STATISTICS IN MEDICINE
Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

A Residuals-Based Transition Model for Longitudinal Analysis with Estimation in the Presence of Missing Data

Tulay Koru-Sengul[1], David S. Stoffer*[2], and Nancy L. Day[3]

[1] *Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[2] *Departments of Statistics and Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA*
and
[3] *Departments of Psychiatry, Pediatrics and Epidemiology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA*

SUMMARY

We propose a transition model for analyzing data from complex longitudinal studies. Because missing values are practically unavoidable in large longitudinal studies, we also present a two-stage imputation method for handling general patterns of missing values on both the outcome and the covariates by combining multiple imputation with stochastic regression imputation. Our model is a time-varying autoregression on the past innovations (residuals), and it can be used in cases where general dynamics must be taken into account, and where model selection is important. The entire estimation process can be carried out using available procedures in statistical packages such as SAS and S-PLUS. To illustrate the viability of the proposed model and the two-stage imputation method, we analyze data collected in an epidemiological study that focused on various factors relating to childhood growth. Finally we present a simulation study to investigate the behavior of our two-stage imputation procedure.

KEY WORDS: Incomplete data; Innovations sequence; Longitudinal analysis; Missing data; Multiple imputation; Stochastic regression imputation; Time-varying autoregression.

*Correspondence to: Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA stoffer@pitt.edu

Contract/grant sponsor: National Science Foundation; contract/grant number: DMS0405038
Contract/grant sponsor: National Institute on Alcohol Abuse and Alcoholism; contract/grant number: AA006666
Contract/grant sponsor: National Institute on Drug Abuse; contract/grant number: DA003874

Received
Revised

1. INTRODUCTION

There are many different modeling approaches for analyzing longitudinal studies and the choice of a model depends, of course, on the desired goal of the study. In this article, we present a transition model that can be useful in longitudinal studies where the data are collected at unequally spaced time intervals and where the dynamics can change at any time point. Second, we present methodology for the estimation of the model parameters when there are missing observations (both responses and covariates). We suggest using a certain type of autoregressive model with time-varying parameters in the case where the general dynamics must be taken into account, but where careful investigation of covariates that change over time and may be periodically missed is also important. In this case, model fitting should be simple enough to concentrate on model selection while still accounting for the longitudinal nature of the data. Other considerations are contending with incomplete records over time and a large number of observations to process. The advantage of our proposed model is that it is general, it allows for time-varying conditional variances, and it can handle observations taken during periods of instability, at irregularly spaced intervals and with missing data. Our proposed method for handling missing at random data is a composite of imputation methods that are described in Little and Rubin [6]. Specifically, we suggest a two-stage procedure using multiple imputation for the covariates and then using stochastic regression imputation for the responses.

We will present the details of our model in Section 2. Because our model is a form of time-varying autoregression, we first give some background on the use of time series models in longitudinal analysis. Autoregressive models with exogenous variables (ARX) are well established and have been used in medicine for some time (e.g., Rosner, Muñoz, Tager, Speizer, and Weiss [7]; Rosner and Muñoz [8, 9]; Zeger and Qaqish [16]; Schmid [11, 12, 13]; Icaza and Jones [2]). For example, Rosner *et al.* [7] studied the stationary transition model for equally spaced longitudinal data of the form

$$y_i(t) = \sum_{\ell=1}^{L} \phi_\ell \, y_i(t-\ell) + \sum_{j=1}^{J} \beta_j \, x_{ij}(t) + \sum_{r=1}^{R} \alpha_r \, z_{ir} + \epsilon_i(t),$$

where $y_i(t)$ is the outcome of subject i at time t, $x_{ij}(t)$ is the j-th time-dependent covariate of subject i at time t, z_{ir} is the r-th time-independent covariate for subject i, and the $\epsilon_i(t)$ are iid (across time and subject) $N(0, \sigma^2)$ errors. Efficient estimation of the model parameters can be accomplised via ordinary least squares, and this is an advantage to using such a model. Rosner and Muñoz [8] extended the model for missing and unequally spaced longitudinal data using weighted nonlinear regression. They suggested using a linear interpolation for missing covariates, however, such interpolation does not include estimation of the variability.

Schmid [11] wrote the stationary transition model in state-space form and suggested using the EM algorithm in the presence of normally distributed outcomes and covariates that are missing at random. For his model, the parameters are estimated using the EM algorithm in conjunction with the Kalman filter and smoother, a technique that was proposed by Shumway and Stoffer [14]. Missing values on discrete or other non-Gaussian variables are difficult to handle using these algorithms. In a somewhat related idea, we also mention that Jones [3] showed how to use the state-space model and Kalman filter recursions to fit random effects model (Laird and Ware [4] to longitudinal data. In the case of normal observations, an explicit EM algorithm can be formulated based on the Kalman filter and smoother (see Shumway and Stoffer [14] or Icaza and Jones [2]). The method is developed for equally spaced longitudinal data with missing values, and unequally spaced with different observation times for different subjects. The same parameter estimates can also be obtained by using the SAS procedure MIXED with special variance-covariance structures on the repeated measurements (Icaza and Jones [2]). It is also claimed that the methods may work for handling missing continuous covariates.

A problem with complex longitudinal studies is the inevitability of missing covariates and missing outcomes. Fitting longitudinal models in the presence of missing data has its own challenges because of the various ways in which data may be missed. For general patterns of missing data, multiple imputation (Little and Rubin [6]) appears to be the best method, and it is the method that we propose for fitting our model. We will discuss the details of our two-stage imputation procedure that combines multiple imputation and stochastic regression imputation in Section 3, after we introduce the model in the following section.

2. OUR MODEL

Before presenting our model, we give some motivation. Suppose we observe outcomes $y_i(t_k)$ on individual i, for $i = 1, \dots, N$, at times t_k for $k = 0, 1, \dots, n$. In addition, suppose we observe one time-varying covariate, $x_i(t_k)$. If the data are sampled at irregular time intervals, or if the stability of the dynamics is in question, then, for $k \geq 1$, we might propose a time-varying regression with autocorrelated errors, say:

$$y_i(t_k) = \beta(t_k)x_i(t_k) + u_i(t_k) \tag{1}$$

$$u_i(t_k) = \rho(t_{k-1})u_i(t_{k-1}) + \epsilon_i(t_k), \tag{2}$$

where initially, $y_i(t_0) = \beta(t_0)x_i(t_0) + \epsilon_i(t_0)$. Here, we can allow $\{\epsilon_i(t_k); k = 0, 1, \dots, n\}$ to be an uncorrelated sequence with time-varying variance, $\text{var}\{\epsilon_i(t_k)\} = \sigma_k^2$. However, because we can write $u_i(t_{k-1}) = y_i(t_{k-1}) - \beta(t_{k-1})x_i(t_{k-1})$, for $k \geq 1$, the model (1)-(2) is

$$\begin{aligned} y_i(t_k) &= \beta(t_k)x_i(t_k) + \rho(t_{k-1})\{y_i(t_{k-1}) - \beta(t_{k-1})x_i(t_{k-1})\} + \epsilon_i(t_k) & (3) \\ &= \rho(t_{k-1})y_i(t_{k-1}) + \beta(t_k)x_i(t_k) + \gamma(t_{k-1})x_i(t_{k-1}) + \epsilon_i(t_k) & (4) \end{aligned}$$

where we have written $\gamma(t_{k-1}) = -\rho(t_{k-1})\beta(t_{k-1})$. Model (4) is in the form of a typical time-varying ARX model.

Now, suppose in (1) we were interested in the effect of the covariate at the previous time point on the current response. That is, suppose in (1) we had

$$y_i(t_k) = \beta(t_k)x_i(t_k) + \delta(t_k)x_i(t_{k-1}) + u_i(t_k). \tag{5}$$

Then, (4) would be

$$\begin{aligned} y_i(t_k) &= \rho(t_{k-1})y_i(t_{k-1}) + \beta(t_k)x_i(t_k) + \big\{\delta(t_k) + \gamma(t_{k-1})\big\}x_i(t_{k-1}) \\ &\quad + \theta(t_{k-1})x_i(t_{k-2}) + \epsilon_i(t_k), & (6) \end{aligned}$$

where $\theta(t_{k-1}) = -\rho(t_{k-1})\delta(t_{k-1})$ This causes an identifiability problem because we cannot distinguish the effect of $x_i(t_{k-1})$ on the current response. For example, if $x_i(t_{k-1})$ were significant in (6), we would not know if it was because the covariate was influencing the outcome at the previous time point ($\gamma(t_{k-1}) \neq 0$), at the current time point ($\delta(t_k) \neq 0$), or both.

To overcome this problem, we suggest using an innovations (residuals) form of the model. That is, in (2), we replace $u_i(t_{k-1})$ by

$$\widehat{u}_i(t_{k-1}) = y_i(t_{k-1}) - \widehat{\beta}(t_{k-1})x_i(t_{k-1}) = y_i(t_{k-1}) - \widehat{y}_i(t_{k-1}),$$

where $\widehat{\beta}(t_{k-1})$ is the estimate obtained from the regression of $y_i(t_{k-1})$ on $x_i(t_{k-1})$. Thus, it is seen from (3) that the model (1)-(2) will now be

$$y_i(t_k) = \beta(t_k)x_i(t_k) + \rho(t_{k-1})\{y_i(t_{k-1}) - \widehat{y}_i(t_{k-1})\} + \epsilon_i(t_k). \tag{7}$$

Hence, including the covariate from a previous time point, as in (5), is no longer a problem. So, for example, using the innovations form of the model, (6) would now be

$$y_i(t_k) = \rho(t_{k-1})\big\{y_i(t_{k-1}) - \widehat{y}_i(t_{k-1})\big\} + \beta(t_k)x_i(t_k) + \delta(t_k)x_i(t_{k-1}) + \epsilon_i(t_k), \tag{8}$$

which separates the effect of $x_i(t_{k-1})$ on the current response.

We are now ready to specify the general model. As in the motivating example, suppose we observe outcomes $y_i(t_k)$ on individual i, for $i = 1, \ldots, N$, at times t_k for $k = 0, 1, \ldots, n$. In addition, suppose we observe time dependent covariates, $x_{ij}(t_k)$, for $j = 1, \ldots, J$ and time independent covariates z_{ir} for $r = 1, \ldots, R$. Then, the basic general model at time t_k, is of the form

$$y_i(t_k) = \sum_{\ell=1}^{p_k} \phi(t_{k-\ell})\big\{y_i(t_{k-\ell}) - \widehat{y}_i(t_{k-\ell})\big\} + \sum_{\ell=0}^{q_k}\sum_{j=1}^{J} \beta_j(t_{k-\ell})x_{ij}(t_{k-\ell}) + \sum_{r=1}^{R} \alpha_{r,k} z_{ir} + \epsilon_i(t_k), \tag{9}$$

where $\widehat{y}_i(t_{k-\ell})$, for $\ell = 1, \ldots, p_k$, denotes the predicted values from previous fits. Note that the orders of the regression on the innovations, $p_k \leq k$, and the regression on the time dependent covariates, $q_k \leq k$, are allowed to change with time. In addition, J or R in (9) may change with k, although we have not shown this fact explicitly. Note that the regression parameters are allowed to change with time. We further assume the $\epsilon_i(t_k)$ are normal errors that are independent across subject i, and time t_k, with time-varying variance, $\text{var}\{\epsilon_i(t_k)\} = \sigma_k^2$. The independence of the errors across time is not restrictive because the innovations will account for serial correlation as discussed in the motivation for the model. A simple example of the utility of this model was given in Shumway and Stoffer ([15], Example 6.23).

The advantage to (9) is that it can be fit sequentially in time using ordinary least squares. Moreover, observations taken at future time points can be processed without having to reanalyze the entire data set. Of course, adjustments will have to be made if any observations are missing, and we discuss this problem in the next section.

3. A TWO-STAGE IMPUTATION METHOD: MI-SRI

In the case of missing (at random) covariates as well as responses, we recommend a two-stage imputation method that combines ideas from multiple imputation (MI) and stochastic regression imputation (SRI); see Little and Rubin ([6], chapters 4 and 5) for details on each technique. Our strategy is as follows:

1. In the first stage, MI is applied to missing covariates by creating M data sets with complete covariates whose outcome values are missing.
2. In the second stage, the M data sets that were created by MI are used to impute the missing outcome values by using model (9) for the SRI.

The first step of the two-stage procedure uses MI for the missing covariates only; note that the outcomes are excluded from this step. The advantage of this step is that it can be performed easily using standard statistical packages, e.g., the SAS procedure MI, or the“missing” library in S-PLUS, to mention two. In the data analysis we present in the next section, we will assume a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance covariance matrix Σ augmented with an informative ridge prior; details on using an informative ridge prior can be found online at `http://support.sas.com/rnd/app/da/new/802ce/stat/chap9/sect20.htm` and also in Schafer ([10], p. 152). We use the multivariate normality assumption in an MCMC method that constructs a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution. By repeatedly simulating steps of the chain, simulated draws from the distribution of interest can be obtained; see Schafer ([10], Section 5.4) for details. Categorical and patently non-normal covariates can be multiply imputed under the normality assumption. For example, imputations can be done to guarantee the range of a variable stays within the support of that variable. After imputation, imputed values can be rounded off to an appropriate value. This technique has been shown to be fairly robust in Schafer ([10], Chapter 6). We note that we are being overly general here, and that some time-varying covariates that are missed may be treated differently if obvious relationships are known. At the end of this step we will have M "completed" (via MI) sets of covariates.

The second step involves SRI wherein missing responses are imputed by their predicted values from the model of the "completed" covariates obtained from the previous step, with the addition of a random term. The random term is obtained from the distribution of the current, "completed" innovations. This technique is performed for each of the M imputed sets of covariates to introduce sampling variability. Specifically, we use model (9) as an imputation model in SRI after imputing missing covariates by MI. For example, suppose $y_i(t_k)$ is missing. Then on the m-th imputed value, for $m = 1, \ldots, M$, we set

$$\begin{aligned} y_i^{(m)}(t_k) &= \sum_{\ell=1}^{p_k} \widehat{\phi}^{(m)}(t_{k-\ell})\, \widehat{\epsilon}_i^{(m)}(t_{k-\ell}) + \sum_{\ell=0}^{q_k} \sum_{j=1}^{J} \widehat{\beta}_j^{(m)}(t_{k-\ell})\, x_{ij}^{(m)}(t_{k-\ell}) \\ &+ \sum_{r=1}^{R} \widehat{\alpha}_{r,k}^{(m)}\, z_{ir}^{(m)} + \widehat{\epsilon^*}_i^{(m)}(t_k), \end{aligned} \tag{10}$$

where $x_{ij}^{(m)}(t_{k-\ell})$ and $z_{ir}^{(m)}$ denote the possibly imputed covariates obtained from the first, or MI, step. The values $\widehat{\epsilon}_i^{(m)}(t_{k-\ell})$ in (10) denote the possibly imputed innovations; these are obtained from previous SRI steps if necessary. The value $\widehat{\epsilon^*}_i^{(m)}(t_k)$ is a random draw from a normal distribution with mean zero and standard deviation $\widehat{\sigma}_k^{(m)}$. We note that in (10), e.g., if $x_{ij}(t_{k-\ell})$ is observed, then $x_{ij}^{(m)}(t_{k-\ell}) = x_{ij}(t_{k-\ell})$ for all m. In general, using SRI alone does not affect the mean level of the estimates but it tends to underestimate the standard errors and inflate the correlations. In Section 5, however, we present evidence to support the case that SRI, when used in conjunction with MI, does not suffer from these problems.

After the two-stage imputation process has been completed, parameter estimation is accomplished using Rubin's method of combining estimates from the multiply imputed data sets (see Little and Rubin [6]).

4. AN APPLICATION

To illustrate the viability of the model (9) and the two-stage MI-SRI imputation method, we analyzed data from an epidemiological study at the University of Pittsburgh that focused on various factors related to childhood growth, as described in Larkby [5]; for design details, see Day, Wagener, and Taylor [1]. As in Larkby [5], we focus on the growth of $N = 694$ children followed from birth to six years of age. In this study, the children were examined at birth ($t_0 = 0$), at eight months ($t_1 = 8$), 18 months ($t_2 = 18$), 36 months ($t_3 = 36$), and 72 months ($t_4 = 72$) of age. The response, $y_i(t_k)$, for $i = 1, ..., 694$ and $k = 0, 1, 2, 3, 4$, is a growth index, which is essentially a standardized score for a child's weight adjusted for that child's age, gender, and height, against the national averages.

Among the 694 children in the study, 353 (51%) are black. Because the growth index is adjusted for age and gender, race [R] is the only time-independent covariate that we consider. The other covariates we focus on are time-dependent; these covariates are current maternal cigarette smoking [S] measured in average number cigarettes per day, current maternal alcohol use [A] measured in average number of drinks per day, the number of illnesses [I] the child had since the last visit, the number of times the child was hospitalized [H] since the last visit, whether or not the child was breastfed [BF], and a measure of child's nutrition [RDA] that was assessed at the final two time points. Maternal cigarette smoking and alcohol use were also assessed prenatally, and they will be used in the initial regression for birthweight. The covariates in the study are also listed in Table I. In particular, note that only maternal smoking [S] and alcohol use [A] were assessed at all five of the time points. The number of illness [I] and number of hospitalizations [H] were not assessed at birth. The nutrition variable [RDA] was assessed at 36 and 72 months, and breastfeeding [BF] was assessed only at 8 months. Race

[R] was assessed once. Thus, there are a total of 22 covariates for the entire study. Of the 694 children, only 225 have complete records with regard to these variables.

There are a number of statistical issues that need to be addressed. For example, the data collection involves unequally spaced time points. As previously indicated, there are a varying number of covariates that are assessed at each time point (e.g., RDA is assessed only at the final two time points, BF will not be a factor by the third time point, and so on). The main interest is in how covariates measured early in the study affect the outcome when the children are older (e.g., the effect of prenatal smoking or alcohol use on growth at 3 years, or 6 years, of age); model (9) can help here. In this particular study there is a general pattern of missing data for both the growth outcome and the covariates. An obvious problem is when a mother-child pair miss an interview, in which case the response as well as all covariates are not assessed. There were many instances in this study, however, when a mother was interviewed but the child was not weighed (no response) or the child was weighed, but certain maternal factors were not assessed.

In our analysis, missing values are handled by the two-stage MI-SRI imputation method discussed in Section 3. The MI step was performed using the SAS procedure MI with an informative ridge prior on Σ of $p = 0.75$ (we found our results to be fairly insensitive to the choice of the ridge prior). The number of imputed values was $M = 10$. Note that because there are 22 covariates for the full study, the multivariate normal distribution used the MI step has 22 dimensions. For each of the imputed data sets, we performed the SRI step as detailed in (10), based on the following models. Initially, the model at birth is a regression of standardized birthweight [Y0] on race [R], prenatal smoking [S0], prenatal alcohol use [A0] and the interactions with race [S0:R and A0:R]. We will denote this model by:

$$\text{Y0} \sim \text{R} + \text{S0} + \text{A0} + \text{S0:R} + \text{A0:R} \tag{11}$$

From the initial regression we obtain the innovations, $\widehat{\epsilon}_i(0) = y_i(0) - \widehat{y}_i(0)$, for $i = 1, \ldots, 694$, where $y_i(0)$ is the standardized birthweight of the i-th child, and $\widehat{y}_i(0)$ is the standardized birthweight predicted from the regression. For the next time point, $t_1 = 8$ months, the model was (using similar notation with Y8 for standardized weight at 8 months and Inn0 for the innovations from the initial fit)

$$\begin{aligned} \text{Y8} \quad \sim \quad & \text{Inn0} + \text{R} + \text{S0} + \text{A0} + \text{S8} + \text{A8} + \text{I8} + \text{H8} + \text{BF8} \\ & + \text{S0:R} + \text{A0:R} + \text{S8:R} + \text{A8:R} + \text{I8:R} + \text{H8:R} + \text{BF8:R} \end{aligned} \tag{12}$$

This modeling procedure was continued in this way to the assessment at 72 months. Each regression included all of the previous innovations and the covariates (and their interactions with race) from each time point.

The results of the analysis are displayed in Table II; only the final models are displayed. Prenatal smoking and race are significant at birth. On average, children born to mothers who smoked were smaller; white children tend to be bigger. There does not appear to be a prenatal alcohol effect at birth, however, prenatal alcohol use affects growth at other assessments to 36 months of age. This result is remarkable because alcohol use in this study refers only to moderate alcohol use. For example, of the women who drank prenatally, the mean alcohol measurement is 0.23 drinks per day, and the maximum is 2.05 drinks per day. Interestingly, the effect of prenatal smoking is gone by 8 months of age. Finally, by the age of 72 months, none of the previously significant predictors of growth are significant.

5. SIMULATION STUDY

In this section we compare the two-stage missing data technique, MI-SRI, described in Section 3 with the technique of multiply imputing all (MI-ALL) the missing data, covariates and outcomes, at the same time.

Instead of creating hypothetical data sets for the simulations, we used the data described in Section 4, but where we had complete records, across all time points, of standardized weight (response), maternal smoking [S] and maternal alcohol use [A] (time-varying covariates), and race [R] (time independent covariate). Of the 694 subjects analyzed in Section 4, only 396 had complete records with regard to these variables. We considered these data as the "complete" case and the parameter estimates obtained from the regressions using the "complete" data are considered the “true” parameter estimates. Then, using the “complete” data, we created 100 hypothetical data sets each with a 5% general pattern of missing data for maternal smoking [S] and for maternal alcohol use [A]. Specifically, for the “complete” data set, let $S_i(t)$ and $A_i(t)$ denote the maternal smoking and maternal alcohol use covariates for subject i at month t, respectively. Using a random number generator, we removed (with probability) 5% of the covariates in the set $\{S_i(t) : i = 1, \ldots, 396; t = 0, 8, 18, 36, 72\}$ and (with probability) 5% of the covariates in the set $\{A_i(t) : i = 1, \ldots, 396; t = 0, 8, 18, 36, 72\}$ from the “complete” data set. This process was repeated 100 times. Then, we repeated this removal scheme in the same manner, with 10% and 20% general patterns of missing observations. In each case, the number of imputations was set to $M = 10$.

The top row of Figure 1 compares the average of the coefficient estimates in the models when data are missing and MI-SRI is used (vertical axis) to the corresponding, actual coefficient estimates, based on the complete data set (horizontal axis). The bottom row of Figure 1 is a similar comparison, but using the MI-ALL method instead. Figure 2 is similar to Figure 1, but shows the estimated standard errors of the estimates based on both techniques. Figures 1 and 2 include all the regression coefficients and their standard errors obtained from five full models, one for each time point, that include race [R], and all the corresponding smoking [S] and alcohol [A] variables, and their interactions. Specifically, these models are the models listed in Table I, but where the covariates related to variables other than race, smoking and alcohol use are not used. In other words, the models we used in the simulations are those that would be obtained by going through Table I and retaining only the variables that begin with the letter R, S, or A. The simulation results for each of the time points are not plotted individually, but they are collapsed into one graph.

From Figure 1, we see that both techniques, MI-SRI and MI-ALL, provide accurate parameter estimates in the presence of missing data. From Figure 2, however, we see that while MI-SRI is accurate in assessing standard errors, the MI-ALL technique tends to overestimate standard errors. Recall that MI-ALL uses a multivariate normal distribution for both the response and the covariates in the imputation model, whereas MI-SRI uses a multivariate normal distribution only for the covariates in the imputation model and a regression model for the response. Since the missing response values are imputed by using a functional relationship among the covariates and the response via a regression model, the MI-SRI method uses more information than the MI-ALL method.

Finally, we mention that, in the simulations, we tried various values for the ridge prior in the MI part of both procedures, MI-SRI and MI-ALL. While the results benefited from the use of a ridge prior, the results were fairly robust to the choice of the actual value chosen. In all our simulations, as in the application discussed in the previous section, we used the $p = .75$ setting in the specification of the ridge prior on Σ in the SAS procedure MI.

6. DISCUSSION

In this paper we introduced a transition model for analyzing data from longitudinal studies. By regressing on the innovation sequence, we are able to separate the effect of past values of a covariate with the current outcome. In addition, because complex longitudinal studies will inevitably be plagued with missing observations, we have devised a two-stage (MI-SRI) procedure for the estimation of the model parameters. In addition, fitting this model to complex longitudinal studies is simple because it can be accomplished using least squares and imputation routines from standard statistical packages such as SAS or S-PLUS with the addition of some simple programming to complete the SRI step. SAS code for the procedure may be obtained from the first author. We showed the viability of the model by using the model and estimation procedure to analyze data from a complex longitudinal study with a large number of missing observations. Finally, we presented a simulation study to show that the proposed estimation method can give reasonable and accurate estimates of both regression parameters and their standard errors.

ACKNOWLEDGEMENTS

The authors thank an Associate Editor and the referees for providing comments and suggestions that led to improvements to the presentation.

REFERENCES

1. Day, N., Wagener, D., and Taylor, P. (1985). Measurement of Substance Use During Pregnancy: Methodological Issues. In *Prenatal Drug Exposure and Consequences of Maternal Drug Use,* TM Pinkert (ed.). NIDA Research Monograph, No.59, 36–40.
2. Icaza, G., and Jones, H. J. (1999). A state space EM algorithm for longitudinal data. *Journal of Time Series Analysis,* **20,** 537–550.
3. Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State Space Approach.* London: Chapman & Hall
4. Laird, N. M., and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics,* **38,** 963–974.
5. Larkby, C. (1998). *Psychosocial Factors Related to Childhood Growth.* Unpublished Ph.D. Dissertation, University of Pittsburgh.
6. Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis of Missing Data.* New York: Wiley and Sons.
7. Rosner, B., Muñoz, A., Tager, I., Speizer, F., and Weiss, S. (1985). The use of an autoregressive model for the analysis of longitudinal data in epidemiology studies. *Statistics in Medicine,* **4,** 457–467.
8. Rosner, B., and Muñoz, A. (1988). Autoregressive modeling for the analysis of longitudinal data with unequally spaced examinations. *Statistics in Medicine,* **7,** 59–71.
9. Rosner, B., and Muñoz, A. (1992). Conditional Linear Models for Longitudinal Data. In *Statistical Models for Longitudinal Studies of Health.* Dwyer, J., Feinleib, M., Lippert, P., Hoffmeister, H. (eds.). Oxford University Press.
10. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman & Hall.
11. Schmid, C. H. (1996). An EM algorithm fitting of first-order conditional autoregressive models to longitudinal data. *Journal of the American Statistical Association,* **91,** 1322–1330.
12. Schmid, C. H. (1999). Regression models for longitudinal data with missing covariate values. *American Statistical Association, Proceedings of the Section on Statistics in Epidemiology,* 25–34.
13. Schmid, C. H. (2001). Marginal and dynamic regression models for longitudinal data. *Statistics in Medicine,* **20,** 3295–3311.
14. Shumway, R. H., and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis,* **3,** 253–264.
15. Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications: With R Examples* (2nd edn). New York: Springer-Verlag
16. Zeger, S. L., and Qaqish, B. (1988). Markov regression models for time series: A quasi- likelihood approach. *Biometrics,* **44,** 1019–1031.

Table I. List of the covariates and interactions that were included in the regression models ($n = 694$).

Assessment Age	Covariates[†]	Interactions
Birth	R, S0, A0	S0:R, A0:R
8 months	R, S0, A0, Inn0, S8, A8, I8, H8, BF8	S0:R, A0:R, S8:R, A8:R, I8:R, H8:R, BF8:R
18 months	R, S0, A0, Inn0, S8, A8, I8, H8, BF8, Inn8 S18, A18, I18, H18	S0:R, A0:R, S8:R, A8:R, I8:R, H8:R, BF8:R, S18:R, A18:R, I18:R, H18:R
36 months	R, S0, A0, Inn0, S8, A8, I8, H8, BF8, Inn8, S18, A18, I18, H18, Inn18, S36, A36, I36, H36, RDA36	S0:R, A0:R, S8:R, A8:R, I8:R, H8:R, BF8:R, S18:R, A18:R, I18:R, H18:R, S36:R, A36:R, I36:R, H36:R, RDA36:R
72 months	R, S0, A0, Inn0, S8, A8, I8, H8, BF8, Inn8, S18, A18, I18, H18, Inn18, S36, A36, I36, H36, RDA36, Inn36 S72, A72, I72, H72, RDA72	S0:R, A0:R, S8:R, A8:R, I8:R, H8:R, BF8:R, S18:R, A18:R, I18:R, H18:R, S36:R, A36:R, I36:R, H36:R, RDA36:R, S72:R, A72:R, I72:R, H72:R, RDA72:R

[†] R is race coded as 0 for black and 1 for white. The time varying covariates (for $t = 0, 8, 18, 36, 72$ months) are S for maternal cigarette smoking (measured as the average number of cigarettes smoked per day), A for maternal alcohol use (measured as the average number of drinks consumed per day), I for the number of child illnesses, H for the number of child hospitalizations, BF for whether or not the child was breastfed (assessed at 8 months only), RDA for a measure of the child's nutrition (ascertained only at 36 and 72 months), and Inn is the innovation (residual). The number after the letter denotes the period (e.g., S8 is the average daily number of cigarettes smoked by the mother during the child's first 8 months of age, and S18 would refer to the maternal consumption of cigarettes from age 8 months to 18 months). The two-way interactions are denoted by a colon (e.g., S0:R is the prenatal smoking–race interaction).

Table II. Final models and combined parameter estimates in the time-varying analysis of data from a longitudinal study focusing on various factors related to childhood growth ($n = 694$).

Age at Assessment	$\widehat{\sigma}_k$	Covariate	$\widehat{\beta}$	se($\widehat{\beta}$)	p-value
Birth	0.453	prenatal smoking	−0.01	0.002	0.0001
		race†	0.27	0.036	0.0001
8 months	0.943	prenatal alcohol	−0.44	0.19	0.024
		innovation(0)	0.20	0.081	0.016
18 months	1.049	prenatal alcohol	−0.50	0.20	0.014
		race	0.18	0.082	0.029
		innovation(8)	0.64	0.043	0.0001
		innovation(0)	0.33	0.089	0.0001
36 months	0.765	prenatal alcohol	−0.36	0.13	0.006
		alcohol(36)	0.22	0.076	0.004
		illness(36)	0.06	0.026	0.010
		race	0.12	0.068	0.074
		alcohol(36)×race	−0.23	0.105	0.030
		innovation(18)	0.26	0.028	0.0001
		innovation(8)	0.38	0.031	0.0001
		innovation(0)	0.39	0.066	0.0001
72 months	1.148	innovation(36)	0.83	0.058	0.0001
		innovation(18)	0.35	0.044	0.0001
		innovation(8)	0.38	0.047	0.0001
		innovation(0)	0.25	0.097	0.011

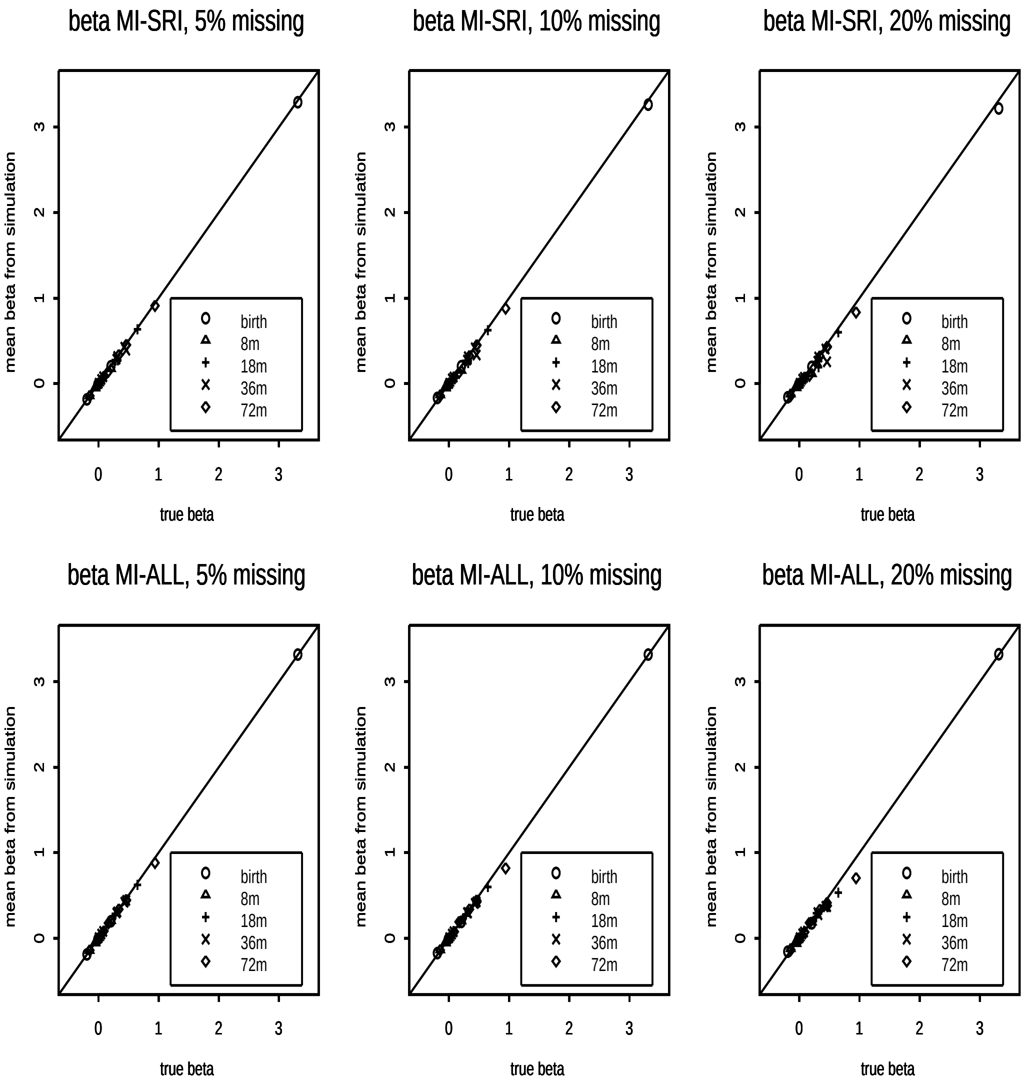
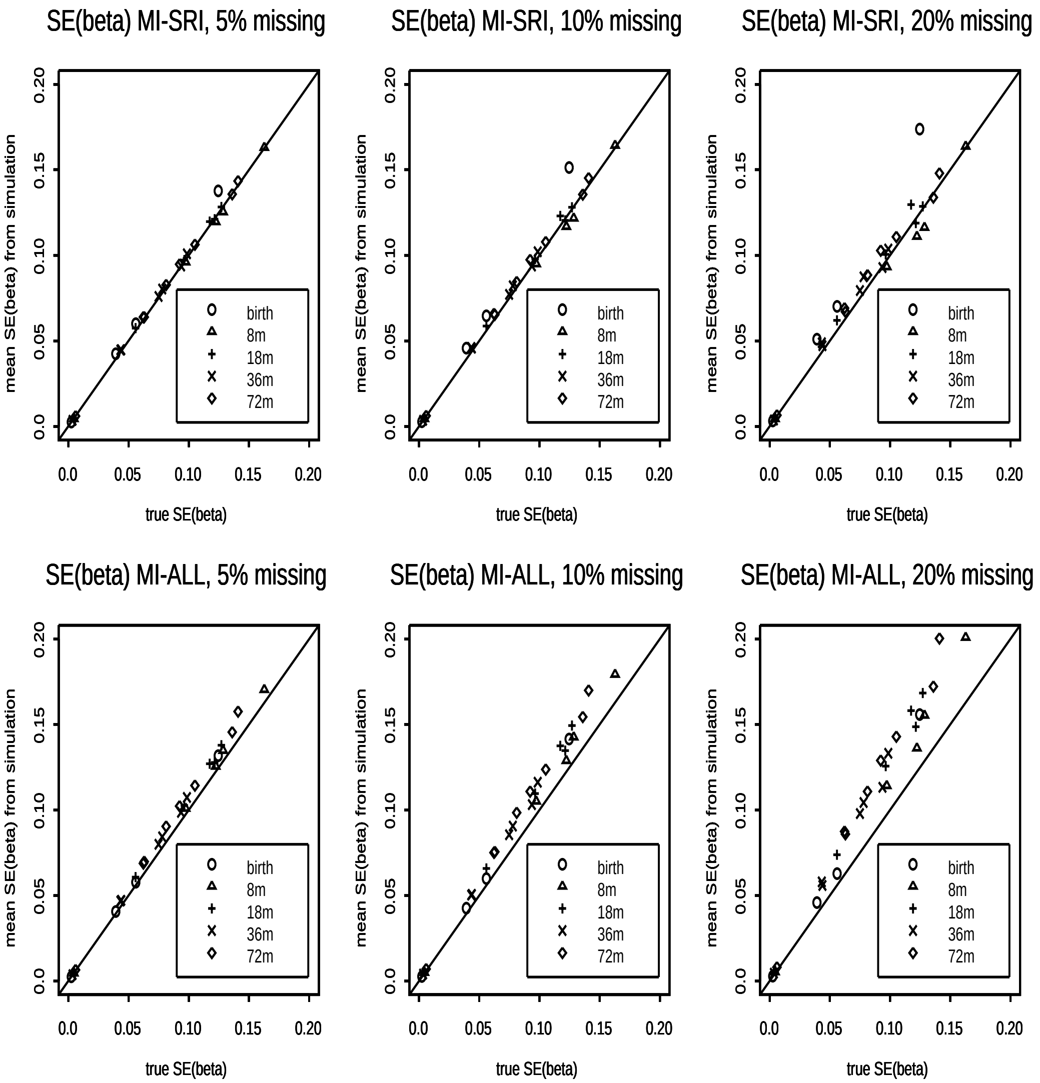


Figure 1. *Top Row:* Comparison of the average coefficient estimates in the model on the vertical axis when data are missing and MI-SRI is used to the corresponding, actual coefficient estimates, based on the complete data set, on the horizontal axis. *Bottom Row:* Similar to the top row but for the MI-ALL technique.



SE(beta) MI-SRI, 5% missing

SE(beta) MI-SRI, 10% missing

SE(beta) MI-SRI, 20% missing

SE(beta) MI-ALL, 5% missing

SE(beta) MI-ALL, 10% missing

SE(beta) MI-ALL, 20% missing

Figure 2. *Top Row:* Comparison of the average estimated standard errors of the coefficient estimates in the model on the vertical axis when data are missing and MI-SRI is used to the corresponding, actual estimated standard errors of the coefficient estimates, based on the complete data set, on the horizontal axis. *Bottom Row:* Similar to the top row but for the MI-ALL technique.