NISS

# Multinomial Logistic Regression with Data from Multi-Cohort Longitudinal Surveys

Iván A. Carrillo and Alan F. Karr

Technical Report 193
July 2014

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

# Multinomial Logistic Regression with Data from Multi-Cohort Longitudinal Surveys

Iván A. Carrillo[1] and Alan F. Karr[2]

July 4, 2014

In this paper we give a detailed explanation of how to estimate a multinomial logistic regression model using data from a longitudinal survey with multiple cohorts. We also show how to estimate the variance of the parameter estimates by using the estimating function bootstrap or any other design variance estimate available in the literature. We argue why it is more appropriate to estimate the autocorrelation matrix by quasi-least squares rather than by the method of moments or the odds ratios parameterization, and we show how to do so. We illustrate the technique by estimating a model for employment sector from the U. S. National Science Foundation's Survey of Doctorate Recipients, and interpret the results. Additionally we present a simulated score test for assessing goodness of fit in general, and conclude that the estimated model for employment sector fits the data well.

*Key Words:* Marginal model parameters; Rotating panel surveys; Replication variance estimation; Weighted Generalized Estimating Equations; Goodness of fit; Hosmer-Lemeshow test.

**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 Stephen Cohen and Nirmala Kannankutty, of the National Center for Science and Engineering Statistics at NSF, for numerous insightful discussions during the research.

[1]National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA. Current affiliation: Statistics Canada, Social Survey Methods Division, Tunney's Pasture, R.H. Coats Building, 15th Floor, Ottawa, ON K1A 0T6, Canada. E-mail: ivan.carrillogarcia@statcan.gc.ca .

[2]National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA. E-mail: karr@niss.org .

# 1 Introduction

Although there are many methods of analysis for single-cohort longitudinal surveys, see for example Carrillo et al. (2010) or Vieira (2009), the estimation of marginal models with data from complex multi-cohort longitudinal surveys is a topic only lately being developed.

In a recent paper, Carrillo and Karr (2013) proposed, under a weighted generalized estimating equation framework, a general way of estimating marginal mean models with data from multi-cohort longitudinal surveys. Their approach permits estimation of the effect of covariates on a response variable, using data from a variety of types of longitudinal survey data. These include, for example, fixed-panel, repeated-panel, rotating-panel, or split-panel survey data. That paper, however, only shows how the method applies to a continuous response. There are no details as of how to proceed with other kinds of responses, for example categorical responses.

Our goal in this paper is to estimate the effect of covariates on a categorical response, namely employment sector (such as academia, government or industry) for Ph.D. recipients in the sciences, engineering and health. We employ data from the Survey of Doctorate Recipients (SDR), which is conducted by the National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation (NSF). The SDR is a rotating-panel/repeated-panel longitudinal survey, and hence the estimation methodology proposed by Carrillo and Karr (2013) is suitable. Here, we present full details of how the technique can be applied to multinomial responses.

Roberts et al. (2009) is another recent work for estimation of covariates' effects on a categorical response using data from longitudinal surveys. However, they only consider binary responses (as opposed to multinomial), and more importantly, they consider only single-cohort data. In other words, in a survey like the SDR, their approach ignores subjects not common to all waves (which can happen either by design or as the result of nonresponse). The method presented in this paper, on the other hand, can incorporate all the available data, as long as any design features or nonresponse have been handled by adjustment of cross-sectional weights.

The paper is organized as follows. In Section 2 we provide the technical details of the application of the methodology of Carrillo and Karr (2013) to multinomial responses. These details include the description of the kinds of survey data to which the method can be applied, the model of interest, the estimation of parameters and of variance, as well as the most suitable way of estimating the autocorrelation matrix used

by the GEE methodology. Section 3 presents a brief description of the sampling design of the SDR, the response variable of interest and covariates, and the estimated effects of the latter on the former. There, we apply the proposed approach to the data at hand, and interpret the results. We close that section by showing how we go about assessing the goodness-of-fit of the proposed model. Finally, in Section 4 we provide some concluding remarks.

# 2 Methodology

In this section, we describe our estimation methodology.

## 2.1 Sample

We consider the case where the data come from a multi-cohort longitudinal survey, of which a "single-panel" longitudinal sample is a special case (see Smith et al., 2009). At a certain point in time a (complex) sample is selected from the target population. For the next wave or cycle, the same set of subjects, or a subset of them plus some new individuals, is interviewed. The original sample is the first cohort, and the new subjects introduced at the second wave comprise the second cohort. This mechanism of "thinning" of previous cohorts and addition of new subjects can be repeated every wave as necessary. Generally, the purposes of such a rotation mechanism are to control costs and to maintain representativeness of the sample cross-sectionally over time.

For the purpose of this paper, the population of interest may be defined as (a) "a static population based on the population at the time the first wave sample is selected," (b) "the intersection of the cross-sectional populations at each wave," or (c) "the union of the cross-sectional populations at each wave" (Smith et al., 2009). The crucial requirement of our approach is that, whatever the target population is, for each subject $i$ interviewed at wave $j$ there is a survey weight $w_{ij}$ to represent the target population at wave $j$. Note that this is applicable even in case (a), where the target population is defined at wave one, as that is exactly the same population of interest at wave $j$. However, in that case, $w_{ij}$ may not depend on $j$ (except perhaps for nonresponse adjustments).

Obviously, the weighting procedure may not be straightforward. First, the original sample, as well as the "new" samples selected at each wave, may be complex. Second, the theory of multiple frames may need to be used; for example where a frame of recent population "births" (e.g., new Ph.D. recipients) is not

available for one or more waves. Finally, adjustments for nonresponse, dropouts, and intermittent patterns are required. However, those topics are treated elsewhere, and are beyond the scope of this paper. For the first, see Särndal et al. (1992), for the second, see for example Rao and Wu (2010), and for the last see Chen et al. (2012).

## 2.2 Model

Assume that the categorical variable of interest $Y$ can take on any one of $K$ values labelled $0,1,2,\ldots,K-1$. Paraphrasing Hosmer and Lemeshow (2000), the goal is to model the odds of category outcome as a function of covariates and to express the results in terms of odds ratios for the different category outcomes.

Let $Y_{ij}$ be the response category for subject $i$ at wave $j$; where $i \in s_j$, the sample at wave $j$, and $j = 1,2,\ldots,J$. For each subject there is a $(p+1)\times 1$ vector of explanatory variables $\boldsymbol{X}_{ij}$ at wave $j$. Interest lies in estimating the $\beta$ regression coefficients in the following model:

$$\log\frac{p_{ijk}}{p_{ij0}} = \log\frac{P(Y_{ij}=k|\boldsymbol{X}_{ij})}{P(Y_{ij}=0|\boldsymbol{X}_{ij})} = \boldsymbol{X}'_{ij}\boldsymbol{\beta}_k, \qquad k=1,2,\ldots,K-1, \tag{1}$$

where $\boldsymbol{\beta}_k = (\beta_{k0},\beta_{k1},\ldots,\beta_{kp})'$, $p_{ijk} \stackrel{\text{def}}{=} P(Y_{ij}=k|\boldsymbol{X}_{ij})$ for $k=0,1,2,\ldots,K-1$, and $\sum_{k=0}^{K-1} p_{ijk} = 1$ for all $i,j$. As Agresti (2002) explains, model (1) permits the simultaneous calculation of the (log) odds for all $\binom{K}{2}$ pairs of response categories.

## 2.3 Estimation

The type of longitudinal survey data described in Section 2.1 can be used to estimate model (1) by means of the approach proposed by Carrillo and Karr (2013). They introduced a methodology suited for estimation of generalized linear model parameters from multi-cohort longitudinal surveys, but did not demonstrate how the method applies to multinomial responses. In this section we present the specifics for applying the technique for estimation of model (1).

Using model (1) with the proposal in Carrillo and Karr (2013), the solution to the estimating equations

$$\Psi_s(\boldsymbol{\beta}) = \sum_{i\in s}\frac{\partial \boldsymbol{p}'_i}{\partial \boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i-\boldsymbol{p}_i) = \boldsymbol{0} \tag{2}$$

is consistent for the $\boldsymbol{\beta}_k$'s as $n_m$ (the minimum of the cross-sectional sample sizes) increases. In expression (2), the sum is over $s$, the entire sample, i.e., all subjects regardless of in what waves they are observed; $\boldsymbol{p}_i = \left(\boldsymbol{p}'_{i1},\boldsymbol{p}'_{i2},\ldots,\boldsymbol{p}'_{iJ}\right)'$; $\boldsymbol{p}_{ij} = (p_{ij1},p_{ij2},\ldots,p_{ij(K-1)})'$; $\boldsymbol{\beta} = \left(\boldsymbol{\beta}'_1,\boldsymbol{\beta}'_2,\ldots,\boldsymbol{\beta}'_{(K-1)}\right)'$; $W_i = \text{diag}\{w_{i1}I_{(K-1)},$

$w_{i2}I_{(K-1)},\ldots,w_{iJ}I_{(K-1)}\}$; $w_{ij}$ is the survey weight for subject $i$ at wave $j$ if that subject is interviewed at wave $j$ and 0 otherwise; $I_{(K-1)}$ is the identity matrix of size $K-1$; $\boldsymbol{y}_i = \left(\boldsymbol{y}'_{i1},\boldsymbol{y}'_{i2},\ldots,\boldsymbol{y}'_{iJ}\right)'$; $\boldsymbol{y}_{ij} = (y_{ij1},y_{ij2},\ldots,y_{ij(K-1)})'$; $y_{ijk}=1$ if $Y_{ij}=k$ and 0 otherwise; $V_i = A_i^{1/2}R(\hat{\alpha})A_i^{1/2}$; $A_i = \text{diag}\{\boldsymbol{\nu}_{i1},\boldsymbol{\nu}_{i2},\ldots,\boldsymbol{\nu}_{iJ}\}$; $\boldsymbol{\nu}_{ij} = \text{diag}\{p_{ij1}(1-p_{ij1}),p_{ij2}(1-p_{ij2}),\ldots,p_{ij(K-1)}(1-p_{ij(K-1)})\}$; and $\hat{\alpha}$ is a $n_m^{1/2}$-consistent estimator of $\alpha$ such that $R(\alpha)$ is a conformable correlation matrix (see Liang and Zeger, 1986, for details). We extend the discussion about this last matrix in the next section.

Based on straightforward calculations, we have that

$$\frac{\partial \boldsymbol{p}_i}{\partial \boldsymbol{\beta}} = \{\mathbf{1}'_{(K-1)} \otimes \boldsymbol{p}_i\} \circ \left\{X_i - \left[\left(\{I_J \otimes \mathbf{1}'_{(K-1)}\}\{\left[\mathbf{1}'_{(K-1)} \otimes \boldsymbol{p}_i\right] \circ X_i\}\right) \otimes \mathbf{1}_{(K-1)}\right]\right\},$$

where $\mathbf{1}_{(K-1)}$ is a column vector of $K-1$ ones, "$\otimes$" denotes the Kronecker product, "$\circ$" denotes element-wise multiplication, $X_i = \left(X'_{i1},X'_{i2},\ldots,X'_{iJ}\right)'$, $X_{ij} = I_{(K-1)} \otimes \boldsymbol{X}'_{ij}$, and $I_J$ is the identity matrix of size $J$.

Expression (2) is a generalization of equation (20.4) in Roberts et al. (2009) in two ways. Firstly, whereas Roberts et al. (2009) consider only the case where the response has two categories, equation (2) is applicable when the response variable can take on any of K categories. More important, expression (2) allows for application to data from multi-cohort surveys, of which the case examined by Roberts et al. (2009), i.e., a single cohort with no dropouts or intermittent patterns, is just a particular example.

With respect to the variance of the estimator $\hat{\boldsymbol{\beta}}$, i.e. the solution to (2), Carrillo and Karr (2013) argue that, if the sampling fraction is small, it can be estimated by

$$\hat{V}(\hat{\boldsymbol{\beta}}) = [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\ \widehat{\text{Var}}[\Psi_s(\boldsymbol{\beta}_N)]\ [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}, \tag{3}$$

where $\hat{H}(\boldsymbol{\beta}) = \sum_{i\in s}(\partial \boldsymbol{p}'_i/\partial \boldsymbol{\beta})V_i^{-1}W_i(\partial \boldsymbol{p}_i/\partial \boldsymbol{\beta})$, and setting $w_{i0}=0$,

$$\text{Var}[\Psi_s(\boldsymbol{\beta}_N)] = \sum_{j=1}^{J}\left\{\text{Var}\Big[\sum_{i\in s_j} w_{ij}B_i 1_U(i)\boldsymbol{e}_{i(j\cdots J)}\Big] - \text{Var}\Big[\sum_{i\in s_{j-1}} w_{i,j-1}B_i 1_U(i)\boldsymbol{e}_{i(j\cdots J)}\Big]\right\}, \tag{4}$$

where $s_j$ is the set of subjects interviewed at wave $j$; $B_i = (\partial \boldsymbol{p}'_i/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}V_i^{-1}$; $\boldsymbol{\beta}_N$ is the "census estimator" that would be obtained as solution to equation (2) if it were applied to the full finite population $U$ instead of to the sample at hand; $1_U(i) = \text{diag}[1_{U_1}(i)I_{(K-1)},\ 1_{U_2}(i)I_{(K-1)},\ \ldots,\ 1_{U_J}(i)I_{(K-1)}]$; $1_{U_j}(i)$ is the indicator of whether or not subject $i$ is part of the finite population at time $j$; $\boldsymbol{e}_{i(j\ldots J)} = \left(\mathbf{0}'_{(j-1)},\mathbf{1}'_{(J-j+1)}\right)' \circ \boldsymbol{e}_i$; $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{p}_i(\boldsymbol{\beta}_N)$; and $\mathbf{0}_{(j-1)}$ is a column vector of $j-1$ zeroes.

To compute $\hat{V}(\hat{\boldsymbol{\beta}})$ from (3), it is necessary to get an estimate of $\text{Var}[\Psi_s(\boldsymbol{\beta}_N)]$ in expression (4); all the terms involved in this expression are design variances of cross-sectional survey design weighted estimators,

for which several estimation methods exist (Wolter, 2007). For estimating these variances in the application in Section 3 we use the estimating function bootstrap proposed by Roberts et al. (2003). Carrillo and Karr (2013) explain the procedure as follows: for estimation of the $j$-th term in (4), "the $r$-th replicate of the first term is $\sum_{i\in s_j} w_{ij}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{ij}^{(r)}$ is the $r$-th replicate weight for subject $i$ at wave $j$, and the $r$-th replicate of the second term is $\sum_{i\in s_{j-1}} w_{i,j-1}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{i,j-1}^{(r)}$ is the $r$-th replicate weight for subject $i$ at wave $j-1$."

### 2.3.1 Estimation of Autocorrelation Matrix

The method of moments proposed by Liang and Zeger (1986) for the estimation of $\alpha$ can produce an estimated matrix $R(\hat{\alpha})$ that is non-positive definite even though $R(\alpha)$ *is* positive definite. According to Chaganty et al. (2012), the fact that $\hat{R}$ may be non-positive definite "can lead to (most harmlessly) convergence problems, but it can also lead to artificially deflated estimator variances for the regression parameters and is thus subject to improper or incorrect inference." Not only that, a non-positive definite $\hat{R}$ may lead to a non-invertible $V_i$, which breaks the iteration procedure, impeding the estimation of $\boldsymbol{\beta}$.

There are several reasons why one may obtain non-positive definite *correlation* matrices in practice (see for example Wothke, 1993). We concentrate on the most important one for our case. In words of Wothke (1993), "missing observations due to nonresponse, response sampling design, morbidity, and various other reasons are a standard occurrence in social science data" and "under pairwise deletion of missing data the estimated sample covariance matrices may become indefinite."

The problem in our situation is that the missing data issue is taken to the extreme. Not only do we have the usual missingness of dropouts and intermittent patterns present in any longitudinal survey, but also the rotation mechanism of multiple cohorts induces missing data *by design*. Subjects in different cohorts enter and leave the survey at different waves. At the times when they are not part of the survey, they are "missing" from the dataset.

In the application presented in Section 3, estimating the matrix $R$ using Liang and Zeger's methodology does generate a non-positive definite $\hat{R}$. Unfortunately, the approach proposed by Lipsitz et al. (1991) and Roberts et al. (2009), which is arguably more appropriate for binary (or multinomial) responses, fails to resolve the issue. This should not be surprising as "existing methods for estimating $\alpha$ sometimes run into problems as there is no guarantee that the estimated value ensures that $R(\alpha)$ is positive definite" (Chaganty and Joe, 2004).

To avoid the non-positiveness issue, Chaganty (1997) proposed a "quasi-least squares" estimation approach, which is a modification of the original GEE methodology of Liang and Zeger (1986). They propose to estimate the parameters by solving equation (2) along with the following equation for the correlation structure (as opposed using to the method of moments):

$$\sum_{i\in s} \boldsymbol{Z}_i' \frac{\partial R^{-1}(\alpha)}{\partial \alpha_q} \boldsymbol{Z}_i = 0, \qquad 1 \le q \le Q,$$

if there are $Q$ correlation parameters (i.e., $R$ is parameterized by $Q$ elements in $\alpha$), where $\boldsymbol{Z}_i = W_i^{1/2} A_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{p}_i)$.

On the other hand, if, as in our case, one prefers not to stipulate any particular structure for the matrix $R$, and let the data "speak for themselves," Chaganty and Shults (1999) recommend the estimator

$$\hat{R} = \begin{cases} \hat{R}_{um} = \hat{R}_m \mathrm{diag}\{(\hat{R}_m \circ \hat{R}_m)^{-1}\mathbf{1}\}\hat{R}_m & \text{if } \hat{R}_m \text{ is positive definite} \\ \hat{R}_{sm} = \left(\mathrm{diag}\{\hat{Z}\}\right)^{-1/2} \hat{Z} \left(\mathrm{diag}\{\hat{Z}\}\right)^{-1/2} & \text{otherwise,} \end{cases} \tag{5}$$

where $\hat{R}_m = \Delta^{-1/2}(\Delta^{1/2}\hat{Z}\Delta^{1/2})^{1/2}\Delta^{-1/2}$; $\Delta$ is the solution to the fixed point equation $\Delta = \mathrm{diag}\{\Delta^{1/2}\hat{Z}\Delta^{1/2}\}^{1/2}$; and $\hat{Z} = \sum_{i\in s} \boldsymbol{Z}_i \boldsymbol{Z}_i'$. They argue that for some longitudinal data, it is possible that $\hat{R}_m$ is not positive definite, which *would indicate that the (multinomial) data are overdispersed* (see Collett, 2003, Ch. 6, for overdispersion). However, in any case, $R_{sm}$ is positive definite, and therefore the $\hat{R}$ in (5) is always positive definite; see Chaganty and Shults (1999) for details. Fortunately, in our application in Section 3, $\hat{R}_m$ is positive definite, and so we can conclude that the data at hand are not overdispersed. Finally, Chaganty and Naik (2002) show that, like GEE estimators, the quasi-least squares estimators obtained by solving equation (2), paired with the correlation matrix in (5), are also consistent for $\boldsymbol{\beta}$.

We end this section sketching an algorithm (adapted from Chaganty and Naik, 2002) for the computation of the quasi-least squares estimator:

1. Set a starting value $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$, can be $\boldsymbol{\beta}_k^{(0)} = \left(\log(\hat{p}_k/\hat{p}_0),\ \mathbf{0}'_{(p)}\right)'$ for $k = 1,2,\ldots,K$, with $\hat{p}_k = (\sum_{i\in s}\sum_{j=1}^{J} w_{ij} y_{ijk})/(\sum_{i\in s}\sum_{j=1}^{J} w_{ij})$ and $\hat{p}_0 = \left(\sum_{i\in s}\sum_{j=1}^{J} w_{ij}(1 - \sum_{k=1}^{K} y_{ijk})\right)/(\sum_{i\in s}\sum_{j=1}^{J} w_{ij})$.
2. With the current value of $\boldsymbol{\beta}^{(0)}$ calculate $A_i^{(0)}$, $\boldsymbol{p}_i^{(0)}$, $\boldsymbol{Z}_i^{(0)}$, and $\hat{Z}^{(0)}$.
3. Set a starting value $\Delta^{(0)}$ for $\Delta$, can be an identity matrix.
4. Calculate $\Delta^{(1)} = \mathrm{diag}\{(\Delta^{(0)})^{1/2}\hat{Z}^{(0)}(\Delta^{(0)})^{1/2}\}^{1/2}$.

5. Repeat step 4 with $\Delta^{(0)} = \Delta^{(1)}$ until convergence.

6. With the current values of $\hat{Z}^{(0)}$ and $\Delta^{(0)}$ calculate $\hat{R}_m^{(0)}$, $\hat{R}_{um}^{(0)}$, $\hat{R}_{sm}^{(0)}$ (if necessary), and $\hat{R}^{(0)}$.

7. Calculate $V_i^{(0)}$ using the current $A_i^{(0)}$ and $\hat{R}^{(0)}$.

8. Compute the updated value

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \left[\sum_{i\in s}\left\{\frac{\partial \boldsymbol{p}_i'}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} \left(V_i^{(0)}\right)^{-1} W_i \frac{\partial \boldsymbol{p}_i}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}}\right\}\right]^{-1} \left[\sum_{i\in s}\left\{\frac{\partial \boldsymbol{p}_i'}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} \left(V_i^{(0)}\right)^{-1} W_i \left(\boldsymbol{y}_i - \boldsymbol{p}_i^{(0)}\right)\right\}\right].$$

9. Stop if $\boldsymbol{\beta}^{(1)} \approx \boldsymbol{\beta}^{(0)}$; otherwise repeat steps 2-8, with $\boldsymbol{\beta}^{(0)}$ replaced by $\boldsymbol{\beta}^{(1)}$.

# 3 Application to the SDR

In this section we illustrate how the general methodology proposed by Carrillo and Karr (2013), and laid out in the previous section for the case of multinomial responses, applies to a real life example. In that paper, the authors only present an application example for a continuous response, namely (log of) salary, but do not show how the method applies to multinomial responses. We briefly describe the survey of interest, the response variable as well as the covariates, and then we discuss the results of the model fitting. Finally we show how the goodness of fit can be ascertained.

The Survey of Doctorate Recipients (SDR) is a National Science Foundation (NSF) longitudinal survey whose design incorporates features of both repeated panels and rotating panels. The purpose of the survey is to study U.S. doctorate recipients in science, engineering, and health fields. It is conducted approximately every two years; in this paper we restrict our attention to the data collected from 1995 through 2008 (7 waves). Subjects are in scope until the age of 75, while living in the U.S. during the survey reference week, and while not institutionalized. A detailed description of the SDR can be found at NSF (2012).

The sampling design of the SDR, which mimics the dynamics of the finite population, is depicted in Figure 1. 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/`). *Not* all the sampled graduates are retained in the SDR sample 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. For details see Carrillo and Karr (2013) or NSF (2012).

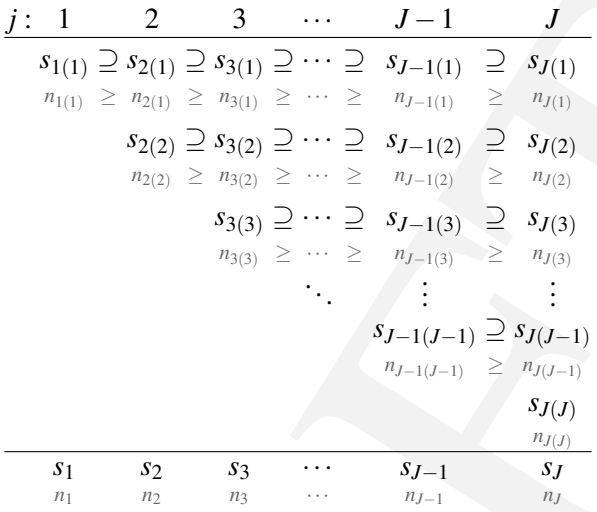


Figure 1: SDR Sample

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 $U_1$ (i.e., Ph.D. holders, either recent or not, who satisfy the requirements of the SDR at wave 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.

At the second wave, the elements in $s_{1(1)}$ who are no longer in scope are simply dropped from the frame (though their observations at wave 1 are kept), and a subsample $s_{2(1)}$, of size $n_{2(1)}$, of those still in scope is selected. Not all the members in $s_{1(1)}$ who are still in scope at wave 2 are retained in the sample. This is in order to be able to make up room for the sample of the new Ph.D. recipients and still maintain more or less the same sample size as in wave 1. A sample $s_{2(2)}$ of size $n_{2(2)}$ is selected from $U_{2(2)}$ (i.e., recent in-scope Ph.D. recipients, who have obtained their degree since wave 1); 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, until the last one ($J$), where a subsample of the remaining subjects from each of the previous $J-1$ cohorts is selected, and a new sample (the new cohort) $s_{J(J)}$ of recent graduates is selected from $U_{J(J)}$. At the last wave, all people in $s_J = \bigcup_{j'=1}^{J} s_{J(j')}$ are interviewed and a design weight $w_{iJ} = 1/\pi_{iJ}$ is created for each person interviewed, so that $s_J$ represents the finite population

$U_J$.

We notice that the SDR sampling scheme is a special case of the general multi-cohort sampling mechanisms described in Section 2.1. The survey weights $w_{ij}$ for cross-sectional analyses of the SDR are available and can be used in equation (2).

Reflecting NSF and societal interest in labor force mobility, the response variable we study is *employment sector*, with seven response categories: "tenured in academics (AcadT)," "non-tenured in academics (AcadNT)," "academics with tenure not applicable (AcadTNA)," "government (Gov)," "business/industry (Bus)," "self-employed (SelfEmpl)," and "unemployed or not in the labor force (Unempl)." The covariates in our model are: age, age squared, and age cubed, all centered at 45 years of age; time in years since 1995; number of years since receiving the doctorate degree (centered at 15 years); indicator for US citizenship; gender; race/ethnicity (with categories non-Hispanic white, Black, Hispanic, Asian, and other/multi-race); marital status/children living at home (married or in a marriage-like relationship with children at home, married or in a marriage-like relationship with no children at home, single or not in a marriage-like relationship with children at home, and single or not in a marriage-like relationship with no children at home); and field of degree (biological sciences, computing and information sciences, mathematics and statistics, physical and astronomical sciences, psychology, social sciences, engineering, and health). The covariates that we considered were suggested either by exploratory analyses or by subject matter experts at the NCSES.

In this analysis, we are asking the question "To what extent, and in what ways, do these covariates affect cross-sector mobility for Ph.D. holders?" Note that this analysis is not able to "see" intra-sector mobility, such as a move from a tenured position at one university to a tenured position at another university. And in the opposite direction, some seeming changes, such as receiving tenure and remaining at the same institution, appear in our analysis as a form of "faux mobility."

The SDR dataset consists of $61,559$ subjects from seven cohorts and $214,103$ observations across seven waves (1995, 1997, 1999, 2001, 2003, 2006, and 2008), distributed as: $n_{95} = 33,571$, $n_{97} = 33,240$, $n_{99} = 29,985$, $n_{01} = 30,115$, $n_{03} = 28,940$, $n_{06} = 29,450$, and $n_{08} = 28,802$. Those data correspond to subjects with non-missing response variable or covariates (whenever they *are* in the sample), and whose position is not a postdoctoral one. The average cross-sectional survey weight for each of those waves are: $\bar{w}_{95} = 15.48$, $\bar{w}_{97} = 16.74$, $\bar{w}_{99} = 20.10$, $\bar{w}_{01} = 21.08$, $\bar{w}_{03} = 23.00$, $\bar{w}_{06} = 23.15$, and $\bar{w}_{08} = 25.16$.

### 3.1 Results for the SDR

In Table 1, we present the parameter estimates, along with the standard errors, 95% confidence intervals, and $p$-values, for model (1) fitted to the SDR dataset. The reference categories of the response variable (tenured in academics) and of the categorical covariates (US citizen, male, non-Hispanic white, married or in a marriage-like relationship with children at home, and biological sciences) are not included in the table as they are set to zero. The estimated autocorrelation matrix $\hat{R}$ appears in the appendix.

Table 1: Parameter Estimates

| Parameter | Response | Estimate | Std.Err | LL | UL | Pr(> \|t\|) | |
|---|---|---|---|---|---|---|---|
| (intercept) | AcadNT | -1.01807 | 0.03804 | -1.09262 | -0.94352 | <0.001 | ** |
| | AcadTNA | -1.26781 | 0.04117 | -1.34851 | -1.18712 | <0.001 | ** |
| | Gov | -0.92726 | 0.04194 | -1.00945 | -0.84506 | <0.001 | ** |
| | Bus | 0.12239 | 0.02990 | 0.06378 | 0.18100 | <0.001 | ** |
| | SelfEmpl | -2.46204 | 0.05988 | -2.57940 | -2.34469 | <0.001 | ** |
| | Unempl | -2.13701 | 0.04223 | -2.21978 | -2.05425 | <0.001 | ** |
| Age-45 | AcadNT | -0.04296 | 0.00371 | -0.05022 | -0.03569 | <0.001 | ** |
| | AcadTNA | 0.00710 | 0.00381 | -0.00036 | 0.01456 | 0.062 | |
| | Gov | 0.02984 | 0.00405 | 0.02191 | 0.03777 | <0.001 | ** |
| | Bus | -0.03496 | 0.00345 | -0.04171 | -0.02821 | <0.001 | ** |
| | SelfEmpl | 0.02718 | 0.00469 | 0.01800 | 0.03637 | <0.001 | ** |
| | Unempl | 0.03258 | 0.00396 | 0.02482 | 0.04034 | <0.001 | ** |
| $(\text{Age-45})^2$ | AcadNT | 0.00445 | 0.00021 | 0.00403 | 0.00486 | <0.001 | ** |
| | AcadTNA | 0.00491 | 0.00023 | 0.00446 | 0.00535 | <0.001 | ** |
| | Gov | 0.00355 | 0.00021 | 0.00313 | 0.00397 | <0.001 | ** |
| | Bus | 0.00374 | 0.00018 | 0.00338 | 0.00411 | <0.001 | ** |
| | SelfEmpl | 0.00248 | 0.00030 | 0.00188 | 0.00307 | <0.001 | ** |
| | Unempl | 0.00971 | 0.00021 | 0.00930 | 0.01012 | <0.001 | ** |
| $(\text{Age-45})^3$ | AcadNT | 0.00004 | 0.00001 | 0.00002 | 0.00006 | <0.001 | ** |
| | AcadTNA | -0.00004 | 0.00001 | -0.00006 | -0.00002 | <0.001 | ** |
| | Gov | -0.00009 | 0.00001 | -0.00011 | -0.00007 | <0.001 | ** |
| | Bus | -0.00006 | 0.00001 | -0.00007 | -0.00004 | <0.001 | ** |
| | SelfEmpl | 0.00002 | 0.00001 | 0.00000 | 0.00004 | 0.063 | |
| | Unempl | -0.00012 | 0.00001 | -0.00013 | -0.00010 | <0.001 | ** |
| Years since 1995 | AcadNT | 0.02269 | 0.00262 | 0.01755 | 0.02783 | <0.001 | ** |
| | AcadTNA | 0.01195 | 0.00281 | 0.00643 | 0.01747 | <0.001 | ** |
| | Gov | 0.00916 | 0.00260 | 0.00407 | 0.01425 | <0.001 | ** |
| | Bus | 0.02007 | 0.00209 | 0.01597 | 0.02417 | <0.001 | ** |
| | SelfEmpl | 0.00876 | 0.00318 | 0.00252 | 0.01500 | 0.006 | ** |
| | Unempl | -0.01528 | 0.00259 | -0.02035 | -0.01022 | <0.001 | ** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Years since degree - 15 | | AcadNT | -0.13430 | 0.00316 | -0.14049 | -0.12811 | <0.001 | ** |
| | | AcadTNA | -0.10000 | 0.00350 | -0.10685 | -0.09315 | <0.001 | ** |
| | | Gov | -0.08574 | 0.00345 | -0.09251 | -0.07898 | <0.001 | ** |
| | | Bus | -0.04354 | 0.00314 | -0.04970 | -0.03738 | <0.001 | ** |
| | | SelfEmpl | -0.04587 | 0.00397 | -0.05364 | -0.03809 | <0.001 | ** |
| | | Unempl | -0.04907 | 0.00333 | -0.05560 | -0.04254 | <0.001 | ** |
| US citizen | No | AcadNT | 0.07641 | 0.04437 | -0.01056 | 0.16338 | 0.085 | |
| | | AcadTNA | 0.09717 | 0.05666 | -0.01388 | 0.20823 | 0.086 | |
| | | Gov | -0.71015 | 0.06418 | -0.83593 | -0.58437 | <0.001 | ** |
| | | Bus | 0.04040 | 0.03833 | -0.03472 | 0.11552 | 0.292 | |
| | | SelfEmpl | -0.12151 | 0.09677 | -0.31118 | 0.06816 | 0.209 | |
| | | Unempl | 0.03649 | 0.06495 | -0.09080 | 0.16378 | 0.574 | |
| Gender | F | AcadNT | 0.23083 | 0.03709 | 0.15814 | 0.30352 | <0.001 | ** |
| | | AcadTNA | 0.40114 | 0.04381 | 0.31528 | 0.48700 | <0.001 | ** |
| | | Gov | 0.01821 | 0.04537 | -0.07072 | 0.10714 | 0.688 | |
| | | Bus | -0.08394 | 0.03512 | -0.15276 | -0.01511 | 0.017 | * |
| | | SelfEmpl | 0.61959 | 0.05667 | 0.50851 | 0.73067 | <0.001 | ** |
| | | Unempl | 0.86145 | 0.04197 | 0.77918 | 0.94372 | <0.001 | ** |
| Race / Ethnicity | Black | AcadNT | -0.10356 | 0.07454 | -0.24964 | 0.04253 | 0.165 | |
| | | AcadTNA | -0.13015 | 0.09268 | -0.31180 | 0.05149 | 0.160 | |
| | | Gov | -0.08207 | 0.08821 | -0.25495 | 0.09082 | 0.352 | |
| | | Bus | -0.37422 | 0.07462 | -0.52048 | -0.22796 | <0.001 | ** |
| | | SelfEmpl | -0.95780 | 0.13388 | -1.22020 | -0.69539 | <0.001 | ** |
| | | Unempl | -0.62594 | 0.09227 | -0.80678 | -0.44510 | <0.001 | ** |
| | Hispa | AcadNT | -0.09656 | 0.07718 | -0.24783 | 0.05472 | 0.211 | |
| | | AcadTNA | -0.17896 | 0.09386 | -0.36293 | 0.00501 | 0.057 | |
| | | Gov | -0.18919 | 0.09347 | -0.37239 | -0.00599 | 0.043 | * |
| | | Bus | -0.26710 | 0.07448 | -0.41308 | -0.12112 | <0.001 | ** |
| | | SelfEmpl | -0.37095 | 0.12641 | -0.61871 | -0.12318 | 0.003 | ** |
| | | Unempl | -0.40395 | 0.09761 | -0.59526 | -0.21264 | <0.001 | ** |
| | Asian | AcadNT | -0.04460 | 0.05188 | -0.14627 | 0.05708 | 0.390 | |
| | | AcadTNA | 0.06682 | 0.05768 | -0.04622 | 0.17986 | 0.247 | |
| | | Gov | 0.10139 | 0.06442 | -0.02488 | 0.22767 | 0.116 | |
| | | Bus | 0.59569 | 0.04646 | 0.50463 | 0.68675 | <0.001 | ** |
| | | SelfEmpl | -0.20763 | 0.08668 | -0.37753 | -0.03773 | 0.017 | * |
| | | Unempl | 0.25760 | 0.05804 | 0.14384 | 0.37137 | <0.001 | ** |
| | Other | AcadNT | -0.04728 | 0.12493 | -0.29214 | 0.19758 | 0.705 | |
| | | AcadTNA | -0.05319 | 0.13485 | -0.31749 | 0.21111 | 0.693 | |
| | | Gov | 0.03621 | 0.15722 | -0.27193 | 0.34435 | 0.818 | |
| | | Bus | 0.05609 | 0.11868 | -0.17652 | 0.28869 | 0.637 | |
| | | SelfEmpl | -0.07491 | 0.16410 | -0.39655 | 0.24672 | 0.648 | |
| | | Unempl | -0.14102 | 0.16015 | -0.45491 | 0.17286 | 0.379 | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MarStat, ChildLivi | MarrLike, NO children | ENT | 0.10541 | 0.02631 | 0.05384 | 0.15699 | <0.001 | ** |
| | | AcadTNA | 0.07533 | 0.02999 | 0.01654 | 0.13412 | 0.012 | * |
| | | Gov | 0.03863 | 0.02757 | -0.01541 | 0.09266 | 0.161 | |
| | | Bus | 0.00988 | 0.01900 | -0.02736 | 0.04712 | 0.603 | |
| | | SelfEmpl | 0.03024 | 0.03180 | -0.03209 | 0.09258 | 0.342 | |
| | | Unempl | 0.08276 | 0.02935 | 0.02524 | 0.14029 | 0.005 | ** |
| | SingLike, NO children | AcadNT | 0.23034 | 0.03446 | 0.16279 | 0.29789 | <0.001 | ** |
| | | AcadTNA | 0.24898 | 0.03973 | 0.17111 | 0.32686 | <0.001 | ** |
| | | Gov | 0.22326 | 0.03953 | 0.14579 | 0.30074 | <0.001 | ** |
| | | Bus | 0.08589 | 0.02827 | 0.03049 | 0.14130 | 0.002 | ** |
| | | SelfEmpl | 0.17895 | 0.05276 | 0.07554 | 0.28237 | <0.001 | ** |
| | | Unempl | 0.13788 | 0.04331 | 0.05299 | 0.22278 | 0.001 | ** |
| | SingLike, Children YES | AcadNT | 0.06347 | 0.05690 | -0.04806 | 0.17499 | 0.265 | |
| | | AcadTNA | 0.10304 | 0.06922 | -0.03264 | 0.23872 | 0.137 | |
| | | Gov | 0.05608 | 0.04647 | -0.03499 | 0.14716 | 0.227 | |
| | | Bus | 0.03314 | 0.03352 | -0.03255 | 0.09883 | 0.323 | |
| | | SelfEmpl | 0.08689 | 0.06837 | -0.04711 | 0.22089 | 0.204 | |
| | | Unempl | -0.17128 | 0.08212 | -0.33224 | -0.01032 | 0.037 | * |
| Degree field | CompInfo | AcadNT | -0.63432 | 0.09768 | -0.82577 | -0.44287 | <0.001 | ** |
| | | AcadTNA | -1.17312 | 0.15323 | -1.47345 | -0.87280 | <0.001 | ** |
| | | Gov | -1.43016 | 0.17706 | -1.77719 | -1.08313 | <0.001 | ** |
| | | Bus | -0.03226 | 0.09856 | -0.22544 | 0.16091 | 0.743 | |
| | | SelfEmpl | -0.37146 | 0.21181 | -0.78660 | 0.04368 | 0.079 | |
| | | Unempl | -0.83341 | 0.15297 | -1.13323 | -0.53358 | <0.001 | ** |
| | MatheSci | AcadNT | -0.56553 | 0.06173 | -0.68651 | -0.44454 | <0.001 | ** |
| | | AcadTNA | -1.08257 | 0.08331 | -1.24585 | -0.91929 | <0.001 | ** |
| | | Gov | -1.17355 | 0.10560 | -1.38053 | -0.96657 | <0.001 | ** |
| | | Bus | -0.70582 | 0.06012 | -0.82366 | -0.58798 | <0.001 | ** |
| | | SelfEmpl | -0.75895 | 0.13058 | -1.01488 | -0.50302 | <0.001 | ** |
| | | Unempl | -0.70505 | 0.07290 | -0.84792 | -0.56218 | <0.001 | ** |
| | PhysicalSc | AcadNT | -0.13701 | 0.05000 | -0.23501 | -0.03901 | 0.006 | ** |
| | | AcadTNA | 0.08017 | 0.05459 | -0.02683 | 0.18717 | 0.142 | |
| | | Gov | 0.25362 | 0.05184 | 0.15201 | 0.35523 | <0.001 | ** |
| | | Bus | 0.71979 | 0.04055 | 0.64031 | 0.79926 | <0.001 | ** |
| | | SelfEmpl | 0.27719 | 0.07711 | 0.12606 | 0.42831 | <0.001 | ** |
| | | Unempl | 0.32687 | 0.04765 | 0.23348 | 0.42026 | <0.001 | ** |
| | Psychology | AcadNT | -0.38277 | 0.05080 | -0.48234 | -0.28319 | <0.001 | ** |
| | | AcadTNA | -0.11102 | 0.05631 | -0.22138 | -0.00066 | 0.049 | * |
| | | Gov | 0.03431 | 0.05961 | -0.08253 | 0.15115 | 0.565 | |
| | | Bus | 0.16309 | 0.04729 | 0.07041 | 0.25576 | <0.001 | ** |
| | | SelfEmpl | 1.80476 | 0.06415 | 1.67903 | 1.93049 | <0.001 | ** |
| | | Unempl | -0.20597 | 0.05609 | -0.31591 | -0.09602 | <0.001 | ** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SocialSci | AcadNT | -0.45801 | 0.04672 | -0.54959 | -0.36644 | <0.001 | ** |
| | AcadTNA | -0.85161 | 0.05749 | -0.96428 | -0.73894 | <0.001 | ** |
| | Gov | -0.70182 | 0.06323 | -0.82575 | -0.57790 | <0.001 | ** |
| | Bus | -0.98835 | 0.05302 | -1.09226 | -0.88444 | <0.001 | ** |
| | SelfEmpl | -0.37867 | 0.08487 | -0.54501 | -0.21233 | <0.001 | ** |
| | Unempl | -0.82988 | 0.05539 | -0.93844 | -0.72133 | <0.001 | ** |
| Engineering | AcadNT | -0.54673 | 0.05526 | -0.65503 | -0.43843 | <0.001 | ** |
| | AcadTNA | -0.54109 | 0.07088 | -0.68001 | -0.40217 | <0.001 | ** |
| | Gov | -0.01633 | 0.06282 | -0.13946 | 0.10681 | 0.795 | |
| | Bus | 0.79038 | 0.04721 | 0.69786 | 0.88290 | <0.001 | ** |
| | SelfEmpl | 0.48493 | 0.08373 | 0.32081 | 0.64904 | <0.001 | ** |
| | Unempl | 0.11307 | 0.05908 | -0.00272 | 0.22886 | 0.056 | |
| Health | AcadNT | -0.05731 | 0.06657 | -0.18779 | 0.07317 | 0.389 | |
| | AcadTNA | -0.59408 | 0.07905 | -0.74902 | -0.43914 | <0.001 | ** |
| | Gov | -0.45342 | 0.08617 | -0.62232 | -0.28452 | <0.001 | ** |
| | Bus | -0.17005 | 0.06295 | -0.29343 | -0.04667 | 0.007 | ** |
| | SelfEmpl | -0.18220 | 0.10826 | -0.39438 | 0.02999 | 0.092 | |
| | Unempl | -0.67445 | 0.08148 | -0.83415 | -0.51475 | <0.001 | ** |

Significance codes: '**' 0.01, '*' 0.05

Since the age factor is included in the model as linear, quadratic, and cubic covariates, the interpretation of the corresponding coefficients is not straightforward. Instead, we ought to combine the three covariate effects into a single "Age" effect. Figure 2 shows the effect of age on the odds of a given category as opposed to a tenured position in academics.

Between the ages of 45 and 54, the odds of being in a non-tenured position in academics instead of in a tenured one are less than one; with the lowest odds being at around 49. For the other ages, the opposite happens: the odds are bigger than one and increase as one moves away from 49. For example, for the ages of 35 and 65 the odds of being in a non-tenured position in academics rather than in a tenured one are about three times the odds at the age of 45.

At 45 years of age the odds of being in a tenure not applicable position in academics instead of tenured in academics are lower than at any other age. The odds increase as one moves away from 45; for example, at the ages of 30 and 60 the odds of being in a non-applicable tenure position in academics rather than tenured are about three times the odds at 45.

With respect to the government sector, something different happens. The lowest odds are at around 41; at that age the odds of being tenured in academics instead of in the government are about 6% higher than

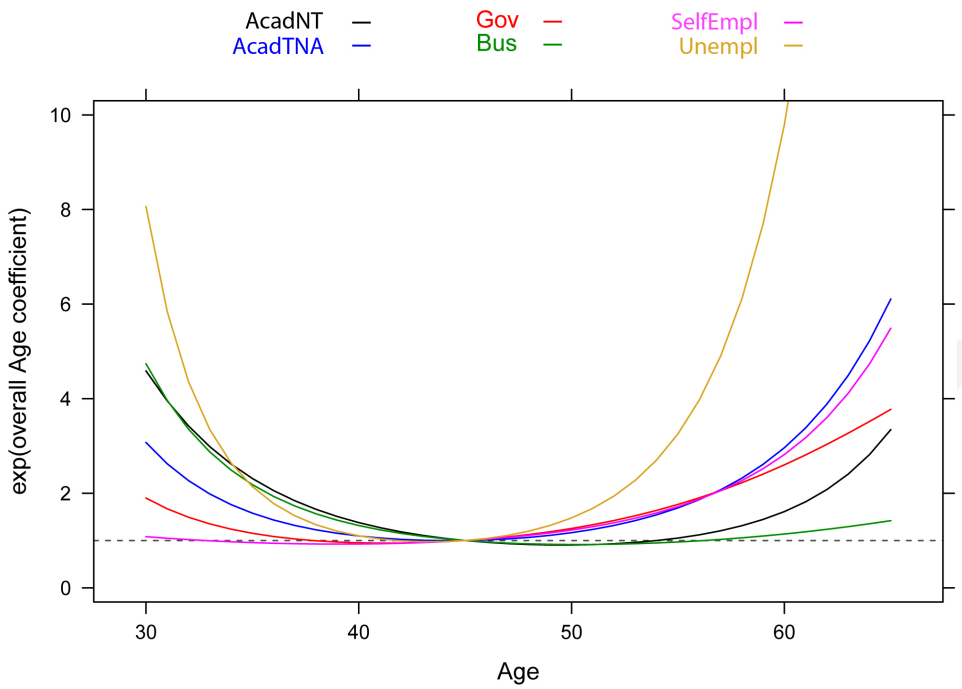


Figure 2: Effect of age on the odds of given category instead of tenured in academics.

at 45 (or at 38 for that matter). Between the ages of 38 and 45 the likelihood of being in the government is lower than that of being in a tenured position in academics. For the other ages, the opposite holds.

We notice that before 45 and after 56 years of age the odds of being in industry instead of tenured in academics are larger than between 45 and 56. The difference in odds is more pronounced the younger the subject is; for example, at 31 years of age the odds of being in industry rather than in a tenured position in academics are four times larger than at 45.

For self-employment, the opposite happens. After 45 years of age the odds of being self-employed instead of tenured in academics are larger than before 45. The difference in odds is more pronounced the older the subject is. For example, at 60 the odds of being self-employed instead of tenured in academics are around three times larger than at 45.

Finally, the odds of being unemployed or not in the labor force rather than tenured in academics are lowest between 40 and 45, with them increasing as one moves away from 43 (when one is the least likely to be unemployed or not in the labor force).

Moving on to the subject of clock time, for every additional year since 1995, the odds of other categories instead of tenured in academics increase by 1–2%. With the unemployed/not in the labor force, the opposite occurs; for every additional year since 1995, the odds of being unemployed versus tenured in academics

(or in any other category, for that matter) decrease. We may interpret this in the following way: as time passes, fewer and fewer people can afford to get by without working. Although the SDR does distinguish being unemployed from not being in the labor force, in our analyses we collapsed the two (for sample sizes purposes). As a result, we may be unable to resolve or interpret some effects. That is, for this population, the unemployment rate reduces with time. On the other hand, as time passes, the odds of being either in industry or non-tenured in academics increase by about 2% every year (compared to the odds of being tenured in academics).

With respect to the experience, the odds of being tenured in academics versus in any other category increase for every year of experience after Ph.D. graduation. The decrease rate is not as marked for industry and self-employed and is the largest for non-tenured in academics. This obviously means that the longer the time since graduation, the less likely it is to find a person in a non-tenured academic position.

Not surprisingly, the only category for which US citizenship is significant is government. The odds of being in a tenured position in academics instead of in a position in the government are twice for non US citizens as they are for citizens.

For women, the odds of being unemployed or not in the labor force instead of tenured in academics are 2.37 times the odds for men. Also, for women the odds of being self-employed instead of tenured in academics are around 86% higher than for their male counterparts. Similarly, the odds for women of being non-tenured in academics or in academics with tenure not applicable instead of tenured are higher than the corresponding odds for men.

The race/ethnicity covariate presents an interesting picture. In the population of interest, i.e., Ph.D. recipients in the fields mentioned, non-Hispanic Whites are *less* likely to be in tenured positions in academics rather than in industry or self-employed, when compared to both Black and Hispanic. However, they are *more* likely to be in tenured positions in academics rather than in industry when compared to Asians.

Note that for single people with no children living at home the odds of being in any other category instead of in a tenured position in academics are between 9% and 28% higher than the odds for married people with children at home. For married people with no children at home the odds of being either in a non-tenured position in academics or unemployed instead of tenured in academics are around 10% higher than the corresponding odds for married people with children at home. (As always, we caution against over-reaching causal interpretations. Children may not impede tenure, but instead people may wait until they receive tenure before having children. Moreover, there is also an age–tenure–children interaction.)

For the degree field covariate we concentrate on the two largest significant effects. The largest effect overall is the effect of having a doctorate degree in psychology on self-employment. For subjects with a degree in psychology the odds of being self-employed instead of in a tenured position in academics are more than six times the odds for people with degrees in biological sciences. Many "practicing" psychologists are, of course, self-employed.

The second largest effect is that of having a degree in computing or information sciences on having a job in the government. For biological science graduates, the odds of having a job in the government instead of being tenured in academics are more than four times the corresponding odds for people with doctorate degrees in computing or information sciences. The other coefficients can be interpreted similarly.

### 3.2 Goodness of Fit

The last step in our analysis is to assess the goodness of fit of the estimated model presented in Table 1. We use a slightly modified version of the GEE score statistic proposed by Horton et al. (1999), which is in turn an "extension of the Hosmer and Lemeshow goodness-of-fit statistic for ordinary logistic regression to marginal regression models for repeated binary responses." Our modification consists of two parts; first, our application requires a modification to survey data, and second, our response is multinomial rather than binary.

The first step is to divide the observations into $G$ (=10 in our case) weighted deciles of risk. We use *weighted* deciles, as our application is for data from a complex survey. The weights are the survey weights. All the observations $i, j$ are sorted by the sum of estimated probabilities $\sum_{k=1}^{K-1} \hat{p}_{ijk} = 1 - p_{ij0}$ (as in Fagerland et al., 2008); then (as in Graubard et al., 1997) $n^{(1)}$ observations with the smallest sums are in the first group, $g_1$, where $n^{(1)}$ is chosen so that $\sum_{i,j\in g_1} w_{ij} / \sum_{i\in s}\sum_{j=1}^{J} w_{ij} \approx 1/G$; the $n^{(2)}$ observations with the next smallest sums are in the second group, $g_2$, where $n^{(2)}$ is chosen so that $\sum_{i,j\in g_2} w_{ij} / \sum_{i\in s}\sum_{j=1}^{J} w_{ij} \approx 1/G$; and so on, until group $G$, which will contain the $n^{(G)}$ observations with the largest sums, where $n^{(G)}$ is chosen so that $\sum_{i,j\in g_G} w_{ij} / \sum_{i\in s}\sum_{j=1}^{J} w_{ij} \approx 1/G$.

Mimicking Horton et al. (1999), define the $G-1$ group indicators:

$$\delta_{ijg} = \begin{cases} 1 & \text{if } \hat{p}_{ij} \text{ is in group } g \\ 0 & \text{otherwise,} \end{cases}$$

for $g = 1, 2, \ldots, G-1$. A test of the goodness-of-fit of model (1) can be based on testing the hypothesis

$H_0 : \gamma_{1,1} = ... = \gamma_{1,G-1} = \gamma_{2,1} = ... = \gamma_{2,G-1} = \cdots = \gamma_{K-1,1} = ... = \gamma_{K-1,G-1} = 0$ in the alternative model:

$$\log \frac{p_{ijk}}{p_{ij0}} = \boldsymbol{X}'_{ij}\boldsymbol{\beta}_k + \gamma_{k,1}\delta_{ij1} + \cdots + \gamma_{k,G-1}\delta_{ij(G-1)}, \qquad k = 1,2,\ldots,K-1.$$

The GEE score statistic for testing $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ is

$$X^2 = \boldsymbol{u}_2(\hat{\boldsymbol{\beta}},\mathbf{0})'\big\{\widehat{\mathrm{Var}}[\boldsymbol{u}_2(\hat{\boldsymbol{\beta}},\mathbf{0})]\big\}^{-1}\boldsymbol{u}_2(\hat{\boldsymbol{\beta}},\mathbf{0}),$$

where

$$\boldsymbol{u}_2(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i\in s}\frac{\partial \boldsymbol{p}_i(\boldsymbol{\beta},\boldsymbol{\gamma})'}{\partial \boldsymbol{\gamma}}V_i^{-1}W_i\big(\boldsymbol{y}_i - \boldsymbol{p}_i(\boldsymbol{\beta},\boldsymbol{\gamma})\big),$$

$$\frac{\partial \boldsymbol{p}_i(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = \{\mathbf{1}'_{(K-1)} \otimes \boldsymbol{p}_i\} \circ \Big\{\Delta_i - \Big[\Big(\big\{I_J \otimes \mathbf{1}'_{(K-1)}\big\}\big\{\big[\mathbf{1}'_{(K-1)} \otimes \boldsymbol{p}_i\big] \circ \Delta_i\big\}\Big) \otimes \mathbf{1}_{(K-1)}\Big]\Big\},$$

$\Delta_i = \big(\Delta'_{i1}, \Delta'_{i2}, \ldots, \Delta'_{iJ}\big)'$, $\Delta_{ij} = I_{(K-1)} \otimes \boldsymbol{\Delta}'_{ij}$, and $\boldsymbol{\Delta}_{ij} = (\delta_{ij1}, \delta_{ij2}, \ldots, \delta_{ij(G-1)})'$. Obviously, to calculate $\widehat{\mathrm{Var}}[\boldsymbol{u}_2(\hat{\boldsymbol{\beta}},\mathbf{0})]$ we can use the same method used to estimate $\mathrm{Var}[\Psi_s(\boldsymbol{\beta}_N)]$ (the estimating function bootstrap in our case). Horton et al. (1999) conclude by saying that "a signicant GEE score statistic indicates that the proposed model leaves a substantial amount of variability in the data not taken into account."

The value of the GEE score test statistic for the multinomial model for sector, in Table 1, is $X^2 = 24.97$. To calculate the $p$-value we follow the procedure used in (Graubard et al., 1997). They use a simulated Wald test, but since we are using the score statistic, our procedure is a "simulated score test."

As in Graubard et al. (1997), the algorithm for calculating the p-value for the simulated score test is as follows: (1) fit model (1) to the true dataset, and calculate the score statistic $X^2$ for that model; (2) generate 999 synthetic datasets by generating for each subject 999 response vectors based on the estimated model in Table 1, along with the autocorrelation matrix in the appendix and his/her own set of covariates; (3) fit model (1) to each of the 999 synthetic datasets taking into account the original survey design characteristics, and also compute the score statistic for each of them (call it $X^2_{(\mathrm{syn})}$); (4) calculate the $p$-value as:

$$\frac{1 + \sum_1^{999}\{X^2_{(\mathrm{syn})} \geq X^2\}}{1000}.$$

In our process, steps (1), (3), and (4) are straightforward; the time-consuming part is step (2). For each subject in the dataset we have to, 999 times, generate a set of *correlated* multinomial responses; for example, for a subject who is part of the survey in the waves 1997, 1999, and 2003, we have to simulate a 3-dimensional 7-category response variable based on his/her (time-dependent) covariates and the appropriate rows and columns from the autocorrelation matrix in the appendix (the 21 rows and 21 columns corresponding to the years 1997, 1999, and 2003 in this case). The method used for generation is the one proposed

by Song (2000), which is a method for generating multivariate dispersion data based on the multivariate Gaussian copula; multinomial variables are a special case of exponential family, which in turn is a special case of dispersion models. Song's method requires computing, for each subject $i$, a $J_i$-dimensional normal distribution function with standardized margins (where $J_i$ is the number of waves that subject $i$ is part of the survey). We used the R package "OpenMx" (see Boker et al., 2011, 2012) for calculating the multivariate normal distribution function.

The results of fitting model (1) to the 999 synthetic datasets and computing the $X^2_{(syn)}$'s yielded 101 values larger than $X^2 = 24.97$; which produces a simulated $p$-value for the score test of $(1+101)/1,000 = 0.102$. Since the $p$-value is larger than 0.05, we cannot reject the null hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$. Therefore, we conclude that the estimated model in Table 1 is a good fit to the data.

# 4 Final Remarks

We have shown how we can use the methodology proposed by Carrillo and Karr (2013) to estimate the effect of covariates on a multinomial response using the data from a survey with multiple cohorts. The only requirement is that there exists a cross-sectional survey weight for each subject, at each wave that the subject is part of the survey, to represent to cross-sectional population of interest at that wave. We showed the details for estimation of effects and variance of the parameter estimates, either by the estimating function bootstrap (as in our case) or by any other of the methods available in the literature.

We also argue why estimating the autocorrelation matrix by the method of quasi-least squares is more appropriate for binomial or multinomial responses and give a general algorithm for point estimation of the parameters of interest by using this approach.

After arguing why the described approach is suitable for application to the SDR, we present the results for a model explaining employment sector as a function of several demographic, time, experience, and degree field variables. Additionally, we give extensive interpretation of the significant effects in the estimated model.

Finally, by using a simulated score test, whose procedure for general application we describe in detail, we argue that the estimated model fits the data well.

# Appendix: Estimated Autocorrelation Matrix

| | | 1995 | | | | | | 1997 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl |
| 1995 | AcadNT | 1 | -0.097 | -0.100 | -0.234 | -0.063 | -0.091 | 0.394 | 0.050 | -0.069 | -0.151 | -0.029 | -0.032 |
| | AcadTNA | -0.097 | 1 | -0.085 | -0.194 | -0.057 | -0.091 | 0.061 | 0.387 | 0.008 | -0.103 | -0.028 | -0.029 |
| | Gov | -0.100 | -0.085 | 1 | -0.212 | -0.070 | -0.091 | -0.052 | -0.027 | 0.596 | -0.134 | -0.033 | -0.038 |
| | Bus | -0.234 | -0.194 | -0.212 | 1 | -0.144 | -0.199 | -0.128 | -0.089 | -0.135 | 0.558 | -0.011 | -0.052 |
| | SelfEmpl | -0.063 | -0.057 | -0.070 | -0.144 | 1 | -0.081 | -0.032 | -0.025 | -0.039 | -0.003 | 0.372 | -0.022 |
| | Unempl | -0.091 | -0.091 | -0.091 | -0.199 | -0.081 | 1 | -0.034 | -0.032 | -0.050 | -0.059 | -0.002 | 0.362 |
| 1997 | AcadNT | 0.394 | 0.061 | -0.052 | -0.128 | -0.032 | -0.034 | 1 | -0.096 | -0.098 | -0.236 | -0.057 | -0.093 |
| | AcadTNA | 0.050 | 0.387 | -0.027 | -0.089 | -0.025 | -0.032 | -0.096 | 1 | -0.081 | -0.183 | -0.053 | -0.090 |
| | Gov | -0.069 | 0.008 | 0.596 | -0.135 | -0.039 | -0.050 | -0.098 | -0.081 | 1 | -0.225 | -0.064 | -0.094 |
| | Bus | -0.151 | -0.103 | -0.134 | 0.558 | -0.003 | -0.059 | -0.236 | -0.183 | -0.225 | 1 | -0.139 | -0.201 |
| | SelfEmpl | -0.029 | -0.028 | -0.033 | -0.011 | 0.372 | -0.002 | -0.057 | -0.053 | -0.064 | -0.139 | 1 | -0.079 |
| | Unempl | -0.032 | -0.029 | -0.038 | -0.052 | -0.022 | 0.362 | -0.093 | -0.090 | -0.094 | -0.201 | -0.079 | 1 |
| 1999 | AcadNT | 0.258 | 0.065 | -0.035 | -0.085 | -0.018 | -0.006 | 0.405 | 0.049 | -0.052 | -0.121 | -0.028 | -0.029 |
| | AcadTNA | 0.028 | 0.319 | -0.027 | -0.066 | -0.026 | -0.011 | 0.050 | 0.410 | -0.032 | -0.095 | -0.025 | -0.019 |
| | Gov | -0.051 | -0.002 | 0.489 | -0.112 | -0.035 | -0.029 | -0.053 | -0.035 | 0.557 | -0.123 | -0.029 | -0.041 |
| | Bus | -0.117 | -0.082 | -0.104 | 0.452 | -0.004 | -0.042 | -0.136 | -0.090 | -0.119 | 0.527 | -0.016 | -0.067 |
| | SelfEmpl | -0.026 | -0.016 | -0.021 | 0.001 | 0.274 | -0.005 | -0.036 | -0.024 | -0.036 | -0.008 | 0.337 | -0.002 |
| | Unempl | -0.026 | -0.021 | -0.019 | -0.024 | 0.007 | 0.226 | -0.039 | -0.034 | -0.034 | -0.046 | -0.013 | 0.337 |
| 2001 | AcadNT | 0.174 | 0.075 | -0.023 | -0.057 | -0.012 | -0.009 | 0.253 | 0.071 | -0.032 | -0.082 | -0.018 | -0.024 |
| | AcadTNA | 0.028 | 0.260 | -0.023 | -0.063 | -0.016 | -0.006 | 0.053 | 0.306 | -0.020 | -0.083 | -0.023 | -0.008 |
| | Gov | -0.045 | 0.001 | 0.414 | -0.097 | -0.029 | -0.019 | -0.044 | -0.022 | 0.473 | -0.108 | -0.026 | -0.033 |
| | Bus | -0.104 | -0.063 | -0.087 | 0.383 | -0.003 | -0.032 | -0.111 | -0.072 | -0.103 | 0.438 | -0.010 | -0.046 |
| | SelfEmpl | -0.022 | -0.020 | -0.022 | 0.007 | 0.220 | 0.003 | -0.032 | -0.019 | -0.028 | 0.001 | 0.281 | -0.011 |
| | Unempl | -0.022 | -0.019 | -0.014 | 0.002 | 0.002 | 0.149 | -0.033 | -0.022 | -0.025 | -0.014 | -0.016 | 0.224 |
| 2003 | AcadNT | 0.120 | 0.056 | -0.016 | -0.042 | -0.009 | -0.002 | 0.158 | 0.068 | -0.024 | -0.055 | -0.016 | -0.009 |
| | AcadTNA | 0.027 | 0.200 | -0.027 | -0.047 | -0.012 | -0.011 | 0.042 | 0.224 | -0.023 | -0.063 | -0.010 | -0.006 |
| | Gov | -0.034 | 0.004 | 0.342 | -0.081 | -0.023 | -0.014 | -0.037 | -0.014 | 0.383 | -0.089 | -0.020 | -0.022 |
| | Bus | -0.085 | -0.050 | -0.066 | 0.315 | -0.009 | -0.014 | -0.094 | -0.051 | -0.082 | 0.362 | -0.021 | -0.028 |
| | SelfEmpl | -0.025 | -0.021 | -0.013 | 0.011 | 0.179 | 0.007 | -0.032 | -0.020 | -0.019 | 0.009 | 0.226 | 0.008 |
| | Unempl | -0.028 | -0.008 | -0.009 | 0.018 | -0.003 | 0.091 | -0.031 | -0.012 | -0.012 | 0.005 | -0.008 | 0.135 |
| 2006 | AcadNT | 0.094 | 0.060 | -0.020 | -0.036 | -0.012 | -0.009 | 0.119 | 0.070 | -0.023 | -0.044 | -0.013 | -0.014 |
| | AcadTNA | 0.021 | 0.120 | -0.013 | -0.040 | -0.007 | -0.001 | 0.042 | 0.140 | -0.020 | -0.052 | -0.012 | 0.008 |
| | Gov | -0.029 | 0.002 | 0.284 | -0.070 | -0.018 | -0.009 | -0.026 | -0.007 | 0.308 | -0.079 | -0.015 | -0.012 |
| | Bus | -0.078 | -0.028 | -0.060 | 0.265 | -0.007 | -0.006 | -0.088 | -0.031 | -0.063 | 0.303 | -0.018 | -0.019 |
| | SelfEmpl | -0.026 | -0.014 | -0.006 | 0.018 | 0.140 | 0.005 | -0.032 | -0.016 | -0.011 | 0.016 | 0.192 | -0.005 |
| | Unempl | -0.017 | -0.001 | 0.002 | 0.018 | -0.001 | 0.060 | -0.020 | -0.004 | 0.001 | 0.008 | -0.014 | 0.099 |
| 2008 | AcadNT | 0.068 | 0.047 | -0.010 | -0.028 | -0.009 | -0.003 | 0.103 | 0.035 | -0.015 | -0.032 | -0.011 | -0.001 |
| | AcadTNA | 0.015 | 0.091 | -0.017 | -0.025 | -0.003 | 0.003 | 0.022 | 0.118 | -0.021 | -0.036 | -0.006 | 0.003 |
| | Gov | -0.026 | -0.003 | 0.243 | -0.059 | -0.016 | -0.010 | -0.024 | -0.010 | 0.258 | -0.065 | -0.013 | -0.014 |
| | Bus | -0.069 | -0.020 | -0.051 | 0.221 | -0.006 | -0.005 | -0.076 | -0.022 | -0.059 | 0.254 | -0.014 | -0.009 |
| | SelfEmpl | -0.018 | -0.009 | -0.011 | 0.021 | 0.120 | -0.002 | -0.027 | -0.010 | -0.012 | 0.016 | 0.151 | -0.005 |
| | Unempl | -0.020 | 0.002 | 0.004 | 0.016 | -0.008 | 0.046 | -0.017 | -0.002 | 0.005 | 0.010 | -0.012 | 0.063 |

| | | 1999 | | | | | | 2001 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl |
| 1995 | AcadNT | 0.258 | 0.028 | -0.051 | -0.117 | -0.026 | -0.026 | 0.174 | 0.028 | -0.045 | -0.104 | -0.022 | -0.022 |
| | AcadTNA | 0.065 | 0.319 | -0.002 | -0.082 | -0.016 | -0.021 | 0.075 | 0.260 | 0.001 | -0.063 | -0.020 | -0.019 |
| | Gov | -0.035 | -0.027 | 0.489 | -0.104 | -0.021 | -0.019 | -0.023 | -0.023 | 0.414 | -0.087 | -0.022 | -0.014 |
| | Bus | -0.085 | -0.066 | -0.112 | 0.452 | 0.001 | -0.024 | -0.057 | -0.063 | -0.097 | 0.383 | 0.007 | 0.002 |
| | SelfEmpl | -0.018 | -0.026 | -0.035 | -0.004 | 0.274 | 0.007 | -0.012 | -0.016 | -0.029 | -0.003 | 0.220 | 0.002 |
| | Unempl | -0.006 | -0.011 | -0.029 | -0.042 | -0.005 | 0.226 | -0.009 | -0.006 | -0.019 | -0.032 | 0.003 | 0.149 |
| 1997 | AcadNT | 0.405 | 0.050 | -0.053 | -0.136 | -0.036 | -0.039 | 0.253 | 0.053 | -0.044 | -0.111 | -0.032 | -0.033 |
| | AcadTNA | 0.049 | 0.410 | -0.035 | -0.090 | -0.024 | -0.034 | 0.071 | 0.306 | -0.022 | -0.072 | -0.019 | -0.022 |
| | Gov | -0.052 | -0.032 | 0.557 | -0.119 | -0.036 | -0.034 | -0.032 | -0.020 | 0.473 | -0.103 | -0.028 | -0.025 |
| | Bus | -0.121 | -0.095 | -0.123 | 0.527 | -0.008 | -0.046 | -0.082 | -0.083 | -0.108 | 0.438 | 0.001 | -0.014 |
| | SelfEmpl | -0.028 | -0.025 | -0.029 | -0.016 | 0.337 | -0.013 | -0.018 | -0.023 | -0.026 | -0.010 | 0.281 | -0.016 |
| | Unempl | -0.029 | -0.019 | -0.041 | -0.067 | -0.002 | 0.337 | -0.024 | -0.008 | -0.033 | -0.046 | -0.011 | 0.224 |
| 1999 | AcadNT | 1 | -0.094 | -0.091 | -0.235 | -0.058 | -0.090 | 0.458 | 0.057 | -0.066 | -0.167 | -0.036 | -0.046 |
| | AcadTNA | -0.094 | 1 | -0.081 | -0.194 | -0.057 | -0.093 | 0.080 | 0.481 | -0.035 | -0.117 | -0.032 | -0.041 |
| | Gov | -0.091 | -0.081 | 1 | -0.221 | -0.063 | -0.094 | -0.049 | -0.037 | 0.681 | -0.155 | -0.040 | -0.051 |
| | Bus | -0.235 | -0.194 | -0.221 | 1 | -0.142 | -0.213 | -0.151 | -0.120 | -0.151 | 0.652 | -0.024 | -0.063 |
| | SelfEmpl | -0.058 | -0.057 | -0.063 | -0.142 | 1 | -0.084 | -0.036 | -0.030 | -0.040 | -0.020 | 0.437 | -0.027 |
| | Unempl | -0.090 | -0.093 | -0.094 | -0.213 | -0.084 | 1 | -0.047 | -0.042 | -0.052 | -0.082 | -0.017 | 0.412 |
| 2001 | AcadNT | 0.458 | 0.080 | -0.049 | -0.151 | -0.036 | -0.047 | 1 | -0.091 | -0.092 | -0.237 | -0.057 | -0.093 |
| | AcadTNA | 0.057 | 0.481 | -0.037 | -0.120 | -0.030 | -0.042 | -0.091 | 1 | -0.083 | -0.196 | -0.057 | -0.096 |
| | Gov | -0.066 | -0.035 | 0.681 | -0.151 | -0.040 | -0.052 | -0.092 | -0.083 | 1 | -0.221 | -0.063 | -0.096 |
| | Bus | -0.167 | -0.117 | -0.155 | 0.652 | -0.020 | -0.082 | -0.237 | -0.196 | -0.221 | 1 | -0.141 | -0.222 |
| | SelfEmpl | -0.036 | -0.032 | -0.040 | -0.024 | 0.437 | -0.017 | -0.057 | -0.057 | -0.063 | -0.141 | 1 | -0.086 |
| | Unempl | -0.046 | -0.041 | -0.051 | -0.063 | -0.027 | 0.412 | -0.093 | -0.096 | -0.096 | -0.222 | -0.086 | 1 |
| 2003 | AcadNT | 0.260 | 0.089 | -0.035 | -0.090 | -0.027 | -0.029 | 0.416 | 0.097 | -0.058 | -0.141 | -0.034 | -0.045 |
| | AcadTNA | 0.045 | 0.339 | -0.033 | -0.086 | -0.025 | -0.018 | 0.072 | 0.442 | -0.044 | -0.118 | -0.031 | -0.025 |
| | Gov | -0.058 | -0.020 | 0.551 | -0.124 | -0.029 | -0.041 | -0.067 | -0.029 | 0.685 | -0.161 | -0.043 | -0.052 |
| | Bus | -0.135 | -0.084 | -0.117 | 0.512 | -0.026 | -0.043 | -0.172 | -0.113 | -0.152 | 0.641 | -0.045 | -0.080 |
| | SelfEmpl | -0.037 | -0.029 | -0.034 | -0.004 | 0.362 | -0.014 | -0.045 | -0.033 | -0.037 | -0.021 | 0.479 | -0.021 |
| | Unempl | -0.036 | -0.024 | -0.034 | -0.013 | -0.024 | 0.254 | -0.052 | -0.040 | -0.049 | -0.047 | -0.028 | 0.381 |
| 2006 | AcadNT | 0.190 | 0.095 | -0.030 | -0.073 | -0.021 | -0.010 | 0.295 | 0.102 | -0.040 | -0.104 | -0.027 | -0.023 |
| | AcadTNA | 0.032 | 0.241 | -0.028 | -0.065 | -0.019 | -0.013 | 0.049 | 0.317 | -0.038 | -0.085 | -0.025 | -0.024 |
| | Gov | -0.046 | -0.009 | 0.440 | -0.103 | -0.022 | -0.028 | -0.050 | -0.009 | 0.538 | -0.133 | -0.030 | -0.039 |
| | Bus | -0.119 | -0.053 | -0.088 | 0.433 | -0.030 | -0.037 | -0.152 | -0.073 | -0.112 | 0.528 | -0.035 | -0.056 |
| | SelfEmpl | -0.039 | -0.023 | -0.023 | 0.014 | 0.288 | -0.008 | -0.041 | -0.034 | -0.025 | 0.010 | 0.349 | -0.014 |
| | Unempl | -0.025 | -0.014 | -0.012 | -0.014 | -0.018 | 0.189 | -0.039 | -0.023 | -0.024 | -0.034 | -0.020 | 0.280 |
| 2008 | AcadNT | 0.143 | 0.074 | -0.016 | -0.057 | -0.012 | -0.008 | 0.173 | 0.092 | -0.023 | -0.067 | -0.015 | -0.012 |
| | AcadTNA | 0.037 | 0.168 | -0.030 | -0.048 | -0.012 | 0.005 | 0.048 | 0.234 | -0.035 | -0.065 | -0.019 | -0.010 |
| | Gov | -0.040 | -0.007 | 0.377 | -0.093 | -0.017 | -0.024 | -0.044 | -0.013 | 0.460 | -0.113 | -0.027 | -0.028 |
| | Bus | -0.101 | -0.040 | -0.082 | 0.356 | -0.016 | -0.026 | -0.122 | -0.051 | -0.103 | 0.432 | -0.019 | -0.042 |
| | SelfEmpl | -0.030 | -0.015 | -0.020 | 0.021 | 0.206 | -0.008 | -0.029 | -0.022 | -0.022 | 0.013 | 0.246 | -0.013 |
| | Unempl | -0.026 | -0.005 | -0.005 | 0.002 | -0.018 | 0.128 | -0.031 | -0.012 | -0.012 | -0.013 | -0.011 | 0.190 |

| | | 2003 | | | | | | 2006 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl |
| 1995 | AcadNT | 0.120 | 0.027 | -0.034 | -0.085 | -0.025 | -0.028 | 0.094 | 0.021 | -0.029 | -0.078 | -0.026 | -0.017 |
| | AcadTNA | 0.056 | 0.200 | 0.004 | -0.050 | -0.021 | -0.008 | 0.060 | 0.120 | 0.002 | -0.028 | -0.014 | -0.001 |
| | Gov | -0.016 | -0.027 | 0.342 | -0.066 | -0.013 | -0.009 | -0.020 | -0.013 | 0.284 | -0.060 | -0.006 | 0.002 |
| | Bus | -0.042 | -0.047 | -0.081 | 0.315 | 0.011 | 0.018 | -0.036 | -0.040 | -0.070 | 0.265 | 0.018 | 0.018 |
| | SelfEmpl | -0.009 | -0.012 | -0.023 | -0.009 | 0.179 | -0.003 | -0.012 | -0.007 | -0.018 | -0.007 | 0.140 | -0.001 |
| | Unempl | -0.002 | -0.011 | -0.014 | -0.014 | 0.007 | 0.091 | -0.009 | -0.001 | -0.009 | -0.006 | 0.005 | 0.060 |
| 1997 | AcadNT | 0.158 | 0.042 | -0.037 | -0.094 | -0.032 | -0.031 | 0.119 | 0.042 | -0.026 | -0.088 | -0.032 | -0.020 |
| | AcadTNA | 0.068 | 0.224 | -0.014 | -0.051 | -0.020 | -0.012 | 0.070 | 0.140 | -0.007 | -0.031 | -0.016 | -0.004 |
| | Gov | -0.024 | -0.023 | 0.383 | -0.082 | -0.019 | -0.012 | -0.023 | -0.020 | 0.308 | -0.063 | -0.011 | 0.001 |
| | Bus | -0.055 | -0.063 | -0.089 | 0.362 | 0.009 | 0.005 | -0.044 | -0.052 | -0.079 | 0.303 | 0.016 | 0.008 |
| | SelfEmpl | -0.016 | -0.010 | -0.020 | -0.021 | 0.226 | -0.008 | -0.013 | -0.012 | -0.015 | -0.018 | 0.192 | -0.014 |
| | Unempl | -0.009 | -0.006 | -0.022 | -0.028 | 0.008 | 0.135 | -0.014 | 0.008 | -0.012 | -0.019 | -0.005 | 0.099 |
| 1999 | AcadNT | 0.260 | 0.045 | -0.058 | -0.135 | -0.037 | -0.036 | 0.190 | 0.032 | -0.046 | -0.119 | -0.039 | -0.025 |
| | AcadTNA | 0.089 | 0.339 | -0.020 | -0.084 | -0.029 | -0.024 | 0.095 | 0.241 | -0.009 | -0.053 | -0.023 | -0.014 |
| | Gov | -0.035 | -0.033 | 0.551 | -0.117 | -0.034 | -0.034 | -0.030 | -0.028 | 0.440 | -0.088 | -0.023 | -0.012 |
| | Bus | -0.090 | -0.086 | -0.124 | 0.512 | -0.004 | -0.013 | -0.073 | -0.065 | -0.103 | 0.433 | 0.014 | -0.014 |
| | SelfEmpl | -0.027 | -0.025 | -0.029 | -0.026 | 0.362 | -0.024 | -0.021 | -0.019 | -0.022 | -0.030 | 0.288 | -0.018 |
| | Unempl | -0.029 | -0.018 | -0.041 | -0.043 | -0.014 | 0.254 | -0.010 | -0.013 | -0.028 | -0.037 | -0.008 | 0.189 |
| 2001 | AcadNT | 0.416 | 0.072 | -0.067 | -0.172 | -0.045 | -0.052 | 0.295 | 0.049 | -0.050 | -0.152 | -0.041 | -0.039 |
| | AcadTNA | 0.097 | 0.442 | -0.029 | -0.113 | -0.033 | -0.040 | 0.102 | 0.317 | -0.009 | -0.073 | -0.034 | -0.023 |
| | Gov | -0.058 | -0.044 | 0.685 | -0.152 | -0.037 | -0.049 | -0.040 | -0.038 | 0.538 | -0.112 | -0.025 | -0.024 |
| | Bus | -0.141 | -0.118 | -0.161 | 0.641 | -0.021 | -0.047 | -0.104 | -0.085 | -0.133 | 0.528 | 0.010 | -0.034 |
| | SelfEmpl | -0.034 | -0.031 | -0.043 | -0.045 | 0.479 | -0.028 | -0.027 | -0.025 | -0.030 | -0.035 | 0.349 | -0.020 |
| | Unempl | -0.045 | -0.025 | -0.052 | -0.080 | -0.021 | 0.381 | -0.023 | -0.024 | -0.039 | -0.056 | -0.014 | 0.280 |
| 2003 | AcadNT | 1 | -0.089 | -0.092 | -0.226 | -0.059 | -0.101 | 0.447 | 0.074 | -0.062 | -0.171 | -0.043 | -0.048 |
| | AcadTNA | -0.089 | 1 | -0.084 | -0.196 | -0.061 | -0.097 | 0.111 | 0.410 | -0.021 | -0.102 | -0.036 | -0.037 |
| | Gov | -0.092 | -0.084 | 1 | -0.216 | -0.066 | -0.095 | -0.059 | -0.048 | 0.643 | -0.126 | -0.035 | -0.046 |
| | Bus | -0.226 | -0.196 | -0.216 | 1 | -0.144 | -0.217 | -0.157 | -0.107 | -0.148 | 0.648 | -0.031 | -0.063 |
| | SelfEmpl | -0.059 | -0.061 | -0.066 | -0.144 | 1 | -0.089 | -0.038 | -0.036 | -0.042 | -0.042 | 0.468 | -0.032 |
| | Unempl | -0.101 | -0.097 | -0.095 | -0.217 | -0.089 | 1 | -0.047 | -0.032 | -0.050 | -0.082 | -0.012 | 0.401 |
| 2006 | AcadNT | 0.447 | 0.111 | -0.059 | -0.157 | -0.038 | -0.047 | 1 | -0.090 | -0.095 | -0.238 | -0.062 | -0.101 |
| | AcadTNA | 0.074 | 0.410 | -0.048 | -0.107 | -0.036 | -0.032 | -0.090 | 1 | -0.080 | -0.189 | -0.060 | -0.101 |
| | Gov | -0.062 | -0.021 | 0.643 | -0.148 | -0.042 | -0.050 | -0.095 | -0.080 | 1 | -0.210 | -0.063 | -0.098 |
| | Bus | -0.171 | -0.102 | -0.126 | 0.648 | -0.042 | -0.082 | -0.238 | -0.189 | -0.210 | 1 | -0.146 | -0.218 |
| | SelfEmpl | -0.043 | -0.036 | -0.035 | -0.031 | 0.468 | -0.012 | -0.062 | -0.060 | -0.063 | -0.146 | 1 | -0.098 |
| | Unempl | -0.048 | -0.037 | -0.046 | -0.063 | -0.032 | 0.401 | -0.101 | -0.101 | -0.098 | -0.218 | -0.098 | 1 |
| 2008 | AcadNT | 0.254 | 0.098 | -0.032 | -0.088 | -0.023 | -0.024 | 0.398 | 0.102 | -0.051 | -0.137 | -0.032 | -0.041 |
| | AcadTNA | 0.044 | 0.316 | -0.039 | -0.080 | -0.026 | -0.020 | 0.085 | 0.406 | -0.048 | -0.104 | -0.036 | -0.035 |
| | Gov | -0.052 | -0.016 | 0.534 | -0.121 | -0.035 | -0.045 | -0.071 | -0.043 | 0.635 | -0.126 | -0.041 | -0.052 |
| | Bus | -0.134 | -0.072 | -0.113 | 0.504 | -0.016 | -0.055 | -0.170 | -0.097 | -0.130 | 0.612 | -0.032 | -0.089 |
| | SelfEmpl | -0.033 | -0.019 | -0.024 | -0.005 | 0.291 | -0.005 | -0.035 | -0.030 | -0.039 | -0.032 | 0.404 | -0.020 |
| | Unempl | -0.031 | -0.025 | -0.030 | -0.028 | -0.015 | 0.264 | -0.049 | -0.036 | -0.044 | -0.066 | -0.031 | 0.392 |

| | | 2008 | | | | | |
|---|---|---|---|---|---|---|---|
| | | AcadNT | AcadTNA | Gov | Bus | SelfEmpl | Unempl |
| 1995 | AcadNT | 0.068 | 0.015 | -0.026 | -0.069 | -0.018 | -0.020 |
| | AcadTNA | 0.047 | 0.091 | -0.003 | -0.020 | -0.009 | 0.002 |
| | Gov | -0.010 | -0.017 | 0.243 | -0.051 | -0.011 | 0.004 |
| | Bus | -0.028 | -0.025 | -0.059 | 0.221 | 0.021 | 0.016 |
| | SelfEmpl | -0.009 | -0.003 | -0.016 | -0.006 | 0.120 | -0.008 |
| | Unempl | -0.003 | 0.003 | -0.010 | -0.005 | -0.002 | 0.046 |
| 1997 | AcadNT | 0.103 | 0.022 | -0.024 | -0.076 | -0.027 | -0.017 |
| | AcadTNA | 0.035 | 0.118 | -0.010 | -0.022 | -0.010 | -0.002 |
| | Gov | -0.015 | -0.021 | 0.258 | -0.059 | -0.012 | 0.005 |
| | Bus | -0.032 | -0.036 | -0.065 | 0.254 | 0.016 | 0.010 |
| | SelfEmpl | -0.011 | -0.006 | -0.013 | -0.014 | 0.151 | -0.012 |
| | Unempl | -0.001 | 0.003 | -0.014 | -0.009 | -0.005 | 0.063 |
| 1999 | AcadNT | 0.143 | 0.037 | -0.040 | -0.101 | -0.030 | -0.026 |
| | AcadTNA | 0.074 | 0.168 | -0.007 | -0.040 | -0.015 | -0.005 |
| | Gov | -0.016 | -0.030 | 0.377 | -0.082 | -0.020 | -0.005 |
| | Bus | -0.057 | -0.048 | -0.093 | 0.356 | 0.021 | 0.002 |
| | SelfEmpl | -0.012 | -0.012 | -0.017 | -0.016 | 0.206 | -0.018 |
| | Unempl | -0.008 | 0.005 | -0.024 | -0.026 | -0.008 | 0.128 |
| 2001 | AcadNT | 0.173 | 0.048 | -0.044 | -0.122 | -0.029 | -0.031 |
| | AcadTNA | 0.092 | 0.234 | -0.013 | -0.051 | -0.022 | -0.012 |
| | Gov | -0.023 | -0.035 | 0.460 | -0.103 | -0.022 | -0.012 |
| | Bus | -0.067 | -0.065 | -0.113 | 0.432 | 0.013 | -0.013 |
| | SelfEmpl | -0.015 | -0.019 | -0.027 | -0.019 | 0.246 | -0.011 |
| | Unempl | -0.012 | -0.010 | -0.028 | -0.042 | -0.013 | 0.190 |
| 2003 | AcadNT | 0.254 | 0.044 | -0.052 | -0.134 | -0.033 | -0.031 |
| | AcadTNA | 0.098 | 0.316 | -0.016 | -0.072 | -0.019 | -0.025 |
| | Gov | -0.032 | -0.039 | 0.534 | -0.113 | -0.024 | -0.030 |
| | Bus | -0.088 | -0.080 | -0.121 | 0.504 | -0.005 | -0.028 |
| | SelfEmpl | -0.023 | -0.026 | -0.035 | -0.016 | 0.291 | -0.015 |
| | Unempl | -0.024 | -0.020 | -0.045 | -0.055 | -0.005 | 0.264 |
| 2006 | AcadNT | 0.398 | 0.085 | -0.071 | -0.170 | -0.035 | -0.049 |
| | AcadTNA | 0.102 | 0.406 | -0.043 | -0.097 | -0.030 | -0.036 |
| | Gov | -0.051 | -0.048 | 0.635 | -0.130 | -0.039 | -0.044 |
| | Bus | -0.137 | -0.104 | -0.126 | 0.612 | -0.032 | -0.066 |
| | SelfEmpl | -0.032 | -0.036 | -0.041 | -0.032 | 0.404 | -0.031 |
| | Unempl | -0.041 | -0.035 | -0.052 | -0.089 | -0.020 | 0.392 |
| 2008 | AcadNT | 1 | -0.097 | -0.092 | -0.234 | -0.056 | -0.097 |
| | AcadTNA | -0.097 | 1 | -0.082 | -0.195 | -0.061 | -0.096 |
| | Gov | -0.092 | -0.082 | 1 | -0.214 | -0.067 | -0.098 |
| | Bus | -0.234 | -0.195 | -0.214 | 1 | -0.146 | -0.222 |
| | SelfEmpl | -0.056 | -0.061 | -0.067 | -0.146 | 1 | -0.103 |
| | Unempl | -0.097 | -0.096 | -0.098 | -0.222 | -0.103 | 1 |

# 5 References


Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition. John Wiley & Sons, Hoboken.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., and Bates, T. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76 (2), 306–317.

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., Estabrook, R., Bates, T. C., Mehta, P., von Oertzen, T., Gore, R. J., Hunter, M. D., Hackett, D. C., Karch, J., and Brandmaier, A. (2012). Openmx 1.2 user guide.

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., and Karr, A. F. (2013). Combining cohorts in longitudinal surveys. *Survey Methodology*, 39 (1), 149–182.

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference*, 63, 39–54.

Chaganty, N. R., and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66 (4), 851–860.

Chaganty, N. R., and Naik, D. N. (2002). Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference*, 103 (1-2), 421–436.

Chaganty, N. R., Sabo, R., and Deng, Y. (2012). Alternatives to mixture model analysis of correlated binomial data. *ISRN Probability and Statistics*, 2012 (Article ID 896082), 10 pages.

Chaganty, N. R., and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference*, 76, 145–161.

Chen, Q., Gelman, A., Tracy, M., Norris, F. H., and Galea, S. (2012). Weighting adjustments for panel nonresponse, Columbia University, unpublished paper.
URL www.stat.columbia.edu/~gelman/research/unpublished/weighting adjustments for panel surveys.pdf

Collett, D. (2003). *Modelling binary data*, 2nd Edition. Chapman & Hall/CRC Texts in statistical sciences series. Chapman & Hall/CRC, Boca Raton.

Fagerland, M. W., Hosmer, D. W., and Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine*, 27 (21), 4238–4253.

Graubard, B. I., Korn, E. L., and Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. In: ASA Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 170–174.

Horton, N. J., Bebchuk, J. D., Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P., and Fitzmaurice, G. M. (1999). Goodness-of-fit for GEE: An example with mental health service utilization. *Statistics in Medicine*, 18, 213–222.

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd Edition. John Wiley & Sons, Inc., New York.

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

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78 (1), 153–160.

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

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.

Roberts, G., Ren, Q., and Rao, J. N. K. (2009). Using marginal mean models for data from longitudinal surveys with a complex design: Some advances in methods. In: Lynn, P. (Ed.), Methodology of Longitudinal Surveys. Wiley, Chichester, Ch. 20, pp. 351–366.

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. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27 (2), 305–320.

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

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

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In: Bollen, K. A., and Long, J. S. (Eds.), Testing Structural Equation Models. Sage Publications Inc, pp. 256–293.