Metodološki zvezki, Vol. 5, No. 2, 2008, 95-111

On Covariance Estimation when Nonrespondents are Subsampled

Wojciech Gamrot[1]

Abstract

The phenomenon of nonresponse in a sample survey reduces the precision of parameter estimates and introduces the bias. Several procedures have been developed to compensate for these effects. An important technique is the two-phase (or double) sampling scheme which relies on subsampling the nonrespondents and re-approaching them in order to obtain the missing data. This paper focuses on the application of double sampling to estimate the finite population covariance. Two covariance estimators using combined data from the initial sample and the subsample are considered. Their properties are derived. Two special cases of the general procedure are discussed.

1 Introduction

The estimation of covariances between individual population characteristics and of the covariance matrix as a whole in the nonresponse situation has enjoyed vivid interest for years. The principle of maximum likelihood (ML) plays a prominent role in constructing estimators for covariance matrices as well as for individual covariances. Finkbeiner (1979) applies Fletcher-Powell optimum-seeking algorithm for obtaining ML estimates of covariance matrix for factor analysis. Lee(1986) proposes the estimation procedure under which missing values are viewed as latent variables and estimators are obtained under normality assumptions via generalized least squares and ML using Fisher scoring (iteratively reweighted Gauss-Newton algorithm). A coherent theory on the use of maximum likelihood principle to estimate the covariance matrix under some model describing the distribution of population values is given by Little and Rubin (1987). The Expectation-Maximization (EM) algorithm by Dempster, Laird and Rubin (1977) is contemplated as a standard method for dealing with nonresponse in a wide range of cases with special emphasis on the normal distribution. This approach is further developed in subsequent papers. Woodruff (1990) considers

[1] Wojciech Gamrot, Department of Statistics, University of Economics, Bogucicka 14, 40-226 Katowice, Poland; gamrot@ae.katowice.pl.

alternative estimator based on the EM algorithm with additional restrictions expressed by the regression superpopulation model. Schneider (2001) proposes a regularized EM algorithm utilizing ridge regression, for the case where number of variables exceeds the sample size. Bentler and Liang (2003) discuss the application of an EM-type gradient algorithm for maximum likelihood estimation in the context of two-level structural equation modelling. Jamshidian (1997) applies EM for confirmatory factor analysis. Jamshidian and Bentler (1999) explore the possibilities of using several complete-data algorithms (EM, Fletcher-Powell and Fisher's scoring) to provide maximum likelihood estimates of the covariance matrix with missing data. Under a competing approach Van Praag, Dijkstra and VanVelzen (1985) construct an asymptotically distribution-free (ADF) estimator of the covariance matrix based on linear regression. Arminger and Sobel (1990) abandon normality constraints and use pseudo-maximum likelihood (PML) method by Arminger and Schoenberg (1989) to construct estimators of covariance matrices while Yuan and Bentler (1995) provide theoretical justification for the use of PML in the context of non-normal distributions. Gold, Bentler and Kim (2003) compare ADF estimators with maximum likelihood estimators in the context of structural equation modelling. Significant attention is also devoted to the use of various imputation methods in covariance estimation. Brown (1994) compares estimators for listwise deletion, pairwise deletion and mean imputation. Kline (1998) compares the properties of estimators computed under mean imputation, regression imputation and pattern matching. Schafer and Olsen (1998) employ a method of multiple imputation invented by Rubin (1987). They develop data augmentation algorithm and provide justification for its use on the grounds of Bayesian theory. Data augmentation is also considered in the paper of Graham and Hofer (2000) and compared with the EM method. The comparison of estimators for listwise and pairwise deletion, mean imputation and full-information maximum likelihood is given by Wothke (2000). In this paper the problem of estimating individual covariance between two population characteristics is discussed under the quasi-randomization approach (Oh and Scheuren (1983)). It is assumed that the nonresponse is a random phenomenon but population values are fixed and they are not subject to modeling as opposed to most of the papers referenced above.

Let us consider some characteristics X and Y in finite population U of size N. Fixed values of these characteristics are respectively denoted by $x_1,...,x_N$ and $y_1,...,y_N$. The aim of the survey is to estimate some functions of these values called population parameters. Let the population be surveyed according to the following general two-phase sampling procedure. In the first phase a random sample s of size n is drawn from U according to some arbitrary sampling design p(s) determining individual inclusion probabilities of the first order $\pi_i = \sum_{s \ni i} p(s)$ and of the second order $\pi_{ij} = \sum_{s \ni i,j} p(s)$, for $i \neq j \in U$. We assume that nonresponse appears in the survey, and consequently some units respond while others do not. The sample s

may consequently be divided into two disjoint subsets s_1 and s_2 such that units from s_1 respond and units from s_2 do not. Following Cassel et al. (1983) we assume that nonresponse is a random event governed by some probability distribution $q(s_1 | s)$ usually referred to as *response distribution* (see e.g. Särndal et al. 1992). In general, it is defined conditionally with respect to s in order to reflect the interactions between sampled units. The response distribution determines individual response probability of any i-th unit $\rho_{i|s} = \sum_{s_1 \ni i} q(s_1 | s)$ and joint response probability of i-th and j-th unit $\rho_{ij|s} = \sum_{s_1 \ni i,j} q(s_1 | s)$ for $i \neq j \in U$. The distribution $q(s_1 | s)$ also determines the behaviour of the random set s_2 which may be expressed as a function of s and s_1, so for any $s_2 = s - s_1$ we have $q(s_1 | s) = q(s_2 | s) = q(s_1, s_2 | s)$. The second phase of the survey is then carried out and in this second phase a subsample s' of size n' is drawn from s_2 according to the sampling design $p'(s'|s,s_2)$ which is characterized by another set of inclusion probabilities of the first order $\pi_{i|s,s_2} = \sum_{s' \ni i} p'(s'|s,s_2)$ and second order $\pi_{ij|s,s_2} = \sum_{s' \ni i,j} p'(s'|s,s_2)$ $i \neq j \in U$. We assume that necessary efforts are undertaken in the second phase that guarantee obtaining complete responses from all subsampled units. In the setup described above three sources of sample randomness were defined, each of them respectively associated with probability distribution p(s), $q(s_1|s)$ and $p'(s'|s,s_2)$. All expectations will be computed jointly with respect to these three probability distributions unless otherwise stated.

2 Estimation of the population total

Let us consider the population total of X (the same may be defined for Y or any other characteristic):

$$t_x = \sum_{i \in U} x_i \tag{2.1}$$

Under complete response it is unbiasedly estimated by the Horvitz-Thompson (1952) statistic:

$$\hat{t}_x = \sum_{i \in s} \frac{x_i}{\pi_i} \tag{2.2}$$

In particular, by putting $x_i = 1$ for $i \in U$ we obtain an unbiased estimator of the population size N in the form:

$$\hat{N} = \sum_{i \in s} \frac{1}{\pi_i} \tag{2.3}$$

Both estimators above are generally biased under nonresponse. As an example consider deterministic nonresponse model, according to which the population is divided into two strata: U_1 and U_2, of sizes N_1 and N_2 such that $\rho_i=1$ for $i\in U_1$ and $\rho_i=0$ otherwise. If the sample is drawn using simple random sampling without replacement (SRSWOR), and hence inclusion probabilities of the first and second order are respectively equal to $\pi_i=n/N$ and $\pi_{ij}=n(n-1)/(N(N-1))$ for $i\neq j\in U$ then the estimator turns out to be biased and its bias is equal to:

$$B(\hat{t}_x) = t_{xU_1} - t_x = -\sum_{i\in U_2} x_i \tag{2.4}$$

where $t_{xU_1} = \sum_{i\in U_1} x_i$. The bias does not depend on the sample size and hence it does not tend to zero with growing n. However, using the data from both phases we can construct an unbiased estimator of t_x in the form (Särndal et al 1992):

$$\hat{t}_x^{\bullet} = \sum_{i\in s_1} \frac{x_i}{\pi_i} + \sum_{i\in s'} \frac{x_i}{\pi_i \pi_{i|s,s_2}} \tag{2.5}$$

Putting $x_i=1$ for $i\in U$ we again obtain an unbiased estimator of N:

$$\hat{N}^{\bullet} = \sum_{i\in s_1} \frac{1}{\pi_i} + \sum_{i\in s'} \frac{1}{\pi_i \pi_{i|s,s_2}} \tag{2.6}$$

These estimators will be used further as building blocks for more complicated estimation procedures.

3 Estimation of the population covariance

Let us consider the population covariance between X and Y defined by the expression:

$$C_U(X,Y) = \frac{1}{N-1} \sum_{i\in U} \left(x_i - \frac{1}{N} \sum_{j\in U} x_j \right) \left(y_i - \frac{1}{N} \sum_{j\in U} y_j \right) \tag{3.1}$$

or by the equivalent formula:

$$C_U(X,Y) = \frac{1}{N-1} t_{xy} - \frac{1}{N(N-1)} t_x t_y \tag{3.2}$$

where $t_{xy} = \sum_{i\in U} x_i y_i$, $t_x = \sum_{i\in U} x_i$ and $t_y = \sum_{i\in U} y_i$. Under complete response the covariance is often estimated by respective statistics:

$$\hat{C}_1(X,Y) = \frac{1}{N-1} \hat{t}_{xy} - \frac{1}{N(N-1)} \hat{t}_x \hat{t}_y \tag{3.3}$$

and

$$\hat{C}_2(X,Y) = \frac{1}{\hat{N}-1}\hat{t}_{xy} - \frac{1}{\hat{N}(\hat{N}-1)}\hat{t}_x\hat{t}_y \tag{3.4}$$

where unknown population totals are replaced with corresponding Horvitz-Thompson estimators. As indicated by Särndal et al (1992), the latter is usually preferred to the former due to better variance properties. These covariance estimators are however biased under nonresponse, As a special case let us consider SRSWOR and deterministic nonresponse. For large sample size approximate biases of both estimators respectively take the form:

$$AB(\hat{C}_1(X,Y)) = C_{U_1}(X,Y) - C_U(X,Y) \tag{3.5}$$

$$AB(\hat{C}_2(X,Y)) = C_{U_{1\#}}(X,Y) - C_U(X,Y) \tag{3.6}$$

where

$$C_{U_1}(X,Y) = \frac{1}{N-1}t_{xyU_1} - \frac{1}{N(N-1)}t_{xU_1}t_{yU_1} \tag{3.7}$$

$$C_{U_{1\#}}(X,Y) = \frac{1}{N_1-1}t_{xyU_1} - \frac{1}{N_1(N_1-1)}t_{xU_1}t_{yU_1} \tag{3.8}$$

while $t_{xU_1} = \sum_{i\in U_1} x_i$, $t_{yU_1} = \sum_{i\in U_1} y_i$ and $t_{xyU_1} = \sum_{i\in U_1} x_i y_i$. Hence the bias does not tend to zero when n grows in an apparent analogy to the expression (2.4). In order to correct for the bias we propose to replace Horvitz-Thompson estimators with their unbiased double sampling counterparts. This leads to two alternative estimators of the population covariance:

$$\hat{C}_{\bullet 1}(X,Y) = \frac{1}{N-1}\hat{t}^{\bullet}_{xy} - \frac{1}{N(N-1)}\hat{t}^{\bullet}_x\hat{t}^{\bullet}_y \tag{3.9}$$

and

$$\hat{C}_{\bullet 2}(X,Y) = \frac{1}{\hat{N}^{\bullet}-1}\hat{t}^{\bullet}_{xy} - \frac{1}{\hat{N}^{\bullet}(\hat{N}^{\bullet}-1)}\hat{t}^{\bullet}_x\hat{t}^{\bullet}_y \tag{3.10}$$

Using Taylor linearization we obtain the approximate variance of $\hat{C}_{\bullet 1}(X,Y)$:

$$AV(\hat{C}_{\bullet 1}(X,Y)) = \frac{1}{(N-1)^2}\left(\sum_{i,j\in U} u_i u_j\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) + E_{pq}\left(\sum_{i,j\in s_2}\frac{u_i u_j}{\pi_i\pi_j}\left(\frac{\pi_{ij|s,s_2}}{\pi_{i|s,s_2}\pi_{j|s,s_2}}-1\right)\right)\right) \tag{3.11}$$

where

$$u_i = (x_i - \overline{X})(y_i - \overline{Y}) - \overline{XY} \tag{3.12}$$

while $\overline{X} = t_x / N$ and $\overline{Y} = t_y / N$. The approximate bias may be expressed in the form:

$$\mathrm{AB}(\hat{\mathrm{C}}_{\bullet 1}(\mathrm{X},\mathrm{Y})) = -\frac{1}{\mathrm{N}^2(\mathrm{N}-1)^2}\left(\sum_{\mathrm{i,j}\in \mathrm{U}} \mathrm{x_i y_j}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right) + \mathrm{E_{pq}}\left(\sum_{\mathrm{i,j}\in s_2}\frac{\mathrm{x_i y_j}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}\left(\frac{\pi_{\mathrm{ij}|s,s_2}}{\pi_{\mathrm{i}|s,s_2}\pi_{\mathrm{j}|s,s_2}}-1\right)\right)\right) \tag{3.13}$$

It is worth noting, that this approximation of the bias is obtained by expanding the estimator in Taylor series including terms up to the second order, as opposed to more crude approximations based only on first-order terms. The symbols AB and AV are used to distinguish linearization-based approximations from the exact bias and variance. Using the same method we obtain the approximate variance of $\hat{\mathrm{C}}_{\bullet 2}(\mathrm{X},\mathrm{Y})$ in the form:

$$\mathrm{AV}(\hat{\mathrm{C}}_{\bullet 2}(\mathrm{X},\mathrm{Y})) = \frac{1}{(\mathrm{N}-1)^2}\left(\sum_{\mathrm{i,j}\in \mathrm{U}} \mathrm{u_i^* u_j^*}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right) + \mathrm{E_{pq}}\left(\sum_{\mathrm{i,j}\in s_2}\frac{\mathrm{u_i^* u_j^*}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}\left(\frac{\pi_{\mathrm{ij}|s,s_2}}{\pi_{\mathrm{i}|s,s_2}\pi_{\mathrm{j}|s,s_2}}-1\right)\right)\right) \tag{3.14}$$

where

$$\mathrm{u_i^*} = (\mathrm{x_i}-\overline{\mathrm{X}})(\mathrm{y_i}-\overline{\mathrm{Y}}) - \mathrm{C_U}(\mathrm{X},\mathrm{Y}) \tag{3.15}$$

and its second-order bias:

$$\mathrm{AB}(\hat{\mathrm{C}}_{\bullet 2}(\mathrm{X},\mathrm{Y})) = -\frac{1}{\mathrm{N}^2(\mathrm{N}-1)^2}\left(\sum_{\mathrm{i,j}\in \mathrm{U}} \mathrm{u_{ij}^*}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right) + \mathrm{E_{pq}}\left(\sum_{\mathrm{i,j}\in s_2}\frac{\mathrm{u_{ij}^*}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}\left(\frac{\pi_{\mathrm{ij}|s,s_2}}{\pi_{\mathrm{i}|s,s_2}\pi_{\mathrm{j}|s,s_2}}-1\right)\right)\right) \tag{3.16}$$

where

$$\mathrm{u_{ij}^*} = \frac{1}{2}\mathrm{N}^2(\mathrm{u_i}+\mathrm{u_j}) + \frac{1}{2}\mathrm{N}(\mathrm{N}-1)\left((\mathrm{x_i}-\mathrm{x_j})(\mathrm{y_i}-\mathrm{y_j}) + \mathrm{u_i} + \mathrm{u_j} + 2\mathrm{C_U}(\mathrm{X},\mathrm{Y})\right) \tag{3.17}$$

The symbol $\mathrm{E_{pq}}(\cdot)$ appearing in expressions above represents the expectation with respect to first-phase sampling design p(s) and to the response distribution $q(s_1|s)$. On assumed level of generality it is impossible to eliminate them from AV formulas but we may achieve this by making additional assumptions concerning the second phase sampling design and response distribution. The example is shown in the next section.

On the other hand, the approximate variances may be estimated, without any assumptions, by employing the approach of Särndal et al (1992) which leads to respective statistics:

$$\hat{\mathrm{V}}(\hat{\mathrm{C}}_{\bullet 1}(\mathrm{X},\mathrm{Y}) = \frac{1}{(\mathrm{N}-1)^2}\left(\sum_{\mathrm{i,j}\in s_1\cup s'}\frac{\hat{\mathrm{u}}_{\mathrm{i}}\hat{\mathrm{u}}_{\mathrm{j}}}{\pi_{\mathrm{ij}}^*}\left(\frac{\pi_{\mathrm{ij}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}}-1\right) + \sum_{\mathrm{i,j}\in s'}\frac{\hat{\mathrm{u}}_{\mathrm{i}}\hat{\mathrm{u}}_{\mathrm{j}}}{\pi_{\mathrm{i}}\pi_{\mathrm{j}}\pi_{\mathrm{ij}\,s,s_2}}\left(\frac{\pi_{\mathrm{ij}|s,s_2}}{\pi_{\mathrm{i}|s,s_2}\pi_{\mathrm{j}|s,s_2}}-1\right)\right) \tag{3.18}$$

and

$$\hat{V}(\hat{C}_{\bullet 2}(X,Y)) = \frac{1}{(N-1)^2}\left(\sum_{i,j\in s_1\cup s'} \frac{\hat{u}_i^*\hat{u}_j^*}{\pi_{ij}^*}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) + \sum_{i,j\in s'} \frac{\hat{u}_i^*\hat{u}_j^*}{\pi_i\pi_j\pi_{ij|s,s_2}}\left(\frac{\pi_{ij|s,s_2}}{\pi_{i|s,s_2}\pi_{j|s,s_2}}-1\right)\right) \tag{3.19}$$

where

$$\hat{u}_i = \left(x_i - \frac{t_x^\bullet}{N}\right)\left(y_i - \frac{t_y^\bullet}{N}\right) - \frac{t_x^\bullet t_y^\bullet}{N^2} \tag{3.20}$$

$$\hat{u}_i^* = \left(x_i - \frac{t_x^\bullet}{\hat{N}^\bullet}\right)\left(y_i - \frac{t_x^\bullet}{\hat{N}^\bullet}\right) - \hat{C}_{\bullet 2}(X,Y) \tag{3.21}$$

and

$$\pi_{ij}^* = \begin{cases} \pi_{ij}\pi_{ij|s,s_2} & \text{for} \quad i,j\in s_2 \\ \pi_{ij}\pi_{i|s,s_2} & \text{for} \quad i\in s_2,\ j\in s_1 \\ \pi_{ij}\pi_{j|s,s_2} & \text{for} \quad i\in s_1,\ j\in s_2 \\ \pi_{ij} & \text{for} \quad i,j\in s_1 \end{cases} \tag{3.22}$$

If the statistics $\hat{u}_i$ and $\hat{u}_i^*$ estimated constants u_i and u_i^* without error, then variance estimators above would respectively be unbiased for $AV(\hat{C}_{\bullet 1}(X,Y))$ and $AV(\hat{C}_{\bullet 2}(X,Y))$. Obviously they do not and some bias appears, but we may hope that it remains modest and tends to zero for large samples. We will now present two important special cases of the general procedure presented above.

4 Equal probability sampling

Let us assume as in papers of Srinath (1971) and Rao (1986) that SRSWOR is used in both phases. Hence inclusion probabilities of the first and second order are respectively equal to $\pi_i=n/N$ and $\pi_{ij}=n(n-1)/(N(N-1))$ for $i\neq j\in U$ in the first phase while $\pi_{i|s,s_2} = n'/n_2$ and $\pi_{ij|s,s_2} = n'(n'-1)/(n_2(n_2-1))$ for $i\neq j\in U$ in the second phase. We also assume that the subsample size is a linear function of the nonrespondent subset size according to formula: $n' = cn_2$ where $0<c<1$ is a constant fixed in advance. Furthermore, we assume deterministic nonresponse model described earlier. Under these assumptions we have $\hat{N}^\bullet \equiv N$. Consequently both estimators $\hat{C}_{\bullet 1}(X,Y)$ and $\hat{C}_{\bullet 2}(X,Y)$ are mutually equivalent and take the common form:

$$\hat{C}_+(X,Y) = \frac{1}{N-1}\hat{t}_{xy}^+ - \frac{1}{N(N-1)}\hat{t}_x^+\hat{t}_y^+ \tag{4.1}$$

where

$$\hat{t}_{xy}^{+} = \frac{N}{n}\left(\sum_{i\in s_1} x_i y_i + \frac{1}{c}\sum_{i\in s'} x_i y_i\right) \tag{4.2}$$

$$\hat{t}_{x}^{+} = \frac{N}{n}\left(\sum_{i\in s_1} x_i + \frac{1}{c}\sum_{i\in s'} x_i\right) \tag{4.3}$$

$$\hat{t}_{y}^{+} = \frac{N}{n}\left(\sum_{i\in s_1} y_i + \frac{1}{c}\sum_{i\in s'} y_i\right) \tag{4.4}$$

The approximate variance of $\hat{C}_{+}(X,Y)$ may be expressed as:

$$AV(\hat{C}_{+}(X,Y)) = \frac{N^2}{(N-1)^2}\left(\frac{N-n}{Nn}S_U^2(u) + \frac{W_2}{n}\frac{1-c}{c}S_{U_2}^2(u)\right) \tag{4.5}$$

where

$$S_U^2(u) = \frac{1}{N-1}\sum_{i\in U}\left(u_i - \frac{1}{N}\sum_{j\in U} u_j\right)^2 \tag{4.6}$$

$$S_{U_2}^2(u) = \frac{1}{N_2-1}\sum_{i\in U_2}\left(u_i - \frac{1}{N_2}\sum_{j\in U_2} u_j\right)^2 \tag{4.7}$$

The second-order approximate bias is

$$AB(\hat{C}_{+}(X,Y)) = -\frac{1}{N^2(N-1)^2}\left(\frac{N-n}{Nn}C_U(X,Y) + \frac{W_2}{n}\frac{1-c}{c}C_{U_2}(X,Y)\right) \tag{4.8}$$

where

$$C_{U_2}(X,Y) = \frac{1}{N_2-1}\sum_{i\in U_2}\left(x_i - \frac{1}{N_2}\sum_{j\in U_2} x_j\right)\left(y_i - \frac{1}{N_2}\sum_{j\in U_2} y_j\right) \tag{4.9}$$

and $W_2=N_2/N$. Both approximate bias and approximate variance decrease when initial sample size n grows. From (3.18) we also obtain the variance estimator:

$$\hat{V}(\hat{C}_{+}(X,Y)) = \frac{N}{(N-1)^2}\frac{N-n}{n(n-1)}\Bigg((n_1-1)S_{s_1}^2(u^{+}) + \frac{N(n_2-1)-cn_2(n-1)+n_1}{c(N-n)}S_{s'}^2(u^{+}) + \frac{n_1 n_2}{n}(\bar{u}_{s_1}^{+} - \bar{u}_{s'}^{+})^2 \Bigg) \tag{4.10}$$

where

$$\bar{u}^{+}_{s_1} = \frac{1}{n_1}\sum_{i\in s_1}\hat{u}^{+}_i \tag{4.11}$$

$$\bar{u}^{+}_{s'} = \frac{1}{n'}\sum_{i\in s'}\hat{u}^{+}_i \tag{4.12}$$

$$S^2_{s_1}(u^{+}) = \frac{1}{n_1-1}\sum_{i\in s_1}(\hat{u}^{+}_i - \bar{u}^{+}_{s_1})^2 \tag{4.13}$$

$$S^2_{s'}(u^{+}) = \frac{1}{n'-1}\sum_{i\in s'}(\hat{u}^{+}_i - \bar{u}^{+}_{s'})^2 \tag{4.14}$$

and

$$\hat{u}^{+}_i = \left(x_i - \frac{t^{+}_x}{N}\right)\left(y_i - \frac{t^{+}_y}{N}\right) - \frac{t^{+}_x t^{+}_y}{N^2}. \tag{4.15}$$

5 Unequal probability sampling

We will now focus on another special case of the general two-phase procedure. The deterministic nonresponse model assuming that response probabilities are either equal to zero or equal to unity is seldom realistic. In practice they are more likely to take any value from the <0,1> interval and depend on the auxiliary variables as well as the variable under study. In particular it may be more reasonable to assume that response probabilities are described by logistic model given by expression:

$$\rho_i = (1+\exp(\boldsymbol{\beta}\mathbf{x}_i))^{-1} \tag{5.1}$$

where x_i is some vector of auxiliary variables corresponding to the i-th population unit while $\boldsymbol{\beta}$ is the parameter vector. Also, more sophisticated sampling designs that make use of auxiliary information to improve the efficiency of parameter estimates are often preferred to the SRSWOR. One of such designs, known as Pareto sampling (Rosén 1997), allows to draw a fixed-size sample with first order inclusion probabilities approximately proportional to the values $z_1,...,z_N$ of the auxiliary characteristic Z. According to this procedure the sample s of the size n is drawn in following two steps (Särndal and Lundström 2005):

1) For any $i\in U$ a realization u_i of a random variable having uniform distribution on the <0,1> interval is generated and the following expression is evaluated:

$$q_i = \frac{u_i(1-\pi_i)}{\pi_i(1-u_i)} \tag{5.2}$$

where the desired inclusion probability π_i is given by expression

$$\pi_i = n\frac{z_i}{\sum_{i\in U} z_i} \tag{5.3}$$

If $\pi_i>0$ then i-th unit is automatically included in the sample and inclusion probabilities for remaining units are recomputed accordingly.

2) The population subset consisting of n units having the largest values of q_k is included in the sample.

Proposed covariance estimators are not equivalent now. Computation of their properties in such a situation using general formulas for approximate variance and approximate bias derived above is possible when some reasonable assumptions are made about response distribution. However, to evaluate exact inclusion probabilities it is necessary to use numerical procedures developed by Aires (2000). This makes the analytical comparison of covariance estimators difficult. Hence, a simulation study was carried out in order to compare both covariance estimators.

During simulation experiments, the population under study was represented by the data obtained from the Polish'1996 agricultural census representing 2420 farms in three boroughs (Bolesław, Gręboszów, Radgoszcz) of the Dąbrowa Tarnowska district. Three variables were used including farm sales (X), farm cattle stock (Y) and farm area (Z). The covariance between X and Y was the estimated parameter. A logistic nonresponse model was arbitrarily assumed stating that population units respond independently with response probabilities respectively equal to $\rho i=(1+\exp(\beta_0+\beta_1 x_i+\beta_2 y_i))^{-1}$ for $i\in U$ with the parameter vector $\boldsymbol{\beta}=[\beta_0,\beta_1,\beta_2]$ chosen arbitrarily. Two simulation experiments were respectively executed for $\boldsymbol{\beta}=[0,0,0]$ (ρi independent on X,Y) and $\boldsymbol{\beta}=[0,1,1]$ (ρ_i dependent on X,Y). The simulations were independently repeated for Pareto sampling utilizing Z as the auxiliary variable in both phases, and for simple random sampling without replacement in both phases. The same sample size n=100, 150, ... , 600 was always assumed for both designs. For each sample size the drawing of 40000 sample-subsample pairs was simulated, and the properties of estimators were assessed on the basis of their empirical distributions. Hence, each point on following graphs corresponds to 40000 estimates. Estimators $\hat{C}_{\bullet 1}(X,Y)$, $\hat{C}_{\bullet 2}(X,Y)$ and $\hat{C}_{+}(X,Y)$ were compared. In the graphs they are respectively denoted by abbreviations (Cov1, Cov2 and Cov_srs).

The mean square error (MSE), the bias and the share of bias in MSE observed in the first experiment are respectively shown on Figures 1, 2 and 3.

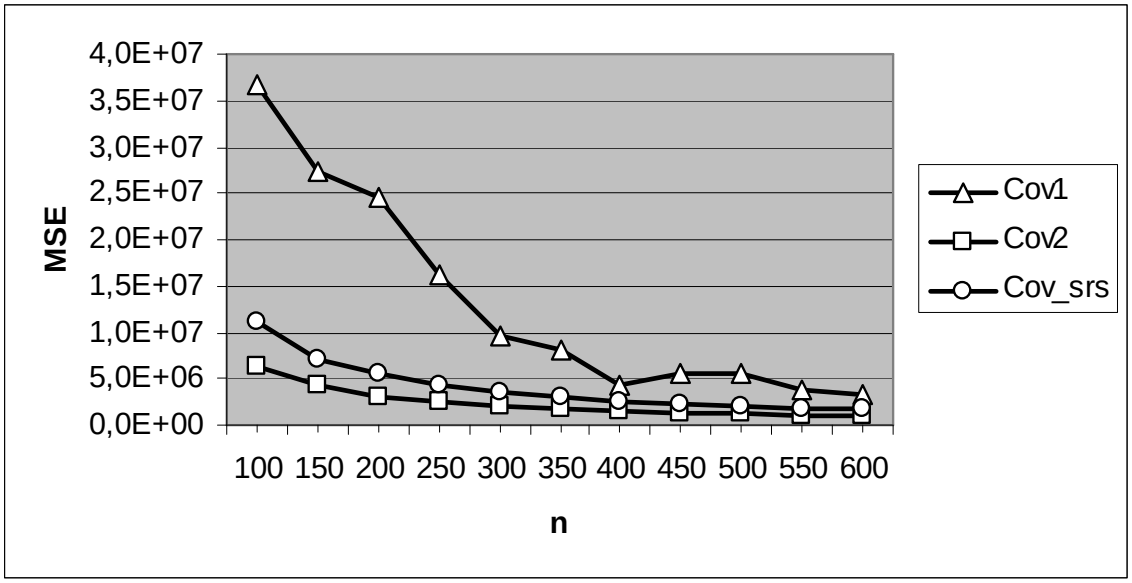

Figure 1: MSE as a function of n for $\boldsymbol{\beta}$=[0,0,0],

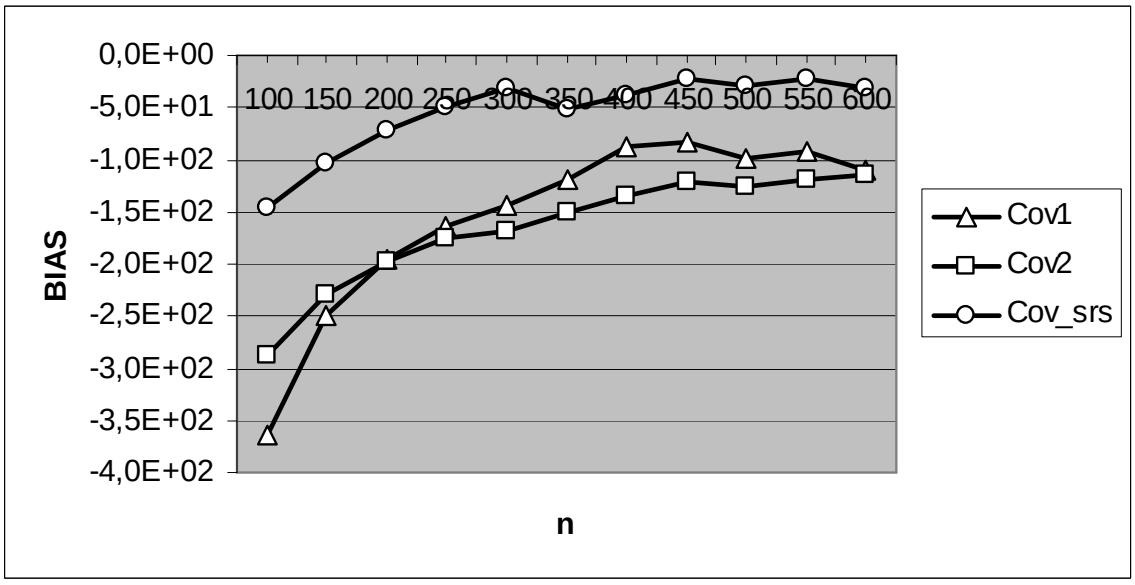

Figure 2: Bias as a function of n for $\boldsymbol{\beta}$=[0,0,0].

The MSE of all three estimators decreases with growing sample size. This observation is consistent with asymptotic results obtained for the SRSWOR case. For any value of n the estimator $\hat{C}_{\bullet 1}(X,Y)$ is much less accurate than others. The

lowest mean square error is observed for $\hat{C}_{\bullet 2}(X,Y)$, which is significantly better than other two in terms of MSE. The bias of all three estimators is negative and its absolute value decreases with growing n. This is also consistent with asymptotic results for SRSWOR. The share of bias in the MSE is approximately constant for each estimator. It does not exceed 1,4% for $\hat{C}_{\bullet 2}(X,Y)$ and 0,4% for $\hat{C}_{\bullet 1}(X,Y)$ and $\hat{C}_{+}(X,Y)$. Hence, each estimator may be treated as approximately unbiased.

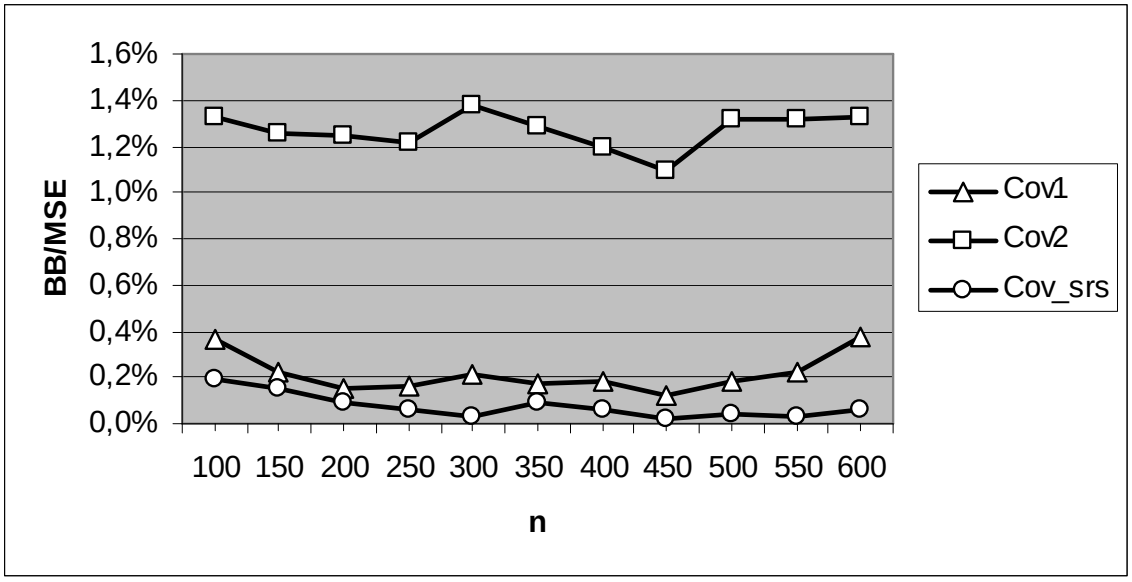

Figure 3: Share of bias in MSE as a function of n for $\boldsymbol{\beta}$=[0,0,0].

The mean square error (MSE), the bias and the share of bias in MSE observed in the second experiment are respectively shown on Figures 4, 5 and 6.

Again, the MSE of all three estimators tends to decrease with growing initial sample size. For any value of n observed MSE of the estimator $\hat{C}_{\bullet 2}(X,Y)$ is the lowest and observed MSE of the estimator $\hat{C}_{\bullet 1}(X,Y)$ is the highest. The bias is negative for $\hat{C}_{\bullet 1}(X,Y)$ and $\hat{C}_{+}(X,Y)$ while oscillating around zero for $\hat{C}_{\bullet 2}(X,Y)$. Its absolute value is lowest for $\hat{C}_{\bullet 2}(X,Y)$ and highest for $\hat{C}_{\bullet 1}(X,Y)$. The observed share of bias in the MSE is again negligible, not exceeding 0,05% for $\hat{C}_{\bullet 2}(X,Y)$ and 0,5% for the other two estimators.

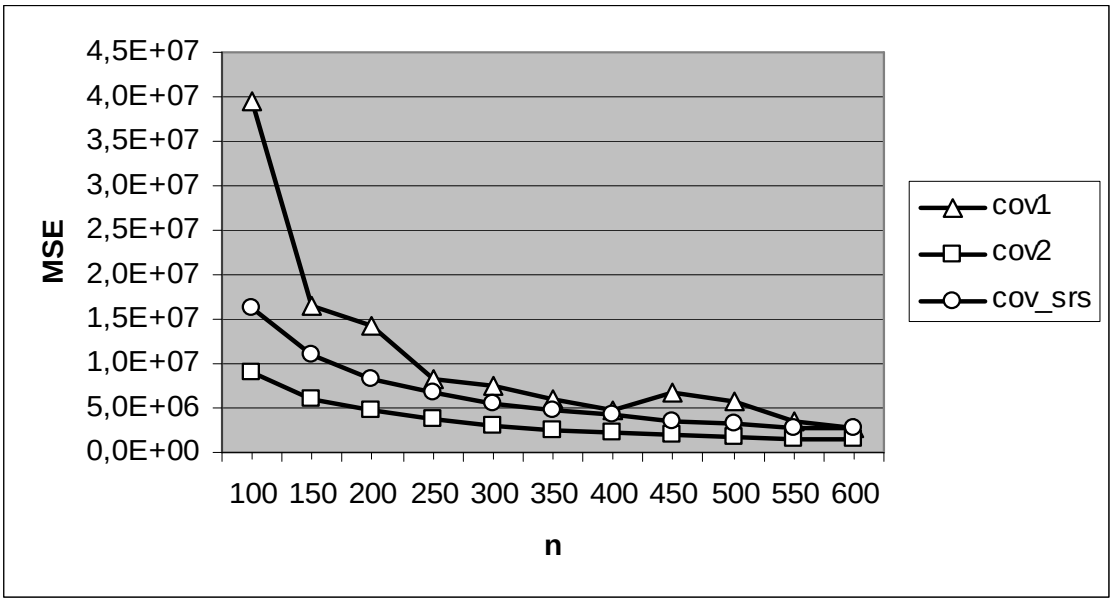

Figure 4: MSE as a function of n for β=[0,1,1].

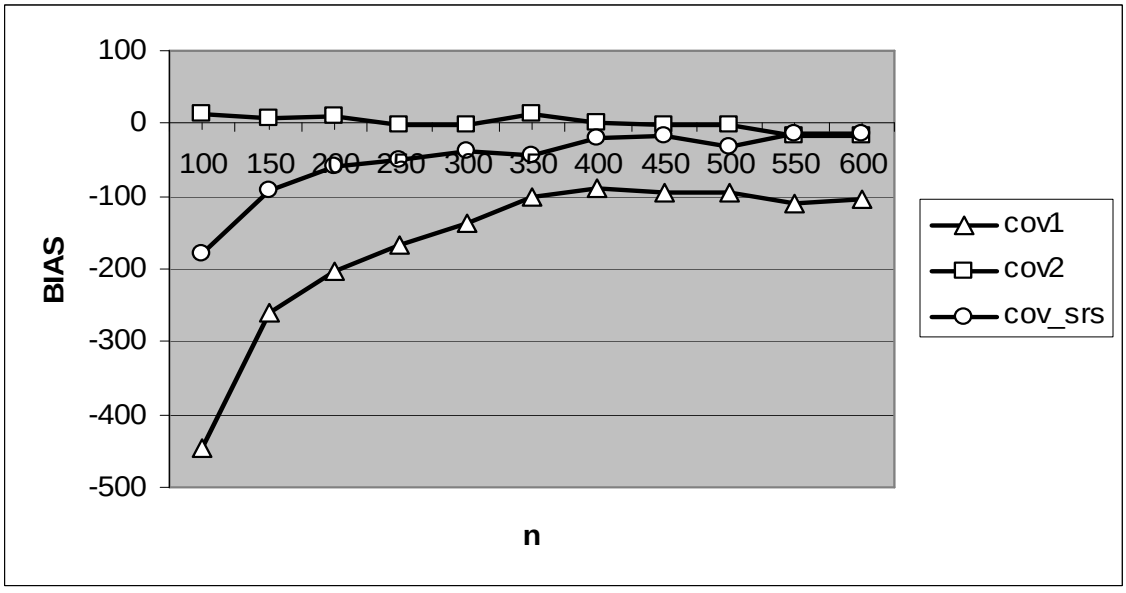

Figure 5: Bias as a function of n for β=[0,1,1].

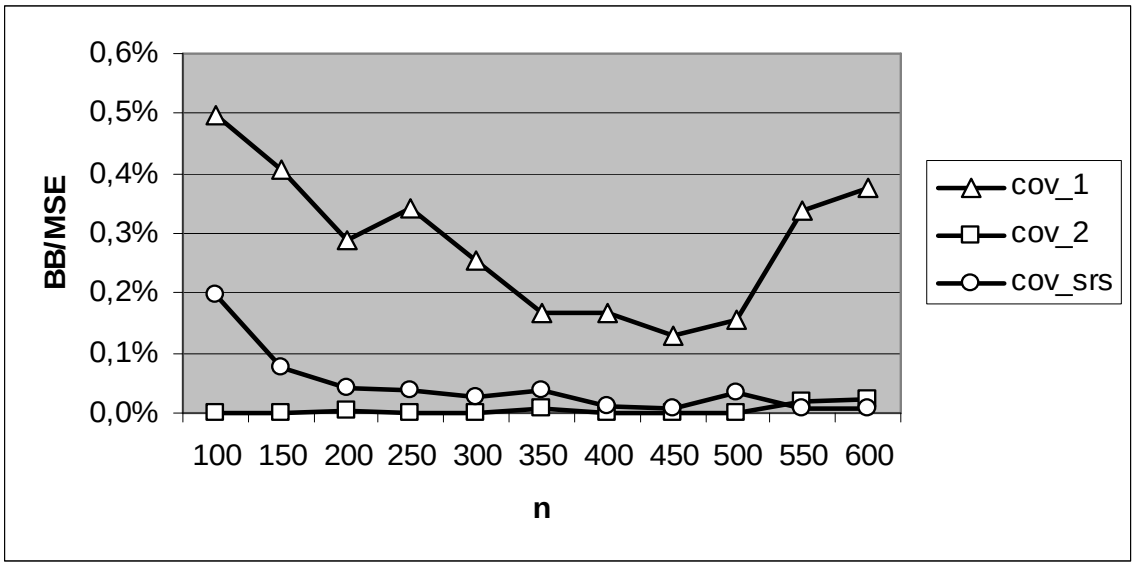

Figure 6: Share of bias in MSE as a function of n for $\boldsymbol{\beta}$=[0,1,1].

6 Conclusions

In this paper two nonresponse-corrected estimators of the finite population covariance are considered under quasi-randomization approach. They are computed using the data obtained using a general two-phase sampling procedure involving arbitrary sampling designs in both phases. Their approximate variances and biases are derived under stochastic nonresponse characterized by arbitrary response distribution. No model assumptions on population distribution are made for their derivation, which alleviates the risk of model misspecification. For the important special case of simple random sampling without replacement and deterministic nonresponse derived formulas suggest that proposed estimators are nearly unbiased. The results of simulation experiments carried out for another special case of stochastic nonresponse and Pareto sampling also seem to support this hypothesis for estimators $\hat{C}_{+}(X,Y)$ and $\hat{C}_{\bullet 2}(X,Y)$. It is worth noting that the properties of proposed estimators have been derived and simulations were executed under the assumption of complete response in the second phase. Further research is needed on their properties if this assumption is not satisfied.

Acknowledgements

The research has been supported by the grant No: 1H02B 022 30 from the Polish Ministry of Education and Science.

References

[1] Aires, N. (2000*): Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs*, PhD Thesis, Chalmers, Göteborg University. Göteborg.

[2] Arminger, G. and Schoenberg, R. (1989): Pseudo-maximum likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika*, **54**, 409-425.

[3] Arminger, G. and Sobel, M.E. (1990): Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, **85**, 195-203.

[4] Bentler, P.M. and Liang, J. (2003): Two-Level mean and covariance structures: Maximum Likelihood via the EM algorithm. In Reise, S.P., Duan, N. (Eds.): *Multilevel Modelling – Methodological Advances, Issues and Applications.* London: Psychology Press.

[5] Brown, R.L. (1994): Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modelling*, **1**, 287-316.

[6] Cassel, C.M., Särdal, C.E., and Wretman, J. (1983): Some uses of statistical models in connection with the nonresponse problem. In Madow, W.G. and Olkin, I. (Eds.): *Incomplete Data in Sample Surveys*. Vol II, New York: Academic Press,.

[7] Dempster, A.P., Laird, N.M., and Rubin, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**, 1-38.

[8] Finkbeiner, C. (1979): Estimation for the multiple factor model when data are missing. *Psychometrika*, **44**, 409-420.

[9] Gold, M.S., Bentler, P.M., and Kim, K.H. (2003): A comparison of maximum likelihood and asymptotically distribution free methods of treating incomplete nonnormal data. *Structural Equation Modelling*, **10**, 47-79.

[10] Graham, J.W. and Hofer, S.M. (2000): Multiple imputation in multivariate research. In Little, T.D., Schnabel, K.U., and Baumert, J. (Eds.): *Modelling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples.* Mahwah, NJ: Lawrence Erlbaum.

[11] Horvitz, D.G. and Thompson, D.J. (1952): A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

[12] Jamshidian, M. (1997): An EM algorithm for ML factor analysis with missing data. In Berkane, M. (Ed.): *Latent Variable Modelling and Applications to Causality*. New York: Springer, 247-258.

[13] Jamshidian, M. and Bentler, P.M. (1999): ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, **24**, 21-24.

[14] Kline, R.B. (1998): *Principles and Practices of Structural Equation Modelling*. New York: Guildford.

[15] Lee, S.-Y. (1986): Estimation for structural equation models with missing Data. *Psychometrika,* **51**, 93-99.

[16] Lessler, J.T. and Kalsbeek, W.D. (1992): *Nonsampling Errors in Surveys*. New York: Wiley & Sons.

[17] Little, R.J.A. and Rubin, D.B. (1987): *Statistical Analysis with Missing Data*. New York: Wiley & Sons.

[18] Oh, H.L. and Scheuren, F.S. (1983): Weighting adjustments for unit nonresponse. In Madow, W.G. and Olkin, I. (Ed.): *Incomplete Data in Sample Surveys*. Vol. 2, New York: Academic Press.

[19] Rao, P.S.R.S. (1986): Ratio estimation with subsampling the nonrespondents. *Survey Methodology*, **12**, 217-230.

[20] Rosén, B. (1997): On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, **62**, 159-191.

[21] Rubin, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

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

[23] Särndal, C.E. and Lundström, S. (2005): *Estimation in Surveys with Nonresponse*, New York: Wiley.

[24] Schafer, J.L. and Olsen, M.K. (1998): Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research*, **33**, 545-571.

[25] Schneider, T. (2001): Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, **14**, 853-871.

[26] Srinath, K.P. (1971): Multiphase sampling in nonresponse problems. *Journal of the American Statistical Association*, **335**, 583-586.

[27] Van Praag, B.M.S., Dijkstra, T.K., and Van Velzen, J. (1985): Least squares theory based on general distributional assumptions with an application to the incomplete observations problem. *Psychometrika*, **50**, 25-36.

[28] Woodruff, S.M. (1990): Model-based estimation of covariance matrices with applications to the EM algorithm. *Proceedings of the Survey Research Methods Section*. American Statistical Association, 425-428.

[29] Wothke, W.(2000): Longitudinal and multi-group modelling with missing data. In Little T.D., Schnabel K.U., and Baumert, J. (Eds.): *Modelling*

Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples. Mahwah, NJ: Lawrence Erlbaum.

[30] Yuan, K.-H. and Bentler, P.M. (1995): Mean and covariance structure analysis with missing data. *UCLA Statistical Series – Report No. 193*, Los Angeles: University of California.