Methodologic Issues in Linked Cross–Sectional Studies: Experiences from the Muscatine Study

William R. Clarke and Robert F. Woolson, University of Iowa
William R. Clarke, 2800 SB, University of Iowa, Iowa City, IA 52242

Abstract

The Muscatine Study is a study of coronary risk factors in school age children. The study consists of a series of biennial surveys of the school age population in Muscatine, Iowa. The study began in 1971 and includes 6 biennial surveys in which information is collected on each child including: height, weight, skin fold thickness, systolic and diastolic blood pressure, and other hemo–dynamic and anthropometric characteristics. One goal of the Muscatine Study is to generate growth curves for blood pressure, height and weight for children of both sexes.

The Muscatine Study is an example of a linked cross–sectional study (Rao and Rao, 1966) where several cross–sectional surveys of the population are taken. Individuals in an earlier survey are often present in later surveys and it is important to link this longitudinal information. There are a number of difficulties and methodologic issues which must be dealt with, both in data analysis and in the design of such studies. When the focus of the investigation is to generate growth curves over time, there are distinct advantages to those approaches of analysis which use the longitudinal information. In this paper we examine and apply various data analytic procedures for the estimation of these growth curves over time. In particular, we discuss a two–stage procedure which may be utilized in the estimation of, and fitting of, such growth curves. In addition, we discuss some considerations related to the design of linked cross–sectional studies.

Introduction

In this paper we describe the Muscatine Study which may be viewed as a linked cross–sectional study in the sense of Rao and Rao (1966). A linked cross–sectional study is an investigation in which a series of cross–sectional surveys are linked together in a manner which incorporates the longitudinal information associated with those individuals who are present in more than one survey. Data analysis for such studies is often complicated by the attempt to recover this longitudinal information, yet it is important to recover this information to provide efficient estimates of growth rates and other quantities which are of primary interest.

First, we discuss some efficiency issues related to the design of linked cross–sectional studies. We then describe the Muscatine Study and discuss a two–stage estimation procedure which we have found of use for the analysis of these data. Following this, we present an example which utilizes the Muscatine data and in particular, focuses on the systolic blood pressures and the heights of the individuals. It is shown that this two–stage procedure is empirically more efficient than various cross–sectional analyses which may be applied to these data. Next, we consider an example using echocardiographic measurements of left ventricular mass, and once again illustrate the manner in which the two–stage procedure might be utilized to estimate mean levels of the response and changes in these mean levels over time.

While linked cross–sectional studies have several advantages compared to purely longitudinal studies, the data analysis can become considerably more complicated. There is a need for further work on procedures of analysis of such data, and on the optimal design of such linked cross–sectional studies.

Design Issues

Linked cross–sectional studies clearly have some practical advantages over purely longitudinal studies. In longitudinal studies in which children are to be followed from say age 6 to age 18 it is necessary to maintain the cohort for a period of 13 years. In some situations it is beneficial to study individuals over a shorter time period then link the longitudinal information from these shorter time periods together. For instance, it might be possible to study individuals in 4 successive years and choose at the outset individuals who are age 6 through 15. In this manner those individuals initially at age 6 would be observed at age 7, 8 and 9, while those initially at age 15 would be observed at ages 15, 16, 17, and 18. Information from these four ages could then be used to construct the longitudinal pattern for each group, then linked together in order to generate the desired growth curves. Such a study design would shorten the time period of follow–up for each child and would decrease the difficulties in maintaining the cohort since it would eliminate the need to follow a single cohort over a 13 year interval of time.

How efficient are such linked cross–sectional designs and in what circumstances might they be used to advantage? These questions have been addressed by several investigators including, Rao and Rao (1966), Machin (1975), and Woolson and Leeper (1980). Rao and Rao consider a very special situation when there are only two ages and where it is desired to estimate age specific means and growth rates, i.e., changes in these means. They examine this problem under the assumption that the variances are equal at the two ages. By letting π denote the fraction of cross–sectional information, they consider studies where a total of $n(1-\pi)$ individuals contribute data at both ages while $n\pi$ contribute data only at the first age, and another $n\pi$ contribute data only at the second age.

The major objective in Rao and Rao's study is to determine the optimal choice of π. Optimum is defined by minimizing the variance of specific estimators. Their method of analysis is called a linked cross–sectional analysis and is comparable to the two–stage method described later in this paper. One begins by transforming the complete pairs by doing a square–root decomposition of the covariance matrix and applying this transform to the complete pairs in the obvious way. The incomplete pairs are simply scaled by the inverse of the standard deviation, σ, and all data are then set into a single linear model in which the parameter vector of interest is (μ_1, μ_2) (i.e., the age–specific means).

Carrying out the linked cross–sectional least squares analysis one then obtains the covariance matrix for the estimator $\hat{\mu}$. This covariance matrix depends on the original variance σ^2, the total sample size n, the correlation ρ, and the quantity π. The choice of π then depends on the objective of interest.

If interest is principally in estimation of the age–specific means then the variance for each of these means may be written in a closed form as $\frac{\sigma^2}{n}(1 - \pi\rho^2)/(1 - \pi^2\rho^2)$, which explicitly represents the contributions of π and σ^2. If one determines the minimum value of this variance with respect to π, it is easy to show that this minimum value is attained when $\pi = 1/(1 + \sqrt{1-\rho^2})$ and depends only on the correlation. Clearly, with high correlation one should take a large fraction of the data to be cross–sectional. For example at ρ=.95, the fraction of the data that should be cross–sectional is .762. On the other hand, if interest is in estimation of growth rates then one must consider the variance of the estimated mean difference which is $\frac{2\sigma^2}{n}\left[\frac{1-\rho}{1-\pi\rho}\right]$. The minimum value of this quantity is obtained at π=0, that is when there is no cross–sectional information. Hence, the fraction of cross–sectional information which should be obtained depends very heavily on the study goals and in this case, whether one is interested primarily in estimation of age–specific means or in growth rates. For most practical problems, one is interested in both of these quantities.

In Table 1 we show the estimated precision that goes along with the different optimal values of π, where these optimal values are determined by minimizing the variance of the estimated age–specific mean. From this very simple example one concludes that the choice of π depends on the study goals (i.e., estimation of norms or rates), and also depends on the underlying covariance matrix. Most importantly, the optimal value of π is usually neither 0 nor 1; therefore, linked cross–sectional studies are generally preferred to either purely longitudinal or purely cross–sectional studies.

Considering more than two age groups, and in general letting p denote the number of time points of interest, it is possible to study and compare the efficiency of various linked cross–sectional designs to a purely longitudinal design. Once again as in the simpler case, when there are simply two ages the underlying variance for the estimators depends on the assumed covariance matrix Σ. In previously published work (Woolson and Leeper, 1980) we have studied such designs and have considered s disjoint subsets of the data, where $m=\frac{p}{s}$ represents the number of ages for each of the subsets. Under the assumptions of a first order auto–regressive covariance structure, one can determine the relative efficiency of the longitudinal and linked cross–sectional study designs by using generalized variance as the comparison criterion. These comparisons have been generated in Table 2 and show the relative efficiencies of various linked cross–sectional designs compared to a purely longitudinal design. The figures in Table 2 are for an assumed linear growth curve model over time and hence, two parameters (i.e., intercept and slope) are being estimated by the two specific designs. In the cases examined these results indicate that for positive correlations the linked cross–sectional study is more efficient than the purely longitudinal and the efficiency is increased by increasing s. For instance, for ρ=0.8 the value of the relative efficiency ranges from 1.4 to 3.63 as s goes from 2 to 6. Very clearly linked cross–sectional studies have some advantage relative to the purely longitudinal study design in the presence of high correlations. When correlations are 0 there is no advantage in one design or the other, but for positive correlation there is generally some advantage in having a linked cross–sectional rather than a purely longitudinal study design. With negative correlations the reverse pattern holds. It is, however, quite unusual to have a first order auto–regressive pattern with a negative correlation, and one would not expect to see such patterns with biologic data. Once again we should emphasize that these results depend on the underlying Σ.

Thus, from a design standpoint linked cross–sectional studies have a number of advantages as compared to purely longitudinal investigations. They offer the possibility of shortening the time period of follow–up for individuals, thereby minimizing the loss due to attrition which usually accompanies longitudinal studies of long duration. In certain select circumstances linked cross–sectional studies also increase the efficiency in estimation of the primary parameters of interest. Hence, these designs are clearly worthy of further study and may be of great use in epidemiologic settings.

A Two–Stage Generalized Least Squares Estimation Procedure

There are a number of general estimation procedures and approaches for the treatment of incomplete longitudinal data. We have found the procedures of Kleinbaum (1970, 1973) quite useful for certain questions in the analysis of the Muscatine study. This project may be characterized as a linked cross–sectional study with 6 surveys, 14 ages represented in each survey, and no person having measurements at more than six ages. In addition, there are a very large number of individuals in the study (10,000+) and numerous missing data patterns. We first describe a general two–stage procedure we have used and programmed for estimating growth norms and growth rates in this study then illustrate with examples in the following sections. The methods are based on the work of Zellner (1962, 1963), Kleinbaum (1973), Fairclough and Helms (1985) and Leeper and Woolson (1982).

We assume that there are p ages for which data have been collected. Some individuals have been measured at one age only, others at exactly two ages, etc. The number of times measured and the ages at measurement uniquely determine a data pattern for each child. For convenience, we assume that there are s such distinct patterns of data. In addition, for persons in pattern k, it is assumed that they have data at exactly p_k of the p ages ($p_k \le p$).

Ordering the p ages from 1 to p, the ages at which data are available for those in pattern k will be denoted by a $p_k$xp indicator matrix $K^{(k)}$ in which each row has one '1' and p–1 '0's. As an example, if a pattern is characterized by individuals having data at ages 1, 2 and 4 then the K matrix is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

We denote the response variable, e.g. systolic blood pressure, by $y_{ij}^{(k)}$ for the i^{th} child at age j in pattern k. Note that $y_{ij}^{(k)}$ is unknown if the child was not measured at age j. For simplicity we assume that the "complete data" (with * for missing y's) are placed into an N x p array Y where the first n_1 rows correspond to missing data pattern 1, the next n_2 to missing pattern 2, etc. Further let $Y = [\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_p]$ where $\underset{\sim}{Y}_j$ is the array of data (complete and incomplete) at age j. Let X_j denote the model matrix corresponding to $\underset{\sim}{Y}_j$ where rows of X_j are assumed to be zeros for the unobserved (missing) individuals at that age. The two–stage estimation procedure follows the lines of Zellner (1962, 1963) where we first estimate Σ from residuals determined by performing an ordinary least squares regression analysis of $\dot{Y}_j$ on $\dot{X}_j$ where $\dot{Y}_j$ and $\dot{X}_j$ are the corresponding observed parts of Y_j and X_j. This is done for each of the p ages. If the data are missing at random then we have the model

$$\dot{Y}_j = \dot{X}_j \beta_j + \dot{e}_j \qquad \text{for } j = 1,2,\ldots,p \quad . \tag{2}$$

The ordinary least squares estimator of $\hat{\beta}_j$

$$\hat{\beta}_j = (\dot{X}_j' \dot{X}_j)^{-1} \dot{X}_j' \dot{Y}_j \qquad \text{for } j = 1,\ldots,p \quad . \tag{3}$$

Residuals from (3) are

$$r_j = Y_j - X_j \hat{\beta}_j \qquad \text{for } j = 1,\ldots,p$$

where it is understood that the residuals are undefined for the incomplete components of Y_j. An estimator of Σ is $S = (s_{jj'})$ where

$$s_{jj'} = r_j' r_{j'} / (N_{jj'} - 1) \tag{4}$$

and $N_{jj'}$ is the number of individuals with observations at both ages j and j′. Also it is assumed that $r_j' r_{j'}$ is computed over only those individuals with data at both ages j and j′.

Expression (4) is but one estimator of Σ and is analogous to the "restricted" estimator discussed by Revankar (1976) for complete data. It should be noted that while Σ is positive definite, it does not follow that S is. For the second stage of the estimation we require that $K^{(k)} S K^{(k)\prime}$ be positive semi–definite for k=1,...,s. An alternative path to follow is to estimate $K^{(k)} \Sigma K^{(k)\prime}$ from each of the patterns of missing data. An estimator of $K^{(k)} \Sigma K^{(k)\prime}$ could be obtained by first performing an OLS regression of $y_j^{(k)}$ on $X_j^{(k)}$, where $y_j^{(k)}$ is the vector of observed values at age j for missing data pattern k and $X_j^{(k)}$ is the corresponding design matrix. Denote the resulting estimator for β_j by $\hat{\beta}_j^{(k)}$. Then residuals for missing data pattern k can be computed by $r_j^{(k)} = (y_j^{(k)} - X_j^{(k)} \hat{\beta}_j^{(k)})$. An estimator $S^{(k)} = (s_{jj'}^{(k)})$ for $K^{(k)} \Sigma K^{(k)\prime}$ can be computed as

$$s_{jj'}^{(k)} = \frac{r_j^{(k)\prime} r_{j'}^{(k)}}{(n_k - 1)} \qquad \text{for } j, j' = 1,\ldots,p \quad . \tag{5}$$

Hence, at least two procedures can be employed to generate an estimator of $K^{(k)} \Sigma K^{(k)\prime}$. While both procedures lead to consistent estimators of $K^{(k)} \Sigma K^{(k)\prime}$, (4) has the advantage of using all the available data, thus being more efficient; however, (4) does not always yield a positive definite estimator of $K^{(k)} \Sigma K^{(k)\prime}$. The second procedure (5) always yields a positive definite estimator of $K^{(k)} \Sigma K^{(k)\prime}$ if $n_k > \max(\text{rank } X_1^{(k)}, \ldots, \text{rank } X_p^{(k)})$; although, this procedure may not be as efficient as (4), and the estimators may have higher variability. If an estimator $\hat{\Sigma}^{(k)}$ is not positive semi–definite, then a smoothed estimate may be obtained through the procedure of Schwertman and Allen (1973).

We should also add that other estimation procedures can also be considered. If Σ is assumed to have a certain pattern, e.g., Σ may be of the repeated measures form of $\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \vdots & & & \vdots \\ \rho & \ldots & & 1 \end{bmatrix}$, or it may be a time series auto–regressive model, then other estimators would be appropriate. These matters will not be addressed in this paper.

The estimator for β that we propose is a generalized least squares estimator and is a natural extension of the usual two–stage estimator of Zellner (1963, 1963). To define this estimator, it is useful to first define the vector $z^{(k)}$ as the column–concatenation of the $n_k p_k$–vector in missing data pattern k. The vector of y's for the first child is stacked in column form on top of that for the second child, and so forth. Hence, $z^{(k)}$ is a vector of length $n_k p_k$. Let $A^{(k)}$ be the model matrix corresponding to $z^{(k)}$ so that $E(z^{(k)}) = A^{(k)} \beta$ and

$$E((z^{(k)} - A^{(k)} \beta)(z^{(k)} - A^{(k)} \beta)') = K^{(k)} \Sigma K^{(k)\prime} \otimes I$$

for k=1,...,s,

where $\otimes$ is the Kronecker (direct) product. Since the s groups of data corresponding to the patterns are independent of one another, it follows that the weighted least squares estimator of β is:

$$\tilde{\beta} = \left[\sum_{k=1}^{s} \left[A^{(k)\prime} (K^{(k)} \Sigma K^{(k)\prime} \otimes I)^{-1} A^{(k)} \right] \right]^{-1} \cdot$$

$$\sum_{k=1}^{s} A^{(k)\prime} (K^{(k)} \Sigma K^{(k)\prime} \otimes I)^{-1} z^{(k)}.$$

Since Σ, and therefore $K^{(k)} \Sigma K^{(k)\prime}$, is unknown, we propose estimating β by $\hat{\beta}$ where $\hat{\beta}$ is $\tilde{\beta}$ with

$K^{(k)} \Sigma K^{(k)\prime}$ estimated by $\hat{\Sigma}^{(k)}$, one of the estimators described earlier.

The estimator of β which we propose is

$$\hat{\beta} = \left[\sum_{k=1}^{s} \left[A^{(k)\prime} (\hat{\Sigma}^{(k)} \otimes I)^{-1} A^{(k)} \right] \right]^{-1} \cdot$$

$$\sum_{k=1}^{s} A^{(k)\prime} (\hat{\Sigma}^{(k)} \otimes I)^{-1} z^{(k)} , \qquad (6)$$

where $\hat{\Sigma}^{(k)}$ is a consistent estimator of $K^{(k)} \Sigma K^{(k)\prime}$.

If $\hat{\Sigma}^{(k)}$ is consistent as an estimator of $K^{(k)} \Sigma K^{(k)\prime}$, it follows that $\hat{\beta}$ is a consistent estimator of β. In addition, if the rows of the original Y matrix arise from a multivariate normal distribution, then $z^{(k)}$ is multivariate normal and β is multivariate normal. Accordingly, $\hat{\beta}$ is asymptotically multivariate normal. In particular, $\hat{\beta}$ is approximately normal with a mean vector of β and a covariance matrix estimated by

$$\left[\sum_{k=1}^{s} A^{(k)\prime} (\hat{\Sigma}^{(k)} \otimes I)^{-1} A^{(k)} \right]^{-1} . \qquad (7)$$

Linear functions of β, say $C\beta$, can be estimated by $C\hat{\beta}$ and would have a covariance matrix estimated by

$$C \left[\sum_{k=1}^{s} A^{(k)\prime} (\hat{\Sigma}^{(k)} \otimes I)^{-1} A^{(k)} \right]^{-1} C' .$$

It may be of interest to model the components of β as a function of age, and C can be chosen to do this. For example, one can model the response variable as a linear function of age and test the hypothesis that this model is adequate. In general, to test hypotheses of the form H_0: $C\beta = 0$ the statistic

$$(C\hat{\beta})' \left[C \left[\sum_{k=1}^{s} A^{(k)\prime} (\hat{\Sigma}^{(k)} \times I)^{-1} A^{(k)} \right]^{-1} C' \right]^{-1} (C\hat{\beta}) \quad (8)$$

may be compared to a chi–square statistic with c degrees of freedom, where c is the row rank of C.

Applications of the Two–Stage Procedure

Muscatine Study Height and Systolic Blood Pressure

The Muscatine Coronary Risk Factor study, a linked cross–sectional study of coronary risk factors in school children, began in 1971 and between 1971 and 1981 six biennial surveys were completed. An additional survey of grades 1 to 3 was conducted in 1974. Only children who were enrolled in school during the year of a survey were eligible to participate and about 70 percent of those eligible actually participated. School–leavers were no longer eligible and two new classes became eligible for each survey; there were many patterns of participation over the 11 years of the study.

Height, systolic blood pressure and other variables were measured on each child participating in a survey. For this illustration, we analyze the height (cm) and systolic blood pressure (mmHg) as a function of age. We restrict attention to females in this analysis.

Over 4,500 girls are included in this analysis. Some of these children participated in only one survey while others participated in all six surveys. The distribution of children by the number of surveys in which they participated is presented in Table 3. As stated earlier, there are a number of reasons why children have participated in fewer than six survey years. Some children did not reach school age until 1981 and were eligible for only one survey year. Others graduated and were no longer eligible. Finally, the 30% non–participation could be regarded as randomly missing data (Clarke, et al. 1978). For the 14 ages considered, there were over 400 different patterns of incomplete data represented in the data set.

Table 4 exhibits the age–survey–year–specific mean and standard errors for height and systolic blood pressure. The final column is a summary collapsed across all survey years; there were approximately 500 observations in this summary ("All") for ages 5, 15, 16, 17 and 18, while there were over 1,000 for each of the ages 6–14 inclusive. The overall cross–sectional summary fails to take into account that the variance–covariance matrix is not estimated properly with such a cross–sectional analysis.

Table 5 displays the two–stage estimates and their standard errors along with the overall cross–sectional and 1981 data only estimates. The "1981 only" analysis is a valid purely cross–sectional analysis while the "All cross–sectional surveys" analysis is not valid since it ignores the correlations between repeated observations on the same individual. The "All surveys" analysis is included for comparison purposes only and indicates what one might expect from a purely cross–sectional study of the same magnitude as the linked cross–sectional study.

The lower standard errors for the two–stage estimators reflect the utilization of the correlation between repeated observations. The correlation estimates between observations separated by two years is given in the last column of Table 5. Correlations for height are quite strong. The correlations for both variables are clearly too large to be ignored. Using all available longitudinal and cross–section information and accounting for the correlation between repeated observations yielded substantial improvements in the precision of the estimates of the growth norms.

We also estimated the two–year growth rates and their standard errors for the "1981 survey only," the "all cross–sectional surveys," and the two–stage method. These results are also displayed on Table 5. As one would predict, even greater improvement in the precision of the estimates was observed for two–year growth rates than for the age specific growth norms.

Muscatine Study Echocardiographic Left Ventricular Mass

It has been hypothesized that cardiac hypertrophy in children (large heart relative to body size) is related to high blood pressure. As part of a study to examine the relationship between left ventricular mass (LV Mass) and blood pressure in children it was necessary to develop growth norms for LV mass during the childhood years. Two groups of children had their LV mass estimated from standard M–mode echocardiograms (Mahoney, et al.). A group of children aged 10 to 18 had echocardiograms taken in 1979, 1981, 1982, and

1983. Those who graduated from high school during this period were no longer considered eligible for study. A second group of children aged 6 to 10 were measured on 4 successive years between 1980 and 1983. We report the results of the analyses only for males.

Notice that this study has allowed the researchers to telescope time. In this case the linked cross–sectional design has compressed a 13 year study into 5 years while taking advantage of the gain in precision from a longitudinal study.

Table 6 displays several characteristics of these data. There were 224 boys who were examined at least once. By design and random nonparticipation there were 73 different sampling patterns. Over one–third of males were sampled four times while a nearly equal number were sampled only once. There were 508 observations in all. Table 7 displays the observed correlations between repeated measures of LV mass. Note that these correlations are mostly high and obviously need to be considered in any analyses.

Table 8 displays the cross–sectional (ignoring correlations) and linked cross–sectional (two–stage) estimates of mean LV mass and their standard errors. Except at ages 6 and 18 (where there is very little data), the standard errors for the two–stage analysis are comparable or smaller than the standard errors for the cross–sectional analisis. In addition, the two–stage analysis has somewhat smoothed the growth curve.

Table 9 displays the estimates of one–year growth rates and their standard errors. As in the previous example there is notable improvement in the precision of the estimates from the two–stage method except where the sample sizes were small and there was very little longitudinal information.

Examination of the results for both growth norms and growth rates suggest that it would be interesting to model both phenomena. The two–stage method accommodates these analyses very easily by either modifying the design matrix in the two–stage procedure or by analyzing the vector of estimates using weighted least squares with the estimated covariance matrix serving as the weighting matrix.

Summary

Linked cross–sectional studies are a practical and efficient method for establishing growth norms and studying changes with time. We have shown that this design is not only practicial but in many real situations it is the design of choice. The two–stage method is also a practical way of analyzing data from these studies. It is computationally feasible even in very large studies, it can be implemented using standard statistical computer packages, and it can yield substantial improvements in the precision of the estimates of growth norms and growth rates.

There are several other methods for analyzing data from linked cross–sectional studies. Comparisons of the properties of these procedures offers a rich field for further statistical investigation.

Acknowledgements

This research supported in part by USPHS Grant #HL14230–17 and USPHS Grant #CA39065–04.

The authors also gratefully acknowledge Ms. Candy Hodges for her patience in typing and editing numerous revisions of this paper.

References:

1. Clarke, W.R., Schrott, H.G., Leaverton, P.E., Connor, W.E., and Lauer, R.M. (1978): Tracking of blood lipids and blood pressures in school age children: The Muscatine study. Circulation, 58(4):626–634.

2. Fairclough, D.L., and Helms, R.W. (1985): A sensitivity analysis of maximum likelihood estimators in a mixed model for analysis of incomplete and mistimed longitudinal data. Presented at the Interface Symposium, Lexington, KY.

3. Kleinbaum, D.G. (1970): Estimation and hypothesis testing for generalized multivariate linear models. Unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill, Department of Statistics.

4. Kleinbaum, D.G. (1973): A generalization of the growth curve model which allows missing data. Journal of Multivariate Analysis, 3:117–124.

5. Leeper, J.D., and Woolson, R.F. (1982): Testing hypothesis for the growth curve model when the data are incomplete. Journal of Statistical Computation and Simulation, 15:97–106

6. Machin, D. (1975): On design problems and growth studies. Biometrics, 31(3):749–755.

7. Mahoney, L.T., Schicken, R.W., Clarke, W.R., Lauer, R.M.: Childhood left ventricular mass and exercise responses predict future blood pressure: The Muscatine study. Hypertension (in press).

8. Rao, M.N. and Rao, C.R. (1966): Linked cross–sectional studies for determining norms and growth rates – pilot survey on Indian school–going boys. Sankhya, Series B, 28:237–258.

9. Revankar, N.S. (1976): Use of restricted residuals in SUR systems: Some finite sample results. Journal of the American Statistical Association, 71:183–188.

10. Schwertman, N.C., and Allen, D.M. (1973): The smoothing of an indefinite matrix with applications to growth curve analysis with missing observations. Technical Report No. 56, University of Kentucky, Department of Statistics.

11. Woolson, R.F. and Leeper, J.D. (1980): Growth curve analysis of complete and incomplete longitudinal data. Communications In Statistics, A9(14):1491–1513.

12. Zellner, A. (1962): An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. Journal of the American Statistical Association, 57:348–368.

13. Zellner, A. (1963): Estimators of seemingly unrelated regression equations: Some exact finite sample results. Journal of the American Statistical Association, 58:977–992.

Table 1: Optimum Value of π for Given ρ for Estimating Age–Specific Means; Precision of Estimated Means and of Growth Rate Estimates

Optimum		Precision of	
ρ	π	Mean*	Growth Rate**
0	.500	1.000	2.000
.1	.501	.997	1.895
.3	.512	.977	1.654
.5	.536	.933	1.366
.7	.583	.857	1.014
.9	.696	.718	.535
.95	.762	.656	.362
.99	.875	.571	.150

*: $(1 - \pi\rho^2)/(1 - \pi^2\rho^2)$
**: $2(1 - \rho)/(1 - \pi\rho)$

Table 2. Relative Efficiencies of Some Incomplete Longitudinal Designs Compared to a Longitudinal Design

		ρ							
p	s	-0.9	-0.5	0.0	0.2	0.5	0.8	0.9	0.95
6	2	0.81	0.88	1.00	1.07	1.20	1.40	1.49	1.54
	3	0.51	0.68	1.00	1.19	1.54	1.89	1.92	1.89
	6	0.01	0.18	1.00	1.64	2.95	3.63	2.72	1.98
12	2	0.91	0.94	1.00	1.04	1.14	1.36	1.52	1.64
	3	0.77	0.84	1.00	1.10	1.36	1.88	2.16	2.32
	4	0.63	0.74	1.00	1.18	1.63	2.53	2.91	3.02
	6	0.38	0.55	1.00	1.34	2.27	4.14	4.70	4.51
	12	0.00	0.14	1.00	1.91	4.89	11.49	12.31	9.54
24	2	0.96	0.97	1.00	1.02	1.08	1.24	1.41	1.57
	3	0.89	0.92	1.00	1.05	1.20	1.62	1.99	2.30
	4	0.82	0.87	1.00	1.09	1.35	2.09	2.71	3.15
	6	0.68	0.77	1.00	1.17	1.68	3.25	4.52	5.22
	8	0.55	0.67	1.00	1.26	2.06	4.70	6.81	7.77
	12	0.32	0.50	1.00	1.44	2.94	8.42	12.84	14.31
	24	0.00	0.12	1.00	2.07	6.54	26.19	42.36	45.40

Woolson & Leeper (1980, Comm. In Statistics)

Table 3: Number of Individuals Sampled by Frequency of Sampling

Number of Times Sampled	Number of Individuals	Number of Samples
1	2095	2095
2	1078	2156
3	618	1854
4	411	1644
5	366	1830
6	116	696
Total	4684	10275

Table 4. Means and Standard Errors For Height and Systolic Blood Pressure by Age and Survey Year For Females

	Survey Year															
	1971		1973		1974		1975		1977		1979		1981		All	
Age	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
a) Height (cm)																
5	126.55	9.65	106.50	8.90	112.81	0.53	111.90	0.66	113.48	0.61	112.69	0.59	113.36	0.58	112.90	0.27
6	119.45	0.47	119.73	0.57	118.52	0.44	116.61	0.40	116.74	0.39	117.04	0.40	116.82	0.37	117.56	0.16
7	123.96	0.41	123.67	0.38	121.78	0.67	123.81	0.47	122.82	0.43	123.94	0.46	122.95	0.46	123.44	0.17
8	129.96	0.32	130.18	0.43	124.57	1.79	128.44	0.44	128.48	0.41	129.06	0.41	129.34	0.45	129.32	0.16
9	135.32	0.34	135.01	0.42	--	--	135.00	0.42	135.34	0.46	135.25	0.48	135.38	0.43	135.22	0.17
10	140.79	0.46	141.08	0.46	--	--	141.72	0.48	140.64	0.48	141.02	0.47	141.14	0.47	141.06	0.19
11	147.72	0.45	148.45	0.50	--	--	147.15	0.48	146.85	0.49	148.49	0.55	148.16	0.59	147.78	0.21
12	154.51	0.50	153.57	0.51	--	--	154.45	0.45	155.28	0.47	152.93	0.57	154.07	0.49	154.19	0.20
13	158.47	0.51	159.02	0.43	--	--	158.68	0.47	158.09	0.49	158.64	0.46	159.23	0.48	158.68	0.19
14	161.41	0.40	161.77	0.45	--	--	161.31	0.45	160.99	0.43	161.91	0.53	161.15	0.52	161.42	0.19
15	162.09	0.47	163.12	0.53	--	--	163.14	0.42	162.30	0.49	162.95	0.51	161.70	0.54	162.58	0.20
16	162.78	0.52	163.33	0.60	--	--	163.93	0.51	163.53	0.54	163.60	0.45	163.34	0.61	163.43	0.22
17	164.11	0.54	162.71	0.72	--	--	164.17	0.64	164.48	0.52	163.31	0.56	163.12	0.71	163.76	0.25
18	162.98	0.79	162.62	1.50	--	--	163.27	1.04	163.94	0.80	164.37	0.71	164.28	1.05	163.75	0.37
b) Systolic Blood Pressure (mmHg)																
5	107.50	0.50	93.00	3.00	91.40	1.25	91.43	1.35	96.39	1.05	96.11	1.13	96.27	1.17	94.30	0.55
6	99.92	0.86	100.51	1.14	93.64	0.93	95.61	0.81	96.97	0.62	94.56	0.76	98.28	0.65	96.76	0.31
7	100.83	0.75	101.12	0.70	96.33	1.43	98.77	0.90	98.32	0.68	95.95	0.82	98.77	0.86	98.96	0.32
8	104.56	0.61	103.11	0.87	91.33	10.73	97.85	0.89	100.01	0.71	96.19	0.61	98.46	0.75	100.52	0.31
9	107.24	0.68	105.64	0.88	--	--	101.71	0.76	102.47	0.84	99.30	0.77	101.41	0.70	103.45	0.33
10	109.49	0.81	108.95	0.79	--	--	102.49	0.81	103.23	0.79	102.15	0.73	100.93	0.65	104.86	0.33
11	114.53	0.85	110.55	0.89	--	--	106.93	0.88	107.54	0.81	107.09	0.88	104.78	0.78	108.87	0.36
12	117.81	1.00	110.89	0.92	--	--	110.07	0.81	108.27	0.76	107.02	0.99	108.51	0.81	110.56	0.37
13	119.68	0.97	115.25	0.82	--	--	112.99	0.99	112.86	0.83	109.61	0.83	111.33	0.91	113.86	0.38
14	120.14	1.00	116.62	0.89	--	--	113.35	0.88	112.99	0.73	110.08	0.84	108.53	0.80	113.98	0.38
15	117.29	1.00	116.27	0.93	--	--	113.73	0.85	114.96	0.79	113.16	0.93	109.06	1.00	114.18	0.38
16	116.05	1.06	114.05	1.24	--	--	114.18	0.86	114.71	0.81	113.23	0.96	111.58	1.06	114.01	0.41
17	116.41	1.20	115.84	1.74	--	--	114.33	1.17	115.70	1.00	113.21	0.93	112.91	1.23	114.71	0.48
18	122.06	2.17	115.20	3.15	--	--	108.88	1.65	114.78	1.32	112.80	1.21	112.62	1.85	114.38	0.73

Table 5. Summary Statistics for Three Analysis: Age-Specific Means, Standard Errors and Two-Year Differences by Height and Systolic Blood Pressure for Females

	1981 Cross-Sectional Survey				All Cross-Sectional Surveys				Two-Stage Estimates				Stage I
			Two-Year Differences				Two-Year Differences				Two-Year Differences		Two-Year Correlations
Age	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
a) Height (cm)													
5	113.36	0.58			112.90	0.27			111.36	0.10			
6	116.82	0.37			117.56	0.16			117.03	0.11			
7	122.95	0.46	9.59	0.74	123.44	0.17	10.54	0.32	123.35	0.10	11.99	0.05	0.87
8	129.34	0.45	12.52	0.58	129.32	0.16	11.76	0.23	129.43	0.11	12.40	0.11	0.88
9	135.38	0.43	12.43	0.63	135.22	0.17	11.78	0.24	135.26	0.12	11.91	0.10	0.92
10	141.14	0.47	11.80	0.65	141.06	0.19	11.74	0.25	140.56	0.14	11.13	0.12	0.91
11	148.16	0.59	12.78	0.73	147.78	0.21	12.56	0.27	148.32	0.16	13.06	0.13	0.90
12	154.07	0.49	12.93	0.68	154.19	0.20	13.13	0.28	153.20	0.14	12.64	0.10	0.91
13	159.23	0.48	11.07	0.76	158.68	0.19	10.90	0.28	158.93	0.15	10.61	0.15	0.86
14	161.15	0.52	7.08	0.71	161.42	0.19	7.23	0.28	161.60	0.13	8.40	0.13	0.82
15	161.70	0.54	2.47	0.72	162.58	0.20	3.90	0.28	162.58	0.19	3.65	0.14	0.89
16	163.34	0.61	2.19	0.80	163.43	0.22	2.01	0.29	163.39	0.17	1.79	0.11	0.95
17	163.12	0.71	1.42	0.89	163.76	0.25	1.18	0.32	163.44	0.27	0.86	0.32	0.96
18	164.28	1.05	0.94	1.21	163.75	0.37	0.32	0.43	164.20	0.45	0.81	0.50	0.99
b) Systolic Blood Pressure (mmHg)													
5	96.27	1.17			94.30	0.55			95.33	0.55			
6	98.28	0.65			96.76	0.31			97.65	0.31			
7	98.77	0.86	2.50	1.45	98.96	0.32	4.66	0.64	99.52	0.32	4.19	0.60	0.42
8	98.46	0.75	0.18	0.99	100.52	0.31	3.76	0.44	101.22	0.32	3.57	0.42	0.38
9	101.41	0.70	2.64	1.11	103.45	0.33	4.49	0.46	103.98	0.34	4.46	0.42	0.41
10	100.93	0.65	2.47	0.99	104.86	0.33	4.34	0.45	105.20	0.35	3.97	0.42	0.73
11	104.78	0.78	3.37	1.05	108.87	0.36	5.42	0.49	108.71	0.38	4.73	0.45	0.73
12	108.51	0.81	7.58	1.04	110.56	0.37	5.70	0.50	110.78	0.38	5.58	0.45	0.41
13	111.33	0.91	6.55	1.20	113.86	0.38	4.99	0.52	113.55	0.38	4.84	0.46	0.48
14	108.53	0.80	0.02	1.14	113.98	0.38	3.42	0.53	113.69	0.38	2.92	0.47	0.43
15	109.06	1.00	-2.27	1.35	114.18	0.38	0.32	0.54	113.85	0.41	0.30	0.49	0.46
16	111.58	1.06	3.05	1.33	114.01	0.41	0.03	0.56	113.83	0.42	0.14	0.48	0.46
17	112.91	1.23	3.85	1.59	114.71	0.48	0.53	0.61	114.50	0.51	0.65	0.59	0.43
18	112.62	1.85	1.04	2.13	114.38	0.73	0.37	0.84	114.34	0.73	0.51	0.76	0.54

Table 6: Numbers of Individuals and Observations for Echocardiographic Left Ventricular Wall Mass Study – Males Only

Number of Individuals	224	
Number of Patterns	73	
Number with		
1 Observation	72	32%
2 Observations	29	13%
3 Observations	42	19%
4 Observations	81	36%
Total	580	

Table 7: Correlations Between Repeated Measures of LV Mass – Males Only

Age	7	8	9	10	11	12	13	14	15	16	17	18
6	0.96	0.59	0.94									
7		0.61	0.62	0.44								
8			0.70	0.83	0.92							
9				0.70	0.75	0.43	0.77	0.91				
10					0.85	0.77	0.64	0.74	0.02			
11						0.72	0.62	0.50	0.64			
12							0.85	0.56	0.63	0.70		
13								0.81	0.64	0.51	0.76	
14									0.76	0.52	0.61	0.90
15										0.75	0.35	0.99
16											0.69	0.92
17												0.92

Table 8: Summary of Cross–sectional and Linked Cross–sectional Analysis of Echocardiographic Left Ventricular Mass – Mass Only

		Cross–Sectional		Linked Cross–sectional	
Age	N	Mean	S.E.	Mean	S.E.
6	7	66.1	3.7	64.1	8.2
7	36	71.8	2.3	69.9	2.5
8	47	76.8	3.0	67.7	2.4
9	65	82.8	2.5	77.1	2.3
10	67	89.6	2.9	87.0	2.4
11	53	96.9	3.5	96.3	2.9
12	50	109.4	3.4	109.2	2.7
13	54	122.9	3.7	122.2	2.8
14	62	138.2	4.1	142.3	3.4
15	48	161.1	5.2	164.9	4.2
16	46	159.8	5.1	169.0	4.3
17	32	173.0	6.3	176.3	4.7
18	13	146.2	8.6	149.5	12.4

Table 9: Estimated Growth Rates for Echocardiographic Left Ventricular Wall Mass Study – Males

	Cross–Sectional		Linked Cross–sectional	
Age Range	Growth	S.E.	Growth	S.E.
6–7	5.7	4.3	5.7	8.9
7–8	5.0	3.7	–2.2	3.2
8–9	6.0	3.9	9.5	2.8
9–10	6.8	3.9	9.9	2.4
10–11	7.3	4.6	9.3	3.3
11–12	12.6	4.9	12.8	3.3
12–13	13.5	5.0	13.0	3.4
13–14	15.3	5.5	20.1	2.9
14–15	22.9	6.6	22.6	3.8
15–16	–1.2	7.3	4.0	5.1
16–17	13.2	8.1	7.4	4.2
17–18	–26.9	10.6	–26.9	13.3