Research Report
Department of Statistics
Göteborg University
Sweden

# Longitudinal methods for analysis of early child health in Pakistan

**Anders Carlquist**

**Research Report 1999:12**
**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 EARLY CHILD HEALTH IN PAKISTAN.

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

Statistical methods to analyse some aspects on early child health in Lahore, Pakistan are discussed. The aim of this thesis is to choose, examine and use statistical methods suitable to analyse different aspects of early child health in a developing country when the child is measured repeatedly over time (longitudinal data). This kind of data are very informative but require care in the choice of statistical method, since seemingly similar analyses can give quite different answers.

This licentiate thesis consists of two parts, which will be referred to in the text by their Roman numerals:

**I**: Anders Carlquist, Valdemar Erling, Rifat Ashraf, Marianne Frisén, Fehmida Jalil, Lars Å Hanson, Shakila Zaman. The impact of season and climate on growth during early childhood in different socio-economic groups in Lahore, Pakistan. Research Report, Department of Statistics, Göteborg University 1999:10.

**II**: Anders Carlquist, Valdemar Erling and Marianne Frisén. Longitudinal Models for Analysis of the Influence of Breastfeeding on Early Child Health in Pakistan Research Report, Department of Statistics, Göteborg University 1999:11.

In report **I**, an explorative cross-sectional analysis is first made to examine the main features of some relations between variables. Then, the method of derived variables is used to investigate the longitudinal patterns. The method of derived variables (also called the method of summary statistics or two-stage method) is a simple and often very effective method. The first step is to summarise the repeated values into a summary statistic, which then, in the second step is analysed. This method has been advocated 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). These methods are used to study the impact of temperature on health, measured as growth. In Lahore, Pakistan temperature was shown to be a factor that influences child health. It is demonstrated that the impact is quite different in different ages and areas of living.

In report **II**, we aim to elucidate the causal effects of breastfeeding on early child health, measured as occurrence or not of diarrhoeal episodes. Here it is not possible to use the method of derived variables. We use generalised linear mixed models to construct a model for the binomial response variable and use 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 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, areas of living the analysis then motivates the conclusion that breastfeeding helps prevent the occurrence of diarrhoea

# ACKNOWLEDGEMENTS

I would like to express my deepest thanks to my supervisor, Prof. Marianne Frisén, without whom this work would not have been possible and for her positive and open attitude – both in scientific and in personal matters.

I would also like to thank Valdemar Erling, M.D., my collaborator and close friend who always had great empathy with the hard living situations that often must have been the reality behind the numbers and for his contagious optimism.

Thank you to all my friends at the Department of Statistics, Göteborg University and to my fellow PhD students Eva, Peter, Magnus, Eric, Christian, Max, Thomas and Abdul, who ought to pay their elder a bit more respect but otherwise were all right.

Also I would like to express my gratitude to my wife Pernilla and my family for making it all worthwhile.

# The Impact of Season and Climate on Growth during Early Childhood in Different Socio-economic Groups in Lahore, Pakistan.

Anders Carlquist [1], Valdemar Erling [2], Rifat Ashraf [3], Marianne Frisén [1]
Lars Å Hanson [2], Fehmida Jalil [3], Shakila Zaman [3]

1. Department of Statistics, Göteborg University, Sweden.
2. Department of Clinical Immunology, Göteborg University, Sweden.
3. Department of Social and Preventive Paediatrics, King Edward Medical College, Lahore, Pakistan

## ABSTRACT

*Aim.* The aim of this study was to analyse the impact of season and climate on early child growth in four socio-economically different groups living in Lahore, Pakistan.

*Methods.* A prospective cohort study was conducted among children living in a village, periurban slum area, urban slum area and a group of children from the upper middle class. Monthly observations were made on 1476 infants from birth to 24 months. Growth in terms of length and weight were analysed in relation to age, socio-economical group, and season of the year and different climate variables such as rain, day temperature and humidity. The longitudinal data was analysed by the derived variable method and regression models.

*Results.* Season of birth did not have any important impact on the birth weight or length. There were differences in weight gain during the first six months as well as between 6-12 months, depending on which time of the year the child was born. This seasonal differences was most prominent in the poorer groups. The climate variable mean day temperature was negatively correlated to weight gain with a p-value<0001 for the regression coefficient in all groups except for the upper middle class. In the poorer groups the impact increased with age up to 14 months and decrease with higher ages. Growth velocity concerning length was not clearly related to season and climate. In the upper middle class neither season nor climate markedly affected weight or length velocity.

*Conclusions.* Season and climate are modifying growth. The socio-economic level and the age of the child influence the impact of climate on growth. Together with age, sex and length of gestation, season is an important factor when comparing growth of children living in poor environments.

## INTRODUCTION

Growth during the first year of life is rapid and can easily be negatively affected by external factors. Studies in developing countries have shown a higher incidence of disturbed growth and development, as well as higher morbidity and mortality among children under

24-months of age compared to other age groups. [1, 2, 3] Genetic predisposition and other intrinsic factors determine growth, but extrinsic factors are important in modelling attained growth. It has repeatedly been shown that the socio-economic situation, such as housing-standard, family income, educational status of the parents [4] and feeding patterns, [5] are of significance for early child growth in developing countries. [6]

Several studies have demonstrated that the pattern of growth of breast-fed infants differs from that of the WHO/CDC reference and of infants fed currently available formulas. [7, 8] The most typical pattern for breastfed infants is to grow as rapidly or more rapidly than the WHO/CDC reference during the first 2-3 months, but thereafter they show a relative deceleration. Currently used growth charts throughout the world for monitoring growth of children are based on the US National Center for Health Statistics reference data. The adequacy of these data has been questioned. It does not take into account infant feeding practices in modelling the growth. [9] Season of the year, which has been shown to influence early child growth, [10] is neither adjusted for in growth charts. The implications of climatic seasonality have not so far been related to growth in length, weight or weight for height. Seasonal variations are not commonly adjusted for in models describing early child growth.

In a community-based prospective cohort study of young children in Lahore, Pakistan, we show that the climate variable temperature affects growth velocity and that the impact of the temperature on growth varies for groups at different socio-economic levels in a developing country.

## MATERIAL

### *Study Population*

Lahore, the second largest city in Pakistan, can be considered a typical expanding city in the third world with migration from villages to slumareas at the outskirts of the city and from there to more organised and steady living in the city slum areas. The study population was selected according to this and consisted of 1476 longitudinally followed infants born between September 1984 and March 1987, composed of four selected groups from a village, a periurban slum area, an urban slum area and an upper middle class group. The urbanised upper middle class group functioned as the control group. The study design is described in detail elsewhere. [11]

The first group was selected from a typical village about 40 km from Lahore. The periurban slum was selected from a mud hut area along a railway track, on the periphery of the city. A typical segment of people living in the old walled city comprised the third group - the urban slum area. The upper middle class group was scattered around the city.

An initial cross-sectional survey, including a population of around 5000 from the three areas, was initiated in March 1984. During this survey 2998 households were identified and basic socio-economic and demographic data were obtained. A total of 1607 pregnancies were registered over a period of 30 months, including 240 pregnancies selected from the upper middle class. Out of these, 117 pregnant mother either refused to participate, or moved. Thirty-six pregnancies ended in stillbirth and thus 1476 newborn infants were included in the cohort. The children from the upper middle class were selected through private obstetric clinics, with the inclusion criteria of having 3 or less than 3 children, an income above US $ 275 per month, of owning a house with more than three bedrooms and of both the parents having at least ten years of schooling.

The new-borns were examined immediately after birth and were then followed by monthly home visits up to the age of 24 months. Attempts were made to record all major diseases during the past month. In addition to this, information on feeding, growth, psychomotor development and childcare practices was collected. A doctor together with a health worker made the examinations. Of the 1476 children entering the study, 485 came from the village, 398 from the periurban slum, 353 from the urban slum and 240 from the upper middle class. Of the live born, 159 (11%) died before reaching the age of 24 months and 289 (20%) moved from the area, or refused to participate in the study. At 24 months of age 1028 children (70%) were still in the study. The total number of examinations performed in our analysis was 21 200. We did not adjust for gestation length and sex in modelling the growth.

*Family Characteristics*

The four different selected groups showed different patterns in the family structure. For example, the birth order of the children included in the study varied between the four selected groups. The mean family size varied from 5.3 in the upper middle class to 6.8 in the village. The mean maternal age was 28 years and similar in the four groups. The families in the urban slum group had better housing standards, and were on a higher parental educational and economic level than in the other two poor areas. The socio-economic levels and housing standards along with family attitudes toward childcare and hygiene have been described elsewhere. [12, 13]

### *Climate*

The changes in climate over the year are substantial. The temperature, humidity and rain showed seasonal variations across the study years. The seasons are commonly divided in a 'temperate/cold' season including November, December, January and February, a 'warm' season, including March, April, September and October and a 'hot' season, including May, June, July and August.

The mean values of humidity at 5 a.m. and 5 p.m., maximum and minimum temperature and amount of rain for each month from January 1985 to December 1987 were obtained from the local meteorological office.

## METHODS

### *Notation.*

The response variable 'weight gain' was the individual gain in weight for the period of interest measured in hectograms (hg). 'Length gain' was the individual gain in body length for the period of interest. 'Weight for length' was 'weight gain' divided by 'length gain'. The climate variable 'temperature' was measured as the monthly mean of the lowest day temperature in degrees Celsius. Group number 1, 2, 3 and 4 refer to the village, the periurban slum, the urban slum and the upper middle class group.

### *Examinations*

Health workers visited the children at home every month, collecting data. The health workers were specially trained in taking body measurements - weight, length and head circumference - in a standardised manner. During the research period the health teams were retrained once every month. The way of taking the body measurements was crosschecked between both the workers and the equipments. A difference of 100 g for weight and less than 1.0 cm for length and head circumference was acceptable. The weighing and measuring equipments were checked weekly and maintained by a trained mechanic. Pretested and standardised questionnaires where used to collect information concerning incidence of diarrhoeal episodes for the children in the study. All cases where the child had had more loose stools, or one ore more watery stools per day during the period since last visited were registered as cases of diarrhoea during the previous month. The duration of the episodes was not used in the analyses.

*Statistical analyses*

The analysis is explorative; thus no tests of hypotheses have been made. However, in some cases p-values are given as descriptive measures of the amount of information.

Our data are longitudinal. The children under study were followed in time with repeated observations as shown in figure Figure 1. At the preliminary modelling stages though, all observations were analysed cross-sectionally without regard to the longitudinal nature of the data in order to achieve a first overview. Linear regression analysis was used to model the relationship between 'weight gain' (dependent variable) and 'temperature' (independent variable).

Since the temperature varies in a cyclic manner through the year a multiple regression model with sine and cosine values of the calendar months as independent variables and 'weight gain' as dependent was evaluated in the following way:

The number of the months {1,12} was transformed into values ($t$) within the range {o, 2π}. Then a graph of the equation A$\cos(t + B)$ was fitted to the seasonal variations in 'weight gain'. B allows the graph of $\cos t$ to be shifted B units horizontally and A allows vertical scaling of the graph, According to the Addition formula:

$$A\cos(t+B) = A\cos t \cos B - A\sin t \sin B = $$

$$= A\cos B\cos t - A\sin B\sin t$$

$$\left|\begin{array}{l} A\cos B = ß_1 \\ A\sin B = ß_2 \Rightarrow \end{array}\right.$$

$$\Rightarrow A\cos(t+B) = ß_1\cos t - ß_2\sin t$$

By adding an intercept ($ß_0$) to the model the graph can be shifted vertically. Our model was then:

$$\text{'weight gain'} = ß_0 + ß_1\cos t - ß_2\sin t$$

This was done as a preliminary stage in order to compare the fit obtained in this model with the one from the model with climate variables as explaining variables.

A longitudinal approach [14] was then adopted for most analysis in order to describe change over time within individuals. One of the reasons for this was to be able to distinguish cohort and ageing effects. The variations between children due to unmeasured variables cause difficulties at the interpretation of the estimate of a cross-sectional regression coefficient. Also, the estimate of variance is a problem in cross-sectional analysis because successive individual measurements are related to each other. Instead the derived variable method was chosen to analyse how the individual growth rate is related to a climate variable. The regression coefficient of the individual multiple regression with 'weight gain' as dependent variable and 'temperature' and age in months as independent variables was considered a good summary statistic as it estimates the average impact of the explaining variables. This first step reduced the repeated measures of each individual to one summary variable. In a second step the summary variable created in the first step was analysed.

The SAS Statistical software was used when analysing the data.

## RESULTS

As will be demonstrated in detail below there is a difference in 'weight gain' per month depending on the 'temperature' exposing the children living in the village, the periurban slum and the suburban slum. No relation between 'weight gain' and 'temperature' could be found among the children from the upper middle class group. Furthermore the impact of 'temperature' on 'weight gain' was different depending on the child's age.

The time of birth did not seem to affect the weight at birth and the pattern did not differ much between the different groups as shown in Figure 2.

The impact of season on ‘weight gain’ during the child’s first six months was studied in the different groups. In Figure 3a all the groups were pooled and no significant differences in weight gain were detected. When the groups were studied separately we found that children born at the end of the ‘hot’ season had a higher weight gain during the first six months of life in group 1, 2 and 3. This was most pronounced in the village. The upper middle class children born after the ‘cold/temperate’ season showed the highest weight gain during the first six-month of life as shown in Figure 3b.

The 'weight gain' from six to twelve months of age is described in Figure 4 and it was clearly different according to birth month. The children born at the end of the 'temperate/cold' season and especially those born during the first months of the year (with an age of 6-12 months during later months of the year) seemed to have the highest weight gain in all groups, except the upper middle class where no obvious pattern was detected.

The impact of 'temperature' on 'weight gain' in different groups was demonstrated by linear regression. At this stage the longitudinal character of the data was not considered. The p-values should thus only be considered as rough indicators. In the village a temperature rise of one degree diminished the 'weight gain' by 0.1 hg per month. The children in the periurban slum had their 'weight gain' diminished by the same amount. In the urban slum the amount was 0,07 hg per month. This pattern was not evident in the upper middle class group. Here a temperature rise of one degree increased the 'weight gain' by 0.006 hg per month. In the village, the periurban slum and the urban slum the regression coefficient had a p-value<0.0001, but for the upper middle class the p-value was 0.55. For the climate variables rain and humidity no relation to 'weight gain' was evident.

The relationship between 'temperature' and 'weight gain' per month was different according to the age of the child. When dividing the children into subgroups according to age in months and socio-economic group, an age specific pattern emerged. A newborn child's weight gain was not strongly influenced by day temperature in any groups. As an example, the linear regression in (Figure 5) uses only the data of children who are 19 months old and comes from the village. They had an impact on 'weight gain' by -0,18 hg per month for a temperature raise of one degree. To highlight the mean pattern of how age affects the relationship between 'temperature' and 'weight gain' the regression coefficient was calculated for each age in the same way as in Figure 5. The regression coefficients are plotted against age in Figures 6a-d. A smoothing spine was fitted to the data. With increasing age the children in the village, periurban slum and urban slum had a more negative relationship between 'temperature' and 'weight gain'. No evident pattern was found among the upper middle class children.

When using 'length gain' as response variable and 'temperature' as regressor no strong relationship was observed in any groups. Division into subgroups according to age did bring forth a pattern in the urban slum where the impact of 'temperature' was increasingly negative for higher values as shown in Figure 7c. In the village, the periurban slum and the upper middle class no clear relation between 'temperature' and 'length gain' could be seen as described in Figure 7a, 7b, and 7d.

'Weight for length' as response variable and 'temperature' as regressor was evaluated in a regression where every age group was treated separately as above. No significant pattern was displayed.

Impact of 'temperature' and 'age' (regressors) on 'weight gain' (regressand) as measured in a derived variable model is examined in (Table 1). The results show a negative relationship between growth rate and temperature in the village, a temperature rise of one degree diminishes the 'weight gain' by 0,13 hg per month (std. err. =0,07), in the periurban slum the relation was –0.16 hg per month (std. err. =0.05) and in the suburban slum -0.25 hg per month (std. err. =0.11). In the upper middle class the relationship was also negative -0.23 hg per month but the standard error was large (std. err. =0.22), hence this regression coefficient was not significant.

The regression with 'weight gain' as dependent variable and the sinus and cosinus values of the transformed calendar month as regressors is presented last. The village, periurban slum and suburban slum show a similar pattern whereas the upper middle class differs in both amplitude and phase as shown in Figure 8. The coefficient of determination, $R^2$ value of this model was 0.05. In the 'cross-sectional' regression model with 'temperature' as regressor this value was 0.04.

The mean incidence of diarrhoeal episodes in relation to calendar month was studied. This reveals a pattern with a higher incidence rate during the warm part of the year. This pattern is common to all groups, although less pronounced in the upper middle class (Figure 9).

## DISCUSSION

In this study we show that variations in seasons affect early child growth in a developing country as described earlier. [2, 10] A longitudinal analysis of our material adds the observation that this effect could be explained by the climate variable minimum day temperature. The impact of the mean day temperature on growth velocity differed according to socio-economic standard, as well as age. Children living in Lahore at a low socio-economic level are more affected in their growth velocity than the children living in the upper middle class. This is most prominent concerning weight gain but to some extent also noticeable in length and weight for length. We propose that mean lowest day temperature of the month can be chosen as a continuous variable describing the seasonal influence on growth.

Growth regulation has in the industrialised world been shown to be affected by external factors, such as nutrition, physical activity or the light and dark cycle, which may vary during the year. It is well known that mean height velocity is higher in the summer than in the winter. [15] A reduced length gain a few months after the summer has been described in developing countries. [10 16] In our study we could not show that length was considerably influenced by the season although there was a tendency for children over 1 year of age to show a declined length velocity at higher temperatures, most clearly shown in group 3. This might be explained by the fact that as the children get older and spend more time outdoors where they may be more vulnerable towards the climate. Lower leg length velocity has been shown to be a sensitive measure to detect seasonal variations, [17] but this measure was not used in our study. Length is a more stable variable than weight and it is not so closely related to nutritional status as weight. [7]

The children in the poorer groups in the study had a low weight gain during the hot season which agrees with several other studies showing a reduced weight gain during the summer months. [16, 18] Seasonal effects of weight gain have been reported to be attenuated over time along with increasing affluence in a developing society, [19] and we show a different seasonal impact at different stages of urban migration. The reduced weight gain during higher temperatures in our study follows the prevalence of diarrhoeal disease during the same time in these groups. Diarrhoeal disease is affecting weight gain for young children and damage to the small intestine has been proposed to be a key feature in the pathogenesis to this faltering in growth. [20] However external factors as high temperature enhances the contamination of food and fluids which might be the reason for the higher prevalence of diarrhoeal disease during the hot season. The impact of temperature on weight gain was higher in children over 1 year of age after the time of weaning. Although little information ia available it has been reported that breast-feeding might attenuate the seasonal variation in diarrhoeal disease. [21] The causal relationship between diarrhoeal disease, breastfeeding and growth is to be further investigated in our material.

Temperature might be used as a variable in studies of short-term growth. Several models have been proposed for modelling seasonal influence when studying different health factors changing over time. [22, 23] The use of sin-cos model is appropriate in detecting seasonal variation [22], although it describes variation over time mechanically, using a mathematical function. Since the annual variation of 'temperature' looks similar to the pattern in the sin-cos regression (peaks in June and December), 'temperature' might be the variable most closley related to the cyclisity in the sin-cos model. The proposal that we put forward is that day temperature can be used as an easy continuous variable when there is a need to adjust for seasonal influence.

The season of the year has a major, but in research often neglected, impact on early child health. When studying determinants of child health in a developing country the implications of climatic seasonality are of major importance. [24, 25, 26] In this study we show that children living in socio-economically poorer groups are more vulnerable towards the effect of climate regarding their weight gain. The impact of the climate on weight gain is age dependent. Together with age, sex and length of gestation season must be considered when comparing growth of children living in poor environments.

## ACKNOWLEDGMENTS

This study was supported by grants from SAREC (the Swedish Agency for Research Cooperation with Developing Countries) King Edward Medical College, Lahore, Pakistan, and the Medical Faculty at Göteborg University.

# REFERENCES

[1] Waterlow JC, Ashworth A, Griffiths M. Faltering in infant growth in less-developed countries. *Lancet* 1980; **2:** 1176-8.

[2] Black RE, Brown KH, Becker S, Alim AR, Huq I. Longitudinal studies of infectious diseases and physical growth of children in rural Bangladesh. II. Incidence of diarrhoea and association with known pathogens. *Am J Epidemiol* 1982;**115:** 315-24.

[3] Mata LJ, Urrutia JJ, Albertazzi C, Pellecer O, Arellano E. Influence of recurrent infections on nutrition and growth of children in Guatemala. *Am J Clin Nutr* 1972; **25:** 1267-75.

[4] Shamebo D, Sandström A, Muhe L, Freij L, Krantz I, Lönnberg G, Wall S. The Butajira project in Ethiopia: a nested case-referent study of under- five mortality and its public health determinants. *Bull World Health Organ* 1993; **71:** 389-96.

[5] Victora CG, Smith PG, Vaughan JP, Nobre LC, Lombardi C, Teixeira AM, Fuchs SM, Moreira LB, Gigante LP, Barros FC. Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil. *Lancet* 1987; **2:** 319-22.

[6] Biddulph J. Child Health in the Third World. *Med J Aust* 1993; **159:** 41-5.

[7] Whitehead R, Paul A, Cole T. Diet and growth of healthy infants. *J Hum Nutr Diet* 1989; **2:** 73-84.

[8] Persson LÅ. Infant feeding and growth -a longitudinal study in three Swedish communities. *Ann Hum Biol* 1985; **12:** 41-52.

[9] Dewey KG, Peerson JM, Brown KH, Krebs NF, Michaelsen KF, Persson LÅ, Salmenpera L, Whitehead RG, Yeung DL. Growth of breast-fed infants deviates from current reference data: a pooled analysis of US, Canadian, and European data sets. World Health Organization Working Group on Infant Growth. *Pediatrics* 1995; **96:** 495-503.

[10] Karlberg J, Ashraf RN, Saleemi M, Yaqoob M, Jalil F. Early child health in Lahore, Pakistan: XI. Growth. *Acta Paediatr* 1993; **390 (Suppl):** 119-49.

[11] Zaman S, Jalil F, Karlberg J, Hanson LÅ. Early child health in Lahore, Pakistan: VI. Morbidity. *Acta Paediatr* 1993; **390 (Suppl):** 63-78.

[12] Zaman S, Jalil F, Karlberg J. Early child health in Lahore, Pakistan: IV. Child care practices. *Acta Paediatr* 1993; **390 (Suppl):** 39-46.

[13] Hagekull B, Nazir R, Jalil F, Karlberg J. Early child health in Lahore, Pakistan: III. Maternal and family situation. *Acta Paediatr* 1993; **390 (Suppl):** 27-37.

[14] Diggle P. Analysis of longitudinal data. Oxford: Oxford University Press, 1995.

[15] Marshall W. Evaluation of growth rate in height over periodes less than one year. *Arch Dis Child* 1971; **46:** 414.

[16] Black RE, Brown KH, Becker S, YunusM. Longitudinal studies of infectious diseases and physical growth of children in rural Bangladesh. I. Patterns of morbidity. *Am J Epidemiol* 1982;**115:** 305-14.

[17] Gelander L, Karlberg J, Albertsson-Wikland K. Seasonality in lower leg length velocity in prepubertal children. *Acta Paediatr* 1994; **83:** 1249-54.

[18] Hauspie R. C, Pagezy H. Longitudinal study of growth of African babies: an analysis of seasonal variations in the average growth rate and the effects of infectious diseases on individual and average growth patterns. *Acta Paediatr* 1989; **350 (Suppl):** 37-83.

[19] Cole, T. J. (1993). Seasonal effects on physical growth and development. In S. J. Ulijaszek & S. S. Strickland (Eds.), Seasonality and human ecology (pp. 89-106). Cambridge: Cambridge Univeristy Press.

[20] Lunn, P. G., Northrop-Clewes, C. A., & Downes, R. M. Intestinal permeability, mucosal injury, and growth faltering in Gambian infants. *Lancet*, 1991 **338**: 907-10.

[21] Bohler, E., Aalen, O., Bergstrom, S., & Halvorsen, S. (1995). Breast feeding and seasonal determinants of child growth in weight in east Bhutan. *Acta Paediatr* **84:** 1029-34.

[22] Karvonen M, Tuomilehto J, Virtala E, Pitkaniemi J, Reunanen A, Tuomilehto-Wolf E, Åkerblom HK. Seasonality in the clinical onset of insulin-dependent diabetes mellitus in Finnish children. Childhood Diabetes in Finland (DiMe) Study Group. *Am J Epidemiol* 1996; **143:** 167-76.

[23] Jones RH, Ford PM, Hamman RF. Seasonality comparisons among groups using incidence data. *Biometrics* 1988; **44:** 1131-44.

[24] Erling V, Jalil F, Hanson LÅ, Zaman S, The impact of the climate on the prevalence of respiratory tract infections in early childhood in Lahore, Pakistan. *Journal of Public Health* 1999 In Press.

[25] Brown K, Black R, Becker S. Seasonal changes in nutritional status and the prevalence of malnutrition in a longitudinal study of young children in rural Bangladesh. *Am J Clin Nutr* 1982 **36:** 303-13.

[26] Chamber R, Longhurs R. Seasonal dimensions to rural poverty: London: Frances Printer Ldt, 1981.

# Legends to figures

*Figure 1*
Weight by age for two boys in the study showing examples of development over time. Boy no 461 was breastfed during the whole study period. Boy no 791 was not breastfed at all. Both boys came from the village.

*Figure 2*
Mean birth weights at time of birth for the different study groups.

*Figure 3a*
Influence of the season on the 'weight gain' during the first six months, all groups pooled.

*Figure 3b*
Influence of the season on the 'weight gain' during the first six months for the different study groups.

*Figure 4*
Influence of the season on the 'weight gain' during six to twelve months for the different study groups.

*Figure 5*
Relation between 'mean weight gain and 'mean monthly minimum temperature' in the village (group 1) at an age of 19 months.

*Figure 6a*
Impact of 'temperature' on 'weight gain' measured by the regression coefficients by age in months for the children living in the village (group 1). Spline with 60% smoothing.

*Figure 6b*
Impact of 'temperature' on 'weight gain' measured by the regression coefficients by age in months for the children living in the periurban slum (group 2). Spline with 60% smoothing.

*Figure 6c*
Impact of 'temperature' on 'weight gain' measured by the regression coefficients by age in months for the children living in the urban slum (group 3). Spline with 60% smoothing.

*Figure 6d*
Impact of 'temperature' on 'weight gain' measured by the regression coefficients by age in months for the children in the upper middle class (group 4). Spline with 60% smoothing.

*Figure 7a*
Impact of 'temperature' on 'gain in length' measured by the regression coefficients by age in months for the children living in the village (group 1). Spline with 60% smoothing.

*Figure 7b*

Impact of 'temperature' on 'gain in length' measured by the regression coefficients by age in months for the children living in the periurban slum (group 2). Spline with 60% smoothing.

*Figure 7c*
Impact of 'temperature' on 'gain in length' measured by the regression coefficients by age in months for the children living in the urban slum (group 3). Spline with 60% smoothing.

*Figure 7d*
Impact of 'temperature' on 'gain in length' measured by the regression coefficients by age in months for the children in the upper middle class (group 4). Spline with 60% smoothing.

*Figure 8*
Sin-cos regression of 'weight gain' on ' month of examination' for the different study groups.

*Figure 9*
Average number of diarrhoeal episodes versus calendar month for children in the different study groups.

# Figures

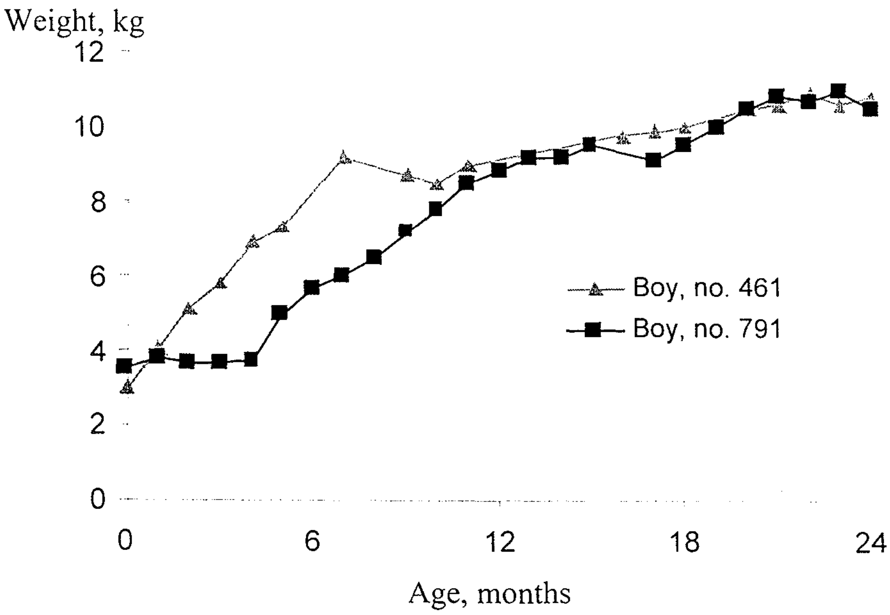

*Figure 1.*

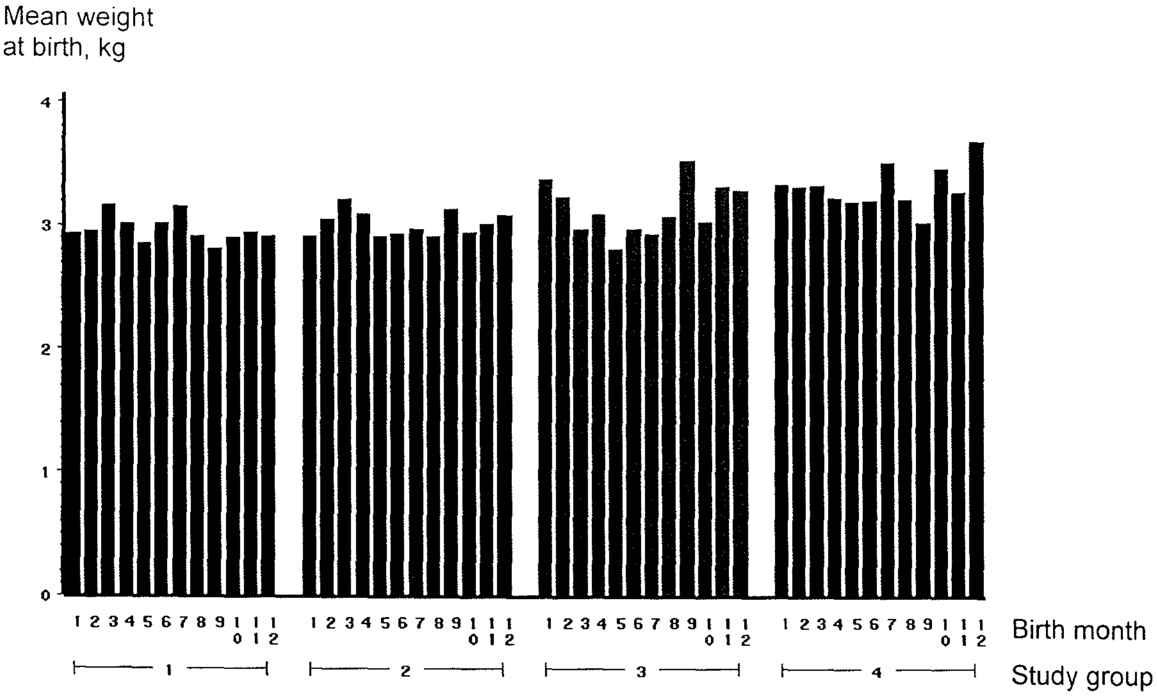

*Figure 2.*

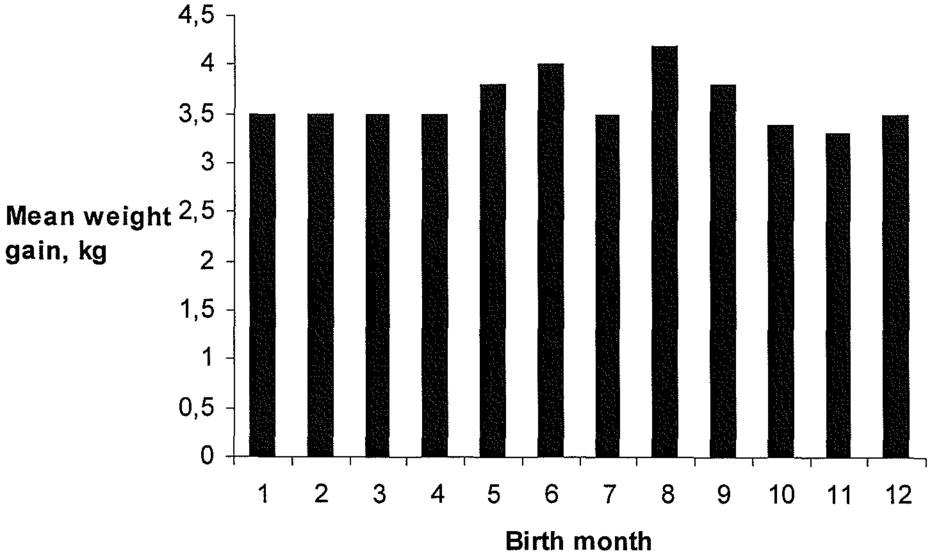

*Figure 3a.*

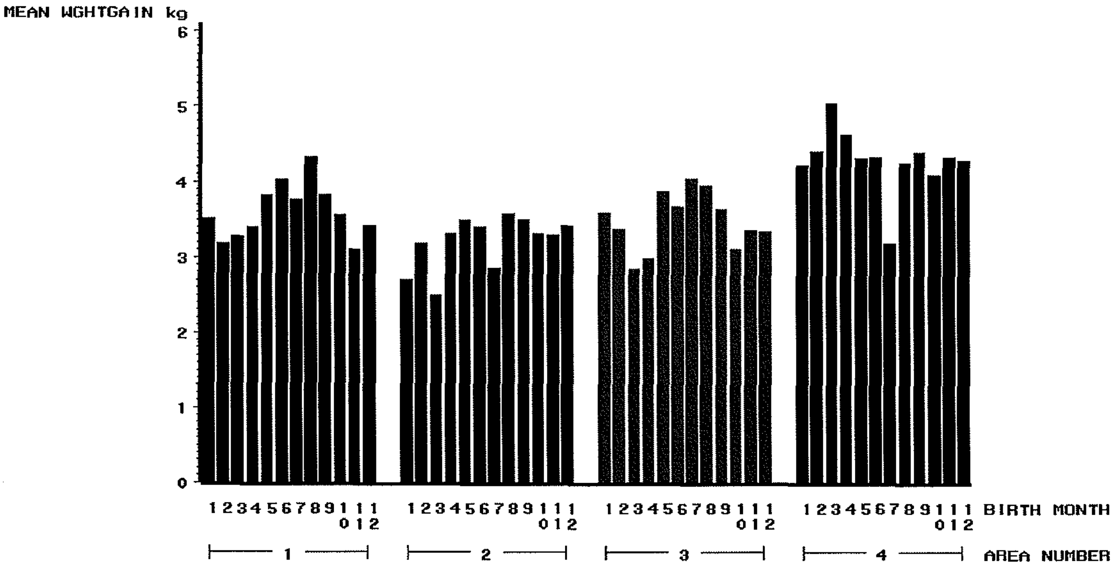

*Figure 3b.*

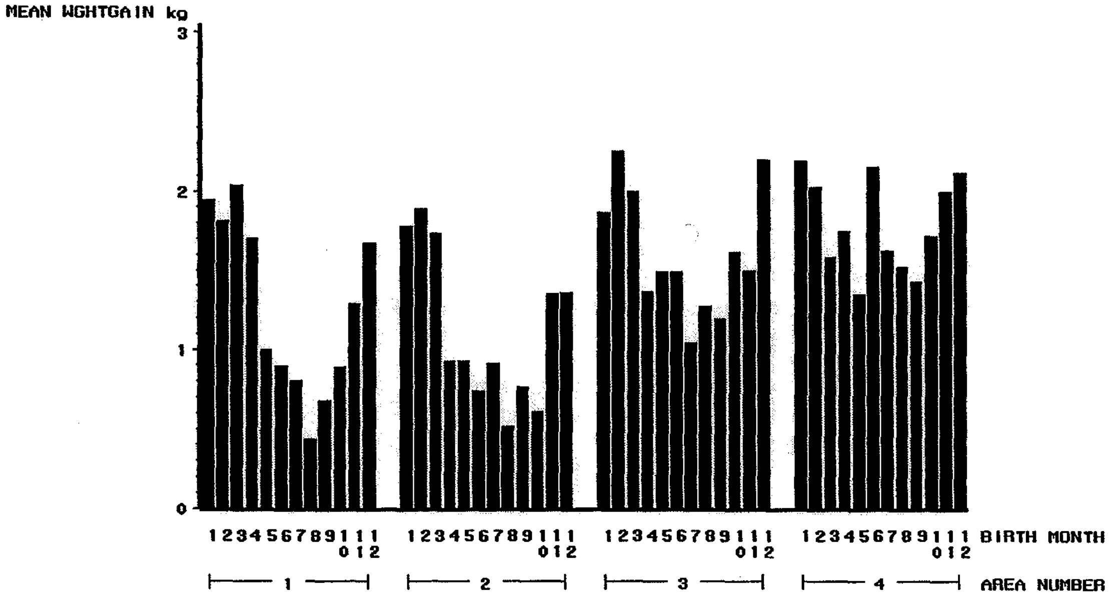

*Figure 4.*

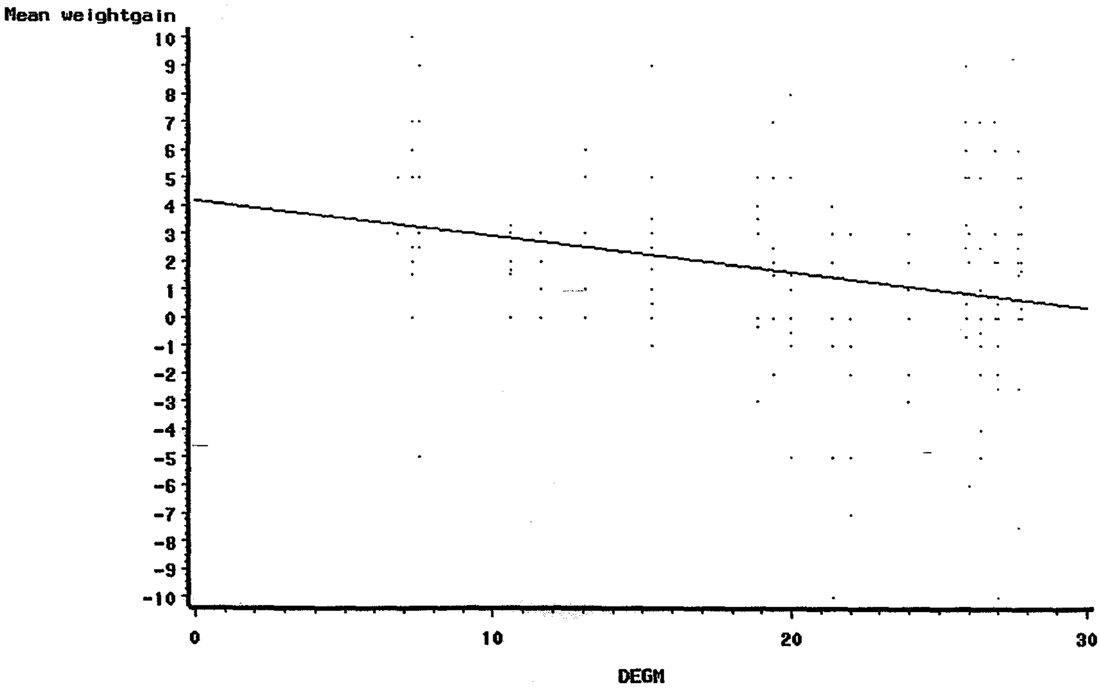

*Figure 5.*

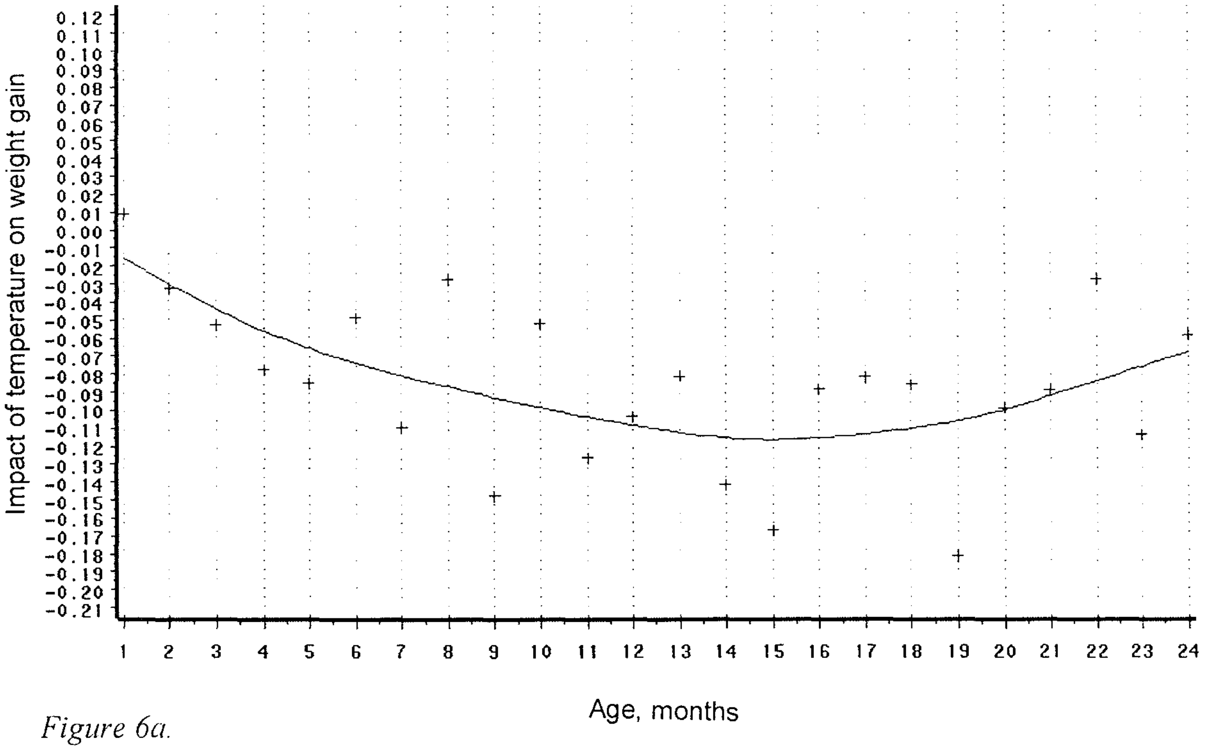

*Figure 6a.*

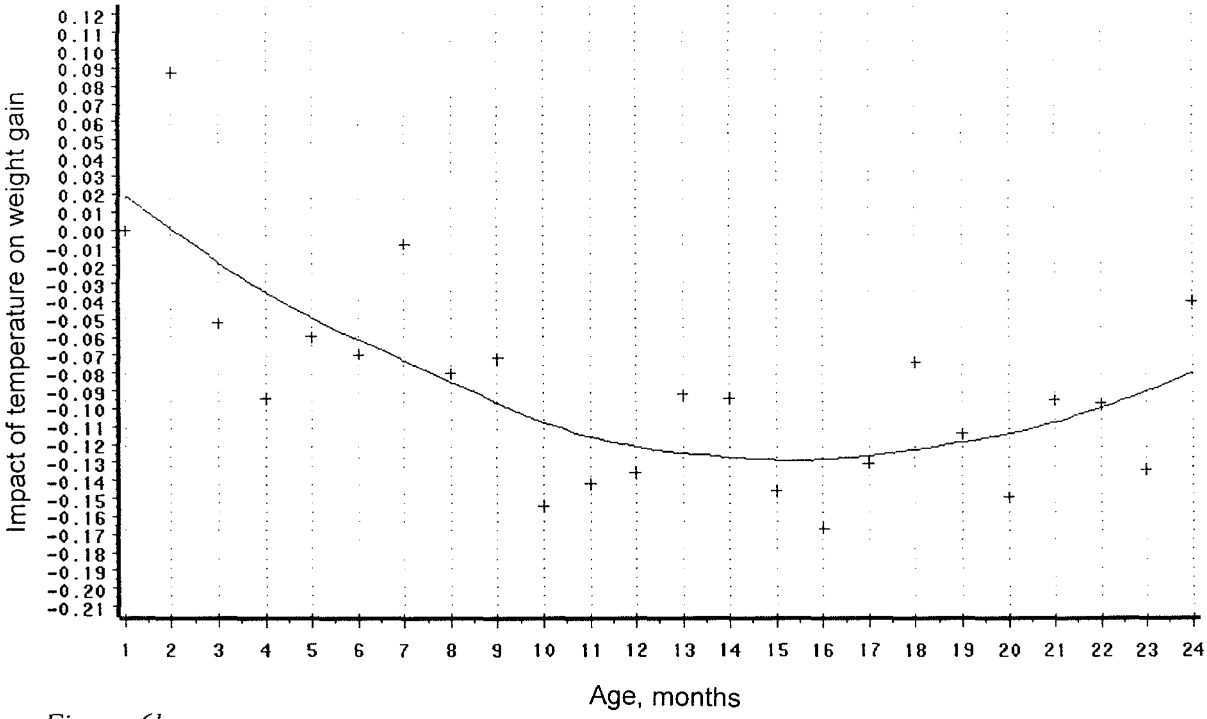

*Figure 6b.*

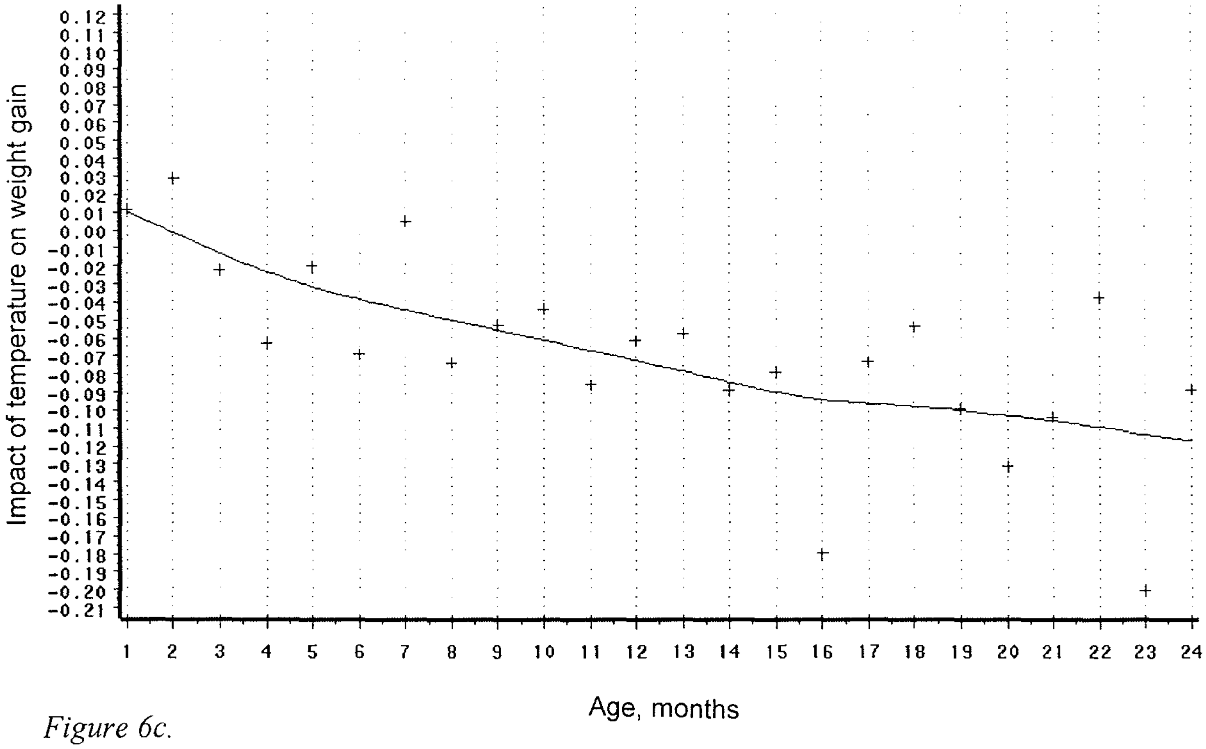

*Figure 6c.*

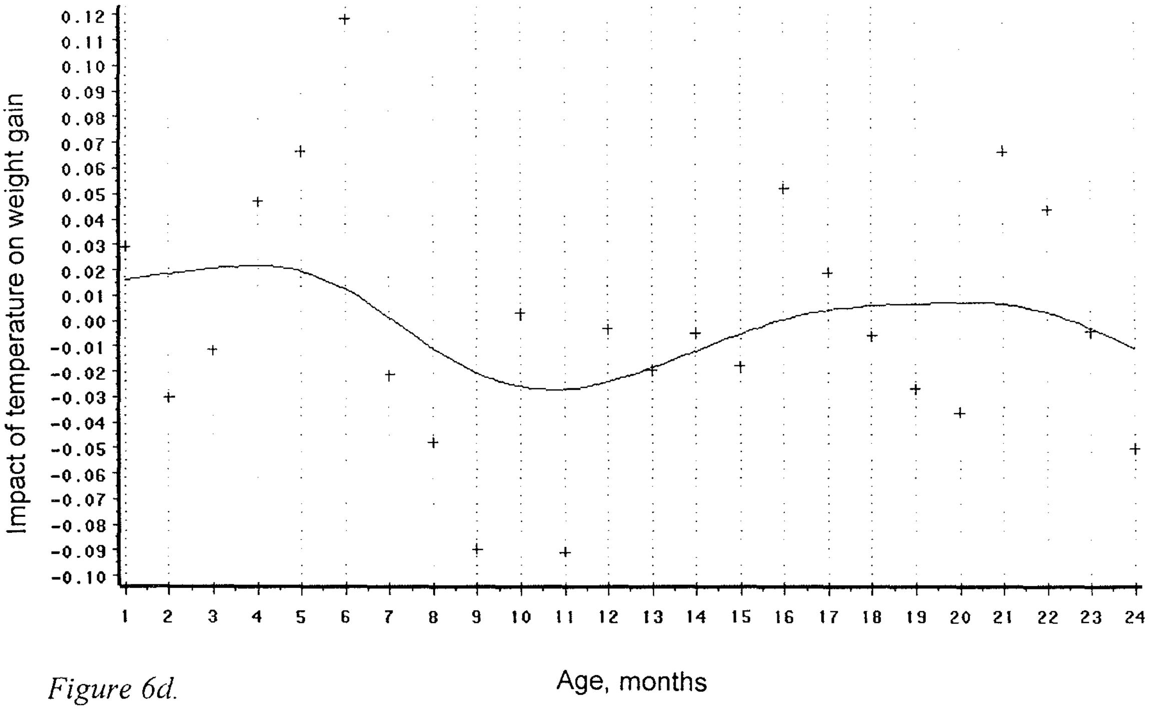

*Figure 6d.*

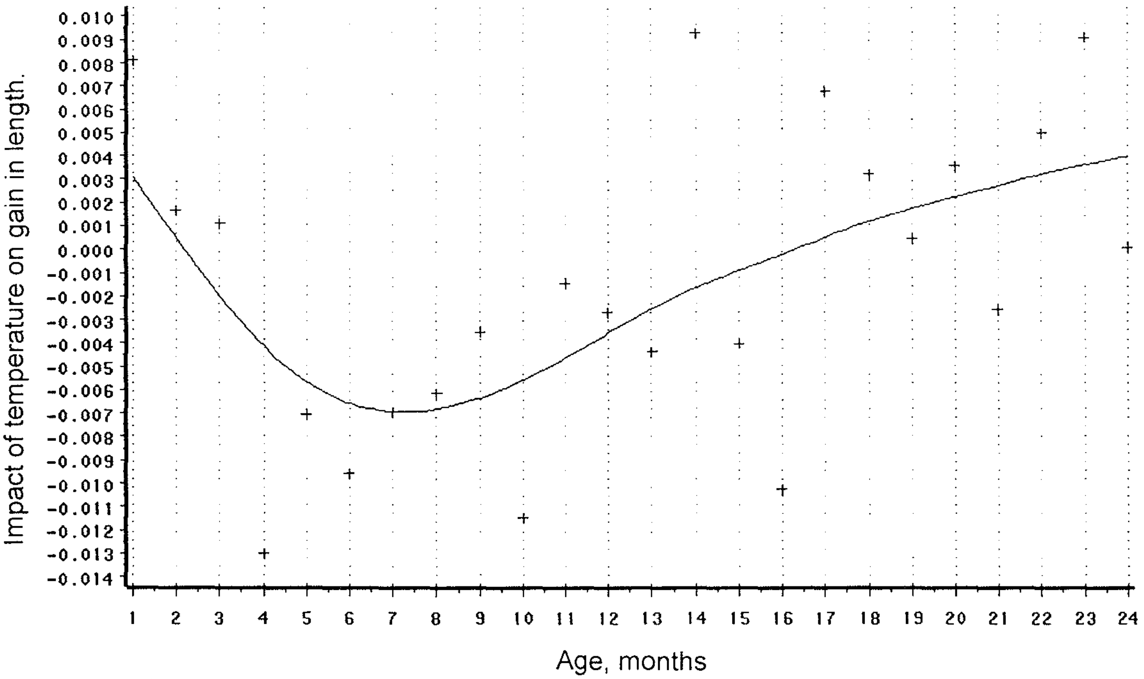

*Figure 7a.*

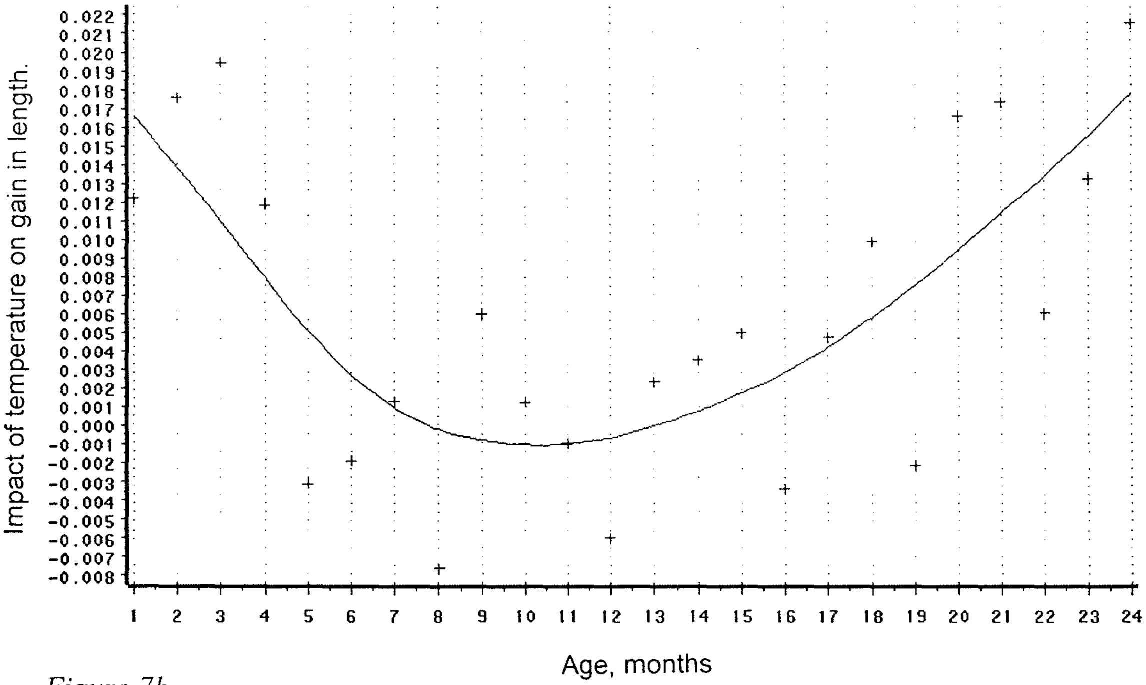

*Figure 7b.*

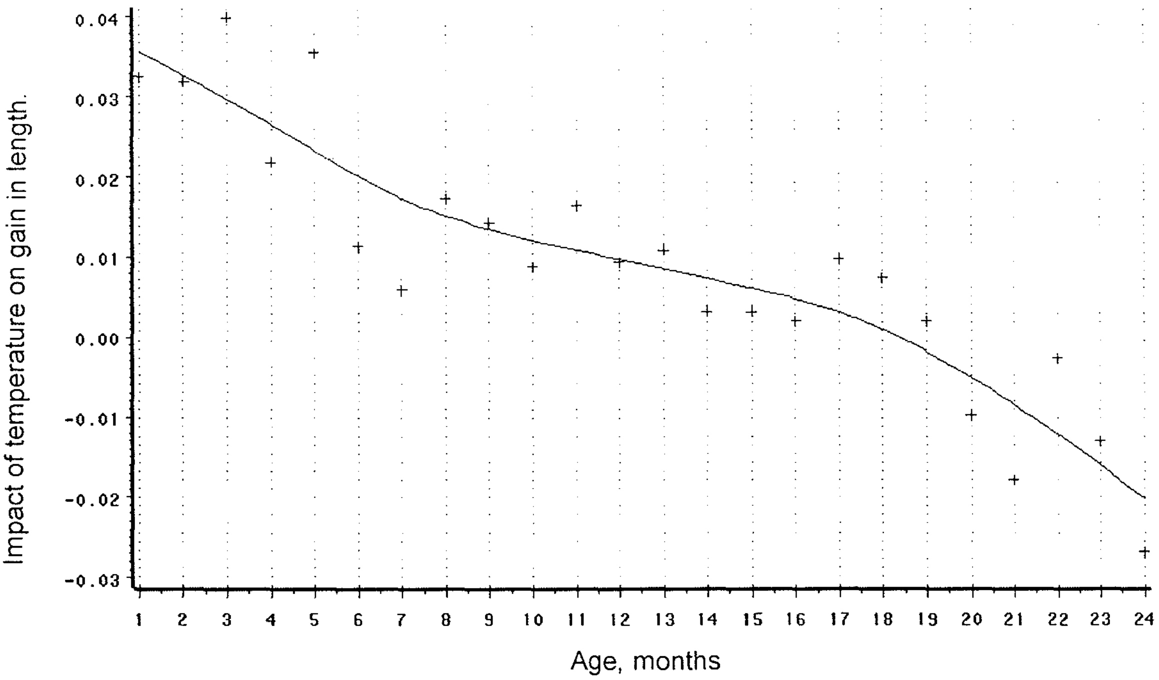

*Figure 7c.*

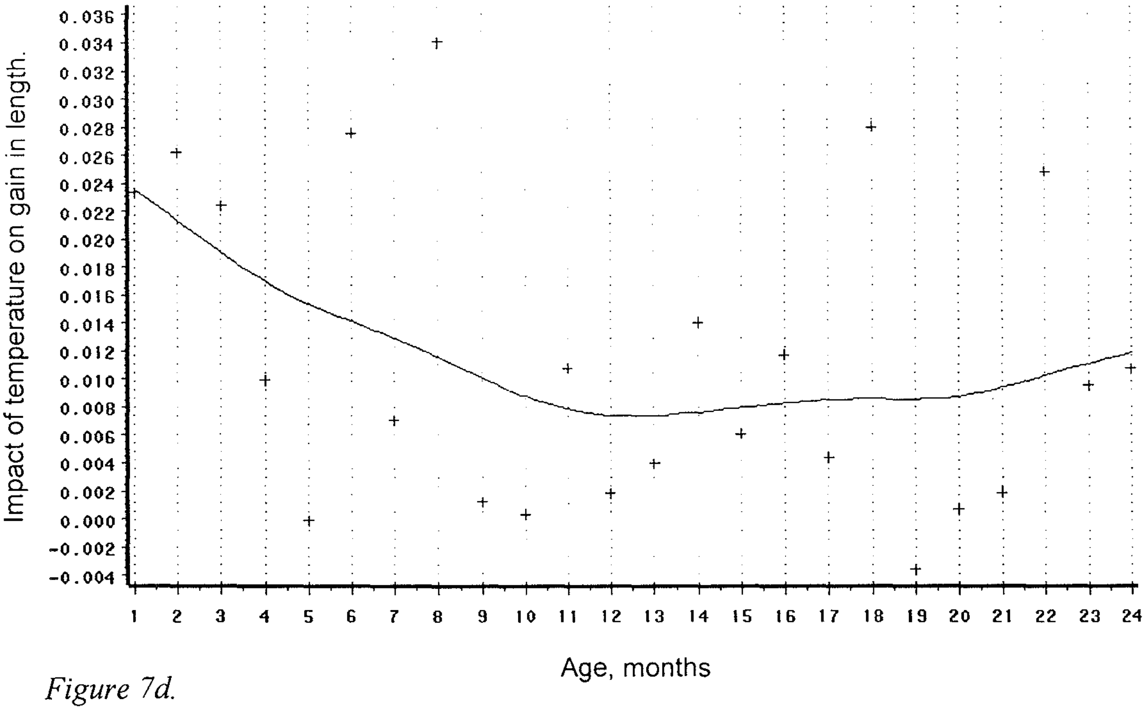

*Figure 7d.*

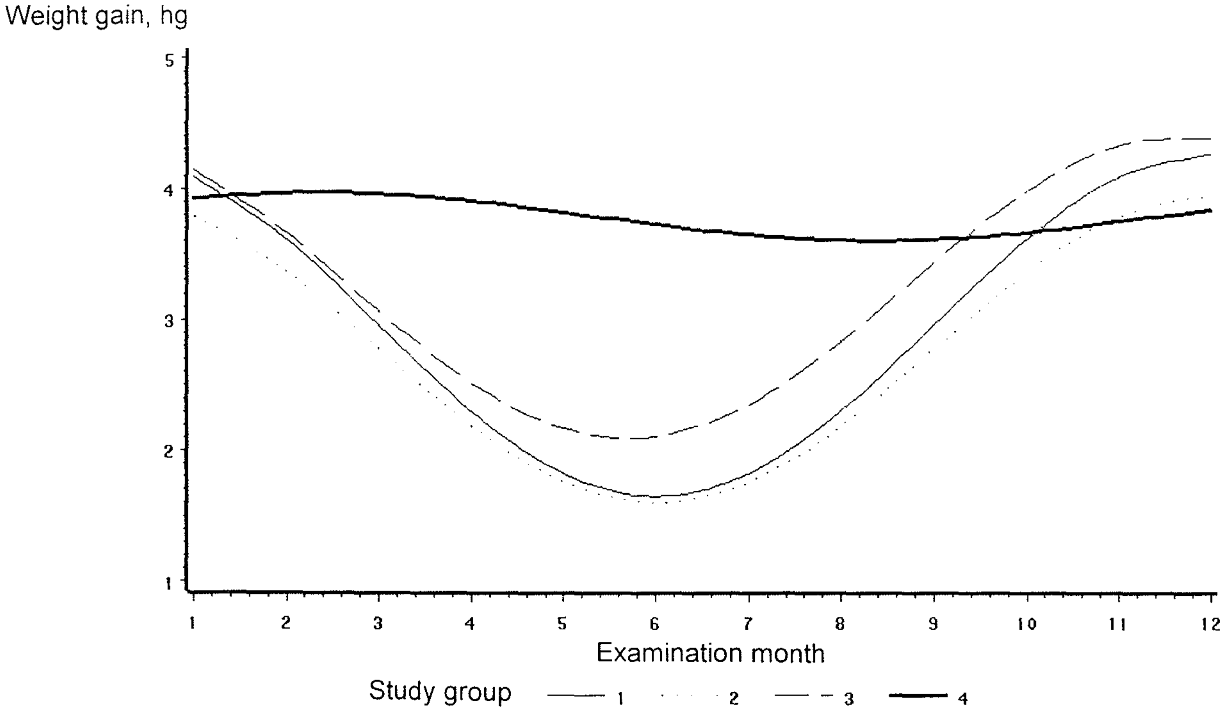

*Figure 8.*

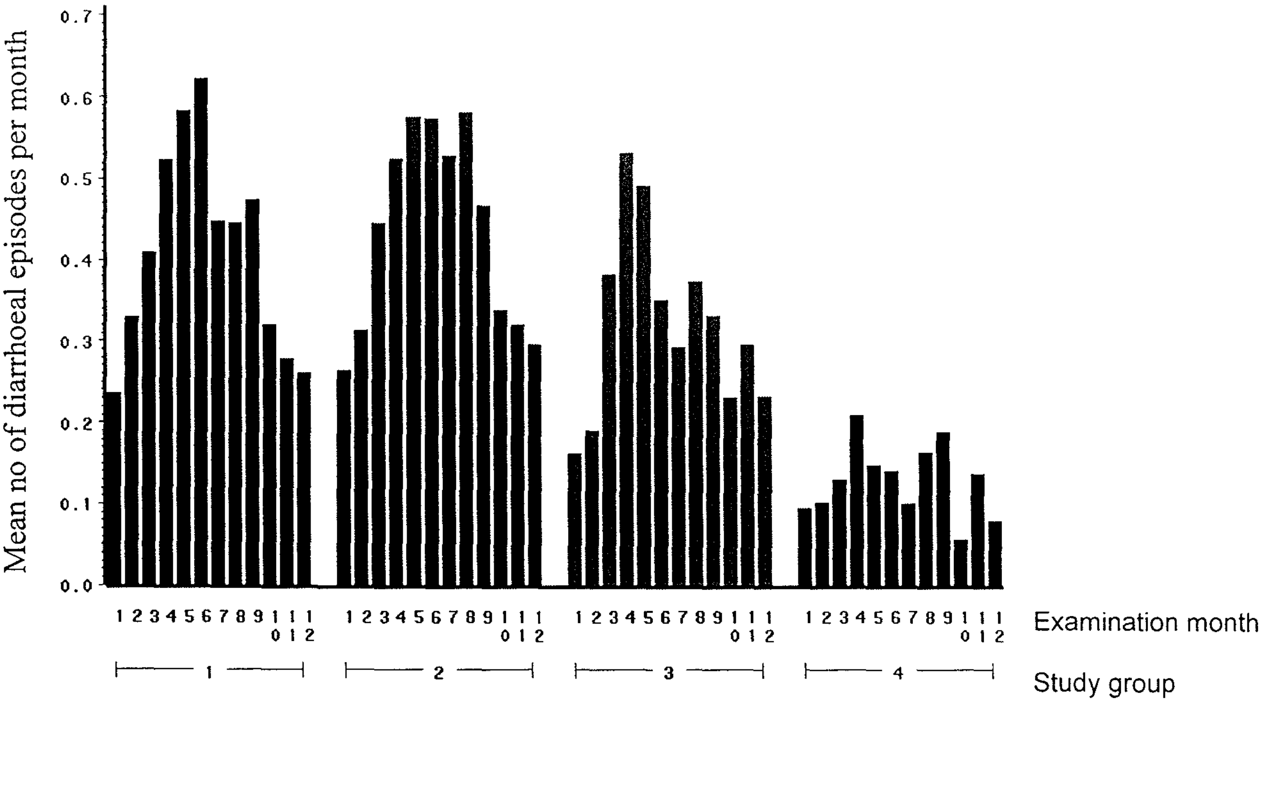

*Figure 9.*

## Legends to table

*Table 1.*
Summary statistics of individual regression analyses for each child in the different areas.

## Table

| | Village | | Periurban slum | | Urban slum | | Upper middle class | |
|---|---|---|---|---|---|---|---|---|
| Variable | Mean | SE | Mean | SE | Mean | SE | Mean | SE |
| Intercept | 9.64 | 2.45 | 8.21 | 0.94 | 10.8 | 2.62 | 13.6 | 5.95 |
| Temperature | -0.125 | 0.069 | -0.157 | 0.046 | -0.245 | 0.112 | -0.226 | 0.224 |
| Age | -0.753 | 0.407 | -0.190 | 0.129 | -0.204 | 0.253 | -0.390 | 0.074 |

*Table 1.*

# 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 healthpromoting 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 $\beta$ is an unknown vector of fixed effects parameters with a known model matrix X, $\gamma$ 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 $\gamma_i$ 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, $\varepsilon_i$, are distributed as;

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

where $W_i$ (within subjects) is a q×$n_i$ 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 $\varepsilon$ 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 $\beta_{DEGM}$ is the main (fixed) effect of temperature, $\beta_{age}$ is the main (fixed) effect of age, $\gamma_i$ 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 $\beta_{BF}$ is the main (fixed) effect of breastfeeding, $\gamma_i$ 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 $\beta_{BF}$. 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).

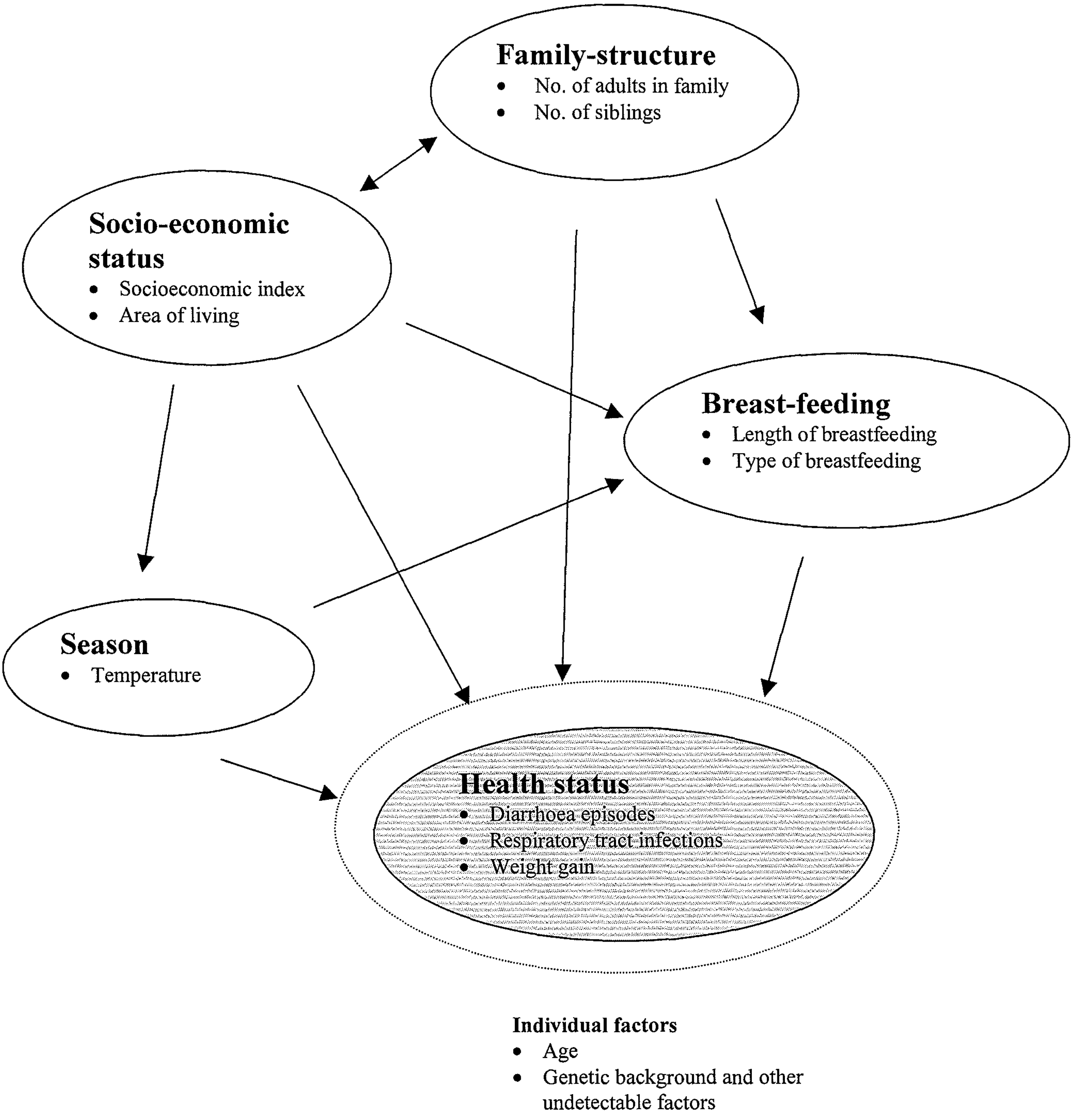

***Figure 1****. Examples of variables that influence breastfeeding and health for children in Lahore Pakistan.*

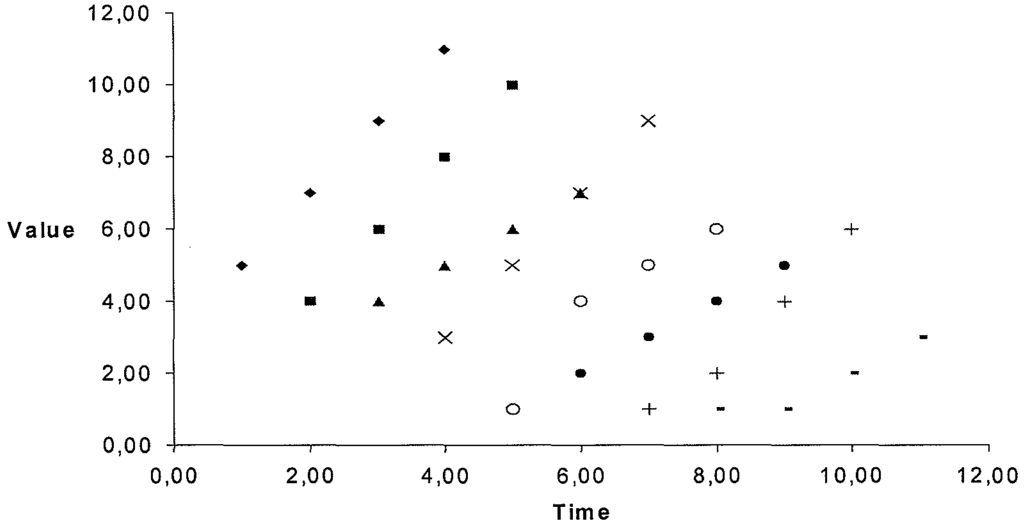

***Figure 2****. Example with 8 individuals measured 4 times each.*

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. |
| 1999:11 | Carlquist, A, Erling, V. & Frisén, M.: | Longitudinal methods for analysis of the influence of breastfeeding on early child in Pakistan. |