Research Report
Department of Statistics
Göteborg University
Sweden

Longitudinal methods for analysis of the influence of breastfeeding on early child health in Pakistan

Anders Carlquist
Valdemar Erling
Marianne Frisén

Research Report 1999:11
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Dept of Statistics	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.handels.gu.se/stat
P.O. Box 660	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
SE 405 30 Göteborg			
Sweden			

LONGITUDINAL METHODS FOR ANALYSIS OF THE INFLUENCE OF BREASTFEEDING ON EARLY CHILD HEALTH IN PAKISTAN

ANDERS CARLQUIST
Department of Statistics, Göteborg University, SE-40530 Göteborg, Sweden.

VALDEMAR ERLING
Department of Clinical Immunology, Göteborg University, SE-41346, Sweden.

AND

MARIANNE FRISÉN
Department of Statistics, Göteborg University, SE-40530 Göteborg, Sweden.

SUMMARY

Statistical methods for analysing aspects of early child health in Lahore, Pakistan are discussed. We construct generalised linear mixed models with a binomial response variable and both fixed and random explaining effects. In order to elucidate the causal effects of breastfeeding on early child health we use the two-step approach recently advocated in the statistical literature, but we modify the procedure to be practicable for the present longitudinal study. The selection effects of breastfeeding are examined, and variables with major effect on the breastfeeding pattern are included in the final model. For some, but not all, social groups the analysis gives enough motivation for the conclusion that breastfeeding prevents the occurrence of diarrhoea

1. INTRODUCTION

The aim of the present study is to choose, examine and use statistical methods suitable for analysing the influence of one factor (breastfeeding) on another (health) when concomitant variables have selection effects. Diarrhoea is a major cause of morbidity and mortality in developing countries[1]. A review of studies on the effect of breastfeeding for the reduction of the morbidity in diarrhoeal disease is given by Jason et al.[2]. Several studies give different kinds of support to the assumption that breastfeeding improves early child health. However, few studies take into account the longitudinal characteristic of these variables, the differences between fixed and random effects, the discrete distribution of the response variable, the selection effects and the seasonal effects. Longitudinal data give much information but require care in the choice of statistical method, since seemingly similar analyses may give quite different answers. To make causal interpretations is an important aim but there are many risks of fallacies.

We try to incorporate the necessary longitudinal characteristics in the analysis and use the suitable statistical techniques to analyse the influence of breastfeeding on the occurrence of diarrhoea in different areas of living in Lahore, Pakistan.

In the rest of this section details about the medical investigation are given. Methods for causal analysis are discussed in Section 2. In Section 3 the analysis of the selection process for breastfeeding is given. In Section 4 models with a mixture of fixed and random effects are presented and results from the final analysis by generalised mixed models are given. The results are discussed in Section 5.

1.2 The Lahore project

The research project "Early Child Health in Lahore Pakistan" aims at describing major health determinants of children in an urbanising poor society to provide an epidemiological basis for health planning. Socio-economic conditions, child care practices, feeding patterns, perinatal events, growth, morbidity and mortality have been described in a material collected between 1984-1987 consisting of 1476 children followed monthly for the first 2 years of life. The children were living in four different socio-economic areas; a village, a periurban slum area, an urban slum area and an upper middle class group. The details are described by Jalil et al.[3].

The main objective of this study is to investigate the impact of breastfeeding on diarrhoeal disease in relation to seasonal and climatic influences affecting different dimensions of early child health in a developing country. The impact of season and climate on respiratory tract infections and growth have been studied earlier in relation to the four different socio-economic areas of living, gender, order among siblings, family size and age of the child [4]. Breast-feeding is an important health-

promoting factor for children living in developing countries[5]. Season and other environmental factors are related to the patterns of breastfeeding[6] and we aim to describe this in relation to the effects of breastfeeding.

1.2 Study population

The cohort consisted of 1476 longitudinally followed infants born between September 1984 and March 1987 of these 485 came from the village, 398 from the periurban slum, 353 from the urban slum and 240 from the upper middle class. Originally a total of 1607 pregnancies were registered over a period of 30 months (September 1984 – March 1987), 117 pregnant mothers either refused, or moved to their parent's house for delivery and during the infant's first months of life according to local customs. Thirty-six pregnancies ended in stillbirths.

A health team consisting of a doctor, a mid-level field-worker (lady health visitor), a vaccinator cum lab-technician and a community health worker (traditional birth attendants) visited each infant at home shortly after birth and then every month during the study period of two years. Attempts were made to record all major diseases in addition to information on mortality, feeding patterns, growth, psychomotor development and childcare practices.

Of the live born 159 died before reaching the age of 24 months and 289 refused to participate in the study, or moved from the area. At 24 months of age 70% (1028 children) were still in the study. There was a higher rate of refusal and/or moving away in the periurban slum and the upper middle class than in the other areas.

Compliance analyses were performed for the study. The percentage of infants participating to 24 months of age and dropping out during this period were similar for the first, middle and the last 500 infants included. Neither body size at 22-24 month of age, nor the duration of breastfeeding was different for the first, the middle or the last third of the infants included in the study [3].

2. CAUSAL ANALYSIS

Most textbooks on epidemiology treat causality seriously but mostly verbally. Rothman and Greenland[7] give a good discussion of different models of causation and the epidemiological practice. It is stated there, that epidemiologists usually focus on testing the negation of the causal hypothesis, that is the null hypothesis that the exposure does not have a causal relation to disease. Then, any observed association can potentially refute the hypothesis, subject to the assumption that biases are absent. Lists of causal criteria have become popular, possibly because they seem to provide a road map through complicated territory. Examples of such criteria, which are commonly listed, are the ones by Hill[8]:

1. Strength of the association. It is argued that strong associations are more likely to be causal than weak ones.
2. Consistency. Observations from different situations strengthen the arguments.
3. Specificity. It is required that a cause lead to a single effect.
4. Temporality. The cause must precede the effect in time.
5. Biological gradient. A monotonic relation between dose and response seems often natural.

The conclusion drawn by Rothman and Greenland is that, apart from the necessity that the cause precedes the effect in time, there is no necessary or sufficient criterion for determining whether an observed association is causal.

Many philosophers have debated the nature of causation. The Scottish philosopher Hume[9] gave a list of causal criteria, which has been the base for many later ones, said that proof is impossible in empirical science. This does not preclude us from trying to formulate causal hypothesis but it should keep us sceptical and critical to our work. It also motivates that all efforts to establish causality should be made in such a way that each step could be debated.

In the statistical literature, there was much discussion by e.g. Rubin[10] during the decade of 1980 about the foundational issues. During the last decade causal inference is a topic that statisticians are addressing vigorously and rigorously[11-13]. An overview of propensity score methods for bias reduction in the comparison between a treatment and a non-randomised control group is given by d'Agostino[14]. The technique consists of two steps. In the first step the conditional probability, given the covariates, of being selected to treatment is modelled by a logistic regression. This propensity score is then used as an explaining variable in the second step, which might be a regression analysis. The similarities to the technique of instrumental variables, used in econometrics, are described in the paper by Angrist et al.[15] and in the discussion following that paper. That comparison gives more insight to the properties of both methods.

We do not believe that the use of a technical procedure without taking advantage of the subject matter knowledge is fruitful. In our analysis medical judgements are incorporated. Also, our situation is more complicated than that treated in most of the theoretical papers on the subject. Therefor we can not follow the exact procedures advocated there. However, we have the ambition to follow the spirit of those papers and describe both the analysis of the selection process and the final model in such a detail that the procedure is open for discussion.

Describing the causal relationship between breastfeeding, morbidity, growth and other environmental and individual factors over time is intricate. In Figure 1, a possible schematic causal dependency structure is described. Besides the arrows indicating probable direct influences there are also higher order interaction. The seasonal variations in many of these variables are substantial and sometimes of a higher order, e.g. the

effect of season on health is modulated by socio-economic status and is thus different in the different areas of living[16].

The duration of breastfeeding depends on socio-economic and other factors, which also have a direct causal influence on health. The direct relations between health and breastfeeding will not only reflect the causal effect of breastfeeding but a mixture of this and the selection effect of the breastfeeding tendency. The technique we use here for analysing if the effect of breastfeeding is mainly causal or mainly a selection bias is in the same spirit as described above. We also proceed in two steps. First, in Section 3, we will model the selection process and then, in Section 4, use this information in the final analysis of causality. As pointed out by d'Agostino[14] this allows a good modelling of the selection effect without burdening the final analysis with over-parameterisation.

In the first step of the causal analysis, we aim to clarify the patterns of breastfeeding by identifying the factors that influence these.

In the second step the selection variables, identified in the first step are put into the model together with the breastfeeding variable itself to see if the breastfeeding variable add any substantial explanation to the model. If so, this indicates that breastfeeding itself contributes to "health" and that it is not only a marker for the selection variables. Rosenbaum and Rubin[17] demonstrated that the technique of including all covariates with selection effects in a regression analysis, in many cases, lead to the same conclusion as the technique of propensity score.

3 VARIABLES THAT HAVE INFLUENCE ON THE PATTERN OF BREASTFEEDING

The variables to be used for explaining breastfeeding patterns is not self-evident. One main issue is whether the child is breastfed or not but in our material breastfeeding was initiated among most of the children and almost all the children were breastfed for at least one period. Thus breastfed or not would be a variable with little information. It is of importance if the child is exclusively breastfed for the first months. However, in our study very few children where exclusively breastfed and in those cases just for the very first months. The children were early given small amounts of additional water or other fluids. Thus, the variable is not informative. The duration of breastfeeding varied considerably among the children and this factor might in a developing country have a direct effect on the child's health situation As indicator of children with different breastfeeding pattern, we choose the age of the child (in months) when ending breastfeeding. We will not model the probability of being breastfed each month since the longitudinal character of that variable involves several complications. Instead we characterise the breastfeeding pattern for each child by the duration of the breastfeeding. Here, we use a regression selection method.

Four different kinds of variables are for medical reasons considered of importance for how long a child was breastfed:

i) *The condition of the child at birth.* This was measured as weight, length and head circumference.

ii) *The family-structure.* Important measures are the total number of persons in the family of the child and what order the child had among its siblings. From these variables we form a new variable (ADULTS) to reflect the number of adults in the family (by subtracting the number of children from the total number of persons in the family).

iii) *Socio-economic factors.* From the variables describing socio-economic background and housing and sanitary conditions, two indices are used, one for family socio-economic level and one for housing standard. Total family income per month is another important variable. The mother's education was expected to enhance breastfeeding.

iv) *Seasonal effects.* As an indicator of different seasonal effects we use the "birth-temperature", which is the average, for the month in which the child was born, of the minimum temperature of each day.

The selection effects of breastfeeding differ in the different areas of living. Thus, each area of living is analysed separately. For each area a forward regression procedure with inclusion criterion $p < 0,05$ is used. In the village birth-weight is found to be the best explaining variable followed by ADULTS and then total family income. In the periurban slum area duration of breastfeeding is best explained by ADULTS followed by family income and then by birth-weight. In the urban slum area ADULTS is the only variable, which has any significant explaining value for the length of breastfeeding. In the upper middle class the birth-temperature is the best explaining variable. To conclude - the better socio-economic standard and housing standard the shorter were the children breastfed. The more adults in the family the longer was the duration of breastfeeding. Low birth-weight and high birth-temperature was associated with short breastfeeding. The variables had different effect in the different area of living.

The variable birth-weight was considered to an important selection variable. However, it is a missing variable for as much as 44% of the children. The missing data cannot be considered to be non-informative since the probability that it will be recorded is dependent on several social factors. Thus, we have done all the analyses only for the category of children for whom the birth-weight was recorded. The conclusions will thus be valid only for that population.

Above, the selection effects for each child are taken into account. However, also the selection effect for each month has to be considered. These additional selection variables are determined by knowledge from earlier studies as well as medical judgements, and the problems with longitudinal modelling of these selection effects are thus avoided. The variables DEGM (the average each month of the minimum temperature during the day) and age which both varies with each month and which

are associated with both the occurrence of breastfeeding and diarrhoea were also included in the model.

4. MODELS FOR THE INFLUENCE OF BREASTFEEDING ON THE OCCURRENCE OF DIARRHOEA

4.1 Mixed linear models

The difference between longitudinal and cross-sectional studies is pointed out by e.g. Diggle et al.[18], and it is concluded that the major advantage of the former is its capacity to separate cohort and age effects. If this separation is not made, the wrong conclusion can be drawn from a material - since it may be an effect of calendar time that is wrongly attributed to age. Longitudinal studies can distinguish changes over time within an individual from differences among people in their baseline levels (cohort effects).

This distinction is made in a linear mixed model where it is possible to distinguish between effects that are constant for an individual but may vary among people from effects that change over time within an individual. Models with both fixed effects and random effects are treated in e.g.[18-21]. This kind of model is becoming an increasingly important statistical tool. In the mixed linear model, mixed stands for a mixture of fixed and random effects in the linear model. The fixed effects describe the population averages while the random effects models stochastic variation between individuals. Mixed and random effects models are also referred to as variance component models. The variances of the random effects are called variance components. In practical applications involving a mixed linear model, the problems of interest usually consist of estimating the fixed effect parameters and the variance components, and testing the significance of the fixed effects and variance components.

The estimation of random effects is sometimes under debate. In the expository overview by Robinson[22] and in the discussion following that paper several inferential logical problems concerning the interpretation of random effects as parameters or stochastic variables are discussed. Also, the properties of the best linear unbiased predictors and the similarities between the mixed models and some shrinking methods are discussed. The individual estimates will be closer to zero than those obtained with the derived variable method. However, in this study the individual estimates of random effects were not used. Instead, the covariance structure of the mixed model is utilised in order to give proper estimates.

A linear mixed model can be written as

$$Y = X\beta + Z\gamma + \varepsilon ,$$

where β is an unknown vector of fixed effects parameters with a known model matrix X, γ is an unknown vector of random effects with a known

model matrix Z and ε is a random error vector. The subject i, $i = 1,\ldots, m$ is measured n_i times and a description of the model that focus on the stochastic components is

$$Y_i = X_i\beta + e_i \quad i = 1,\ldots,m.$$

where Y_i is the $n_i \times 1$ vector of responses for the i:th individual and X_i is a $n_i \times p$ covariate matrix and $\beta = (\beta_1,\ldots,\beta_p)$ is a p-dimensional vector of unknown regression coefficients, called fixed effects, describing the population averages. The $n_i \times 1$ vector e_i is a random variable representing all remaining variability, and is assumed normally distributed with mean zero, and to be independent across individuals. The vectors of error components, e_i can be decomposed as

$$e_i = Z_i\gamma_i + \varepsilon_i$$

in which the first term models subject specific effects in e_i. Now Z_i is a $n_i \times q$ dimensional covariate matrix, and γ_i is a q×m-dimensional vector of subject specific regression coefficients, modelling stochastic variation between individuals.

The vector γ is assumed to independently distributed across subjects with the distribution, $N(0,\sigma^2 B)$, where B (B for between subjects) is a p×m dimensional covariance matrix. The within subject errors, ε_i, are distributed as;

$$\varepsilon_i \sim N(0,\sigma^2 W_i)$$

where W_i (within subjects) is a $q \times n_i$ dimensional matrix with few parameters because the random effects have removed many of the variance components. Often W_i is assumed to equal the identity matrix.

In this very general model, subjects can have different number of observations and different observation times. This generalisation of the standard linear model provides the possibility not only to model the mean of Y but also to model the variance of Y.

To see how the mixed linear model can distinguish between individual and group effects the following example can be of use. Let us assume that we have collected 4 measurements on each of 8 individuals where the individual effects have opposite signs compared to the group effect (Figure 2).

With the random effect assumptions the intercept was estimated as –3.39 with a slope of 1.49, which well estimates an average of the intercept and slope of the individuals. Without any random effect assumption the intercept was estimated as 7.2 and the slope to –0.39, which are not estimates of the individual but the group effects. When not taking the longitudinal aspect into consideration (as with the negative slope) time disguises the results leading to not desired conclusions.

4.2 Generalised linear mixed models

In a generalised linear mixed model a linear function of a mixture of fixed and random explaining variables is used just as in the linear mixed model. One generalisation is the link function, which gives the link between the linear expression and the response variable. In the linear model the link between the expectation of the response variable and the linear expression is the identity function. Another possibility, which will be used below, is the logit link. The other generalisation by the generalised linear mixed model is the possibility to use other stochastic assumptions than the normal distribution. In the next section the binomial distribution will be used.

4.3 Mixed models for early child health in Pakistan

The statistical model is useful if it contains the important factors without disguising them by irrelevant details. Models with different degree of details can thus be useful for different purposes.

In order to analyse the effect of breastfeeding on health we need all selection variables besides breastfeeding itself. We use all variables that were shown in Section 3 to have major influence on the duration of the breastfeeding as explaining factors, besides the variable BREASTFEEDING (which is the occurrence of breastfeeding or not each month), to explain the occurrence of diarrhoea each month. Also age and the temperature variable DEGM, which is the average each month of the minimum temperature during the day, are used as explaining variables. The variable age is a selection effect for the binary variable BREASTFEEDING and should thus be included to separate the effect of breastfeeding from that of age (see Section 3). The effect of temperature has been shown to be important in this material in earlier studies [4, 16] and of special concern since the way of measuring the seasonal effect by temperature has demonstrated important effects. Besides these fixed effects we also use a random component for each child. This component reflects effects that are not incorporated in the model but that are associated with the tendency of each child to get diarrhoea. Each of the four areas of living are analysed separately as the breastfeeding pattern differed much. The models used are exemplified for the village.

We start with a linear model for the growth in weight. A linear model for the analysis of the main effect on growth, Y, by temperature, X_{DEGM} and age X_{age} could be

$$Y_{ij} = \beta_0 + X_{DEGM,ij}\beta_{DEGM} + X_{age,ij}\beta_{age} + \varepsilon_{ij}$$

where $E(\varepsilon_i) = 0$ and the covariance matrix of ε contains the dependency structure due to repeated observations on the individuals. However the following type of model is more useful for our purposes

$$Y_{ij} = \beta_0 + X_{DEGM,ij}\beta_{DEGM} + X_{age,ij}\beta_{age} + Z_{individual,i}\gamma_i + \varepsilon_{ij}$$

where β_{DEGM} is the main (fixed) effect of temperature, β_{age} is the main (fixed) effect of age, γ_i is a remaining individual effect with $E(\gamma) = 0$ and the covariance is $\sigma^2 W$, assuming W to be the identity matrix.

When the binary variable Y= "occurrence of diarrhoea" is used as response variable we use the logit link, exemplified by

$$\log\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)} = \beta_0 + X_{BF,ij}\beta_{BF} + Z_{individual,i}\gamma_i$$

where β_{BF} is the main (fixed) effect of breastfeeding, γ_i is a remaining individual effect.

In Section 4.5 we use the logit link to several explaining variables to analyse the size and significance of β_{BF}. The purpose is to examine if the effect of breastfeeding is large even when the selection effects are included.

4.4. Methods for estimation in longitudinal analyses

4.4.1 The derived variable method

The method of derived variables (also called the method of summary statistics or two-stage method) is a simple and often very effective method[18]. The first step is to summarise the repeated values into a summary statistic, which then, in the second step is analysed as a function of x_i.

This method has been used by Carlquist et al.[16] and it has been advocated by e.g. Frison[23] for it's many attractive features. The results are readily interpretable. No assumptions are needed about the covariance structure among the repeated measures (but it is useful to take it into account when choosing the summary statistic).

The full maximum likelihood estimation in a mixed linear model is more efficient than the method of derived variables since the information of several observations on each child is taken full advantage of. Also, the possibility in the full likelihood method to easily use fixed effects, common for all children increases the efficiency. In a linear model the derived variable method is a useful but not fully efficient method for estimation. In a generalised linear derived model with binary data, the lack of efficiency is more serious. Either a very large number of observations for each individual or a probability of occurrence near 50% is necessary to get information enough with the derived variable method. Another disadvantage with the derived variable method is that the fixed and random effects cannot be estimated simultaneously.

4.4.2 Maximum likelihood estimation

Even though the computer programs have been more and more efficient the computational burden of the full maximum likelihood estimation is enormous. Also, the complicated structure of the models makes it common with near singularities and convergence problems. Thus, different variants are used. We used the restricted maximum likelihood estimation (REML). This method copes with the near-singular variance matrix much more effectively than does the ordinary maximum likelihood estimation[18]. The computational problems with the generalised linear models are even harder and the properties of the estimates worse than in the linear case[24]. The procedures provided by the SAS package and described by Wolfinger[25] are used here.

4.5 Results on the causal influence of breastfeeding on the occurrence of diarrhoea

The variables, which were found (Section 3) important for explaining the duration of breastfeeding, are used in our mixed generalised linear model with a health indicator as response variable. In this report the occurrence of diarrhoea each month is used as the dependent variable.

Only variables constant for the individual were considered when modelling the duration of breastfeeding and consequently a cross-sectional approach was used in the analysis of the selection process. Then, a longitudinal approach is used for assessing the importance of breastfeeding for 'occurrence of diarrhoea'. The breastfeeding variable in the model, which describes the health each month, is the binary variable BREASTFEEDING (breastfed or not that month). Since BREASTFEEDING and age are not independent, age was included as an explaining variable in order to test the effect of BREASTFEEDING itself.

Results of the analysis by the generalised linear models described above are reported for each area of living in Tables 1 – 4, as given by the SAS macro GLIMMIX.

The village					
Effect	Estimate	Std Error	DF	t	Pr>|t|
Intercept	-1.0658	0.3242	296	-3.29	0.0011
BREASTFEEDING	-0.3024	0.0955	4402	-3.17	0.0016
Age	-0.0065	0.0057	4402	-1.14	0.2531
DEGM	0.0458	0.0042	4402	12.29	0.0001
Birth-weight	-0.0521	0.0982	4402	-0.34	0.7316

Table 1. Estimates of the parameters in the model for each factor included and also the standard error, degrees of freedom, the t-statistic and the p-value for the analysis of data from the village

The periurban slum area					
Effect	Estimate	Std Error	DF	t	Pr>|t|
Intercept	-0.8891	0.5676	122	-1.57	0.1199
BREASTFEEDING	-0.2319	0.1502	1802	-1.54	0.1229
Age	-0.0075	0.0087	1802	-0.86	0.3909
DEGM	0.0631	0.0067	1802	9.38	0.0001
Birth-weight	-0.1216	0.1678	1802	-0.73	0.4685
ADULT	-0.0074	0.0564	1802	-0.13	0.8960
Total income	-0.0003	0.0002	1802	-1.56	0.1195

Table 2. Estimates of the parameters in the model for each factor included and also the standard error, degrees of freedom, the t-statistic and the p-value for the analysis of data from the periurban slum area

The urban slum area					
Effect	Estimate	Std Error	DF	t	Pr>|t|
Intercept	-0.9333	0.2068	163	-4.51	0.0001
BREASTFEEDING	-0.3572	0.1156	2450	-3.09	0.0020
Age	-0.0309	0.0073	2450	-4.24	0.0001
DEGM	0.0374	0.0059	2450	6.32	0.0001
ADULT	-0.0039	0.0331	2450	-0.12	0.9051

Table 3. Estimates of the parameters in the model for each factor included and also the standard error, degrees of freedom, the t-statistic and the p-value for the analysis of data from the urban slum area

The upper middle class					
Effect	Estimate	Std Error	DF	t	Pr>|t|
Intercept	-1.8868	0.2907	141	-6.49	0.0001
BREASTFEEDING	-0.2729	0.1697	1856	-1.61	0.1079
Age	-0.0238	0.0101	1856	-2.37	0.0180
DEGM	0.0260	0.0087	1856	2.98	0.0029
Birth-temperature	0.0047	0.0103	1856	0.46	0.6451

Table 4. Estimates of the parameters in the model for each factor included and also the standard error, degrees of freedom, the t-statistic and the p-value for the analysis of data from *the upper middle class group*

In the village and in the urban slum, breastfeeding gives a significant contribution to the explanation of the event of diarrhoea in addition to the effects of the selection variables. This effect is in addition to the effects of age and the seasonal effect (measured by DEGM). In the periurban slum and the upper middle class group the effect of breastfeeding is not significant. However, the 'lack of evidence' for an effect is no evidence for a 'lack of effect', as will be discussed below.

5. DISCUSSION

A technique, which first identifies the variables influencing the duration of breastfeeding and then uses them as concomitant variables when analysing the effect of breastfeeding on health, is used. The models are considered as approximations and simplifications useful for structuring the main features. It is never possible to be absolutely sure that all selection bias is eliminated and that the observed effects are causal. However, any reduction of the selection bias will make interpretations easier. The modest but important aim here is that known major fallacies, which are present in many investigations of this kind, are avoided.

Variables associated with each child and influencing the duration of the breastfeeding are analysed with a cross-sectional analysis, since the model would be too complicated if also time-dependent variables were included.

Birth weight can in many ways affect the length of breastfeeding. The initiation of breastfeeding is vulnerable to circumstances around the child. A child of low birth weight has a larger risk of becoming ill during the first month of life, which then later can affect further breastfeeding. A child of low birth weight may have more difficulties to establish a good suckling pattern, which might shorten the breastfeeding. The mother of a child with low birth weight may herself be undernourished and therefore have more difficulties with breastfeeding. A child of low birth weight has a tendency of shorter life and this will influence the possible time for breastfeeding. However, the pattern is strong also for those children who survived the whole period of study.

The structure of the family seems to predict the length of breastfeeding. The number of adults in the family has a positive effect on the duration of breastfeeding. This might be due to a better situation for the mother in a family with more adults, giving the mother a hand in the household. There might be more time to breastfeed the child.

Social status is a difficult parameter when predicting duration of breastfeeding since a woman that breastfeeds her child shorter shows that she can afford to bottle-feed the child. If she has a good education and a corresponding job, she will also leave home earlier and in that way shorten the duration of breastfeeding. On the other hand education generally favours a positive view towards breastfeeding. However, in this study education has no substantial effect on the duration of breastfeeding. This could be due to lack of education that focuses on benefits of a long

breastfeeding period. High social level is associated to a short duration of breastfeeding in the society of this study. This could also be the explanation of why high family income had a negative effect on the duration of breastfeeding.
The selection effects influencing which children are breastfed for a long duration, is not the same in the different areas of living. The effect of many adults in the household was the most important variable in the periurban and urban slum. It might reflect the need of a more stable family situation in order to cope with urban life. The effect of birth weight on the duration of breastfeeding in the village might be due to rural customs trying to give the child other foods when having a low birth weight. This might inhibit breastfeeding. The effect of birth-temperature, which was seen in the upper middle class, is difficult to interpret. One possible explanation might be that in the upper middle class substitutes for breast-milk are available and that this is used more commonly during the hot season.

The variables found to be important for the duration of breastfeeding were put into a mixed linear model with 'occurrence of diarrhoea' as dependent variable. For the final analysis it was necessary to follow the longitudinal pattern. Even though the selection effects for each child is taken care of, the selection effect for each month has to be considered. Two more variables, DEGM and age, which varies with each month and which are associated with both the occurrence of breastfeeding and diarrhoea were included in the model.

The final analysis by the SAS-macro GLIMMIX must be interpreted with care. The approximation with the t-distribution might not be very good. Also, the significance analysis is for the case of one analysis and is not adjusted for the two-step procedure. However, the results by Rosenbaum and Rubin[17] for a similar situation indicates that the conclusion from the two-step procedure leads to the same conclusions as one simultaneous analysis for the case when the same variables are used in both steps. The use of more variables in the first step will in most cases have a conservative effect.

The results from models that include selection effects and also models the longitudinal pattern is different for the four living areas. In the village and the urban slum a significant effect of breastfeeding on the occurrence of diarrhoeal disease is demonstrated as can be expected. The effect is not only statistically significant but the size of the effect is also large enough to be of medical significance. For example, in the urban slum the estimate of –0.36 corresponds to an odds ratio of 0.70. Thus, the odds of diarrhoea when breastfed is 0.70 times less than for children who are not breastfed and who have the same values of the selection variables. In the upper middle class the effect is less and is not significant on the 5% level. This is in agreement with a less vulnerable status for a child in this group. Among children living in the periurban slum under extremely poor circumstances the selection mechanisms are complicated and several variables had a significant effect on the feeding pattern. These variables are also associated with the occurrence of diarrhoea. Because of the

strong selection effects too little information was left to give significance to the contribution of breastfeeding by itself.

The new techniques proposed in the statistical literature are important of several reasons. One is that the steps in the procedure are clear and thus open for discussion. In practical applications complications arise and we have demonstrated a way to handle whose.

ACKNOWLEDGEMENTS

The authors wish to express thanks to Associated Professor Agnes Wold and Professor Lars Å Hanson for their helpful comments and valuable suggestions on the manuscript. Professor Lars Å Hanson, Professor Femida Jalil, Professor Rifat Ashraf and Associate Professor Shakila Zaman has kindly made it possible for us to use the data from the Lahore project.

REFERENCES

1. Black, R. E. 'Epidemiology of diarrhoeal disease: implications for control by vaccines', *Vaccine*, **11**, 100-106 (1993).

2. Jason, J. M., Nieburg, P. and Marks, J. S. 'Mortality and infectious disease associated with infant-feeding practices in developing countries', *Pediatrics*, **74**, 702-727 (1984).

3. Jalil, F., Lindblad, B. S., Hanson, L. A., et al. 'Early child health in Lahore, Pakistan: I. Study design', *Acta Paediatr Suppl*, **82 Suppl 390**, 3-16 (1993).

4. Erling, V., Jalil, F., Hanson, L. and Zaman, Z. 'The impact of the climat on the prevalence of respiratory tract infections in early childhood in Lahore, Pakistan', *Journal of Public Health Medicin*, **In Press** (1999).

5. Victora, C. G., Smith, P. G., Vaughan, J. P., et al. 'Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil', *Lancet*, **2**, 319-322 (1987).

6. Bohler, E., Aalen, O., Bergstrom, S. and Halvorsen, S. 'Breast feeding and seasonal determinants of child growth in weight in east Bhutan', *Acta Paediatr*, **84**, 1029-1034 (1995).

7. Rothman, K. J. and Greenland, S. *Modern Epidemiology*, Little, Brown, Boston, 1998.

8. Hill, A. B. 'The Environment and Disease: Association or Causation?', *Proc R Soc Med*, **58**, 295-300 (1965).

9. Hume, D. *A Treatise of Human Nature*, 2 ed., Oxford University Press, Oxford, 1978.

10. Holland, P. W. 'Statistics and causal inference', *Journal of the American Statistical Association*, **81**, 945-960 (1986).

11. Rubin, D. B. 'Practical implications of modes of statistical inference for causal inference and the central role of the assignment mechanism.', *Biometrics*, **47**, 1213-1234 (1991).

12. Greenland, S., Robins, J. M. and Pearl, J. 'Confounding and Collapsibility in Causal Inference', *Statistical Science*, **14**, 29-46 (1999).

13. Keiding, N. and Eerola, M. Discussion on Statistics and the Assessment of Causality, International Statistical Institute, Helsinki, (1999).

14. D'Agostino, R. B., Jr. 'Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group', *Statistics in Medicine*, **17**, 2265-2281 (1998).

15. Angrist, J., Imbens, G. and Rubin, D. 'Identification of causal effects using instrumental variables (with Discussion)', *Journal of the American Statistical Association*, **91**, 444-469 (1996).

16. Carlquist, A., Erling, V., Frisén, M., Hanson, L., N., A. R. and Zaman, S. *The Impact of Season and Climate on Growth during Early Childhood in Four Different Socio-Economical Groups in Lahore, Pakistan.*,. Research report 1999:10, Department of Statistics, Göteborg University,1999.

17. Rosenbaum, P. R. and Robin, D. B. 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, **70**, 41-55 (1983).

18. Diggle, P., Liang, K.-Y. and Zeger, S. *Analysis of Longitudinal Data*, Oxford University Press,1994.

19. Hand, D. J. and Crowder, M. J. *Practical Longitudinal Data Analysis*, Chapman & Hall,1996.

20. Khuri, A. I., Mathew, T. Sinha, B.K. *Statistical Tests for Mixed Linear Models*, John Wiley,1998.

21. Verbeke, G. and Molenberghs, G. *Linear mixed models in practice : a SAS-oriented approach*,. Lecture notes in statistics ; 126 Springer, New York, 1997:XIII, 306 s.

22. Robinson, G. K. 'That BLUP is a Good Thing: The estimation of Random effects', *Statistical Science*, **6**, 15-51 (1991).

23. Frison, L. *Analysis of repeated measures in clinical trials using summary statistics*,. Medical Statistics Unit. London School of Hygien and Tropical Medicine. University of London, London, 1994.

24. Engel, B. 'A Simple Illustration of the Failure of PQL, IRREML and APHL as Approximate ML Methods for Mixed Models for Binary Data', *Biometrical Journal*, **40**, 141-154 (1998).

25. Wolfinger, R. and O'Connell, M. 'Generalized linear mixed models: A pseudo-likelihood approach.', *Journal of Statistical Computation and Simulation*, **48**, 233-243 (1993).

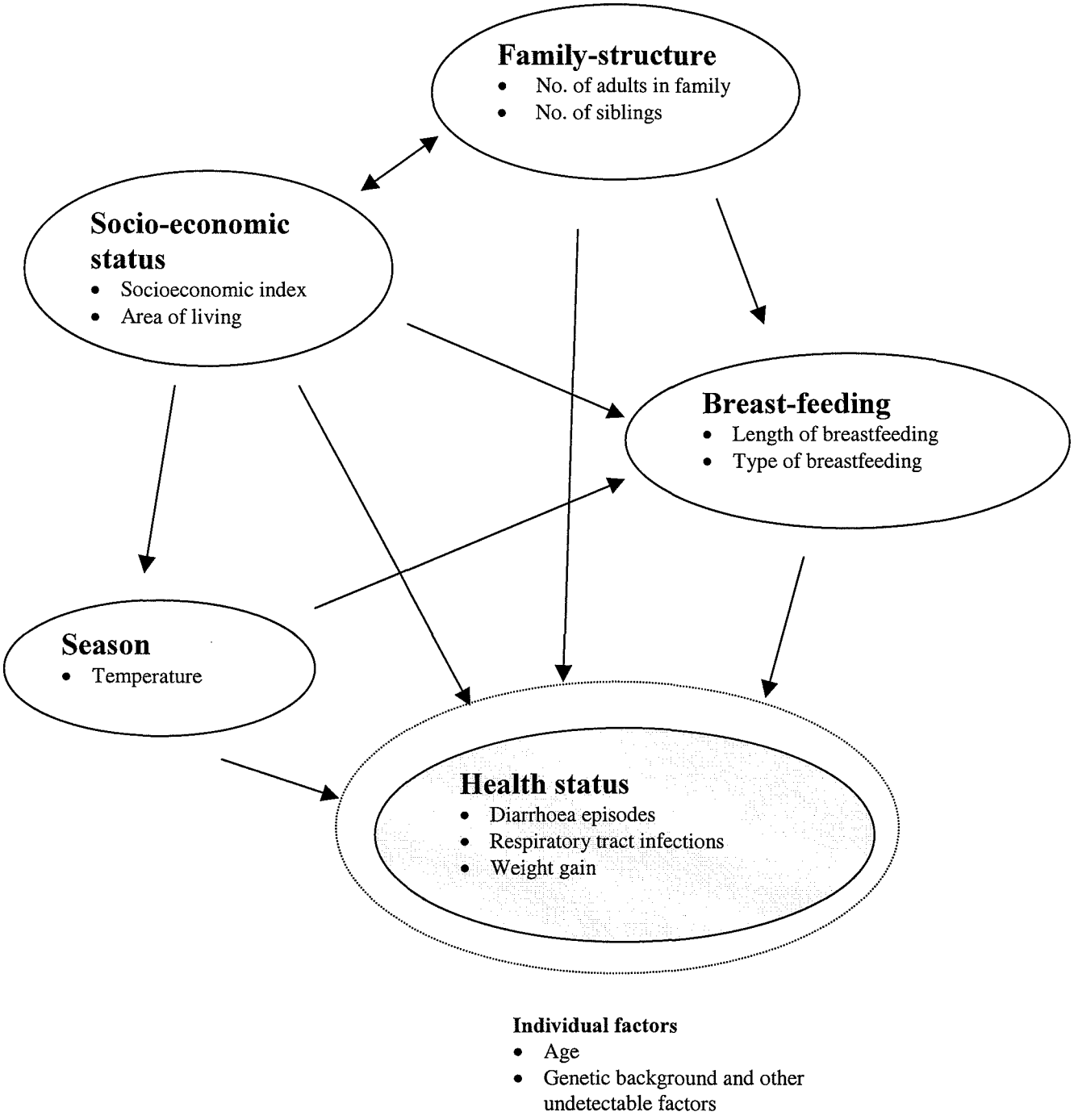

***Figure 1**. Examples of variables that influence breastfeeding and health for children in Lahore Pakistan.*

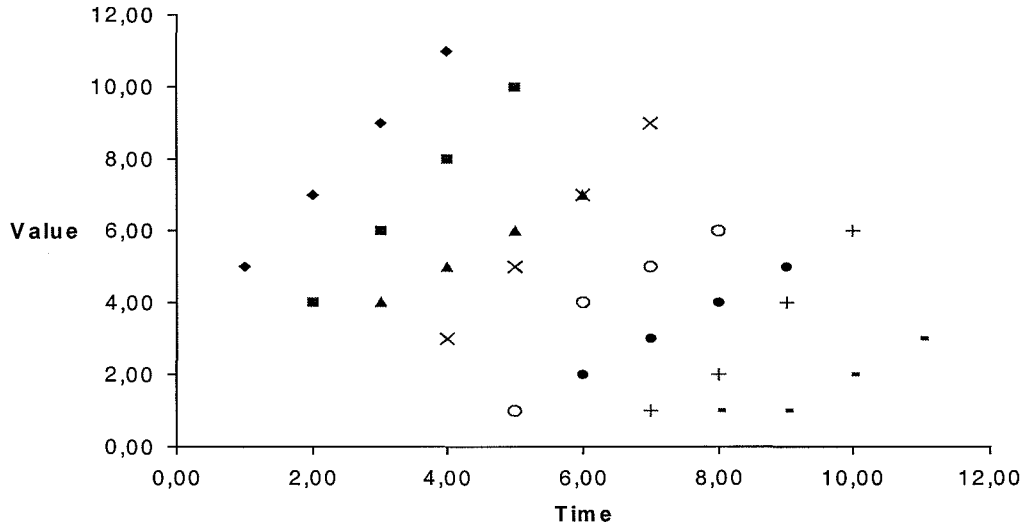

Figure 2. *Example with 8 individuals measured 4 times each.*

Research Report

1999:1	Andersson, E.:	On monotonicity and early warnings with applications in economics.
1999:2	Wessman, P.:	The surveillance of several processes with different change points.
1999:3	Andersson, E.:	Monotonicity aspects on seasonal adjustment.
1999:4	Andersson, E.:	Monotonicity restrictions used in a system of early warnings applied to monthly economic data.
1999.5	Mantalos. P. & Shukur, G.:	Testing for cointegrating relations- A bootstrap approach.
1999:6	Shukur, G.:	The effect of non-normal error terms on the properties of systemwise RESET test.
1999:7	Järpe, E. & Wessman, P.:	Some power aspects of methods for detecting different shifts in the mean.
1999:8	Johnsson, T.:	On statistics and scientific thinking.
1999:9	Afsarinejad, K.:	Trend-free repeated measurement designs.
1999:10	Carlquist, A. m.fl.	The impact of season and climate on growth during early childhood in different socio--economic groups in Lahore, Pakistan.