*Survey Nonresponse.* (eds. R. Groves, D. Dillman, J. Eltinge, and R. Little), New-York: Wiley, 2001, pp. 417-429

Chapter 28

# Imputation for Wave Nonresponse - Existing Methods and a Time Series Approach

***Danny Pfeffermann and Gad Nathan,*** *Hebrew University, Israel*

## 28.1 INTRODUCTION AND LITERATURE REVIEW

Longitudinal studies have recently become a mainstay of sample survey practice (Binder, 1998). Following Duncan and Kalton (1987), we include in this term any type of survey for which at least some of the units are measured more than once. These include traditional panel surveys, with fixed or rotating panels, retrospective longitudinal studies and cohort follow-ups. The imputation methods discussed in the following are, in general, equally applicable to data obtained from administrative sources and to non-survey data, as in the medical and biological sciences, as to sample surveys proper.

We focus primarily on missing data resulting from wave nonresponse where data are available for some points in time and missing for others. Different patterns of wave nonresponse to be considered are attrition (no observations from some time point onwards), missing for a single time or for a continuous period and intermittent dropout. For all these patterns the existence of observations for some points in time for the same unit suggests the consideration of plausible relationships over time between individual measurements for more efficient imputation. This is in contrast to the treatment of missing data in cross-sectional settings and requires more elaborate modeling efforts.

When studying the properties of imputation methods, the relationships between the missing data mechanism and the missing and observed data need to be examined - see, e.g., Rubin (1976), Little (1982), Little and Rubin (1987) Little (1995) and chapter 1 of this monograph. An important distinction is between the mechanisms of missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) or informative missingness. Imputation methods have been proposed for dealing with both ignorable and non-ignorable mechanisms. The applications are in general to cross-sectional data, though Little and Rubin (1987, pp. 161-168) also consider missing data in univariate and bivariate time series. Little (1995) considers complete dropout in a longitudinal study.

The specific treatment of wave nonresponse in panel surveys has been addressed in a randomization framework in a series of papers by Kalton, Lepkowski and Lin (1985), Kalton (1986), Kalton and Miller (1986) and Lepkowski (1989). The methods proposed use imputation and weighting based on regression models. The models incorporate known auxiliary data, including response to other waves, and take into account cross-sectional and longitudinal interrelationships.

The analysis of longitudinal data has received widespread attention in the medical sciences. See, e.g., Laird and Ware (1982) and Jenrich and Schluchter (1986). Recently the analysis of longitudinal data has been largely influenced by the introduction of generalized estimating equations (GEE) by Zeger and Liang, (1986). See, e.g., Laird (1988), Diggle and Kenward (1994) ,Diggle, Liang and Zeger (1994), Murphy and Li (1995), Paik (1997) and Schafer and Schenker (2000). In addition, Rotnizky, Robins and Scharfstein (1998) consider the use of semiparametric regressions for the treatment of informative nonresponse. This estimation procedure can be viewed as an extension of the GEE method that allows for informative nonresponse.

In this chapter we propose to use time series structures with hierarchical modeling to take into account time series relationships between lower level observations and higher-level group effects (e.g. household effects). The models combine standard multi-level (mixed linear) models operating at given time points with time series state-space models for the random group effects and the individual measurements. The use of unit level time series models for the analysis of unequally

spaced longitudinal data has been considered by Jones and Boadi-Boateng (1991), Jones and Vecchia (1993) and Jones (1993). Chi and Reinsel (1989) consider first-order autoregressive models for within-individual errors and develop a score test for autocorrelation, on which they base an explicit maximum likelihood estimation procedure. However, none of these studies considers explicitly the effects of missing data and they do not account for hierarchical population structures. Goldstein, Healy and Rasbash (1994) consider the analysis of repeated measurements using a two-level hierarchical model, with individuals as second levels and the repeated measurements as the first levels. The model permits the first level measurements to be correlated over time.

The following section reviews possible mechanisms for wave nonresponse and the methods considered for imputation of the missing data. Section 28.3 sets up the proposed multilevel time series modeling approach. Section 28.4 reports the results of a simulation study and an analysis of empirical data. The final section mentions possible ramifications and extensions.

## 28.2 NONRESPONSE MECHANISMS AND IMPUTATION METHODS

We consider longitudinal data as obtained from retrospective longitudinal studies, cohort studies, or from panel surveys with fixed or rotating panels. Our interest is in wave nonresponse whereby data are missing from one or more waves of a multi-wave survey either according to design, or due to common reasons for nonresponse such as 'not contacted','refusal',etc. For modeling purposes, wave nonresponse is regarded as resulting from a random mechanism, which may be related to the outcome variables. If it can be shown that no such relationship exists, the nonresponse mechanism is MCAR. Even if such a relationship does exist, it may still be due to a mechanism which is MAR, whereby the probability of the observed response pattern, given the missing and observed data, does not depend on the missing values. If the nonresponse probabilities depend also on the values of the missing data, then they define an informative missing or NMAR data mechanism. Precise definitions of these terms can be found in chapter 1.

In what follows we consider several imputation methods for wave nonresponse with emphasis on methods that utilize the time series structure of longitudinal data and the hierarchical structure of the population. An important outcome of the present study is that by considering the interrelationships among observations related to the same natural group, the imputation bias induced by informative nonresponse can be largely reduced, even when ignoring the response process. The methods we consider assume the existence of (fully observed) auxiliary variables, which bear some relationship to the response variable. Below is a brief description of the methods considered (the first three serve as benchmarks). Exact specifications under the proposed model are given in section 28.3.

**1)** Mean imputation.

Homogeneous imputation groups are created on the basis of values of the auxiliary variables and/or outcomes from other waves. Missing values are imputed as the mean of the reported values in the corresponding imputation group.

**2**) Nearest neighbor.

A distance measure is defined in the auxiliary variable space and the missing value is imputed as the reported value nearest to the missing point. In longitudinal studies the outcome obtained in another wave for the same individual may be defined as the 'nearest neighbor' after appropriate modifications to account for fixed time effects.

**3)** Simple regression imputation.

A regression relationship is estimated between the response variable and the auxiliary variables. The regression coefficients are estimated by Ordinary Least Squares (OLS). The regression predicted value serves as the imputed value. A variation of this method is to add a random residual to the imputed value so as to better reflect the variation in the data.

**4)** Augmented regression imputation.

The regression prediction is extended by adding a correction term that accounts for the existing correlations between the observed and the missing data. The unknown regression coefficients are estimated by Generalized Least Squares (GLS). The imputation of missing data is

based on all the observations for all the time periods. Two variants are considered. (a) Only the observed data for the individual with the missing data are used for the imputation, (b) All the observed data for all individuals in the same hierarchical group are used for the imputation. As in (3), random residuals can be added to the imputed values.

**5)** State-space model imputation

A state-space model combining multi-level models operating at given points in time is postulated. Predictions obtained under the model are used as the imputed values. The unknown parameters of the combined model are estimated by MLE.

## 28.3 TIME-SERIES MODELS FOR LONGITUDINAL DATA

### 28.3.1 Model Specification

Following Feder et al. (2000), we seek a model that encompasses the hierarchical nature of many human populations and the time series relationships between repeated measurements and between the random effects of higher-level groups. The proposed model combines separate two-level mixed linear models (Goldstein, 1986, 1995), operating at given points in time by a state-space model that represents the time series relationships of the random group effects and the individual measurements. Basic notation and assumptions follow. We refer for convenience to the higher-level groups as 'households' and to the lower level units as 'individuals'. Let $y_{hjt}$ define the value of the response variable at time $t = 1,\ldots,T$, for individual $j = 1,\ldots,n_h$, belonging to household $h = 1,\ldots N$. The measurements $y_{hjt}$ are assumed to follow the hierarchical two level linear model:

$$y_{hjt} = \mathbf{x}'_{hjt}\mathbf{b}_t + \mathbf{z}'_{ht}\mathbf{v}_t + \mathbf{z}'_{ht}\mathbf{u}_{ht} + e_{hjt}, \qquad (3.1.1)$$

where $\mathbf{x}_{hjt}$ is a p-dimensional vector of individual level explanatory variables values; $\mathbf{z}_{ht}$ is a q-dimensional vector of household level explanatory variables; $\mathbf{b}_t$ and $\mathbf{v}_t$ are fixed vector coefficients of appropriate orders; $\mathbf{u}_{ht}$ is a $(q \times 1)$ vector of household level random effects and $e_{hjt}$ is an individual level random error. The random household effects represent specific household characteristics not represented by the fixed effects.

The individual and household level random errors are assumed to follow independent first order autoregressive models,

$$\mathbf{u}_{ht} = \mathbf{A}\mathbf{u}_{ht-1} + \mathbf{d}_{ht}; \quad \mathbf{d}_{ht} \sim \mathrm{N}\left(\mathbf{0}_q, \mathbf{D}\right); \tag{3.1.2}$$

$$e_{hjt} = \rho e_{hjt-1} + \varepsilon_{hjt}; \quad \varepsilon_{hjt} \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right). \tag{3.1.3}$$

In the following, we assume for convenience that A and D are diagonal, implying independence of the random household effects. We also assume $\left|\mathbf{A}_{ii}\right| < 1$ and $\left|\rho\right| < 1$ to ensure stationarity. More elaborate models could be considered, depending on the number of observations per unit. It follows from (3.1.2) and (3.1.3) that for any given time t,

$$\mathbf{u}_{ht} \sim \mathrm{N}\left(\mathbf{0}_q, \mathbf{D}^*\right); \ \mathbf{D}^* = \left(\mathbf{I}_q - \mathbf{A}^2\right)^{-1}\mathbf{D}; \tag{3.1.4}$$

$$e_{hjt} \sim \mathrm{N}\left(0, \sigma_e^2\right); \ \sigma_e^2 = \left(1 - \rho^2\right)^{-1}\sigma_\varepsilon^2. \tag{3.1.5}$$

Thus, the models operating at various time points are standard multilevel models with fixed variances for the random first and second level effects.

### 28.3.2 Model-based imputation methods

In this section we illustrate the application of the imputation methods described in section 28.2 under the model defined in section 28.3.1. We assume for convenience that the longitudinal measurements are taken over a fixed period t=1,…,T with some of the measurements possibly missing, and that all the model parameters are known. The unknown

model parameters are replaced in practice by sample estimates. Parameter estimation is considered in section 28.3.3.

**1)** Mean imputation As described in Section 28.2

**2)** Nearest neighbor

Here we restrict to the case where the nearest neighbor is defined by the nearest observation in time obtained for the same individual. If *t* is the time point with missing value and t* is the nearest time with an observation, the imputed value is defined as:

$$y_{hjt}^{(nn)} = \mathbf{x}'_{hjt}\mathbf{b}_t + \mathbf{z}'_{ht}\mathbf{v}_t + (y_{hjt^*} - \mathbf{x}'_{hjt^*}\mathbf{b}_{t^*} - \mathbf{z}'_{ht^*}\mathbf{v}_{t^*}) \qquad (3.2.1)$$

**3)** Simple regression imputation

The imputed values are obtained as the simple regression predictions:

$$y_{hjt}^{(P)} = \mathbf{x}'_{hjt}\mathbf{b}_t + \mathbf{z}'_{ht}\mathbf{v}_t, \qquad (3.2.2)$$

**4)** Augmented regression

Let $\tilde{\mathbf{Y}}_{hj} = (y_{hj1},\ldots,y_{hjT})'$ represent the generic vector of complete values (observed and missing) for individual j in household h, with variance-covariance (V-C) matrix, $\mathbf{S}_h$ ( defined by the parameters contained in $\mathbf{A},\mathbf{D},\rho,\sigma_e^2$ ). Let $\mathbf{Q}_{hj}$ define the response indicator matrix of size $t_{hj}\times T$ corresponding to unit hj, ($t_{hj}$ is the number of times that unit hj is observed), such that the observed values are $\mathbf{Y}_{hj} = \mathbf{Q}_{hj}\tilde{\mathbf{Y}}_{hj}$. Similarly, denote by $\overline{\mathbf{Q}}_{hj}$ the indicator matrix for the missing values, of size $\bar{t}_{hj}\times T$, $(\bar{t}_{hj} = T - t_{hj})$, such that the missing values are $\mathbf{Y}_{hj}^{(m)} = \overline{\mathbf{Q}}_{hj}\tilde{\mathbf{Y}}_{hj}$.

The imputed values, based only on data for the same individual (method 4a) are the augmented (BLU) regression predictions (Pfeffermann 1988):

$$\hat{\mathbf{Y}}_{hj}^{(m)} = \overline{\mathbf{Q}}_{hj}\tilde{\mathbf{Y}}_{hj}^{(p)} + \overline{\mathbf{Q}}_{hj}\mathbf{S}_h\mathbf{Q}'_{hj}\left(\mathbf{Q}_{hj}\mathbf{S}_h\mathbf{Q}'_{hj}\right)^{-1}\left(\mathbf{Y}_{hj} - \mathbf{Q}_{hj}\tilde{\mathbf{Y}}_{hj}^{(p)}\right), \qquad (3.2.3)$$

where $\tilde{\mathbf{Y}}_{hj}^{(p)} = \left(y_{hj1}^{(p)},\ldots,y_{hjT}^{(p)}\right)$ is the complete vector of regression predictions defined by (3.2.2); $\overline{\mathbf{Q}}_{hj}\mathbf{S}_h\mathbf{Q}'_{hj} = \mathrm{Cov}\left(\mathbf{Y}_{hj}^{(m)},\mathbf{Y}_{hj}\right)$ and

$\mathbf{Q}_{hj}\mathbf{S}_h\mathbf{Q}'_{hj} = \mathrm{V}(\mathbf{Y}_{hj})$. Similarly, we denote by $\tilde{\mathbf{Y}}_h = (\tilde{\mathbf{Y}}'_{h1},\ldots,\tilde{\mathbf{Y}}'_{hn_h})'$ the generic vector of complete values for all the individuals belonging to household h, with V-C matrix, $\tilde{\mathbf{S}}_h$. Let $\mathbf{Q}_h$ and $\overline{\mathbf{Q}}_h$ define the household indicator matrices for observed and missing values of orders $\sum_j t_{hj} \times n_h T$ and $\sum_j \bar{t}_{hj} \times n_h T$, respectively, so that $\mathbf{Y}_h = \mathbf{Q}_h\tilde{\mathbf{Y}}_h$ and $\mathbf{Y}_h^{(m)} = \overline{\mathbf{Q}}_h\tilde{\mathbf{Y}}_h$ are corresponding vectors of observed and missing values.

The imputed values, based on all the observed data for all the individuals in the household (method 4b) are the augmented regression predictions:

$$\hat{\mathbf{Y}}_h^{(m)} = \overline{\mathbf{Q}}_h\tilde{\mathbf{Y}}_h^{(P)} + \left(\overline{\mathbf{Q}}_h\tilde{\mathbf{S}}_h\mathbf{Q}'_h\right)\left(\mathbf{Q}_h\tilde{\mathbf{S}}_h\mathbf{Q}'_h\right)^{-1}\left(\mathbf{Y}_h - \mathbf{Q}_h\tilde{\mathbf{Y}}_h^{(P)}\right) \qquad (3.2.4)$$

The imputations defined by (3.2.3) and (3.2.4) are the conditional expectations of the missing values given the observed values. They are the best linear unbiased predictors even without normality of the residual terms.

**5)** State-space model imputation

The model (3.1.1)-(3.1.3) is written in state-space form with *observation equation*;

$$\left[\mathbf{Q}_{ht}\mathbf{Y}_{ht}\right] = \left[\mathbf{Q}_{ht}\tilde{\mathbf{X}}_{ht}\right]\underline{\tilde{\beta}}_t + \left[\mathbf{Q}_{ht}\tilde{\mathbf{Z}}_{ht}\right]\underline{\alpha}_{ht}, \qquad (3.2.5)$$

and *transition equation*;

$$\underline{\alpha}_{ht} = \mathbf{T}_h\underline{\alpha}_{h,t-1} + \underline{\nu}_{ht}, \qquad (3.2.6)$$

where $\left[\mathbf{Q}_{ht}\mathbf{Y}_{ht}\right]$ denotes the observed values for household h at time t ($\mathbf{Y}_{ht}$ defines the generic vector at time t of complete values for all the individuals in household h, of order $n_h$ and $\mathbf{Q}_{ht}$ is the corresponding response indicator matrix), $\left[\mathbf{Q}_{ht}\tilde{\mathbf{X}}_{ht}\right]$ and $\left[\mathbf{Q}_{ht}\tilde{\mathbf{Z}}_{ht}\right]$, with $\tilde{\mathbf{X}}_{ht} = \begin{pmatrix} \mathbf{x}'_{h1t} & \mathbf{z}'_{ht} \\ \vdots & \vdots \\ \mathbf{x}'_{hn_ht} & \mathbf{z}'_{ht} \end{pmatrix}$ and $\tilde{\mathbf{Z}}_{ht} = \begin{pmatrix} \mathbf{z}'_{ht} & \\ \vdots & \mathbf{I}_{n_h} \\ \mathbf{z}'_{ht} & \end{pmatrix}$, are the design matrices of the explanatory variables; $\underline{\tilde{\beta}}_t = (\mathbf{b}'_t, \mathbf{v}'_t)'$ is the (p+q)×1 vector of fixed

parameters; $\underline{\alpha}_{ht} = \left(\mathbf{u}'_{ht}, \mathbf{e}'_{ht}\right)'$ is the $(q+n_h)\times 1$ state vector with $\mathbf{e}_{ht} = \left(e_{h1t}, \ldots, e_{hn_h t}\right)'$; $\mathbf{T}_h = \mathbf{A} \oplus \rho \mathbf{I}_{n_h}$ is the transition matrix (a block-diagonal matrix with A and $\rho \mathbf{I}_{n_h}$ as the two blocks) and $\underline{\nu}_{ht} = (d'_{ht}, \underline{\varepsilon}'_{ht})'$ is a vector of random errors with V-C matrix: $\mathrm{V}\left(\underline{\nu}_{ht}\right) = \mathbf{R}_h = \mathbf{D} \oplus \sigma_\varepsilon^2 \mathbf{I}_{n_h}$.

Under the model (with known parameters), the random components $\mathbf{u}_{ht}$ and $\mathbf{e}_{ht}$ can be predicted either by application of the Kalman filter, if only current and past observations are available, or by an appropriate smoothing filter if data for subsequent time periods are known, see Harvey (1989) and De Jong (1989) for details. Starting values for the filters at time t=1 are defined by (3.1.4) and (3.1.5). The imputed values under the model are obtained as

$$\hat{y}_{hjt}^{(M)} = \hat{y}_{hjt}^{(P)} + \mathbf{z}'_{ht}\hat{\mathbf{u}}_{ht} + \hat{e}_{hjt}, \qquad (3.2.7)$$

where $\hat{y}_{hjt}^{(P)}$ is defined by (3.2.2) and $\hat{\mathbf{u}}_{ht}$ and $\hat{e}_{hjt}$ are the predicted values obtained from the Kalman filter or the smoothing algorithm.

It should be noted that the difference between the augmented regression imputations defined by (3.2.4) and the corresponding state-space imputations defined by (3.2.7) is only in the estimation of the unknown parameters (see below). Both procedures use the same data, the same model and the same imputation criterion.

### 28.3.3 Estimation of model parameters

The use of the proposed imputation methods requires the estimation of all the model parameters. For the simple regression imputation (method 3), the coefficients are ordinarily estimated by OLS. For the augmented regression and state-space modeling approaches we consider MLE under the model (3.1.1)-(3.1.3). Direct maximization of the likelihood under augmented regression (methods 4 and 5) is complicated due to the complex structure of the V-C matrix $\tilde{\mathbf{S}}_h = \mathrm{V}[\tilde{\mathbf{Y}}_h]$. This matrix takes the

form $\tilde{\mathbf{S}}_h = \mathbf{I}_{n_h} \otimes \sigma_e^2 \tilde{\mathbf{R}} + \mathbf{J}_{n_h} \otimes \left(\mathbf{Z}_h \tilde{\mathbf{A}} \mathbf{Z}_h'\right)$, where $\otimes$ denotes the Kronecker product; $\mathbf{J}_{n_h} = \mathbf{1}_{n_h} \mathbf{1}_{n_h}'$ with $\mathbf{1}_{n_h}$ defining the unit vector of order $n_h$; $\mathbf{Z}_h = [\mathbf{z}_{h1}, \ldots, \mathbf{z}_{hT}]$; $\tilde{\mathbf{R}}$ is $T \times T$ with $\tilde{\mathbf{R}}_{t,t'} = \rho^{|t-t'|}$; and $\tilde{\mathbf{A}}$ is a $qT \times qT$ block matrix, whose (t,t')-th block is $\mathbf{A}^{|t-t'|}\mathbf{D}^*$, $t,t' = 1..T$. The V-C matrix for the observed data in household h is $\mathbf{Q}_h \tilde{\mathbf{S}}_h \mathbf{Q}_h'$.

For the simulation study in the next section we used the method of Iterative Generalized Least Squares (IGLS) as proposed by Anderson (1973). The iterations alternate between estimation of the fixed parameters, $\underline{\tilde{\beta}} = \left(\underline{\tilde{\beta}}_1', \ldots, \underline{\tilde{\beta}}_T'\right)'$, where $\underline{\tilde{\beta}}_t = (\mathbf{b}_t', \mathbf{v}_t')'$, and estimation of the elements of $\tilde{\mathbf{S}}_h$.

To simplify the computations and also obtain estimators that are more robust to possible model misspecifications, we use a more flexible definition of the matrix $\tilde{\mathbf{A}}$, whereby the (t,t')-th block of $\tilde{\mathbf{A}}$ is $\mathbf{D}^*_{|t-t'|}$, where $\mathbf{D}^*_0, \ldots, \mathbf{D}^*_{T-1}$ are arbitrary diagonal matrices of order $q$. Note that the number of unknown parameters in $\tilde{\mathbf{A}}$ is increased this way from 2*q* to *Tq*.

For the state-space model imputations (method 5), we maximize the likelihood by use of the method of scoring. Let $\mathbf{Y}_{ht}^{(r)} = \mathbf{Q}_{ht} \mathbf{Y}_{ht}$ denote the vector of $r_{ht}$ observed values for household h at time t. Let $\tilde{\mathbf{X}}_{ht}^{(r)} = \mathbf{Q}_{ht} \tilde{\mathbf{X}}_{ht}$ and $\tilde{\mathbf{Z}}_{ht}^{(r)} = \mathbf{Q}_{ht} \tilde{\mathbf{Z}}_{ht}$ be the corresponding $r_{ht} \times (p+q)$ and $r_{ht} \times (q+n_h)$ system matrices. The state vector estimate at time t, $\underline{\hat{\alpha}}_{ht}$, and its V-C matrix, $\mathbf{P}_{ht}$, are computed by the Kalman filter (Harvey 1989), with initial values set as $\underline{\hat{\alpha}}_{h0} = 0$ and $\mathbf{P}_{h0} = D^* \oplus \sigma_e^2 \mathbf{I}_{n_h}$, where $\mathbf{D}^*$ and $\sigma_e^2$ are defined by (3.1.4) and (3.1.5). The likelihood is computed as follows: at time t, the predicted state vector, given all data till time t-1, is: $\underline{\hat{\alpha}}_{ht|t-1} = \mathbf{T}_h \underline{\hat{\alpha}}_{ht-1}$ with prediction error V-C matrix $\mathbf{P}_{ht|t-1} = \mathbf{T}_h \mathbf{P}_{ht-1} \mathbf{T}_h' + \mathbf{R}_h$, where $\mathbf{R}_h = \mathbf{D} \oplus \sigma_e^2 \mathbf{I}_{n_h}$. The predicted value of $\mathbf{Y}_{ht}$ at time (t-1) is $\hat{\mathbf{Y}}_{ht|t-1}^{(r)} = \tilde{\mathbf{X}}_{ht}^{(r)} \underline{\beta}_t + \tilde{\mathbf{Z}}_{ht}^{(r)} \underline{\hat{\alpha}}_{ht|t-1}$ with prediction error V-C matrix

$\mathbf{F}_{ht} = \tilde{\mathbf{Z}}_{ht}^{(r)} \mathbf{P}_{ht|t-1} \tilde{\mathbf{Z}}_{ht}^{(r)\prime} = \mathrm{V}(\tilde{\mathbf{e}}_{ht})$ where $\tilde{\mathbf{e}}_{ht} = \mathbf{Y}_{ht}^{(r)} - \hat{\mathbf{Y}}_{ht|t-1}^{(r)}$ is the prediction error. The contribution to the log-likelihood from household h is:

$$\mathrm{L}_h(\theta) = -\left\{\sum_t \ln\left(|\mathbf{F}_{ht}|\right) + \sum_t \tilde{\mathbf{e}}_{ht}' \mathbf{F}_{ht}^{-1} \tilde{\mathbf{e}}_{ht}\right\}. \qquad (3.3.1)$$

### 28.3.4 Estimation of biases and mean square errors

Though imputation methods are aimed primarily at completing the sample data and improving point estimation, the variance of estimators that use the observed and imputed values is needed for inference and interval estimation. The model considered in the present article permits model-based estimation of imputation bias and mean square error (MSE). The appropriate expressions are presented in the appendix, assuming known parameter values and that that the model holds for both the missing and the observed data. The contribution to the variance from parameter estimation can be ignored in practice since it is ordinarily of lower order than the imputation variance.

## 28.4 SIMULATIONS AND EMPIRICAL EXAMPLE

To evaluate and compare the performance of the various imputation methods, we generated populations of size N=1000 (500 for the state-space imputation method) from the model (3.1.1)-(3.1.3) for 4 time points. Household sizes, $n_h$, were randomly assigned the values 2 or 3. The parameters and regressors are: $\mathbf{b}_t = (1,2)'$; $\mathbf{v}_t = (3,4)'$; A=diag[0.7,0.7]; D=diag[0.5,0.5]; $\rho = 0.4$; $\sigma_\varepsilon^2 = 4$;, $z_{ht1} \equiv 5$;. The values $x_{hjtl}$ ($l$=1,2) and $z_{ht2}$ were selected independently from the uniform distributions $x_{hjtl} \sim \mathrm{U}(1,10)$, $z_{ht2} \sim \mathrm{U}(1,10)$ for each set of indices.

We simulated three different response mechanisms with an expected nonresponse rate of $P_0$=0.2 for each time point. The mechanisms are defined by the distribution of the response indicator:

**1)** MCAR: $R_{hjt} \overset{ind}{\sim} \text{Bernoulli}(1-P_0)$.

**2)** MAR:

For t=1; $R_{hj1} \overset{ind}{\sim} \text{Bernoulli}(1-P_0)$;

For t=2; $R_{hj2} = \text{I}\left(R_{hj1}e_{hj1} \geq \sigma_e Z_q\right)$, where I(.) is the indicator variable, $e_{hj1}$ is the residual defined by (3.1.1), $q = P_0/(1-P_0)$ and $Z_q$ is the 100q percentile of the standard normal distribution. Units missing at time 1 are observed at time 2, since $Z_q < 0$ for $P_0 < 0.5$.

For t>2; $R_{hjt} = \begin{cases} \text{I}\left(R_{hj2}e_{hj2} \geq \sigma_e Z_{P_0}\right); & if \quad R_{hj1} = 0 \\ \text{I}\left(R_{hj1}e_{hj1} \geq \sigma_e Z_{P_0}\right); & if \quad R_{hj1} = 1 \end{cases}$

**3)** NMAR: $R_{hjt} = \text{I}\left(e_{hjt} \geq \sigma_e Z_{P_0}\right)$.

Under the second and third mechanisms, non-response occurs for large negative values of the outcome variable. In the case of MAR, non-response is determined by a previously observed outcome whereas in the case of NMAR, non-response depends on the missing outcome.

The five imputation methods (and their variants) specified in section 28.3 were applied to each simulated sample. For method 1 (mean imputation), eight imputation groups were formed by the cubes created by dividing the range [1,10] of each of the auxiliary variables, $x_{hjt1}, x_{hjt2}$ and $z_{ht2}$ into two equal parts (separately for each t). The group mean of the reported values was used for the imputation.

The imputed values, $y_{hjt}^{(I)}$, were compared to the true missing values, $y_{hjt}$, to obtain empirical estimates of bias and root mean square error (RMSE), as averages over all the missing values in the four time points. The bias and RMSE are defined as:

$$BIAS = \Sigma_{hjt}\left(1-R_{hjt}\right)\left(y_{hjt}^{(I)} - y_{hjt}\right) \Big/ \Sigma_{hjt}\left(1-R_{hjt}\right);$$

$$RMSE = \left[\Sigma_{hjt}\left(1-R_{hjt}\right)\left(y_{hjt}^{(I)} - y_{hjt}\right)^2 \Big/ \Sigma_{hjt}\left(1-R_{hjt}\right)\right]^{1/2}.$$

One hundred samples were generated for each method. Table 28.4.1 shows the average relative bias and RMSE over the 100 samples. The values in parentheses are the corresponding empirical standard errors.

**Table 28.4.1: Simulation results - relative bias and RMSE under different imputation methods (%). Standard errors in parentheses.**

| | | RELATIVE BIAS | | | RELATIVE RMSE | | |
|---|---|---|---|---|---|---|---|
| No. | Imputation method | MCAR | MAR | NMAR | MCAR | MAR | NMAR |
| 1 | Mean imputation | 0.12 | 0.92 | 6.14 | 20.13 | 20.11 | 20.63 |
| | | (0.05) | (0.05) | (0.05) | (0.72) | (0.80) | (0.77) |
| 2 | Nearest neighbor | 0.07 | -3.44 | 6.15 | 14.79 | 16.53 | 15.96 |
| | | (0.08) | (0.09) | (0.09) | (0.33) | (0.38) | (0.35) |
| 3 | Simple regression | 0.06 | 0.94 | 6.10 | 16.22 | 16.17 | 16.81 |
| | | (0.02) | (0.03) | (0.02) | (0.12) | (0.22) | (0.19) |
| 4a | Augmented regression - individual | 0.06 | -1.13 | 6.17 | 12.14 | 13.58 | 13.84 |
| | | (0.02) | (0.03) | (0.02) | (0.03) | (0.03) | (0.21) |
| 4b | Augmented regression - household | 0.03 | -0.05 | 7.10 | 6.21 | 6.73 | 8.66 |
| | | (0.03) | (0.04) | (0.03) | (0.02) | (0.03) | (0.02) |
| 5a | State space - unsmoothed | 0.03 | -0.01 | 7.31 | 6.79 | 7.05 | 9.09 |
| | | (0.03) | (0.03) | (0.02) | (0.04) | (0.05) | (0.03) |
| 5b | State space - smoothed | 0.01 | -0.03 | 7.13 | 6.20 | 6.75 | 8.67 |
| | | (0.03) | (0.04) | (0.02) | (0.07) | (0.08) | (0.09) |

The main outcomes emerging from Table 28.4.1 are as follows. All the methods yield unbiased imputations under MCAR and the last three methods are also unbiased under MAR. The bias of the first three methods under MAR is explained by the fact that the MAR response is

determined by the observed outcome on a previous occasion and this outcome is not used for the first three imputation methods (except in some cases for the nearest neighbor method).

Next consider the relative RMSE that show much more pronounced differences between some of the methods. As expected, the methods 4b and 5b, that use all the observed household data for all the time points perform best. However, method 5a, which likewise uses the data observed for other household members also performs well, despite the fact that it only employs past and current data. On the other hand, method 4(a) which uses only the data observed for the same individual performs much worse although it still dominates the first 3 methods. These results illustrate the potential benefits from accounting for the relationships between individual measurements within households and the relationships between individual measurements over time.

Table 28.4.2 compares the average imputation MSE estimators presented in the appendix (with true parameter values replaced by the sample estimates), with the corresponding empirical MSE (same as in table 28.4.1). The values shown indicate a very close agreement between the estimated and empirical RMSE under the MCAR and MAR response mechanisms. This is true for all the imputation methods. The RMSE estimators underestimate the true RMSE for NMAR since they fail to account for the imputation bias inherent under this meachnism.

Finally, we applied the imputation methods to data extracted from the Israeli Labour Force Survey (ILFS) for Jerusalem during the years 1990-1994. This survey uses a rotation pattern of two quarters in the sample, two quarters out of the sample and two quarters in again. The data contain complete records for 567 individuals in 475 households, with each individual observed for four quarters according to the rotation pattern described above. The dependent variable is the number of hours worked during the week preceding the interview ($\bar{y} = 39.8$; $\mathrm{sd}(y) = 14.8$, calculated over all individuals and all time periods). The household level auxiliary variables are $z_1 = 1$ and $z_2$ = the number of employed persons in the household ($\bar{z}_2 = 1.48$; $\mathrm{sd}(z_2) = 0.56$); the individual level auxiliary variables are $x_1$ = years of education ($\bar{x}_1 = 13.4$; $\mathrm{sd}(x_1) = 4.8$) and $x_2$ = gender (41% females). The estimation methods described in section 28.3.3 are easily

modified to handle the missing values implied by the two quarters interview gap. We generated missing values using the same response mechanisms as used for the simulation study (a single set of missing data under each of the three response mechanisms).

**Table 28.4.2: Empirical and estimated relative RMSE (%)**

| | | MCAR | | MAR | | NMAR | |
|---|---|---|---|---|---|---|---|
| No. | Imputation method | Estimate | Empirical | Estimate | Empirical | Estimate | Empirical |
| 1 | Mean imputation | 20.13 | 20.13 | 20.13 | 20.11 | 20.11 | 20.63 |
| 2 | Nearest neighbor | 14.77 | 14.79 | 16.24 | 16.53 | 15.19 | 15.96 |
| 3 | Simple regression | 16.20 | 16.22 | 16.19 | 16.17 | 16.18 | 16.81 |
| 4a | Augmented regression - individual | 12.18 | 12.14 | 13.56 | 13.58 | 12.85 | 13.84 |
| 4b | Augmented regression - household | 6.29 | 6.21 | 6.75 | 6.73 | 6.09 | 8.66 |
| 5a | State space - unsmoothed | 6.79 | 6.79 | 7.10 | 7.05 | 6.05 | 9.09 |
| 5b | State space - smoothed | 6.28 | 6.20 | 6.87 | 6.75 | 5.89 | 8.67 |

The model parameters were estimated by use of only the state-space modeling approach. We encountered convergence problems and negative variance estimators with the IGLS approach. For this reason we omit method 4b from the table (yields the same imputations as method 5b when using the same parameter values), whereas method 4a uses the parameter estimates obtained for method 5. The average relative bias and RMSE over all the missing values are shown in Table 28.4.3.

**Table 28.4.3: ILFS data set- relative bias and RMSE (%)**

| | | RELATIVE BIAS | | | RELATIVE RMSE | | |
|---|---|---|---|---|---|---|---|
| No. | Imputation method | MCAR | MAR | NMAR | MCAR | MAR | NMAR |
| 1 | Mean imputation | -1.83 | 21.95 | 74.18 | 32.57 | 45.50 | 81.61 |
| 2 | Nearest neighbor | 1.39 | -14.41 | 57.65 | 33.94 | 43.18 | 73.33 |
| 3 | Simple regression | -2.23 | 22.51 | 74.58 | 32.79 | 46.86 | 82.13 |
| 4a | Augmented regression – individual | -1.57 | 14.73 | 49.84 | 31.14 | 41.57 | 60.41 |
| 5a | State space – unsmoothed | 1.20 | 15.38 | 56.60 | 35.39 | 43.99 | 65.08 |
| 5b | State space – smoothed | 1.38 | 15.03 | 56.77 | 35.80 | 43.68 | 65.20 |

The results here are much more erratic than the results obtained for the simulated data. The large RMSE's can be explained by three main reasons: a) the fit of the model is no longer "perfect"; b) in almost all the households there is only one member so that the clear advantages noted in Table 28.4.1 from borrowing information within households cannot be utilized here; and c) the average sample size available for parameter estimation is almost 80% smaller than the sample size used for the simulation study ($567 \times 0.8$ compared to $2500 \times 0.8$). We mention in this regard that when estimating the model parameters for method 5b from all the data and not just the responses, the RMSE's are already reduced by about 40% under NCAR, 25% under MAR and 5% under NMAR.

## 28.5 RAMIFICATIONS AND EXTENSIONS

The model defined by (3.1.1)-(3.1.3) refers to the population measurements and the question arising is whether it holds equally for the sample data. As is often the case, the first and/or second level units are selected with unequal probabilities and when the selection probabilities are related to the values of the outcome variable, even after conditioning on the model explanatory variables, the model holding for the sample data could be distorted by the sampling process. For the two level model operating at given time points (equations 3.1.1, 3.1.4 and 3.1.5), Pfeffermann et *al.* (1998) propose a weighting procedure that accounts for the sampling effects and guarantees consistent estimators for all the model parameters. Feder et *al.* (2000) extend this procedure to the model underlying the present study by weighting the time series likelihoods defined by (3.3.1). Finally, we mention possible extensions to discrete or multivariate outcomes; and the problem of the robustness of the proposed imputation methods to model mis-specification.

### APPENDIX Model-based estimation of bias and MSE

Under the model, $\mathrm{V}(y_{hjt}) = \mathrm{V}(\mathbf{z}'_{ht}\mathbf{u}_{ht} + e_{hjt}) = C_{ht} + \sigma_e^2 = \mathbf{V}_{yht}$ where $C_{ht} = \mathbf{z}'_{ht}\mathbf{D}^*\mathbf{z}_{ht}$ $= \mathrm{Cov}(y_{hjt}, y_{hj't}),\ (j \neq j')$. Let the missing outcome be $y_{h_0 j_0 t}$.

1) Mean imputation: Define by $g_t$ the group to which unit $(h_0, j_0)$ belongs at time t. Let $m_{ght} (\leq n_h)$ be the number of individuals $j$ in household $h$ belonging to $g_t$ with mean value $\bar{y}_{ght}$. The group mean for imputation is $\bar{y}_{gt} = \sum_{h=1}^{N} m_{ght}\bar{y}_{ght} / m_{g.t}$ where $m_{g\cdot t} = \sum_{h=1}^{N} m_{ght}$ and $\mathrm{V}\left(\bar{y}_{gt}\right) = [\sum_{h=1}^{N} m_{ght}^2 C_{ht} / m_{g\cdot t}^2] + [\sigma_e^2 / m_{g\cdot t}]$. Let $\bar{\mathbf{x}}^*_{ght} = \sum_{j=1}^{n_h} \tilde{x}_{hjt} / m_{ght}$ and $\bar{\mathbf{x}}^*_{g\cdot t} = \sum_{h=1}^{N} m_{ght}\bar{\mathbf{x}}^*_{ght} / m_{g\cdot t}$ where $\tilde{x}_{hjt} = (x'_{hjt}, z'_{ht})'$. The imputation bias is:

$$\mathrm{B}_1(h_0, j_0, t) = \left(\tilde{\mathbf{x}}_{h_0 j_0 t} - \bar{\mathbf{x}}^*_{g\cdot t}\right)' \underline{\tilde{\beta}}_t; \qquad \text{(A.1)}$$

$$\mathrm{MSE}_1(h_0 j_0 t) = \mathbf{V}_{yh_0t} + \mathrm{V}(\bar{y}_{gt}) + \mathrm{B}_1^2(h_0 j_0 t). \qquad \text{(A.2)}$$

The variance computation assumes given imputation groups and $y_{h_0 j_0 t}$ independent of $\bar{y}_{gt}$.

For the remaining imputation methods there is no imputation bias under the model .

2) Nearest Neighbor: (imputed value defined by nearest observed value $y_{h_0 j_0 t^*}$ )

$$\mathrm{V}_2(h_0, j_0, t) = \mathbf{V}_{yh_0t^*} + \mathbf{V}_{yh_0t} - 2\mathbf{z}'_{h_0t}\mathbf{D}^*\mathbf{A}^{|t-t^*|}\mathbf{z}_{h_0t^*} - 2\sigma_e^2\rho^{|t-t^*|} \qquad \text{(A3)}$$

3) Simple regression:

$$\mathrm{V}_3(h_0, j_0, t) = \mathrm{E}(y_{h_0 j_0 t} - \tilde{\mathbf{x}}'_{h_0 j_0 t}\underline{\tilde{\beta}}_t)^2 = \mathbf{V}_{yh_0t} \qquad \text{(A.4)}$$

4) Augmented regression:

4a) Based on observed data for the same individual:

Let $\Lambda_{h_0 j_0} = \mathbf{Q}_{h_0 j_0}\mathbf{S}_{h_0}\mathbf{Q}'_{h_0 j_0}$ where $\mathbf{S}_{h_0} = \mathrm{V}[\tilde{\mathbf{Y}}_{h_0 j_0}]$. Denote by $\overline{\mathbf{q}}'_{h_0 j_0 t}$ the row of the missing indicator matrix $\overline{\mathbf{Q}}_{h_0 j_0}$ with 1 in column $t$ and define $\lambda'_{mt} = \overline{\mathbf{q}}'_{h_0 j_0 t}\tilde{\mathbf{S}}_{h_0 j_0}\mathbf{Q}'_{h_0 j_0}$.

$$\mathbf{V}_4^a(h_0, j_0, t) = \mathbf{V}_{yh_0t} + \lambda'_{mt}\Lambda^{-1}_{h_0 j_0}\lambda_{mt} - 2\overline{\mathbf{q}}'_{h_0 j_0 t}\mathbf{S}_{h_0 j_0}\mathbf{Q}'_{h_0 j_0}\Lambda^{-1}_{h_0 j_0}\lambda_{mt} \qquad \text{(A.5)}$$

4b) Based on all the household data:

Let $\Lambda_{h_0} = \mathbf{Q}_{h_0}\tilde{\mathbf{S}}_{h_0}\mathbf{Q}'_{h_0}$ where $\tilde{\mathbf{S}}_{h_0} = \mathrm{V}[\tilde{\mathbf{Y}}_{h_0}]$. Denote by $\overline{\mathbf{q}}^{*'}_{h_0 j_0 t}$ the row of $\overline{\mathbf{Q}}_{h_0}$ (indicator matrix for the missing household data) with 1 in the column corresponding to the missing value $y_{h_0 j_0 t}$ and zeroes elsewhere, and define $\lambda^{*'}_{mt} = \overline{\mathbf{q}}^{*'}_{h_0 j_0 t}\tilde{\mathbf{S}}_{h_0}\mathbf{Q}'_{h_0}$.

$$\mathrm{V}_4^b(h_0, j_0, t) = \mathbf{V}_{yh_0t} + \lambda^{*'}_{mt}\Lambda^{-1}_{h_0}\lambda^{*}_{mt} - 2\mathbf{q}^{*'}_{h_0 j_0 t}\tilde{\mathbf{S}}_{h_0}\mathbf{Q}'_{h_0}\Lambda^{-1}_{h_0}\lambda^{*}_{mt} \qquad \text{(A.6)}$$

5) State-space model imputation:

5a) Based only on current and past data (unsmoothed)

By (3.2.6): $\hat{y}^{(M)}_{h_0 j_0 t} - y_{h_0 j_0 t} = (\mathbf{z}'_{h_0 t}, 1)(\underline{\hat{\alpha}}_{h_0 t} - \underline{\alpha}_{h_0 t}) = \tilde{\mathbf{z}}'_{h_0 t}(\underline{\hat{\alpha}}_{h_0 t} - \underline{\alpha}_{h_0 t})$ where $\underline{\alpha}_{h_0 t} = (u'_{h_0 t}; e_{h_0 j_0 t})$. Hence,

$$\mathrm{V}_5^a(h_0, j_0, t) = \tilde{\mathbf{z}}'_{h_0 t}\mathbf{P}_{h_0 t}\tilde{\mathbf{z}}_{h_0 t}, \qquad \text{(A.7)}$$

where $\mathbf{P}_{h_0 t} = \mathrm{E}[(\underline{\alpha}_{h_0 t} - \underline{\hat{\alpha}}_{h_0 t})(\underline{\alpha}_{h_0 t} - \underline{\hat{\alpha}}_{h_0 t})']$ is obtained from the Kalman filter.

5b) Based on all the data for all the time periods (smoothed)

Same as in (A.8), with appropriate modficiations. The smoothed predictor predictor $\underline{\hat{\alpha}}^{sm}_{h_0 t}$ and the corresponding V-C matrix $\mathbf{P}^{sm}_{h_0 t}$ are obtained from the smoothing algorithm.

References

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure **.***The Annals of Statistics* **1,** 135-141.

Binder, D. A. (1998). Longitudinal surveys: why are these surveys different from all other surveys**?** *Survey Methodology* **24**, 101-108.

Chi, Eric M. and Reinsel, Gregory C., (1989). Models for longitudinal data with random effects and AR(1) errors **.***Journal of the American Statistical Association* **84**, 452-459.

Diggle, P. J., Liang, K-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Diggle, P. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis **.** *Applied Statistics* **93,** 43-49

Duncan, G. J. and Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review* **55**, 97-117.

Feder, M., Nathan, G. and Pfeffermann, D. (2000). Multilevel modelling of complex survey longitudinal data with time varying random effects. *Survey Methodology* **26,** 53-65.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43-56

Goldstein, H., Healy, M. J. R. and Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data **.** *Statistics in Medicine* **13**, 1643-1655.

Goldstein **,**H. (1995). *Multilevel Statistical Models* (2nd edition). NY: Halstead.

Harvey, A. C. (1989). *Forecasting, structural time series models, and the Kalman filter*. Cambridge: Cambridge University Press.

Harville, D.A. (1977). Maximum likelihood approaches to variancecomponent estimation and to related problems (with discussion), *Journal ofthe American Statistical Association* **72**, 320-340.

Jennrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices **,** *Biometrics* **42**, 805-820.

Jones, R. H. and Boadi-Boateng (1991). Unequally spaced longitudinal data with AR(1) serial correlation **.***Biometrics* **47,** 161-175.

Jones, R. H. and Vecchia, Aldo V. (1993). Fitting continuous ARMA models to unequally spaced spatial data. *Journal of the American Statistical Association* **88**, 947-954.

Jones, R. H. (1993). *Longitudinal Data with Serial Correlation A State-space Approach*. NY: Chapman

Kalton, G. (1986). Handling wave nonresponse in panel surveys **.***Journal of Official Statistics* **2,** 303- 314.

Kalton, G., Lepkowski, J. and Lin, T. (1985) Compensating for wave nonresponse in the 1979 ISDP research panel**.** *American Statistical Association, Proceedings of the Survey Research Methods Section*, 372- 377.

Kalton, G. and Miller, M. E. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *American Statistical Association, Proceedings of the Survey Research Methods Section*, 194- 199.

Laird, N. M. (1988). Missing data in longitudinal studies, *Statistics in Medicine* **7***,* 305-315.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data **,***Biometrics* **38**, 963-974.

Lepkowski, J. (1989). Treatment of wave nonresponse in panel surveys, **in** D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh (eds.), *Panel Surveys*. New York: Wiley, 348-374.

Little, R. J. A. (1982). Models for nonresponse in sample surveys **.** *Journal of the American Statistical Association* **77,** 150-237

Little, R. J. A. (1995). Modeling the dropout mechanism in repeated measures studies **.***Journal of the American Statistical Association* **90**, 1112-1121.

Little, R. J. A. and Rubin, D. B. (1987) .*Statistical analysis with missing data* .New York: Wiley.

Murphy, S. and Li, B. (1995). Projected partial likelihood and its application to longitudinal data. *Biometrika* **82**, 399-406.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random .*Journal of the American Statistical Association* **92**, 1320–1329.

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based prediction .*Journal of the American Statistical Association* **83**, 824–833.

Pfeffermann, D., Skinner, C., Goldstein, H., Holmes, D. J. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical Society* **B 60,** 23-40.

Rotnizky, A., Robins, J. M. and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

Schafer, J.L. and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association* **95**, 144-154.

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

Rubin, D. B. (1987). *Multiple imputation* .New York: Wiley.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes .*Biometrics* **42,** 121-130.