_{i\in U_{2(2)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(2)}\Big]\\
=&\ \mathrm{Var}\Big\{E\big[\sum_{i\in U_{2(2)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(2)}\ \big|\ s_{2(2)}\big]\Big\}+E\Big\{\mathrm{Var}\big[\sum_{i\in U_{2(2)}} A_iW_i\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(2)}\ \big|\ s_{2(2)}\big]\Big\}\\
=&\ N^2D_{(2)2}+N^2D_{(2)3}\ ,
\end{aligned}
$$

with:

$$
N^2D_{(2)2}=\mathrm{Var}\Big\{\sum_{i\in s_{2(2)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots 3)}\Big\}
$$

and

$$
\begin{aligned}
N^2D_{(2)3}=&E\Big\{\mathrm{Var}\big[\sum_{i\in s_{3(2)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\ \big|\ s_{2(2)}\big]\Big\}\\
=&\ \mathrm{Var}\Big[\sum_{i\in s_{3(2)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big]-\mathrm{Var}\Big\{\sum_{i\in s_{2(2)}} w_{i2}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big\}.
\end{aligned}
$$

Finally,

$$
N^2D_{(3)}=\ N^2D_{(3)3}\ =\ \mathrm{Var}\Big\{\sum_{i\in s_{3(3)}} w_{i3}A_i\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(3\cdots 3)}\Big\}.
$$

Appendix B - Tables

Table 1: Variable/Category Labels

Label	**Definition**
YrsSince95	Number of years since 1995
YrsSinceDe	Number of years since doctoral degree
YrsSinceDe2	Square of number of years since doctoral degree
Male, Female	Gender male, female
UScit	U.S. citizen
NotUScit	Non-U.S. citizen
AGEatGrad	Age at doctoral graduation
NotPostDoc	Principal job is not a postdoc
PostDoc	Principal job is a postdoc
HRSWK	Number of hours worked per week
HRSWK2	Square of number of hours worked per week
NotAdjFac	Position is not as adjunct faculty
AdjFac	Position is as adjunct faculty
JobCloselyRel	Principal job closely related to doctoral degree
JobSomewhaRel	Principal job somewhat related to doctoral degree
JobNotRelOthe	Principal job not related to doctoral degree for other reasons
JobNotRelCarr	Principal job not related to doctoral degree for change in career or professional interests
White	Non-Hispanic white
Asian	Non-Hispanic Asian
NatAm	Non-Hispanic American Indian/Alaska Native
Black	Non-Hispanic black
Hispa	Hispanic, any race
Other	Non-Hispanic Native Hawaiian/Other Pacific Islander ONLY and multiple race
C4TenPu	Sector: Tenured in public 4-year college
C4NTePu	Sector: Not tenured in public 4-year college
C4NATPu	Sector: Tenure N/A in public 4-year college
C4TenPr	Sector: Tenured in private 4-year college
C4NATPr	Sector: Tenure N/A in private 4-year college
C4NTePr	Sector: Not tenured in private 4-year college
Coll2yr	Sector: Two-year college
IndProf	Sector: Industry for profit
IndSelf	Sector: Industry self-employed
IndNoPr	Sector: Industry non-profit organization
Federal	Sector: Federal government

StLocGv	Sector: State or local government
D-BioloSci	Doctoral degree in biological sciences
D-AgriFood	Doctoral degree in agricultural and food sciences
D-EnvirSci	Doctoral degree in environmental life sciences
D-CompInfo	Doctoral degree in computing and information sciences
D-MatheSci	Doctoral degree in mathematics and statistics
D-PhyAstro	Doctoral degree in physical and astronomical sciences
D-ChemNoBi	Doctoral degree in chemistry (except biochemistry)
D-EarAtmOc	Doctoral degree in earth, atmospheric, and ocean sciences
D-Psycholo	Doctoral degree in psychology
D-Economic	Doctoral degree in economics
D-PolitSci	Doctoral degree in political sciences
D-OtherSoc	Doctoral degree in other social sciences
D-AerosEng	Doctoral degree in aerospace, aeronautical and astronautical engineering
D-ChemiEng	Doctoral degree in chemical engineering
D-CivilEng	Doctoral degree in civil engineering
D-ElecComp	Doctoral degree in electrical and computer engineering
D-OtherEng	Doctoral degree in other engineering
D-MechaEng	Doctoral degree in mechanical engineering
D-HealthSc	Doctoral degree in health
J-Biological	Job category: biological sciences
J-Computer	Job category - computing and information sciences
J-Math	Job category - mathematics and statistics
J-AgriFood	Job category - agricultural and food sciences
J-EnvEarthAt	Job category: environmental, earth, atmospheric, and ocean sciences
J-Chemistry	Job category - chemistry (except biochemistry)
J-Physical	Job category - physical and astronomical sciences
J-Economics	Job category: economics
J-Political	Job category: political sciences
J-Psychology	Job category: psychology
J-OtherSoc	Job category: other social sciences
J-EngArcTec	Job category: engineers, architects, engineering technicians
J-Health	Job category: health occupations
J-Manager	Job category: managers
J-NonSandE	Job category: non-science and engineering
J-Other	Job category: other
RegionCPI	Average Consumer Price Index of the region of employment for the half-year including the survey reference week
FT/PTotherReaNNW	Principal job 35+ hours/week or <35 for other reasons than not needing/wanting more hours

PTNotNeedWant	Principal job <35 hours/week b/c did not need or want to work more hours
FT/PTotherReaRet	Principal job 35+ hours/week or <35 for other reasons than (semi-)retired
PTRET0	Principal job <35 hours/week b/c (semi-)retired <1 year ago
PTRET1	Principal job <35 hours/week b/c (semi-)retired 1 year ago
PTRET2	Principal job <35 hours/week b/c (semi-)retired 2 years ago
PTRET3	Principal job <35 hours/week b/c (semi-)retired 3 years ago
PTRET4pl	Principal job <35 hours/week b/c (semi-)retired 4+ years ago
FT/PTotherReaFTNA	Principal job 35+ hours/week or <35 for other reasons than full-time job not available
PTFullNA	Principal job <35 hours/week b/c full-time job not available
NuMonSinSTRT	Number of months since started principal job
STRTJan – STRTDec	Month of start of principal job: January – December
SamEmpSamJob	Same employer and same type of job during previous wave's reference week
SamEmpDifJob	Same employer but different type of job during previous wave's reference week
DifEmpSamJob	Different employer but same type of job during previous wave's reference week
DifEmpDifJob	Different employer and different type of job during previous wave's reference week
NOWorkPrevRW	Not working for pay during previous wave's reference week
SamEmJo/CHotherReaLay	Same employer/type of job or changed employer/type of job for reasons other than laid off/job terminated
CHLayTerm	Change employer or job (from previous wave) b/c laid off or job terminated
Married	Married
MarrLik	Living in a marriage-like relationship
Widowed	Widowed
Separat	Separated
Divorce	Divorced
NevMarr	Never married
NoChild	No children living in family
ChUnd02	At least on child <2 in family
Ch02_05	At least on child 2-5 in family
Ch06_11	At least on child 6-11 in family
Ch12plu	At least on child 12+ in family
NoSpou/SpouNotWk	Widowed/Separated/Divorced/Never married or Spouse not working
SpouFT	Spouse working full-time
SpouPT	Spouse working part-time

Alabama – Wyoming	State of employment: 50 states plus Washington D.C.
Puerto Rico	State of employment: Puerto Rico
Terr/Abroad	State of employment: Other

Table 2: Parameter Estimates

Parameter	**Estimate**	**SE**	**LL95**	**UL95**	**p.value**	
Intercept	9.40456	0.05489	9.2970	9.5121	0.0000	*
YrsSince95	0.02136	0.00137	0.0187	0.0240	0.0000	*
YrsSinceDe	0.03038	0.00065	0.0291	0.0317	0.0000	*
YrsSinceDe2	-0.00055	0.00002	-0.0006	-0.0005	0.0000	*
Male	0	0	0	0	.	
Female	-0.06800	0.01358	-0.0946	-0.0414	0.0000	*
UScit	0	0	0	0	.	
NotUScit	-0.01898	0.00523	-0.0292	-0.0087	0.0003	*
AGEatGrad	-0.00569	0.00039	-0.0065	-0.0049	0.0000	*
NotPostDoc	0	0	0	0	.	
PostDoc	-0.29812	0.00599	-0.3099	-0.2864	0.0000	*
HRSWK	0.03848	0.00090	0.0367	0.0402	0.0000	*
HRSWK2	-0.00031	0.00001	-0.0003	-0.0003	0.0000	*
NotAdjFac	0	0	0	0	.	
AdjFac	-0.52241	0.03378	-0.5886	-0.4562	0.0000	*
JobCloselyRel	0	0	0	0	.	
JobSomewhaRel	-0.06921	0.01692	-0.1024	-0.0360	0.0000	*
JobNotRelOthe	-0.28027	0.04413	-0.3668	-0.1938	0.0000	*
JobNotRelCarr	-0.19284	0.03381	-0.2591	-0.1266	0.0000	*
White	0	0	0	0	.	
Asian	-0.00325	0.00628	-0.0156	0.0090	0.6042	
NatAm	-0.00107	0.02568	-0.0514	0.0492	0.9666	
Black	0.00281	0.00921	-0.0152	0.0209	0.7606	
Hispa	0.01659	0.00844	0.0000	0.0331	0.0493	.
Other	-0.03544	0.02027	-0.0752	0.0043	0.0803	
C4TenPu	0	0	0	0	.	
C4NTePu	-0.07451	0.00486	-0.0840	-0.0650	0.0000	*
C4NATPu	-0.12011	0.00602	-0.1319	-0.1083	0.0000	*
C4TenPr	-0.02019	0.00740	-0.0347	-0.0057	0.0064	*
C4NATPr	-0.11777	0.00817	-0.1338	-0.1018	0.0000	*
C4NTePr	-0.09621	0.00669	-0.1093	-0.0831	0.0000	*
Coll2yr	-0.13495	0.01084	-0.1562	-0.1137	0.0000	*
IndProf	0.20468	0.00580	0.1933	0.2160	0.0000	*
IndSelf	0.07374	0.01264	0.0490	0.0985	0.0000	*

IndNoPr	0.05115	0.00850	0.0345	0.0678	0.0000	*
Federal	0.13489	0.00723	0.1207	0.1491	0.0000	*
StLocGv	-0.00783	0.00960	-0.0266	0.0110	0.4148	
D-BioloSci	0	0	0	0	.	
D-AgriFood	-0.07480	0.01330	-0.1009	-0.0487	0.0000	*
D-EnvirSci	-0.03586	0.01653	-0.0683	-0.0035	0.0301	.
D-CompInfo	0.19753	0.01543	0.1673	0.2278	0.0000	*
D-MatheSci	0.02077	0.01197	-0.0027	0.0442	0.0825	
D-PhyAstro	0.03981	0.01131	0.0176	0.0620	0.0004	*
D-ChemNoBi	0.00877	0.00911	-0.0091	0.0266	0.3354	
D-EarAtmOc	-0.03736	0.01603	-0.0688	-0.0059	0.0197	.
D-Psycholo	0.00615	0.01145	-0.0163	0.0286	0.5911	
D-Economic	0.15835	0.01864	0.1218	0.1949	0.0000	*
D-PolitSci	0.07451	0.01703	0.0411	0.1079	0.0000	*
D-OtherSoc	-0.04433	0.01076	-0.0654	-0.0232	0.0000	*
D-AerosEng	0.10661	0.02297	0.0616	0.1516	0.0000	*
D-ChemiEng	0.09342	0.01328	0.0674	0.1195	0.0000	*
D-CivilEng	0.05277	0.01579	0.0218	0.0837	0.0008	*
D-ElecComp	0.16121	0.01098	0.1397	0.1827	0.0000	*
D-OtherEng	0.08402	0.01132	0.0618	0.1062	0.0000	*
D-MechaEng	0.07870	0.01540	0.0485	0.1089	0.0000	*
D-HealthSc	0.08850	0.01035	0.0682	0.1088	0.0000	*
J-Biological	0	0	0	0	.	
J-Computer	0.02962	0.00908	0.0118	0.0474	0.0011	*
J-Math	0.01153	0.01141	-0.0108	0.0339	0.3121	
J-AgriFood	0.00556	0.01047	-0.0150	0.0261	0.5958	
J-EnvEarthAt	0.02963	0.01152	0.0071	0.0522	0.0101	.
J-Chemistry	-0.02075	0.00858	-0.0376	-0.0039	0.0156	.
J-Physical	0.00145	0.00967	-0.0175	0.0204	0.8805	
J-Economics	0.01261	0.01685	-0.0204	0.0456	0.4544	
J-Political	-0.09633	0.01945	-0.1344	-0.0582	0.0000	*
J-Psychology	-0.00517	0.01001	-0.0248	0.0144	0.6053	
J-OtherSoc	-0.01645	0.00970	-0.0355	0.0026	0.0898	
J-EngArcTec	0.03292	0.00814	0.0170	0.0489	0.0001	*
J-Health	0.10233	0.00788	0.0869	0.1178	0.0000	*
J-Manager	0.18488	0.00652	0.1721	0.1977	0.0000	*
J-NonSandE	-0.03081	0.00846	-0.0474	-0.0142	0.0003	*
J-Other	-0.23034	0.01704	-0.2637	-0.1969	0.0000	*
RegionCPI	0.00240	0.00028	0.0019	0.0029	0.0000	*
FT/PTotherReaNNW	0	0	0	0	.	
PTNotNeedWant	-0.05529	0.01518	-0.0850	-0.0255	0.0003	*
FT/PTotherReaRet	0	0	0	0	.	

PTRET0	-0.10097	0.04615	-0.1914	-0.0105	0.0287	.
PTRET1	-0.16652	0.03616	-0.2374	-0.0956	0.0000	*
PTRET2	-0.21385	0.03988	-0.2920	-0.1357	0.0000	*
PTRET3	-0.29363	0.04860	-0.3889	-0.1984	0.0000	*
PTRET4pl	-0.31368	0.02787	-0.3683	-0.2591	0.0000	*
FT/PTotherReaFTNA	0	0	0	0	.	
PTFullNA	-0.19218	0.02189	-0.2351	-0.1493	0.0000	*
NoMonSinSTRT	0.00023	0.00002	0.0002	0.0003	0.0000	*
STRTJan	0.01768	0.00497	0.0079	0.0274	0.0004	*
STRTFeb	0.03100	0.00744	0.0164	0.0456	0.0000	*
STRTMar	0.03782	0.00690	0.0243	0.0513	0.0000	*
STRTApr	0.03794	0.00717	0.0239	0.0520	0.0000	*
STRTMay	0.04195	0.00655	0.0291	0.0548	0.0000	*
STRTJun	0.02299	0.00567	0.0119	0.0341	0.0001	*
STRTJul	0.05065	0.00521	0.0404	0.0609	0.0000	*
STRTAug	0	0	0	0	.	
STRTSep	0.00301	0.00424	-0.0053	0.0113	0.4781	
STRTOct	0.03655	0.00647	0.0239	0.0492	0.0000	*
STRTNov	0.04083	0.00728	0.0266	0.0551	0.0000	*
STRTDec	0.03931	0.00663	0.0263	0.0523	0.0000	*
SamEmpSamJob	0	0	0	0	.	
SamEmpDifJob	0.03309	0.00451	0.0242	0.0419	0.0000	*
DifEmpSamJob	0.04057	0.00493	0.0309	0.0502	0.0000	*
DifEmpDifJob	-0.00994	0.00453	-0.0188	-0.0010	0.0284	.
NOWorkPrevRW	-0.10712	0.00714	-0.1211	-0.0931	0.0000	*
SamEmJo/CHotherReaLay	0	0	0	0	.	
CHLayTerm	-0.04758	0.00532	-0.0580	-0.0372	0.0000	*
Married	0	0	0	0	.	
MarrLik	-0.06368	0.01890	-0.1007	-0.0266	0.0008	*
Widowed	-0.06456	0.01216	-0.0884	-0.0407	0.0000	*
Separat	-0.06628	0.01024	-0.0863	-0.0462	0.0000	*
Divorce	-0.09140	0.00750	-0.1061	-0.0767	0.0000	*
NevMarr	-0.11255	0.00971	-0.1316	-0.0935	0.0000	*
NoChild	0	0	0	0	.	
ChUnd02	-0.00246	0.00430	-0.0109	0.0060	0.5680	
Ch02_05	0.01049	0.00366	0.0033	0.0177	0.0041	*
Ch06_11	0.02092	0.00362	0.0138	0.0280	0.0000	*
Ch12plu	0.01751	0.00375	0.0102	0.0249	0.0000	*
NoSpou/SpouNotWk	0	0	0	0	.	
SpouFT	-0.05595	0.00463	-0.0650	-0.0469	0.0000	*
SpouPT	-0.03224	0.00502	-0.0421	-0.0224	0.0000	*
Alabama	-0.10595	0.01616	-0.1376	-0.0743	0.0000	*

Alaska	-0.11938	0.03388	-0.1858	-0.0530	0.0004	*
Arizona	-0.12386	0.01657	-0.1563	-0.0914	0.0000	*
Arkansas	-0.15199	0.02340	-0.1979	-0.1061	0.0000	*
California	0	0	0	0	.	
Colorado	-0.16204	0.01409	-0.1897	-0.1344	0.0000	*
Connecticut	-0.03439	0.01439	-0.0626	-0.0062	0.0169	.
Delaware	-0.02527	0.01765	-0.0599	0.0093	0.1522	
Washington D.C.	0.01504	0.01288	-0.0102	0.0403	0.2430	
Florida	-0.13093	0.01241	-0.1553	-0.1066	0.0000	*
Georgia	-0.09078	0.01284	-0.1160	-0.0656	0.0000	*
Hawaii	-0.08002	0.02831	-0.1355	-0.0245	0.0047	*
Idaho	-0.19638	0.02956	-0.2543	-0.1384	0.0000	*
Illinois	-0.06628	0.00924	-0.0844	-0.0482	0.0000	*
Indiana	-0.11406	0.01346	-0.1404	-0.0877	0.0000	*
Iowa	-0.15171	0.01829	-0.1876	-0.1158	0.0000	*
Kansas	-0.17994	0.02280	-0.2246	-0.1353	0.0000	*
Kentucky	-0.17654	0.01936	-0.2145	-0.1386	0.0000	*
Louisiana	-0.12316	0.01470	-0.1520	-0.0944	0.0000	*
Maine	-0.24040	0.03333	-0.3057	-0.1751	0.0000	*
Maryland	-0.03444	0.00917	-0.0524	-0.0165	0.0002	*
Massachusetts	-0.05635	0.00930	-0.0746	-0.0381	0.0000	*
Michigan	-0.07467	0.00939	-0.0931	-0.0563	0.0000	*
Minnesota	-0.10286	0.01492	-0.1321	-0.0736	0.0000	*
Mississippi	-0.10886	0.02459	-0.1571	-0.0607	0.0000	*
Missouri	-0.12262	0.01357	-0.1492	-0.0960	0.0000	*
Montana	-0.25572	0.03636	-0.3270	-0.1844	0.0000	*
Nebraska	-0.16338	0.02185	-0.2062	-0.1205	0.0000	*
Nevada	-0.05641	0.03107	-0.1173	0.0045	0.0694	
New Hampshire	-0.17370	0.02234	-0.2175	-0.1299	0.0000	*
New Jersey	-0.01083	0.01081	-0.0320	0.0104	0.3165	
New Mexico	-0.10454	0.01763	-0.1391	-0.0700	0.0000	*
New York	-0.02956	0.00872	-0.0467	-0.0125	0.0007	*
North Carolina	-0.08689	0.01029	-0.1071	-0.0667	0.0000	*
North Dakota	-0.18912	0.03479	-0.2573	-0.1209	0.0000	*
Ohio	-0.13808	0.00915	-0.1560	-0.1201	0.0000	*
Oklahoma	-0.16905	0.02032	-0.2089	-0.1292	0.0000	*
Oregon	-0.16912	0.01340	-0.1954	-0.1428	0.0000	*
Pennsylvania	-0.11581	0.00940	-0.1342	-0.0974	0.0000	*
Rhode Island	-0.11207	0.02565	-0.1624	-0.0618	0.0000	*
South Carolina	-0.12120	0.01733	-0.1552	-0.0872	0.0000	*
South Dakota	-0.19963	0.03694	-0.2720	-0.1272	0.0000	*
Tennessee	-0.13393	0.01296	-0.1593	-0.1085	0.0000	*

Texas	-0.06889	0.00732	-0.0832	-0.0545	0.0000	*
Utah	-0.14044	0.02164	-0.1829	-0.0980	0.0000	*
Vermont	-0.24691	0.02608	-0.2980	-0.1958	0.0000	*
Virginia	-0.03924	0.00774	-0.0544	-0.0241	0.0000	*
Washington	-0.11082	0.01196	-0.1343	-0.0874	0.0000	*
West Virginia	-0.15164	0.02746	-0.2055	-0.0978	0.0000	*
Wisconsin	-0.15275	0.01285	-0.1779	-0.1276	0.0000	*
Wyoming	-0.15591	0.04627	-0.2466	-0.0652	0.0008	*
Puerto Rico	-0.26112	0.02378	-0.3077	-0.2145	0.0000	*
Terr/Abroad	-0.21240	0.05858	-0.3272	-0.0976	0.0003	*
Male / 0YrsSince95	0	0	0	0	.	
Fem*YrsSince95	0.00068	0.00065	-0.0006	0.0019	0.2913	
Male / 0YrsSinceDe	0	0	0	0	.	
Fem*YrsSinceDe	-0.00316	0.00051	-0.0042	-0.0022	0.0000	*
NotAdjFac / 0HRSWK	0	0	0	0	.	
AdjFac*HRSWK	0.00852	0.00073	0.0071	0.0100	0.0000	*
JobCloselyRel / 0HRSWK	0	0	0	0	.	
JobSomewhaRel*HRSWK	0.00128	0.00034	0.0006	0.0020	0.0002	*
JobNotRelOthe*HRSWK	0.00379	0.00098	0.0019	0.0057	0.0001	*
JobNotRelCarr*HRSWK	0.00301	0.00068	0.0017	0.0043	0.0000	*
Male / Married	0	0	0	0	.	
Fem*MarrLik	0.08722	0.03345	0.0217	0.1528	0.0091	*
Fem*Widowed	0.04184	0.02291	-0.0031	0.0867	0.0678	
Fem*Separat	0.05907	0.01737	0.0250	0.0931	0.0007	*
Fem*Divorce	0.07054	0.01475	0.0416	0.0994	0.0000	*
Fem*NevMarr	0.07495	0.01745	0.0407	0.1092	0.0000	*
Male / NoChild	0	0	0	0	.	
Fem*ChUnd02	0.00761	0.00853	-0.0091	0.0243	0.3723	
Fem*Ch02_05	-0.00674	0.00719	-0.0208	0.0073	0.3484	
Fem*Ch06_11	-0.02766	0.00707	-0.0415	-0.0138	0.0001	*
Fem*Ch12plu	-0.00779	0.00806	-0.0236	0.0080	0.3339	
Male / NoSpou/SpouNotWk	0	0	0	0	.	
Fem*SpouFT	0.03216	0.01096	0.0107	0.0536	0.0033	*
Fem*SpouPT	0.01514	0.01319	-0.0107	0.0410	0.2511	
NoMarrLik / NoSpou/SpouNotWk	0	0	0	0	.	
MarrLik*SpouFT	0.04734	0.02060	0.0070	0.0877	0.0215	.
MarrLik*SpouPT	0.02542	0.03715	-0.0474	0.0982	0.4939	
FT/PTotherReaNNW / 0MonSinSTRT	0	0	0	0	.	
PTNotNeedWant*NoMonSinSTRT	0.00034	0.00007	0.0002	0.0005	0.0000	*

FT/PTotherReaNNW / PTotherReaRet	0	0	0	0	.	
PTNotNeedWant*PTRET0	-0.10434	0.06991	-0.2414	0.0327	0.1356	
PTNotNeedWant*PTRET1	0.05585	0.05171	-0.0455	0.1572	0.2801	
PTNotNeedWant*PTRET2	0.03425	0.04640	-0.0567	0.1252	0.4605	
PTNotNeedWant*PTRET3	0.12503	0.05914	0.0091	0.2409	0.0345	.
PTNotNeedWant*PTRET4pl	0.00877	0.03432	-0.0585	0.0760	0.7984	
FT/PTotherReaRet / SamEmJo/CHotherReaLay	0	0	0	0	.	
PTRET0*CHLayTerm	0.20425	0.19539	-0.1787	0.5872	0.2959	
PTRET1*CHLayTerm	0.06206	0.15226	-0.2364	0.3605	0.6836	
PTRET2*CHLayTerm	0.09143	0.10861	-0.1215	0.3043	0.3999	
PTRET3*CHLayTerm	0.30352	0.14646	0.0165	0.5906	0.0382	.
PTRET4pl*CHLayTerm	0.26258	0.12391	0.0197	0.5054	0.0341	.
FT/PTotherReaFTNA / 0MonSinSTRT	0	0	0	0	.	
PTFullNA*NoMonSinSTRT	0.00058	0.00014	0.0003	0.0009	0.0001	*
SamEmpSamJob / FT/PTotherReaRet	0	0	0	0	.	
SamEmpDifJob*PTRET0	-0.20041	0.07602	-0.3494	-0.0514	0.0084	*
SamEmpDifJob*PTRET1	-0.23828	0.07638	-0.3880	-0.0886	0.0018	*
SamEmpDifJob*PTRET2	-0.16398	0.08670	-0.3339	0.0059	0.0586	
SamEmpDifJob*PTRET3	-0.08353	0.10852	-0.2962	0.1292	0.4415	
SamEmpDifJob*PTRET4pl	-0.12142	0.10231	-0.3219	0.0791	0.2353	
DifEmpSamJob*PTRET0	-0.15913	0.14136	-0.4362	0.1179	0.2603	
DifEmpSamJob*PTRET1	-0.05197	0.07589	-0.2007	0.0968	0.4935	
DifEmpSamJob*PTRET2	-0.07256	0.12517	-0.3179	0.1728	0.5621	
DifEmpSamJob*PTRET3	-0.32880	0.17586	-0.6735	0.0159	0.0615	
DifEmpSamJob*PTRET4pl	-0.05116	0.11117	-0.2690	0.1667	0.6454	
DifEmpDifJob*PTRET0	-0.22100	0.09540	-0.4080	-0.0340	0.0205	.
DifEmpDifJob*PTRET1	-0.14962	0.07145	-0.2897	-0.0096	0.0362	.
DifEmpDifJob*PTRET2	-0.17286	0.07168	-0.3134	-0.0324	0.0159	.
DifEmpDifJob*PTRET3	-0.28605	0.14911	-0.5783	0.0062	0.0551	
DifEmpDifJob*PTRET4pl	-0.18792	0.10055	-0.3850	0.0092	0.0616	
NOWorkPrevRW*PTRET0	0.04092	0.17089	-0.2940	0.3759	0.8107	
NOWorkPrevRW*PTRET1	-0.10666	0.13375	-0.3688	0.1555	0.4251	
NOWorkPrevRW*PTRET2	-0.30248	0.08783	-0.4746	-0.1303	0.0006	*
NOWorkPrevRW*PTRET3	-0.20368	0.07911	-0.3587	-0.0486	0.0100	.
NOWorkPrevRW*PTRET4pl	-0.14042	0.04983	-0.2381	-0.0428	0.0048	*
SamEmpSamJob / 0MonSinSTRT	0	0	0	0	.	

SamEmpDifJob*NoMonSinSTRT	-0.00016	0.00005	-0.0003	-0.0001	0.0011	*
DifEmpSamJob*NoMonSinSTRT	-0.00021	0.00011	-0.0004	0.0000	0.0646	
DifEmpDifJob*NoMonSinSTRT	0.00004	0.00014	-0.0002	0.0003	0.7947	
NOWorkPrevRW*NoMonSinSTRT	0.00020	0.00008	0.0000	0.0004	0.0105	.
NoMarrLik / Male / NoSpou/SpouNotWk	0	0	0	0	.	
MarrLik*Fem*SpouFT	-0.06740	0.03630	-0.1385	0.0037	0.0633	
MarrLik*Fem*SpouPT	-0.02805	0.05674	-0.1393	0.0832	0.6210	

Signif. codes: '*' 0.01, '.' 0.05

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 / La Revue Canadienne de Statistique*, 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 / La Revue Canadienne de Statistique*, 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-García, I. A. (2008). Analysis of longitudinal surveys with missing responses. Ph.D. thesis, University of Waterloo, ON, Canada.

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 / La Revue Canadienne de Statistique*, 28 (3), 449–481.

Larsen, M. D., Qing, S., Zhou, B., and Foulkes, M. A. (2011). Calibration estimation and longitudinal surey weights: Application to the NSF Survey of Doctorate Recipients. Topic Contributed Paper. 2011 Joint Statistical Meetings, Miami, FL.

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: ASA Proceedings of the Joint Statistical Meetings. American Statistical Association, pp. 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.

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*. Springer-Verlag, New York.

Smith, P., Lynn, P., and Elliot, D. (2009). Sample design for longitudinal surveys. In: Lynn, P. (Ed.), Methodology of Longitudinal Surveys. Wiley, Chichester, Ch. 2, pp. 21–33.

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

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

Wolter, K. M. (2007). *Introduction to Variance Estimation*, 2nd Edition. Springer, New York.