Modeling Diarrhea Disease in Children less than 5 years old: A GAM and GLM approach

Sharif Mahmood
Institute of Statistical Research and Training
University of Dhaka, Bangladesh.
sharif@isrt.ac.bd

Abstract

This paper presents the application of generalized additive model (GAM) and generalized linear model (GLM) as an exploratory tool for analyzing the factors that affect the occurrence of diarrhea of Bangladeshi child. The relation between the factors that are related with occurrence of diarrhea can be obtained by modeling parametric approach (GLM). But in practice the relation is not straight forward and we require elaborate explanations which incline semiparametric regression (GAM). We present a unified approach for analyzing factors affecting diarrhea via GLM and GAM. We applied Akaike's information criterion to select the best model for our data. Our study analyzes nonlinear resolution of covariate not available with traditional parametric models and the results provide some evidence on how to reduce occurrence of diarrhea by improving socio-economic and public health conditions.

Keywords: Generalized linear model, Generalized additive model Diarrhea, Smoothing, Akaike's information criterion.

1. Introduction

Child illness and malnutrition are among the most serious socio-economic and demographic problems in developing countries, and they have great impact on future development.Demographic and Health Surveys (DHS) are designed to collect data on health and nutritionof children and mothers as well as on fertility and family planning.The discovery and use of the rehydration solution, during the last decade, has reduced morbidity and mortality in most cases. Despite this, diarrheic disease is still the major cause of death in childhood. In this paper, we focuson occurrence of Diarrhea among childrenunder five years of age,using data from the Bangladesh Demographic and Health Survey (BDHS), 2007.

The success of any policy or health care intervention depends on a correctunderstanding of the socioeconomic, environmental and cultural factors thatdetermine the occurrence of diseases and deaths. Until recently, any morbidityinformation available was derived from clinics and hospitals. Information on theincidence of diarrhea, obtained from hospitalsrepresents only a small proportion of all illnesses, because many cases do notseek medical attention. Thus, the hospital records may not beappropriate for estimating the incidence of diseases for program developments(Woldemical, 2001, D'Souza et al., 2002). Diarrhea disease is a common symptom, patient with diarrhea disease defecates more frequentlythan in normal time, and stool is loose and there is more water, the quantity of defecation is morethan 200g or the quantity of defecation is lower than 200g but the defecation is more than 3 timesassociated with mucus, bloody pus or undigested food.According to the estimate by Streatfield (2001), in Bangladesh, the disease burden caused by diarrhea is as following: 11% of total casualty is caused by

diarrhea, 12.1% of disease to analysis disease burden loss is also caused bydiarrhea.In the Bangladesh Overall, 10 percent of children under five were reported to have had diarrhea in the two-week period before the survey (BDHS, 2007). In 2009 diarrhea was estimated to have caused 1.1 million deaths in people aged 5 and over and 1.5 million deaths in children under the age of 5 over the world.

The use of DHS data in the understanding of childhood morbidity hasexpanded rapidly in recent years (Ryland, 1998, Walter, 2001).However, few attempts have been made to addressexplicitly the problems nonlinear effects ofmetrical covariates in the interpretation of results. This study shows how the GAM model (Hastie et al.,1990) can be adapted to extend the analysis of GLM (Nelder et al.,1972, Dobson, 1990) to provide an explanation of nonlinear relationship of the covariate.Incorporation of nonlinear term in the model improves the estimates in terms of goodness of fit. The GLM model is explicitlyspecified by giving a symbolic description of the linear predictor and a description of the error distribution and the GAM model is fit using the local scoring algorithm, which iteratively fits weighted additive modelsby backfitting. The backfitting algorithm is a Gauss-Seidel method for fitting additive models,by iteratively smoothing partial residuals. The algorithm separates the parametric from the nonparametricpart of the fit, and fits the parametric part using weighted linear least squares within thebackfitting algorithm.

The rest of the paper is organized as follows. Section 2 describes the data and methodology that used in this article Section 3 Model description and estimation procedure applied based on Generalized Additive Models (GAM). Section 4 presents the outcomes obtained and compares the results based on GLM and GAM. Finally, Section 5 summarizes and concludes.

2. Source of Data

The BDHS (2007) was implemented through a collaborative effort of the National Institute of PopulationResearch and Training (NIPORT), Macro International, USA, and Mitra & Associates. Thesurvey is based on a two-stage stratified sample of households. At the first stage of sampling, 361Primary Sampling Units (PSUs) were selected among them 227 rural PSUs and 134 urban PSUs.On average, 30 households were selected from each PSU, using an equal probability systematicsampling technique. In this way, 10,819 households were selected for the sample. However, someof the PSUs were large and contained more than 300 households. Large PSUs were segmented,and only one segment was selected for the survey, with probability proportional to segment size.Households in the selected segments were then listed prior to their selection. Thus, BDHS (2007)sample cluster is either an enumeration area (EA) or segments of an EA. Interviews were successfullycompleted in 10,400 households, or 99.4 % of households.

The main objective of the BDHS,2007 is to provide up-to date information on child illness and treatment and childhood mortality levels, on awareness, health caring methods, and on nutritional level. This isintended to assist policymakers and administrators in evaluating and designing programs,and to develop strategies for improving health facility or a medicallytrained provider for treatment in Bangladeshwhichin turn should reduce occurrence of diarrhea. The 2007 BDHS asked

mothers if each childunder age five had experienced an episode of diarrheain the two weeks before the survey. If the child hadhad diarrhea during this period, the mother was askedwhat she did to treat the diarrhea. Because theprevalence of diarrhea varies seasonally, the surveyresults pertain only to the period from March throughAugust when the fieldwork took place.

A number of demographic and socioeconomic factors can influence occurrence of diarrhea and among these factors sex of child, source of drinking water, toilet facility, mother's education, wealth quintile, place of residence, division and mother's age and child age are considered in the analysis to examine their role in influencing occurrence of diarrhea in Bangladesh. Mother's age and child age are considered as nonlinear relationship with occurrence of diarrhea. Sources of drinking water and toilet facility are factorized in three categories and other variables are considered as it is given in 2007, BDHS.

3. Model Description and Application

To extend the additive model to a wide range of distributionfamilies, Hastie and Tibshirani (1990) proposedgeneralized additive models. These modelsassume that the mean of the dependent variable dependson an additive predictor through a nonlinearlink function. Generalized additive models permit theresponse probability distribution to be any member ofthe exponential family of distributions. Many widelyused statistical models belong to this general class,including additive models for Gaussian data, nonparametriclogistic models for binary data, and nonparametric log-linear models for Poisson data.

In GLM the dependent variable values are predicted from a linear combination of predictor variables, which are "connected" to the dependent variable via a link function.Let Y be a response random variable and $X_1, \dots, X_p$be a set of predictor variables. In generalized linear model a response variable Y scan be viewed as a method for estimating howthe value of Y depends on the values of $X1, \dots, Xp$. The generalized linear model is assumed to be

$$E(Y) = f(X_1, \dots, X_p) = g(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)$$

where $g(.)$ is known as link function. Given a sample of values for Y and$X$, estimates of are $\beta_0, \beta_1, \dots, \beta_p$ often obtained by the least squares method or maximum likelihood method.

The additive model generalizes the linear model bymodeling the expected value of Y as

$$E(Y) = f(X_1, \dots, X_p) = s_0 + s_1(X_1) + \dots + s_p(X_p)$$

where$s_i(X), i = 1, \dots, p$ are smooth functions. The usual linear function of a covariate$\beta_j X_j$is replaced with $s_i(X)$, an unspecified smooth function.These functions are not given a parametric form butinstead are estimated in a nonparametric fashion.

In addition, the additive models require specification of the smooth function using a scatter plot smoother such as loess (a locally weighted regression smoother), running mean, or a smooth spline. The scatter plot smoother used in this application of the additive model is the cubic B-spline. The degree of smoothing in a scatter plotsmoother, for example in a loess, is controlled by the span, which is the proportion of points contained in each neighborhood (the set of X values within a defined distance to X_j). The resulting 'smooth' characterizes the trend of the response variable as a function of the predictor variables.

The algorithm for generalized additive models is a little more complicated. Generalized additive models (GAM) extend generalized linear models in the same manneras additive models extend linear regression models, that is, by replacing the linear form $\alpha + \sum_j x_j\beta_j$ withthe additive form $\alpha + \sum_j s_j(\beta_j)$.

The fitting of the GAM is an iterative looping process involving the scatter plot smooth, the backfittingalgorithm, and the local scoring algorithm, a generalization of the Fisher scoring procedure in a GLM. Each iterations of the local scoring algorithm produces a new working response and weights that are directedback to the backfitting algorithm which produces a new additive predictor using the scatterplot smoother(Hastie, 1992, Hastie and Tibshirani, 1990,Swartzman et al., 1992).

The back fitting and local scoring algorithms consider the estimation of the smoothing term $S_k$the additive model. Many ways are available to approach the formulation and estimation of additive models. The back fitting algorithm is a general algorithm that can fit an additive model using any regression-type smoothers. Define thej^{th} set of partial residuals as

$$R_j = Y - s_0 - \sum_{k\neq j} s_k(x_k).$$

The partial residuals remove the effects of all the other variables from y; therefore they can be used to model the effects against x_j.This is the foundation for the back fitting algorithm, providing a way for estimating each smoothing function $s_j(.)$ given estimates $\{s_i(.), i \neq j\}$; for all the others. The back fitting algorithm is iterative, startingwithinitialfunctions$s_0, \ldots, s_p$and an iterationcyclingthroughthepartialresiduals,fittingtheindividualsmoothingcomponents to its partial residuals. Iteration proceeds until the individual components do not change.The algorithm so far described fits just additive models.

In the same way, estimation of the additive terms for generalized additive models is accomplished by replacing the weighted linear regression for the adjusted dependent variable by the weighted back fitting algorithm, essentially fitting a weighted additive model. The algorithm used in this case is called the local scoring algorithm. It is also an iterative algorithm and starts with initial estimates of $s_0, \ldots, s_p$. During iteration, an adjusted dependent variable and a set weight are computed, and then the smoothing

components are estimated using a weighted back fitting algorithm. The scoring algorithm stops when the deviance of the estimates ceases to decrease.

Overall, then, the estimating procedure for generalized additive models consists of two loops. Inside each step of the local scoring algorithm (outer loop),a weighted back fitting algorithm (inner loop) is used until convergence. Then, based on the estimates from this weighted back fitting algorithm, a new set of weights is calculated and the next iteration of the scoring algorithm starts. Any nonparametric smoothing method can be used to obtain $s_j(x)$. The GAM procedure implements the B-spine and local regression methods for univariate smoothing components and the thin-plate smoothing spline for bivariate smoothing components.

A unique aspect of generalized additive models is the non-parametric functions of the predictor variables. Specifically, instead of some kind of simple or complex parametric functions, Hastie and Tibshirani (1990) discuss various general scatter plot smoothers that can be applied to the X variable values, with the target criterion to maximize the quality of prediction of the (transformed) Y variable values. One such scatter plot smoother is the cubic smoothing spines smoother, which generally produces a smooth generalization of the relationship between the two variables in the scatter plot. Computational details regarding this smoother can be found in Hastie and Tibshirani (1990, see also Schimek, 2000).

A step-wise GAM is performed to determine the best fitting model based on the criteria of the lowestAkaike Information Criterion (AIC) test statistic which is a function of both the log likelihood function and theeffective number of parameters being estimated. The AIC in the step-wise GAM (Hastie, 1992) is calculatedas:

$$AIC = D + 2df\varphi$$

where$D$= Deviance (residual sums of squares),

df = effective degrees of freedom, and

φ= dispersion parameter (variance).

The model with the lowest $\boldsymbol{AIC}$ is considered to have the best number of parameters to include in the final model. The deviance estimated in the model, analogous to the residual sums of squares, is a measure of the fit of the model. A pseudo coefficient of determination,$\boldsymbol{R^2}$, is estimated as 1.0 minus the ratio of the deviance of the model to the deviance of the null model (Swartzman et al., 1992).

4. Analysis of Data

It is believed that the diarrhea cause degradation in the nutritional stateand that successive episode may compromise physical development ininfants, leading to malnutrition. However, the risk that undernourishedchildren are more likely to develop diarrhea is as yet inconclusive. In thesechildren, however, an episode of diarrhea is more serious due to its longerduration.Diarrhea affects mainly children in their first year of life,

but especially atweaning age. During this period a higher mortality rate is observed, and thenutritional consequences are more serious.

In BDHS 2007, Bangladesh is considered have six divisions.Diarrhea situation in each division is not same. From the diarrhea situation analysis division is one of most independent variable for this study.The table given below shows an overall scenario of diarrhea in Bangladeshi child by division.

Table 1: Total number and percentage of diarrhea in Bangladesh by Division

			No diarrhea	Haddiarrhea	Total
Division	Barisa	Count(% within Division)	682 (91.2%)	66 (8.8%)	748 (100%)
	Chittagong	Count(% within Division)	1070 (89.1%)	131(10.9%)	1201 (100%)
	Dhaka	Count(% within Division)	1096 (89.5%)	129(10.5%)	1225 (100%)
	Khulna	Count(% within Division)	623 (91.5%)	58 (8.5%)	681 (100%)
	Rajshahi	Count(% within Division)	846 (92.3%)	71 (7.7%)	917 (100%)
	Sylhet	Count(% within Division)	911 (89.7%)	105 (10.3%)	1016 (100%)
Total		Count(% within Division)	5228 (90.3%)	560 (9.7%)	5788 (100%)

From the above table we see that Dhaka and Chittagong division are more affected area than other four divisions in Bangladesh.Rajshahiand Khulna are less affected area from other divisions. Again percentage of occurring diarrhea in rural area is higher than urban area. To get an overall scenario of diarrhea with different covariates we need to explore these by modeling.

In this study, three different models are used for analyzing occurrence of diarrhea in Bangladesh. Model1 is a generalized linear model where we consider sex, sources of drinking water (SDW), Total toilet facility (TTF), mother's education, wealth index of the family, residence and division with diarrhea. In Model2 we added two more independent variable Mother's age and Child age with model1. In Model3 we consider Mother's age and Child age as nonlinear smoothing function.

Table2: A comparison of parametric and semiparametric models of the diarrhea disease in children less than 5 years old in Bangladesh

	Model 1	Model 2	Model 3
INTERCEPT	-1.470[a]	-0.858[c]	-1.449[c]
SEX			
Female	-0.135	-0.136	-0.138
Male	-	-	-
SDW			
Piped	-0.382[c]	-0.371[c]	-0.367[c]
Tube well	-0.527	-0.527	-0.505
Others	-	-	-
TTF			
Flush	-0.330[c]	-0.350[c]	-0.357[c]
Pit toilet	0.033	0.001	-0.009
Others	-	-	-
EDUCATION			
Higher	-1.061[a]	-1.103[a]	-1.104
Secondary	-0.085	-0.161	-0.164
Primary	-	-	-
WEALTH			
Rich	-0.181	-0.177	-0.189
Middle	0.228	0.222	0.217
Poor	-	-	-
RESIDENCE			
Rural	-0.044	-0.034	-0.032
Urban	-	-	-
DIVISION			
Chittagong	0.185	0.166	0.166
Dhaka	0.089	0.072	0.077
Khulna	-0.119	-0.123	-0.107
Rajshahi	-0.234	-0.267	-0.267
Sylhet	0.043	0.036	0.045
Barisal	-		
MOTHER'S AGE	-	-0.013	n
CHILD AGE	-	-0.124[a]	n[a]
[illegible]	3423.184	3407.634	3397.144

[a]= pvalue< .001, [b]= pvalue< .01, [c]= pvalue< .05, n=nonparametric estimate.

In this analysis we see sources of drinking water (SWD) and total toilet facility (TTF) have significant association with occurrence of diarrhea because many people of our country use pit toilet(means hanging toilet,bush, ponds,fields, etc.) and other toilet and free from pure water.That is why these are most important factorsof causing diarrhea.The occurring diarrhea at Chittagong, Dhaka and Sylhet division is higher than Barisal division. The probability of occurring diarrhea for rural and urban area has no significant

difference. Education of motheris another significant factor of occurring diarrhea because most of people of our country are illiterate they do not know how the life is free from diarrhea.

We see that Residual degrees of freedom and Residual Deviance for smooth analysis is less than without smooth analysis and AIC for Model1 is greater than $AIC$of model3 which means Model3 interprets the data quite well and generalized additive model fits well and explain more information than generalized linear models.

We estimated a logistic GAM with smoothing applied to the measures of mothers age and child age. At this stage, we could either conduct a series of likelihood ratio tests or plot the two nonparametric estimates and inspect them for nonlinearity. Visual inspection of the plots may be enough to understand which terms are nonlinearly related and nonparametric estimates. The visual test is quite clear that child age is nonlinearly related whereas mother's age is almost linearly related hence obviously linear functional forms. Figure 1 contains plots of the two nonparametric estimates. The reader may have noticed that the scale for the y-axis in Figure 1is unusual. There are two reasons for this. First, all the variables are mean deviated by the estimation algorithm to increase numerical stability. This is why all the effects are centered at 0 on the y-axis. Second, we plotted the nonparametric estimates in the untransformed scale of the linear predictor, and therefore,the predictions for dependent variable are not on a probability scale. It is standard practice toconvert logistic regression parameter estimates to odds ratios or predicted probabilities,and the same can be done for the nonparametric estimates, which wewill do shortly. However, for visual diagnosis of nonlinearity, it is best to plotin the linear scale, since applying the link function can obscure the form of thenonlinearity.

Next, we use a likelihood ratio test to decide whether the effect of mother's age and child ageare significantly nonlinear altogether. The test indicates that the spline estimate for the effect of mother's age and child age is superior to using a parametric term. The final specification uses a spline fit for these variables but the other terms are modeled parametrically.We tested this specification against a model which was fully parametric.

Converting the nonparametric estimate to the scale of the link function is donein an identical fashion converting to the coefficients to this scale.Holding other covariatesin the model constant at some appropriate category or value, we calculate thepredicted value of dependent variable for each value of the nonparametric estimate.We convert these predicted values that are in the scale of the linear predictor to a probability scale using the link function.

Model3 is estimated with smoothing splines for Mother's age and Children age and the results are presented in the Figure1.

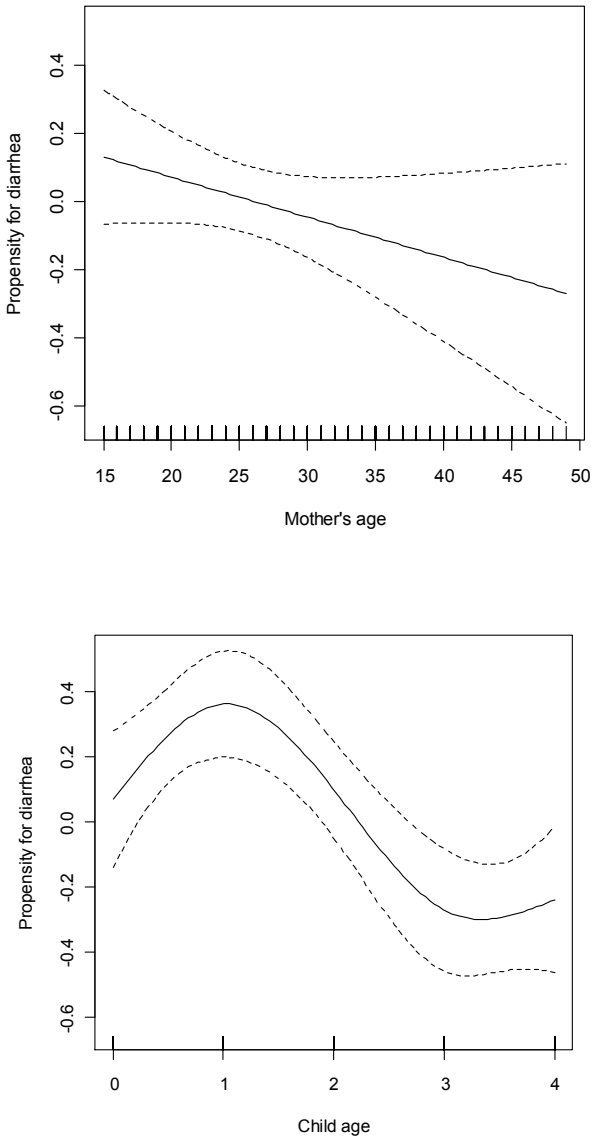

Figure1: Generalized additive model for diarrhea disease in children less than 5 years in Bangladesh as a function of Mother's age and Children age.

Generalized additive models are very flexible, and can provide an excellent fit in the presence of nonlinear relationships and significant noise in the predictor variables. However, note that because of this flexibility, you must be extra cautious not to over-fit the data, i.e., apply an overly complex model (with many degrees of freedom) to data so as to produce a good fit that likely will not replicate in subsequent validation studies. In other words, evaluate whether the added complexity (generality) of generalized additive models (regression smoothers) is necessary in order to obtain a satisfactory fit to the data. Often, this is not the case, and given a comparable fit of the models, the simpler generalized linear model is preferable to the more complex generalized additive model.

Conclusion

Diarrhea remains a leading cause of childhood morbidity and mortality in developing countries. Dehydration caused by severe diarrhea is a major cause of illness among young children. The prevalence of diarrhea is highest at age low aged child, a period during which solid foods are first introduced into the child's diet. This pattern is believed to be associated with increased exposure to illness as a result of both weaning and the greater mobility of the child, as well as with the immature immune system of children in this age group. The prevalence of diarrhealiving in Chittagong, Sylhet, and Dhaka divisions are quite high. Again children whose source of drinking water are not improved and living in households with on-improved or shared toilet facilities than among other children. The chance of occurring diarrhea is lowest among children of mothers who had completed secondary or higher education and children living in the wealthiest households. However, only about one in four children who received ORT also received zinc. Children living in urban areas and in Dhaka and Khulna divisions are more likely to have received both ORT and zinc. Children whose mothers completed secondary or higher education and those in the highest wealth quintile were more likely to receive both ORT and zinc than children of mothers with no education and children in the lowest wealth quintile.

Despite high prevalence, rotavirus diarrhea can successfully and confidently be managed at home and in the oral rehydration corner of small hospitals.The results of this study are expected to reduce the number of referral of diarrhea patients from rural centers to secondary/tertiary level hospitals which almost always remain over-occupied. Moreover, treatment of diarrhea at home and nearby hospital can save many working hours of the parents, who must accompany the ailing children.

Persistent diarrhea is both uncomfortable and dangerous to the health, as it can indicate an underlying infection. It may also mean that the body is not able to absorb some nutrients due to a problem in the bowels. Treatment includes drinking plenty of fluids to prevent dehydration, over-the-counter remedies in most cases, and medical examination if diarrhea persists for more than a couple of days, particularly in small children or elderly people.

Acknowledgments

I would like to acknowledge permission granted by MEASURE DHS to use the Bangladesh DemographicHealth Survey, 2007.

Bibliography

1. Dennehy PH (2000). Transmission of rotavirus and other enteric pathogens in the home. Pediatr. *Infect. Dis. J.* 19 (10 Suppl): S103–5.
2. D'Souza A.L., Rajkumar, C., Cooke, J., Bulpitt, C. J. (2002). Probiotics in prevention of antibiotic associated diarrhoea: meta-analysis. *British Medical Journal* 324: 1361 doi: 10.1136/bmj.324.7350.1361
3. Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

4. Estes MK, Palmer EL and Obijeski JF (1983). Rotaviruses: a review. *Current Topics in microbiology and Immunology* 105: 123-184.

5. Guerrant, R.L., Mcauliffe, J.F. (1986). *Special problems in developing countries*. In: Gorbach, S.L. ed. Infections diarrhea, Boston:Blacwekk, cap. 19, 287-308.

6. Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, New York: Chapman and Hall.

7. Hastie, T. J. (1992). *Generalized additive models In: Statistical models in S*. Chambers, J. M. and T. J. Hastie (eds). Wadsworth and Brooks, Pacific Grove.

8. Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A 135: 370-84.

9. Ryland, S. and Raggers, H. (1998). Child Morbidity and Treatment Patterns. DHS Comparative Studies, No.27, Claverton, Maryland: Macro International.

10. Swartzman, G., Huang, C. and Kaluzny, S. (1992). Spatial analysis of Bering Sea ground fish survey data using generalized additive models. Can. J. Fish. *Aquat. Sci*., 49: 1366–1378.

11. Snyder, J.D. Merson, M.H. (1982). The magnitude of the global problem of acute diarrhoeal disease: a review of active surveillance data. *Bulletin of the World Health Organization*, 60(4):605-613,

12. Stone, C.J. (1986). Comment: Generalized Additive Models. *Statistical Science* 2: 312-314.

13. Streatfield K, Persson LA, Chowdhury HR, Saha KK. (2001). *Disease Patterns in Bangladesh: Present and Future Needs. Dhaka, Bangladesh*: International Centre for Diarrhoeal Disease Research, Bangladesh.

14. Walter D.R.O. (2001). *Child Morbidity in Kenya: Does Women's Status Matter?, Paper Presented at the Canadian Population Society 2001 Annual General Meeting*, Laval University, Quebec City, May 27-29.

15. Woldemical, G. (2001). Diarrhoeal Morbidity among Young Children in Erithrea: Environmental and Socioeconomic Determinants. Jour. *Health Pop. Nutr*., 19(2):83-90.