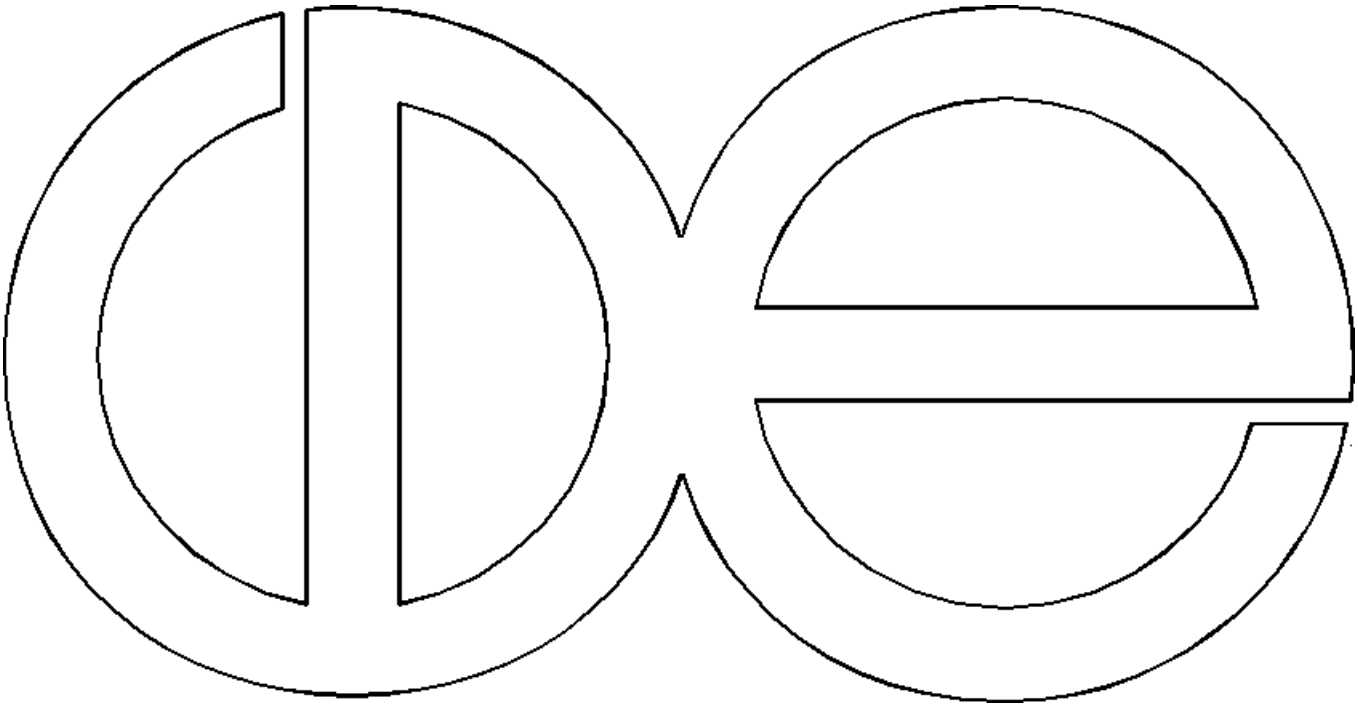
Center for Demography and Ecology

University of Wisconsin-Madison

Analysis of Child Mortality with Clustered Data: a Review of Alternative Models and Procedures

Cibele Comini Cesar
Alberto Palloni
Hantamalala Rafalimanana

CDE Working Paper No. 97-04



Analysis of Child Mortality with Clustered Data:

a Review of Alternative Models and Procedures

Cibele Comini Cesar[1]
Alberto Palloni[2]
Hantamalala Rafalimanana[2]

[1]Departamento de Estatistica, Universidade Federal de Minas Gerais.
[2]Center for Demography and Ecology, University of Wisconsin

I. ANALYSIS OF CHILD MORTALITY AND CLUSTERING

1. Objective

In this work we evaluate the performance of alternative methods to estimate parameters of models with data on child mortality that are clustered. Most commonly, clustering occurs as a result of sample design that focuses on mothers or households and on all the children associated with them. To the extent that the risks of mortality of children belonging to the same household are correlated, standard estimation procedures can produce faulty results. Clustering can also occur at higher levels of aggregation (neighborhood, villages, communities), but the nature of the problem remains unchanged. Irrespective of the level at which clustering occurs, the problems of estimation can be viewed as part of more general estimation difficulties related to problems created by omitted covariates.

The 'clustering' problem was identified and dealt with in the literature on child mortality using simple and very crude approaches that, as we now know in retrospect, could lead to misinterpretations (Cleland and Sathar, 1984; Pebley and Stupp, 1987; DeSweemer, 1984; Gribble, 1993; Hobcraft, McDonald and Rutstein, 1985; Koening et al., 1990; Lantz, Partin and Palloni, 1992; Miller, 1989; Miller et al., 1992; Retherford et al., 1989, Winikoff, 1983; Palloni and Tienda, 1986; Palloni and Millman, 1986). More recently, estimation of mortality models with data gathered under clustered designs has been confronted with more sophisticated procedures (Curtis, 1991; Curtis, Diamond and McDonald, 1991; Curtis and Steele, 1994; Madise and Diamond, 1995; Guo and Grummer-Strawn, 1993; Guo and Rodriguez, 1991; Guo, 1993; Sastry, 1995a, b). These methods have clear merits over their predecessors because they explicitly and purposively model what was ignored before, namely, the correlation between siblings' mortality risks. However, their nature remains somewhat obscure to the mainstream of demographers, and their performance has not been evaluated in realistic demographic settings.

Furthermore, most of these methods are suitable for binary response models and have not been extended to and tested in the context of hazard models, now commonly applied in the study of child mortality. In this paper, we partially fill these gaps and evaluate five methods for the study of child mortality with correlated sibling data using both binary response models and hazard models. The paper is by design more thorough in its review of the problem as it presents itself in binary response models than in hazard models. The plan of the paper is as follows: Section II reviews basic estimation approaches for the analysis of correlated sibling survival data; Section III uses simulation to assess the performance of the approaches introduced in Section II. In Section IV we discuss the application of these methods to data from Brazil and, finally, in Section V we address some issues of interpretation. The results presented in the paper can be easily extended to empirical problems involving outcomes other than death/survival.

2. Omitted variables and clustering

During the past fifteen years, a large set of studies—principally based on analyses of data from developing countries that carried out World Fertility Surveys (WFS) and Demographic and Health Surveys (DHS)—show a strong negative association between short birth intervals and infant and child survival. There are few empirical relations in the demographic literature as consistent as this one. It appears in disparate cultural and demographic settings and persists even when controlling for confounding effects of relevant covariates such as prematurity, breastfeeding, socioeconomic status, and other bio-demographic correlates of child mortality (for a review see Hobcraft, 1992).

Despite the surprising invariance of these findings, their validity has been questioned due to concerns about the influence of confounding factors. The literature on the role of confounding factors in these models has two different strands. In the first the problem is conceptualized as one of omission of **individual characteristics** that affect the hazards or odds of a child's death. The

omitted covariate is **specific** to the child and possibly distinct from those affecting other children in the sample. This is analogous to the classic problem of omitted variables in linear models. It is known in the literature on hazard models as the problem of unmeasured heterogeneity (Vaupel, Manton and Stallard, 1979; Heckman and Singer 1982; Trussell and Rodriguez, 1990; Manton, Singer and Woodbury, 1992). As is well known, unmeasured heterogeneity in hazard models leads to inconsistent estimates of effects.

In a second strand of the literature the focus is on unmeasured **family characteristics** affecting mortality risks of children born to a mother or living in the same household. One or more covariates shared by children of the same mother or living in the same household will induce a correlation between their mortality risks. An empirical manifestation of this mechanism found in developing countries is the concentration of child deaths in a few families (Das Gupta, 1990; Potter, Das Gupta, and Wyshak, 1989). More commonly, in a typical demographic survey individual child mortality data are gathered from households and include all children of (some) mothers living in such households. There surely are unobserved familial factors (such as shared genetic endowments, common environmental risks, and behaviors to which siblings are exposed) that could lead to correlation between survival outcomes of siblings. Examples of these factors are unmeasured genetic propensity to certain diseases, household pollution, family diet, levels of hygiene and cleanliness (Klein 1992). Also, unmeasured parental competence, such as a mother's natural (in)ability to care for her children, is a possibly influential factor to explain the concentration or clustering of child deaths within "high-risk" families (Das Gupta 1990; Potter, Das Gupta, and Wyshak, 1989). In this literature the problem is referred to as the ‘clustering’ problem.

Although not exclusively, the statistical literature on clustering has focused on binary response models that, unlike hazard models, ignore the effects of time. The most distinctive

marker of clustering in binary response models is overdispersion of responses: because the mortality risks of children belonging to the same household or family tend to be more alike than those of children from different families, the data contains greater variation than what is expected from a conventional model that assumes independence between outcomes. Ignoring clustering and treating the information on child survival as if it came from independent observations can result in attenuated estimates and loss of efficiency (Breslow and Day, 1980; Liang and Zeger, 1989; Liang and Zeger, 1993; Neuhaus, 1992; Liang et al., 1995; Pendergast et al., 1996). As a consequence, hypothesis testing with estimates obtained via conventional likelihood models will be faulty. Furthermore, we will show that clustering also generates interpretational difficulties heretofore ignored in the literature on child mortality. In Section II we discuss models suitable for clustered data both in the context of binary response and hazard analysis.

II. MODELLING CLUSTERED SURVIVAL DATA

1. The case of a continuous response (linear models)

To fix ideas we start with the simplest of cases, one where the dependent variable is continuous and can be expressed as a linear function of a set of predictors. We have observations for child j ($j=1,2,...,n_i$) of a household i ($i=1,...,N$) which contains n_i children. In the absence of any unmeasured characteristics the responses can be modelled as:

$$Y_{ij} \mid X_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij} \qquad (1)$$

We assume that ε_{ij}'s are independent and possibly (but not necessarily) normally distributed with mean 0 and variance σ_ε. One then proceeds to apply conventional OLS procedures to retrieve the estimates of α and β.

Suppose, however, that there is an unmeasured factor common to children living in one

household. We represent those effects assuming that the constant in (1) is actually given by

$$\alpha_i = \alpha_o + \upsilon_i$$

where υ_i is a random variable with a suitable distribution. **Conditional on the value of the unmeasured variable**, the responses within each household are given by

$$Y_{ij} \mid \alpha_i, X_{ij} = \alpha_i + \beta X_{ij} + \varepsilon_{ij} \qquad (2)$$

where α_i, for example, is distributed normally with mean α_o and variance σ_α, and the ε_{ij}'s are independent, (possibly normal) variates with mean 0 and variance σ_ε. Model (2) is conditional or cluster-specific. The **population** or **marginal** or **unconditional** model for responses is

$$Y_{ij} \mid X_{ij} = \alpha_o + \beta X_{ij} + \varepsilon_{ij}^* \qquad (3)$$

where the new error term, ε_{ij}^* term has a non-diagonal covariance matrix. In fact, its variance is $\sigma_\alpha + \sigma_\varepsilon$, and the covariance of the error terms of any two children belonging to the same household is σ_α whereas the covariance of the error terms for any two children who do not belong to the same household is 0.

This simple example shows the consequence of clustering due to unmeasured characteristics at the household level: the variance of the error term in the marginal model is larger than in the conditional model (overdispersion) and, furthermore, the correlation between the error terms invalidates the utilization of conventional OLS. However, the **interpretation of the vector of effects, β, remains unaltered**, that is, it is both the population averaged and

cluster-specific effect. To compute estimates we generally resort to GLS procedures that remove biases in the estimated covariance matrix of coefficients.

2. Binary response models

We now turn our attention to models with binary response for mortality of children. These models focus on a dichotomous response (dead or alive) and ignore the role of time.[1] Assume again that clustering occurs at the level of households only and that each household contributes a variable number of observations (children).[2] Further, suppose that a cross-sectional study of infant and child mortality provides information on N households and on n_i children within the i^{th} family (i = 1, ..., N). For each child, we have a binary response Y_{ij} (j = 1, ..., n_i)=1 or 0 associated with death or survival during the first year of life and a vector of relevant explanatory variables, X_{ij}. Some of these covariates have the same values for all children in the household whereas others may vary within households. We will assume that some elements of X_{ij} belonging to the class of shared characteristics in households are not observed by the investigator. Thus $\mathbf{Y}_i = (Y_{i1}, ..., Y_{ik}, ..., Y_{ini})$ is a n_ix1 vector of correlated responses for the i^{th} household and $\mathbf{Y}=(\mathbf{Y}_1, ..., \mathbf{Y}_N)$ is the population vector of responses. Conditional on the observed characteristics X_{ij} only, a residual correlation of responses remains; this is what generates the clustering problem.

There are two broad classes of models for clustered observations, **conditional** (or cluster-specific) and **marginal** (or population averaged) models.

---

[1] Time is explicitly included in hazard models which are examined in the next section of the paper.

[2] Multiple levels of clustering presents similar problems and has been addressed in the literature (for example, Sastry, 1995a, b). This is a particular case of a more general problem belonging to multilevel modelling.

2.1. Conditional models

Conditional models[3] are models for intra-cluster responses that condition on the unmeasured characteristic(s). In these models we introduce a term to capture household-specific conditions that have a suspected influence on the mortality level of children in the household. The model for responses of children in a household is then **conditional** on the (unmeasured) household-specific covariate. There are two different strategies to formulate these models.

2.1.1. Random effects models

We assume there is a household-specific unmeasured 'background' level of mortality which is continuously distributed, and that the distribution is known except for its main parameters. In the case of a logit model for child mortality, where $Y_{ij}=1$ if the death of child j in household i is observed, we start with:

$$\text{logit } P(Y_{ij}=1 \mid \alpha_i, X_{ij}) = \alpha_i + \beta X_{ij} \qquad (4)$$

where α_i is the 'background' mortality level for household i and is assumed to have a known distribution dependent on one or two parameters. The quantity α_i can be interpreted as a household-specific 'frailty' factor that affects equally the probability of dying among children living in it (Babiker and Cuzick, 1994). Note that in this model the vector β refers to effects of covariates for individuals **in each of the underlying risk groups or strata captured by α_i and not to the average effects over the entire population**.

Expression (4) can be manipulated to yield the probability that the event will occur to

[3] Conditional models are sometimes referred to as random effects models in the epidemiological literature. In the econometric literature, however, random effects models refer to models where the vector β is a random vector.

individual j in household i. Assuming a distribution governing for α_i and integrating over all possible values of α_i yields the **marginal probability** of response:

$$P(Y_{ij}=1 \mid X_{ij}) = \int (1+\exp(-\alpha_i-\beta X_{ij}))^{-1} \, dG(\alpha) \qquad (5)$$

where $G(\alpha)$ is the assumed distribution function of α and integration is over all possible values of α. Clearly, the resulting model, the marginal (population average) model for the response, is **not** of a logistic form. This is an important lesson to note at the outset: a marginal logistic model is not consistent with a conditional logistic model and the estimated effects from the marginal model when the conditional model applies are downwardly biased (Neuhaus, Hauck and Kalbfleisch, 1992; Neuhaus, Kalbfleisch and Hauck, 1991, 1994; Neuhaus and Jewell, 1990).

The parameters of (5) can be estimated via maximum likelihood procedures which are laborious and computationally difficult. In addition to numerical problems, however, these models face a potentially serious problem since information about α is rarely available. It then becomes important to assess the extent to which estimated parameters are sensitive to misspecification of G. This problem was studied in the biostatistical literature by Neuhaus and colleagues (Neuhaus, Kalbfleisch, and Hauck, 1994) who show numerically and analytically that consistent estimation of β is possible even when the mixing distribution used by the investigator **is misidentified** provided it belongs to a family satisfying some very general restrictions. We address the sensitivity problem in a demographic context in Section III.

A random effects model such as (4) can be generalized to the case when the effects in β, and not just the intercept, are also random variables. These models are what the econometric literature refers to as 'random effects models'. Without sufficiently restrictive distributional

assumptions, however, such models are not identifiable.[4]

2.1.2. Stratified (or conditional) models.

There is an alternative modelling strategy for the intracluster (conditional) response which circumvents the need to identify the distribution of α_i. This considers α_i as nuisance parameters, that are of no interest to the investigator and that reflect clusters' background levels of mortality. Using simple argumentation one can derive an expression for the probability of observing any given combination of deaths and survivors among siblings in a household **conditional on the number of children exposed to death and the number of deaths observed in the household**. Such an expression is independent of α_i and can be used to estimate β. For example, suppose that in cluster i there are $n_i=2$ children and that $d_i=1$ are observed to die during infancy over the period of observation. Let us say that the death occurs to child 1. After some simplification, the expression for the **conditional probability** of observing the outcome in a household with vectors of covariates X_{i1} and X_{i2} for child 1 and 2 respectively, is given by

$$_i=P(Y_{i1}=1,Y_{i2}=0 \mid n_i=2,d_i=1)=\{\exp(\alpha_i+\beta X_{i2})\}/(((\exp(\alpha_i+\beta X_{i1}))+(\exp(\alpha_i+\beta X_{i2})))$$

All terms containing α_i cancel out and it is therefore possible to construct a likelihood function (the product of all $_i$) for the sample of households only depending on β. The same occurs when a household experiences more a complicated set of outcomes. The likelihood is conditional and cannot, of course, be used for estimation of α or of any parameters associated

[4] In the econometric literature 'random effects models' refer to those where the vector either α or β or both are random variables.

with covariates that do not vary within clusters.[5] The biostatistical and epidemiological literature contains numerous applications of stratified logistic models but, to our knowledge, they have not been used in demography.

2.2. Marginal models

The second class of models for clustered data are **marginal** models. In these models one represents the marginal probability of dying (or its logit transformation) so that:

$$\text{logit } P(Y_{ij}=1 \mid X_{ij}) = \alpha' + \beta' X_{ij} \qquad (6)$$

Just as model (3) in the linear case can be estimated by specifying an estimate of the matrix of variance-covariance of ε_{ij}^{*} — this is what, after all, GLS does very well — so model (6) can be estimated via (quasi) likelihood procedures as long as we specify a guessed value for the matrix of correlation of intra-cluster responses. If the matrix reflects no correlation within clusters, these techniques will yield identical results to conventional logistic estimation. This procedure, a particular type of more general procedures known as Generalized Estimation Equations (GEE), was developed by Liang and Zeger (Liang and Zeger, 1986). It has been shown to perform well even when the intracluster correlation matrix is incorrectly specified. Models estimated with GEE are commonly used in epidemiological research but they have been seldom used by demographers. Only Zenger (1993) applied it to study siblings' neonatal mortality risks in Bangladesh and found strong intrafamily correlations.

2.3. Marginal versus conditional models: the issue of interpretation

Unlike the marginal representation (5) implied by a conditional model like (4), model (6) **is**

[5] Any term containing covariates that are cluster-invariant will cancel out of the expressions for the conditional probabilities.

logistic and the effects β' have the conventional population averaged interpretation, namely, they represent the changes in the odds of a positive(negative) response conditional on a change in the covariate. Thus β' and β **do not** measure the same thing and cannot generally attain the same values; they will be identical **only** when the intracluster correlation is nil (when α_i is a constant) or, more trivially, when β=0. In any other situation β and β' will differ by an amount that is roughly proportional to the mean intracluster correlation of responses. Neuhaus and colleagues (Neuhaus, Hauck, and Kalbfleisch, 1992) show that when β is small,

$$\beta' = \beta(1- \quad)$$

where is the intracluster correlation of responses calculated at β=0 (Neuhaus et al., 1991, 1994). Put otherwise, if a conditional model applies, estimation of β' from a marginal model underestimates the true (conditional) coefficient.

In principle at least, a marginal model can be estimated using only one observation per cluster. If, however, the investigator decides to use the entire sample, a suitable estimation procedure must be adopted. GEE is perhaps the better known and more broadly used. It is based on the formulation of two separate models, one for the dependence of the outcome variables on covariates and another to represent the intracluster covariance structure. Instead, a conditional model can **only** be estimated with multiple observations per cluster and to do so we must formulate a model that simultaneously represents the dependence on covariates and the intracluster correlation. The decision about which model to use should depend not on statistical but on theoretical considerations, e.g., on the type of questions one wishes to answer. If, for example, our interest is in making inferences about the effects of birth spacing within the context of other characteristics of the mother (or household), then a conditional model is more

appropriate. Note that a conditional model cannot provide much useful information about the influence of characteristics that are invariant within households. Indeed, estimation procedures that rest on stratification of the sample by households will provide no information about them.

3. The case of hazard models

When the objective is to model the time-dependent dynamics of mortality in infancy and early childhood, the most appropriate tool for the analysis is some form of hazard model, most commonly a proportional hazard model.

Suppose again we wish to model survival with data on children clustered by households. Since we are interested in how the process evolves over time, we model the risk of dying at time t for a child j in household i, rather than the probability of dying during a time interval as we did before. A conventional proportional hazard model is:

$$\mu_{ij}(t \mid X_{ij}) = \mu_o(t)\exp(\beta X_{ij}) \qquad (7)$$

where $\mu_o(t)$ is a (possibly unspecified) baseline hazard and β are population effects. As in the case of binary response models, the analysis of longitudinal clustered data allows two broad classes of models, **conditional and marginal** models. Conditional models include **random effects models**, which are extensions of conventional hazard models with unmeasured heterogeneity, as well as **stratified proportional hazard** models. We now briefly review these two types of models.[6]

3.1. Conditional hazard models

To the extent that children living in a household share unmeasured characteristics, their mortality risk will be correlated and a better rendition of the process is the following model:

[6] We are preparing a more extensive review of clustered hazard models.

$$\mu_{ij}(t \mid X_{ij}, \upsilon_i)=\mu_o(t)\upsilon_i \exp(\beta X_{ij}) \qquad (8)$$

where υ_i is a household-specific effect on the baseline hazard.[7] Note that this expression corresponds to a **conditional** hazard model, just as (4) does for binary responses, and, consequently, β are effects that apply relative to the (unknown) baseline hazard $\mu_o(t)\upsilon$: they are not population averaged effects and cannot be interpreted as relative risks in the conventional way.

3.1.1. Random effects models: classic formulation

Assume that the unmeasured effect is invariant over time and is randomly distributed at t=0 according to an arbitrary distribution. Since the composition of the sample relative to υ_i will change over time—those with high values will be more likely to experience the event first—the expression for the **marginal hazard**, $\mu_{ij}(t \mid X_{ij})$ will change over time. This will violate the assumption of proportionality and some consequences will follow. **First**, as in the case of binary responses, the marginal model that results from the (true) conditional model **does not** preserve the original functional form and the estimates corresponding to each of them will be different. **Second,** unlike the case of binary responses, a hazard marginal model estimated with a sample of only one member per cluster will yield incorrect results as a result of omitting the (mother/household) characteristic. In fact, when each household contributes only one (child) observation, the problem of estimation in (7) or (8) is the so-called problem of unmeasured heterogeneity. As is well-known, when this issue is not appropriately addressed, unmeasured heterogeneity will result in inconsistent estimates. The same consequences will occur when households contribute more than one observation.

[7] The random term ought to be strictly positive and sometimes it is expressed as an exponentiated quantity.

Estimation of these models is implemented assuming a distribution for $f_{\upsilon}(\upsilon \mid t=0)$, and then maximizing the corresponding marginal likelihood. This has been done both in the context of individual (unclustered) observations (for example: Vaupel, Manton and Stallard, 1979; Heckman and Singer 1982, 1984) as well as in samples where children are clustered in households. Applications to clustered data have been more common in epidemiology (Clayton, 1978; Clayton and Cuzick, 1985; Hougaard, 1986; McGilchrist, 1993) but, more recently, they have also been applied to demographic studies involving mortality of children living in the same household (Guo, 1993; Guo and Rodriguez, 1991; Sastry, 1995a, b). While we know something about the robustness of models with unmeasured heterogeneity for non-clustered data, we know very little about their performance with clustered data.

3.1.2. Stratified hazard models

Suppose again one considers the household-specific mortality level, $\mu_o(t)\upsilon_i$, as nuisance parameters. There are then conditional procedures, analogous to those deployed in the case of binary response which lead to estimation of β. Two of these procedures are worth mentioning. The **first** is the so-called **paired failure model** formulated by Kalbfleisch and Prentice (1980) which allows estimation of hazard models in which a pair of members in each cluster are allowed to have their own baseline hazards. This model has been used in a variety of contexts and in one demographic application (Mare and Palloni, 1988). The **second** conditional procedure is an extension of the paired failure model as it can handle multiple observations per cluster. It is known as the **stratified Cox model** and was first suggested by Prentice (1988). Although it is potentially very useful in studies of siblings' mortality, we know of no application in the area.

Both the paired failure model and the stratified Cox model have important shortcomings, however. In the first place, as in all stratified procedures, they produce neither estimates of quantities measuring background levels of mortality nor estimated effects of variables that are

constant within cluster. Secondly, as we show in our simulations, their performance is seriously impaired when clusters have to be removed from analysis because one does not observe the event under investigation.

3.2. Marginal hazard models

As in the case of binary response models, if one wishes to estimate hazard models where estimated effects have the more conventional population average interpretation, it is necessary to formulate and estimate marginal models. But while marginal models for binary responses, such as those of the GEE type, are quite popular in epidemiology, if not in demography, marginal models for hazards are not widely used. This is mostly due to the fact that they require the formulation of multivariate survival distributions for the observed marginal distributions of potentially censored survival times. Model formulation and numerical procedures required to estimate parameters are cumbersome and do not lead to easy and clear-cut interpretations.[8]

4. Comparison of binary response and hazard models for clustered data

Is the analyst better off modelling mortality with a binary response model than with a hazard model? The answer depends in part on the objectives of the analyses. If the trajectory over time of the mortality process is of interest, one is better off using hazard models. But note that there is an important asymmetry in terms of costs/benefits when the analysis of child mortality uses household (clustered) information. Clustering can only be produced by the existence of some variable(s) that affects the outcome of at least some of the children born in a household. If one does not know what those factors are but one believes they are unrelated to measured

[8] There is an exception to this. In one application to repeated events occurring to the same individual—a case that can be interpreted as clustered failure survival times—the authors formulate a Cox-proportional hazard model with correlated survival times. Estimates of effects are simply obtained but additional calculations are required to derive estimates of standard errors. This formulation is, therefore, analogous to a GEE approach for binary response models (see Wei, Lin and Weissfeld, 1989).

covariates, a reasonable strategy would be to use GEE based procedures, for they would yield consistent estimates with adjusted variances. Alternatively, one could simply estimate conventional logit models using only one child per household. In either case no information about changes over time would be possible unless multiple conditional logistic models are employed, and then only with some effort. If, however, the analyst insists on making inferences about changes over time, a hazard model is more appropriate. But, in this case unless one addresses explicitly the existence of the unmeasured characteristics generating clustering, the estimates of coefficients will be inconsistent whether or not one chooses to use one child per household. This is what is at stake then: with a clustered design a hazard model could produce erroneous results whereas a logit model (conventional or adjusted such as GEE) may save the day albeit with a somewhat cruder knowledge return.

III. ROBUSTNESS OF ESTIMATES: A SIMULATION

In this section we formulate the problem as one of estimating effects on mortality of siblings living in the same household. We simulate mortality data during the first year of life and then apply selected estimation procedures covering a broad range of sophistication. We then evaluate the performance of these procedures using alternative numerical criteria.

1. The problem

For the sake of clarity we re-state the main problem: we wish to assess the effects of selected covariates on the risk of mortality during the first year of life for a group of children grouped by mother or household. The response, died/not died at the time of survey, is a function of a series of characteristics some of which are child specific and others could be either mother, household or community specific. In order to simplify presentation we only distinguish between child and household specific characteristics. We are particularly interested in evaluating the extent to which infant mortality in influenced by the pace of childbearing: is it the case that mortality is

higher for children in a household who were conceived or born very shortly after the birth of an older sibling? To answer the question we propose to estimate effects of pertinent covariates using binary response models and hazard models that include a series of relevant controls for potentially disturbing factors. The estimation problem is that the observed responses belonging to the same household are correlated due to the influence of household level unmeasured characteristics.

The estimation problem was recognized right from the start, when data from household based survey began to be massively used to study determinants of mortality. Unfortunately, at the time only very crude solutions were proposed. One of them, utilized quite liberally and sometimes for the wrong reasons, involves the inclusion among the control variables of a dummy indicator for mortality of the previous sibling or, more generally, some measure of mortality among older siblings. Several authors used this solution (Cleland and Sathar, 1984; Hobcraft, McDonald and Rutstein, 1985; Miller et al. 1992; Palloni and Millman, 1986; Palloni and Tienda, 1986; Retherford et al., 1989) only to find that very short preceding and following birth intervals remain strongly detrimental to the survival of infants and young children and are unlikely to be artifacts of clustering.[9] These rather crude solutions, however, are now considered woefully inadequate and are challenged by work that introduces more sophisticated methods (Curtis, Diamond and McDonald, 1993; Guo 1993; Guo and Rodriguez, 1992; Zenger, 1993; Madise and Diamond, 1995; Sastry, 1995a, b).

[9] The original motivation for including PDIED in these models was to eliminate spuriousness created by the fact that a short birth interval could be the result of intentionally accelerating the pace of childbearing to replace a dead child. This is thought to be of more significance in families with high levels of mortality. Thus the motivation to control for PDIED, or similar variables, had little or nothing to do with a design to correct for clustering effects. In the demographic literature the use of PDIED has also been invoked as a means of 'controlling' for unmeasured characteristics that cause clustering. Although this is an intuitively plausible approach to deal with clustering, it leads to conceptual difficulties that we document below.

Our next task consists of evaluating the performance of various methods to deal with clustered data.

2. The simulation

To evaluate the performance of various methods described before, we design a Monte Carlo simulation. The simulations closely replicate data sets of the type commonly used by demographers to study mortality determinants. The simulated data sets have two important characteristics:

1) Each data set contains mothers as primary units of observations. Mothers represent households to which we assign a fixed number of children ever born and the values of a small set of characteristics thought to be relevant for children's mortality, and

2) Given a known functional form for the risk of dying during the first year of life, we simulate the survival process during the first 12 months of life for the children assigned to a mother (household). That is, for each child we calculate a survival time and determine whether or not the child died during infancy.

The functional form for the hazard is defined as follows:

$$\mu_{ij}(t \mid X_{ij}, \upsilon_i)=\mu_o(t) \exp(\beta X_j + Z_{ij}) \qquad (9)$$

where X_j are characteristics of mother j, Z_{ij} are characteristics of child i and mother j. $\mu_o(t)$ is the baseline hazard, β is a vector of effects of mothers' characteristics and are the effects of children's characteristics. The variables included in vector X_j are categorical variables for region of residence, education, and rural-urban residence. Variables included in vector Z_{ij} are child-specific categorical variables including length of previous interval, sex of child, and death of the previous sibling which indicates whether the sibling born immediately before child j died before

child j had been conceived (PDIED). In the simulated data the effects of PDIED are always set to 0, namely, the death of the preceding sibling does not have a direct influence on a child's mortality risks. However, to replicate what has been done in the demographic literature, we also will estimate models that control for PDIED.

Mothers and their children together with their characteristics were taken from the DHS samples of Brazil. The only variable that was not drawn from the DHS data files was PDIED. Instead, the value for this variable was directly generated by the simulation: we begin simulating the survival of the oldest child (for whom PDIED is always 0) and then proceed upwards in the birth order sequence. In each case we check the survival status of the previous child and thus determine the value of the variable PDIED for child j.

Note that the underlying mortality regime is determined by $\mu_o(t)$: once this is fixed we have a mortality regime that can be taken as a baseline for simulations. The baseline hazard model is set to reproduce infant mortality levels in Brazil. These levels are of moderate magnitude and result in relatively few events. This fact should be kept in mind when evaluating the performance of various models since they are differentially sensitive to the frequency of the event of interest.

To induce intra-family correlation we generate a household-specific random variable υ_i using 8 different distributions: two gamma, two normal, two exponential gamma and two non-parametric. In each pair of distributions we include one with small and the other with large variance. The distributions are described in Table 1. When the random component is added, the expression for the hazard becomes:

$$\mu_{ij}(t \mid X_{ij}, \boldsymbol{\upsilon_i}) = \mu_o(t)\upsilon_{ij} \exp(\beta X_{ij}) \qquad (10)$$

The range of variances obtained from the assumed distributions of υ_i is wider than the

range of values of unobserved characteristics responsible for clustering estimated in the literature. The intracluster correlation coefficient associated with each of them is listed in Table 1.[10] Table 2 displays the values assigned to the coefficients and descriptive statistics for the independent variables used in the simulation.[11] In all we generated 100 samples or realizations of the time of death of each child for each of the 8 chosen distributions. Each set contains a combined total of 1,000 children.

Since our data are generated using a **conditional model**, the estimated effects ought to be interpreted as conditional rather than population averaged effects.

3. Performance of estimation methods

We use the simulated data to estimate the parameters with methods listed in Table 3. These methods are widely used in biostatistics and epidemiology though not all of them are routinely applied in demographic research. To simplify the discussion that follows we will only report results for 5 procedures. These are classified in two groups:

3.1. Hazard models

Among these we include a conventional Cox model (CC) and the stratified Cox model (SC). To enhance comparability with methods for binary response mentioned in (3.2) we restrict attention to estimation of effects and mortality for the age range 1 to 11 months (completed).[12]

---

[10] The data we simulate does not incorporate very large intrafamily correlation but substantially more than what is found in empirical cases (Guo, 1993; Sastry 1995a,b). Here we need to write down values of empirically estimated association that should be comparable to those obtained from our distributions.

[11] All characteristics contained in vector X_j and all but one characteristic contained in vector Z_{ij} have the same values in all simulated data sets. The exception is PDIED whose features change across simulated sets since it is dependent of the parameters used to generated the simulated cases.

[12] Elsewhere we pursue a more complete evaluation of semi-parametric and fully parametric hazard methods.

3.2. Binary response models

We include tests for the conventional logit model (CL), random logit model (RL), and marginal logit model (GEE). In all cases we present results corresponding to the response (death/survival) evaluated between age 1 month and 11 (completed) months.[13]

All models are estimated with and without control for the variable reflecting mortality of previous sibling. This is an exercise that will enable us to assess the extent to which one can avoid application of specially designed models to correct estimates of effects and variances. It is also useful to highlight the interpretational difficulties caused by the inclusion of this variable.

To evaluate each method's performance we use the mean and standard distribution of the estimates as well as their mean relative errors and mean squared errors. Although this is done for each one of the variables included in the model, we only report results for maternal education, length of previous birth interval and, when applicable, death of previous child.

4. Results

The full results appear in Table 4 in panels a, b, and c. They correspond, respectively, to the cases where the true distribution of the error term is normal, gamma, and non-parametric. The results for each distribution correspond to the case with the parameters producing maximum intra-cluster correlation (see Table 1). Since the table contains large amounts of information, it is is cumbersome to read. A more condensed yet useful display of results appears in Figures 1a and 1b. The box plots in these figures are drawn using the three mean estimated values for PINT1 obtained from each method in each of the three simulated scenarios. These three values define the

[13] Binary response models of the logit type can be accommodated to estimate effects within any arbitrarily defined age segment. Thus we could have estimated effects in the age segment 0 to 1 month also. We chose not to do so since it would have added very little to our assessment of the method's performance under conditions with no time dependent covariates and no time dependent effects. Estimation of effects in the age segment 1-11 months has been the focus of the recent literature.

boundaries of the boxes. The boundaries of the whiskers for each box correspond to the arithmetic average of the three means plus (minus) the largest standard deviation associated with each procedure. One can think of these boxes as a portrait of the bias and precision of each method when the real world relations are unknown: the closer the box is to the line for the true value of the effect (1.57), the less the amount of biases expected from the corresponding method. Similarly, the larger the span of the box's whiskers, the smaller the precision associated with the procedure. Figure 1a corresponds to estimates in models not including a control for PDIED whereas Figure 1b corresponds to estimates from models including a control for PDIED. Figures analogous to 1a and 1b for the other variables reveal similar features and we will not show them.

The figures reveal some features of importance that are confirmed by the more general results in Table 4. Note that, as expected, the marginal models always produce downwardly biased estimates of the true value. The average biases are very close to what we would expect from the approximation formula discussed before and the intracluster correlation coefficients displayed in Table 2. These biases virtually disappear once a control for PDIED is included. In models not including a control for PDIED, the estimates produced by GEE and CC, followed closely by CL, are the most precise whereas those retrieved using RL or SC procedures are the least precise of all.

More generally, an examination of Table 4 reveals the following noteworthy results:

a) CC models without a control for PDIED produce downward biases in estimates of PINT1, PINT2 and PINT3 regardless of the distribution of the error term. The absolute and relative biases and the mean relative and quadratic errors are largest for PINT1, the variable with the largest of the true effects, and smallest for maternal education. These biases are attenuated in CC models that include a control for PDIED. Since the same pattern appears with other models, we conclude that variables that change **within** cluster will be more affected by intracluster correlation that those

that do not.

b) SC models without the inclusion of PDIED attenuate downward biases particularly in PINT1. However, these models do not yield estimates of effects for variables that are invariant within clusters and tend to produce larger quadratic errors than CC models. Standard deviations associated with SC estimates are large since one cannot use the entire set of clusters for estimation but only those where at least one event is observed. This leads to losses of information reflected in larger standard errors. A paradoxical feature of the SC model that includes a control for PDIED is that the estimated effect of this variable bears a **negative** (instead of positive) sign. This result is due in part to the loss of information referred to above and to the fact that inclusion of the variable violates a critical assumption of the model, namely, that the unmeasured characteristic is uncorrelated with measured variables.

c) More generally, the estimated effects of PDIED **always** contain very large biases, irrespective of the true error distribution and method of estimation. The lesson is that in the analysis of child mortality, the estimated effect of PDIED should be interpreted with extreme caution. The same should apply to other areas of study where control variables of similar nature are used.

d) The CL models including a control for PDIED produce estimates that are close to the true ones and with mean relative and mean quadratic errors that are comparable to those associated with other procedures. CL with no control yields estimates for PINT1, PINT2 and PINT3 with downward bias and larger mean relative and quadratic errors.

e) By all metrics GEE models with no control for PDIED perform better than CL models with similar specification, but the estimates contain downward biases. This agrees with the expectation that marginal models — as GEE and CL — should underestimate conditional effects. Also as expected, GEE's standard errors are lower and a control for PDIED improves their performance much as it did for CL models.

f) RL models perform less well than GEE in terms of quadratic errors even when the assumed distribution (normal) corresponds to the true one. Generally RL estimates exaggerate the effects of at least one of the birth interval variables (PINT1) and of maternal education. The inclusion of PDIED as a control has less of an effect in RL models than elsewhere , thus suggesting that controlling for a condition that reflects intra-cluster correlation in Rl models is redundant. The estimated effects of PDIED are lower than those obtained from the other models but they still would lead to the rejection of the (true by design) null hypotheses that the population effects are equal to zero. Estimates in models that include PDIED, however, have smaller quadratic errors. [14]

These results suggests the following two conclusions:

- if the analyst chooses to apply a non-parametric hazard model, it is better to use CC models with a control for intracluster correlation. SC models should only be used when the fraction of clusters that are useful for estimation is sizeable (higher than 80 percent) and never includes a variable that could proxy the intracluster correlation.
- if the analyst chooses a CL or GEE it is preferable to include a control for variable(s) proxying the intracluster correlation. Without such controls both CL and GEE models produce estimates that are biased relative to the **true conditional effects**. On the other hand, RL models are laborious to estimate and tend to be more imprecise.

[14] The lackluster performance of the random effects model is somewhat puzzling. One explanation points toward numerical precision. In order to estimate these models we first used the program EGRET and a discrete approximation to a normal distribution requiring 5 points of support. The results in Table 4 correspond to this approximation. However, we also estimated the same models using a normal distribution approximated with 20 points of support and obtained results that are almost indistinguishable from those corresponding to a procedure with less numerical precision. Thus, the conclusions about the upward bias in the estimates of effects and the the larger quadratic errors of the RL model are not accounted for by issues pertaining to numerical precision. An alternative but somewhat unlikely possibility is that in the case of RL models at least, one requires a larger number of simulated data sets to produce a well calibrated distribution of estimates. Finally, it is also possible that RL does not perform well when, as it is the case in our data set, the event of interest is rare.

The idea of including a control for a measure of intracluster correlation deserves some qualifications. In the child mortality example, the inclusion of PDIED is equivalent to controlling not only for a measure of intracluster correlation but for the response of one (an older) individual in it. It has been shown that models (marginal or conditional) that condition on responses of other members of the cluster lead to downward biases in the estimated effects of the other variables (Rosner, 1984; Glynn and Rosner, 1994; Neuhaus and Jewell, 1990). Our simulations show that although, indeed, some of the estimated coefficients are attenuated, the biases are modest and, at any rate, smaller than those that are obtained if nothing is done to address intracluster correlation.

IV. APPLICATION TO DATA FROM BRAZIL

The results of the simulation suggest that alternative procedures perform quite differently and that in an empirical case we should expect to observe some differences in the estimates. In order to illustrate the application of these methods to an empirical case we used the DHS data from Brazil that was also used by Curtis and colleagues (Curtis, Diamond and McDonald, 1993) in their original application of random effects models. After applying a series of exclusions and restrictions we settle on a sample of about 4,761 births during the five years preceding the survey. We then estimated models including the same six categorical variables described before and, in addition, some control variables. The results which we display in Table 5 only refer to the estimates of effects of the former. The comparison includes CL, RL and GEE as binary response models, and CC and SC as hazard models.

The first feature apparent from the table is that the three binary response models produce very similar results. Not only are the signs and magnitude of estimated coefficients quite close to each other but the tests of significance are, with one exception, identical. The exception is the effect of PINT1 for GEE which is significantly different from zero but only at a higher α-level.

Second, the effects of PDIED are strong and, as predicted by the simulation, they are

considerably attenuated in the case of a RL model. As was emphasized before, these estimates are very likely to reflect pure clustering effects.

Third and finally, a comparison of estimates from the CC model (including a control for PDIED) and binary response models suggests a high level of similarity.[15]

The only notorious outlier is the SC model which produces lower effects for the birth interval variables and a negative estimate for PDIED. The relative magnitude of the estimates for PINT1, PINT2 and PINT3 is quite different from that obtained with the other models: a null hypothesis testing the equality of coefficients could not be rejected. Despite this discrepancy, an SC model leads to identical outcomes in the testing of null hypotheses regarding the effects of pace of childbearing and maternal education.

V. ISSUES OF INTERPRETATION

We now address the distinction between marginal and conditional effects. We will do so with an illustration that highlights misinterpretations likely to occur when the contrasts between them are blurred.

The central distinction is between marginal (or population averaged) and conditional (or cluster-specific) models (compare models (5) and (6) above). Whereas the estimated effects of the latter are interpreted in the conventional way, those from a conditional model refer to measured effects **within a strata induced by the unmeasured characteristics: they are effects that apply holding constant measured and cluster-specific unmeasured characteristics**. As suggested by the relation derived by Neuhaus and colleagues (Neuhaus, Hauck, and Kalbfleisch, 1992) the numerical difference between the two depends on the strength of the intracluster

[15] A direct comparison is possible since the levels of mortality in the interval examined are relatively low. This ensures that the effects on the odds (estimated by binary response models) are equivalent to the effects on the hazards.

correlation. We show briefly that this difference could lead to errors of interpretation and misleading implications.

Suppose the data are generated by a conditional model such as (4) and that the analyst correctly posits a random effects logit model, estimates parameters, and then attempts to draw policy relevant consequences. But in doing so he/she overlooks the distinction between a marginal and a conditional model and ends up utilizing the more conventional and better known interpretation related to marginal effects. Let us assume that there is only one dichotomous covariate, X, and one associated effect. Further, the analyst is interested in calculating the relative change in the prevalence of the outcome of interest (say Y=1) if all individuals in the population were to have characteristic X=1. In a hazard context this relative change is measured by the so-called population attributable risk:

$$PAR=(w(\exp(\beta)-1))/(1+w(\exp(\beta)-1))$$

which is interpreted as the proportionate change in the prevalence of the outcome if the behavior indexed by X=0 is eliminated. Here w is the fraction of individuals with X=0.

A quantity analogous to PAR can be calculated in a binary response context. This quantity depends on the estimated effects on the log-odds of the outcome. In particular, if the appropriate model is a marginal model and there is no clustering, PAR is defined as:

$$PAR_m= (1-w)(D'-D)/D$$

where

$$D= (\exp(\alpha+\beta)/(1+\exp(\alpha+\beta))) \text{ and } D'= \exp(\alpha)/(1+\exp(\alpha))$$

Instead, if a random effects models such as (4) is appropriate, PAR is given by:

$$PAR_c = (1-w)\,(\int_\alpha E'_i\, f(\alpha)d\alpha \; - \; \int_\alpha E_i\, f(\alpha)d\alpha)/(\int_\alpha E_i\, f(\alpha)d\alpha)$$

where

$$E_i = \exp(\alpha_i+\beta)/(1+\exp(\alpha_i+\beta)) \text{ and } E'_i = \exp(\alpha_i)/(1+\exp(\alpha_i))$$

and where $f(\alpha)$ is the density function for the random variable α_i.

If the analyst ignores the differences between the conditional and marginal model and interprets β as an estimate of population averaged effects, he/she will calculate PAR_m and use it to draw policy implications. Instead, since the data were generated by a conditional model, PAR_c ought to be used. Most of the time the differences between these indicators may not amount to much. But on occasion they will be considerable and the erroneous implications for policy could be quite significant. Table 6 displays the results from four numerical exercises. In one case we assumed that the distribution of α_i was approximately normal with 200 points of support with a mean of -2.50 and variance equal to 1. In the other we utilized a rougher distribution for α_i with four points of support, mean -2.13 and standard deviation 1.80. In each of the two cases we alternatively assumed that β was .50 and 1.50.

Table 6 shows that when β=.50 the values of PAR_m overestimate the true population attributable risk by about 17 percent. But when β = 1.50 the overestimation is larger than 100-200 percent. Thus, under certain conditions, misinterpretation of effects can have damaging consequences for the integrity or adequacy of policy formulations.

VI. CONCLUSION

Clustered data are fairly common in demography and are certainly central to the analysis of child mortality data from household surveys. Many important conclusions about determinants of infant and child mortality are based on models estimated with clustered data which violate a

number of important model assumptions. Ignoring the problem can lead to difficulties of all sorts, not the least of which is that estimates may be inconsistent and/or inefficient depending on whether we formulate a binary response or a hazard model. Yet solving the problem is not a straightforward matter either. Although there are a number of procedures suited for either binary response models or hazard models, they require discussion about the nature of the true model. In particular, the analyst must decide between conditional and marginal models and choose among a number of alternative procedures to estimate the relevant parameters. In turn, these decisions involve choosing numerical procedures and making assumptions about distribution of unobservables. Procedures differ in terms of their sensitivity to violations of assumptions.

Our simulation exercise suggests that the best choice in quadratic and relative error terms is a procedure of the GEE type. Since GEE procedures are very easily implemented by various software packages (Stata, SAS), they are a convenient tool to deal with the clustering problem as it presents itself in the area of child mortality. The random effects model tends to produce larger quadratic and relative errors, particularly when the distribution of the error term is not normal. Although in theory, at least, the most robust hazard procedure should be the stratified Cox model, its performance in the context of mortality data leaves something to be desired since the precision of the estimates suffers when part of the sample of clusters has to be removed from consideration due to lack of events. Using a conventional (uncorrected) logit model leads to surprisingly good results **when one controls for the death of the previous sibling**. This occurs despite the fact that estimated effects for the other covariates should contain downward biases (citations). In any case, the effects of this control variable are not easily interpretable.

Finally, we end with a note of caution. Marginal and conditional models in the world of non-linear links are not identical and the estimated effects cannot be interpreted in the same way. When choosing to estimate a conditional model, the analyst should refrain from interpreting the

coefficients as if they were measures of population averaged effects like those derived from marginal models are. If this caution is not observed, erroneous interpretation and policy relevant conclusions will follow.

REFERENCES

Babiker, A. and J. Cuzick. 1994. "A simple model for family studies with covariates." *Statistics in Medicine* 13: 1679-1692.

Breslow, N. E. and N. E. Day. 1980. *Statistical Methods in Cancer Research 1: The Analysis of Case-Control Studies*. Lyon: I.A.R.C.

Clayton, D. G. and J. Cuzick. 1985. "Multivariate generalizations of the proportional hazards model." *Journal of the Royal Statistical Society* Series A, 148(2): 82-117.

Clayton, D. G. 1978. "A model for association in bivariate life tables and its applications in epidemiological studies of family tendency in chronic diseases incidence." *Biometrika* 65, 1:141-151.

Cleland J. G. and Z.A. Sathar. 1984. "The effect of birth spacing on childhood mortality in Pakistan" *Population Studies* 38:401-418

Curtis, S. L., J. W. McDonald, and I. Diamond. 1991. "Birth interval effects on healthy families in Brazil." *Proceedings of Demographic and Health Surveys World Conference*, Washington, DC, Vol. II: 1207-1227.

Curtis, S. L., I. Diamond, and J. W. McDonald. 1993. "Birth interval and family effects on postneonatal mortality in Brazil." *Demography* 30, 1:33-43.

Curtis, S. L. and F. Steele. 1994. "Variations in family neonatal mortality risks in four countries." Paper presented at the annual meeting of the Population Association of America, Miami, May 5-7, 1994.

Das Gupta, M. 1990. "Death clustering, mother's education and the determinants of child mortality in rural Punjab, India." *Population Studies* 44:489-505.

DeSweemer, C. 1984. "The influence of child spacing on child survival." *Population Studies* 38:47-72.

Glynn R. J. and B. Rosner. 1994. "Comparison of alternative regression models for paired binary data." *Statistics in Medicine* 13: 1023-1036.

Gribble, J. N. 1993. "Birth intervals, gestational age, and low birth weight: Are the relationships confounded?" *Population Studies* 47:133-146.

Guo, G. 1993. "Use of sibling data to estimate mortality effects in Guatemala." *Demography* 30:15-32.

Guo, G. and L. M. Grummer-Strawn. 1993. "Child mortality among twins in less developed countries." *Population Studies* 47:495-510.

Guo, G. and G. Rodriguez. 1992. "Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala." *Journal of the American Statistical Association* 87:969-976.

Heckman, J. and B. Singer. 1982. "Population heterogeneity in demographic models." Pp. 567-594 in K. Land and A. Rogers (eds.) *Multidimensional Mathematical Demography*. New York: Academic Press.

———. 1984. "Econometric duration analysis." *Journal of Econometrics* 24:63-132.

Hobcraft, J. 1992. "Fertility patterns and child survival: a comparative analysis." *Population Bulletin of the United Nations* 33:1-31.

Hobcraft, J., J. W. McDonald, and S. Rutstein. 1985. "Demographic determinants of infant and early child mortality: A comparative analysis." *Population Studies*, 39:363-385.

Hougaard, P. 1989. "A class of multivariate failure time distributions." *Biometrika* 73:671-678.

Kalbfleisch, J. D. and R. L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

Klein J. P. 1992. "Semiparametric estimation of random effects using the cox model based on the EM algorithm." *Biometrics* 48:795-806.

Koenig, M. A., J. F. Philips, O. M. Campbell, and S. D'Souza. 1990. "Birth intervals and childhood mortality in rural Bangladesh." *Demography* 27:251-265.

Lantz, P., M. Partin, and A. Palloni. 1992. "Using retrospective surveys for estimating the effects of breast-feeding and child spacing on infant and child mortality." *Population Studies* 46:121-139.

Liang, K.Y., S. G. Self, K. J. Bandeen-Roche and S. L. Zeger. 1995. "Some recent developments for regression analysis of multivariate failure time data." *Lifetime Data Analysis* 1:403-415.

Liang, K. Y., and S. L. Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73:13-22.

———. 1989. "A class of logistic regression models for multivariate binary time series." *Journal of the American Statistical Association* 84:447-451.

———. 1993. "Regression analysis for correlated data" *Annual Review of Public Health* 14:43-68.

Madise, N. J. and I. Diamond. 1995. "Determinants of infant mortality in Malawi: an analysis to control for death clustering within families" *Journal of Biosocial Sciences* 27:95-106.

Manton, K. G., B. Singer, and M. A. Woodbury. 1992. "Some issues in the quantitative characterization of heterogeneous populations, in J. Trussell et al., *Demographic Applications of Event History Analysis*. Oxford: Clarendon Press.

Mare, R. D., and A. Palloni. 1988. "Couple methods for socioeconomic effects on the mortality of old persons." Working Paper No. 88-7. Madison, Wisconsin: University of Wisconsin Center for Demography and Ecology.

McGilchrist, C. A. 1993. "REML estimation for survival models with frailty." *Biometrics* 49:221-225.

Miller, J. E. 1989. "Is the relationship between birth intervals and perinatal mortality spurious?" Evidence from Hungary and Sweden." *Population Studies* 43:479-495.

Miller, J. E., J. Trussell, A. R. Pebley, and B. Vaughan. 1992. "Birth spacing and child mortality in Bangladesh and the Philippines." *Demography* 29, 2:305-318.

Neuhaus, J. M. 1992. "Statistical methods for longitudinal and clustered designs with binary responses." *Statistical Methods in Medical Research* 1:249-273.

Neuhaus, J. M., W. W. Hauck, and J. D. Kalbfleisch. 1992. "The effects of mixture distribution misspecification when fitting mixed-effects logistic models." *Biometrika* 79:755-762.

Neuhaus, J. M., and N. P. Jewell. 1990. "Some comments on Rosner's multiple logistic model for clustered data." *Biometrics* 46:523-534.

Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1991. "A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data." *International Statistical Review* 59:25-35.

———. 1994. "Conditions for consistent estimation in mixed-effects models for binary matched pair data." *The Canadian Journal of Statistics* 1: 139-148.

Palloni, A., and S. Millman. 1986. "Effects of inter-birth intervals and breast-feeding on infant and early childhood mortality." *Population Studies* 40:215-236.

Palloni, A., and M. Tienda. 1986. "The effects of breast-feeding and pace of childbearing in mortality at early ages." *Population Studies* 40:215-236.

Pebley, A. R. and P. W. Stupp. 1987. "Reproductive patterns and child mortality in Guatemala." *Demography* 24:43-60.

Pendergast, J. F., S. J. Gange, M. Lindstrom, M. Newton, M. Palta, and M. Fisher. 1996. "A survey of methods for analyzing clustered binary response data." *International Statistical Review* 64:1-30

Potter, J. E., M. Das Gupta, and G. Wyshak. 1993. "Death clustering: A quantitative exploration of the distribution of child deaths among mothers." In *The Young and the Elderly: Issues on Morbidity and Mortality*, eds. D. O. Sawyer and S. D. McCracken. Belo Horizonte, Brazil: CEDEPLAR/UFMG.

Prentice, R. T. 1988. "Correlated binary regression with covariates specific to each binary observation." *Biometrics* 44:1033-1048.

Retherford, R. D., M. K. Choe, S. Thapa, and B. B. Gubhaju. 1989. "To what extent does breast-feeding explain birth-interval effects on early childhood mortality?" *Demography* 26(1):439-450.

Rosner, B. 1984. "Multivariate methods in ophthalmology with application to other paired-data situations." *Biometrics* 40:1025-1035.

Sastry, N. 1995a. "Community characteristics, individual and household attributes, and child survival in Brazil." Working Series Paper Series 95-13. Santa Monica, California: Rand Corporation Labor and Population Program.

———. 1995b. "A multilevel hazards model for hierarchically clustered data: model estimation and an application to the study of child survival in Northeast Brazil." Working Paper Series No. 95-15. Santa Monica, California: Rand Corporation Labor and Population Program.

Trussell, J., and G. Rodriguez. 1990. "Heterogeneity in demographic research," Pp. 111-132 in J. Adams et al., *Convergent Questions in Genetics and Demography*. New York: Oxford University Press.

Vaupel, J. P., K. G. Manton, and E. Stallard. 1979. "The impact of heterogeneity in individual frailty on the dynamics of mortality." *Demography* 16:439-454.

Wei, L. J., D.Y. Lin, and L. Weissfeld. 1989. "Regression analysis of multivariate incomplete failure data by modeling marginal distributions." *Journal of the American Statistical Association* 84:1065-1073.

Winikoff, B. 1993. "The effects of birth spacing on child and maternal health." *Studies in Family Planning* 14, 10:231-245.

Zenger, E. 1993. "Siblings' neonatal mortality risks and birth spacing in Bangladesh." *Demography* 30:477-488.

Table 1: Description of variables included in the simulation exercise

a. Baseline model: gompertz function defined as:

$\mu(t) = \exp(\alpha+\beta t)$
α=-6.0
β=-.04

b. Independent variables and parameters [1]

variable name	definition	parameter
Pint1	Previous birth to conception interval shorter than 11 (completed) months	1.57
Pint2	Previous birth to conception interval between 12 and 17 (completed) months	.93
Pint3	Previous birth to conception interval between 18 and 23 (completed) months	.57
Residual	Previous birth to conception interval longer than 224 (completed months)	-
Pdied	Previous child died before conception of index child	0
Residual	No death of previous child	-
educ1	Mother's education between 5 and 8 years of schooling	-.55
educ2	Mother's education higher than 8 years of schooling	-2.34
residual	Mother's education less than 5 years	-

[1] The simulated data reflects a hazard function dependent on the variables listed here plus other relevant controls that are of no immediate interest to us for this paper.

Table 2: Properties of distributions for the error term

Distribution	Parameters	Intra-cluster correlation
Normal 1	mean=.9, st.dev=.3	.02
Normal 2[1]	mean=.3, st.dev=1.5	.31
Gamma 1[2]	a=.8, b=5.5	.28
Gamma 2	a=.3, b=7.0	.52
Non Parametric 1	2 clusters representing .25 and .75 of families	.04
Non-Parametric 2	3 clusters representing .25, .25 and .50 of families	.23

Notes

[1] Mean and st.dev correspond to the mean and standard deviation of the normal distribution.

[2] For the gamma distribution we generate a gamma distributed random variable but the parameters shown here correspond to the log form of such random variable.

Table 3: Procedures for estimation

1. Binary Response Models

- o Conditional models
  - oo random effects logit model (RL)
  - oo conditional logit model[1]
- o Marginal models
  - oo conventional logistic models (CL)
  - oo GEE

2. Hazard Models

- o Conditional models
  - oo conventional Cox regression model (CC)
  - oo stratified Cox regression model (SC)
  - oo paired failure model[2]
  - oo random effects model (RL)
- o Marginal models
  - oo GEE models for Cox based approach [3]
  - oo multivariate survival distribution models [3]

Notes:
[1] The conditional logit model produces estimates that are very similar to the SC model and are not discussed in the paper.
[2] The paired failure model requires to discard information since it is applicable to pairs and is not applied to our data.
[3] These type of procedures have not been widely tested and are not easily implemented.

Table 4: Main results from simulations[a/]

(a) Distribution of Error Term: Normal 2							
		Variable					
Procedure		Pint1	Pint2	Pint3	Pdied	Educ1	Educ2
CC							
	M	1.49	.66	.38	1.60	-.21	-2.18
	S	.30	.26	.29	.40	.24	1.47
	Q	.24	.30	.27	-	.36	1.02
	R	.15	.32	.48	-	.65	.43
CC							
	M	1.14	.46	.29	-	-.27	-2.36
	S	.27	.24	.28	-	.27	1.38
	Q	.50	.53	.49	-	.41	1.37
	R	.25	.51	.55	-	.60	.39
SC							
	M	1.26	.75	.51	-1.21	-	-
	S	.71	.55	.52	.50	-	-
	Q	.77	.57	.52	-	-	-
	R	.39	.47	.71	-	-	-
SC							
	M	1.67	.99	.63	-	-	-
	S	.66	.54	.49	-	-	-
	Q	.66	.54	.48	-	-	-
	R	.36	1.07	2.02	-	-	-

a/ Meaning of symbols for tables 4a-4f:

M=mean
S=standard deviation
Q=mean quadratic error
R=mean relative error

(b) Distribution of Error Term: Gamma 2

Procedure		Variable: Pint1	Pint2	Pint3	Pdied	Educ1	Educ2
CC							
	M	1.38	.98	.51	1.35	-.30	-2.14
	S	.30	.21	.27	.20	.22	1.10
	Q	.35	.21	.28	1.36	.33	1.11
	R	.18	.19	.37	-	.51	.33
CC							
	M	1.07	.84	.46	-	-.38	-2.23
	S	.29	.20	.25	-	.22	1.07
	Q	.57	.21	.27	-	.28	1.06
	R	.33	.18	.18	-	.41	.31
SC							
	M	1.09	.80	.46	-1.11	-	-
	S	.66	.38	.47	.38	-	-
	Q	.81	.40	.48	1.18	-	-
	R	.43	.33	.66	-	-	-
SC							
	M	1.58	.99	.58	-	-	-
	S	.63	.35	.46	-	-	-
	Q	.62	.36	.46	-	-	-
	R	.31	.30	.65	-	-	-

	(c) Distribution of Error Term: Non-parametric 2					
	Variable					
Procedure	Pint1	Pint2	Pint3	Pdied	Educ1	Educ2
CC						
M	1.50	.89	.52	.46	-.60	-2.21
S	.22	.25	.31	.37	.34	.52
Q	.20	.24	.31	.58	.33	.53
R	.20	.21	.43	-	.49	.19
CC						
M	1.12	.87	.51	-	-.62	-2.26
S	.28	.25	.31	-	.34	.49
Q	.33	.26	.31	-	.35	.49
R	.16	.17	.43	-	.50	.18
SC						
M	1.21	.75	.50	-1.25	-	-
S	.56	.41	.45	.48	-	-
Q	.66	.45	.46	1.34	-	-
R	.34	.39	.65	-	-	-
SC						
M	1.62	.99	.57	-	-	-
S	.50	.37	.48	-	-	-
Q	.50	.36	.42	-	-	-
R	.25	.31	.60	-	-	-

(d) Distribution of Error Term: Normal 2

Procedure		Variable Pint1	Pint2	Pint3	Pdied	Educ1	Educ2
CL							
	M	1.65	.73	.43	1.80	-.22	-1.84
	S	.36	.30	.32	.34	.27	.60
	Q	.37	.36	.35	1.83	.42	.77
	R	.18	.31	.47	-	.66	.28
CL							
	M	1.27	.48	.31	-	-.31	-2.05
	S	.31	.26	.30	-	.28	.60
	Q	.42	.52	.40	-	.37	.66
	R	.22	.49	.55	-	.57	.24
GEE							
	M	1.66	.72	.43	1.70	-.23	-1.85
	S	.33	.30	.33	.33	.28	.60
	Q	.33	.36	.36	1.73	.42	.76
	R	.17	.32	.48	-	.66	.28
GEE							
	M	1.30	.53	.35	-	-.37	-2.05
	S	.31	.26	.31	-	.29	.59
	Q	.32	.48	.37	-	.33	.65
	R	.21	.44	.50	-	.52	.24
RL							
	M	1.87	.80	.49	1.43	-.36	-2.11
	S	.42	.35	.36	.48	.33	.70
	Q	.52	.37	.37	2.27	.37	.73
	R	.27	.32	.50	-	.58	.26
RL							
	M	1.91	.78	.50	-	-.63	-2.67
	S	.55	.40	.41	-	.39	.73
	Q	.64	.42	.41	-	.40	.79
	R	.33	.36	.56	-	.55	.27

(e) Distribution of Error Term: Gamma 2

Procedure		Variable Pint1	Pint2	Pint3	Pdied	Educ1	Educ2
CL							
	M	1.51	1.10	.57	1.56	-.34	-2.06
	S	.34	.24	.30	.24	.24	.68
	Q	.35	.28	.30	1.57	.32	.73
	R	.18	.26	.42	-	.46	.27
CL							
	M	1.16	.92	.50	-	-.42	-2.27
	S	.52	.27	.27	-	.29	.70
	Q	.52	.22	.28	-	.26	.69
	R	.28	.28	.39	-	.37	.25
GEE							
	M	1.50	1.09	.58	1.47	-.35	-2.06
	S	.34	.24	.30	.27	.24	.68
	Q	.35	.28	.30	1.49	.30	.73
	R	.18	.25	.42	-	.45	.23
GEE							
	M	1.17	.91	.52	-	-.45	-2.26
	S	.33	.21	.22	-	.24	.69
	Q	.51	.22	.28	-	.35	.25
	R	.28	.19	.19	-	.35	.25
RL							
	M	1.72	1.23	.69	1.11	-.49	-2.35
	S	.43	.29	.36	.39	.28	.73
	Q	.45	.42	.39	1.12	.40	.26
	R	.22	.37	.53	-	.40	.26
RL							
	M	1.72	1.27	.75	-	-.67	-2.84
	S	.50	.33	.39	-	.33	.83
	Q	.52	.47	.44	-	.35	.96
	R	.26	.42	.61	-	.51	.33

(f) Distribution of Error Term: Non parametric 2

Procedure		Variable					
		Pint1	Pint2	Pint3	Pdied	Educ1	Educ2
CL							
	M	1.64	.96	.54	.51	-.65	-2.34
	S	.29	.27	.33	.42	.36	.52
	Q	.30	.21	.33	.66	.37	.51
	R	.19	.22	.45	-	.59	.19
CL							
	M	1.20	.92	.53	-	-.67	-2.38
	S	.34	.27	.33	-	.36	.51
	Q	.34	.26	.33	-	.37	.50
	R	.23	.23	.45	-	.54	.19
GEE							
	M	1.62	.96	.54	.47	-.65	-2.34
	S	.33	.27	.33	.44	.36	.52
	Q	.32	.28	.33	.63	.37	.51
	R	.32	.23	.45	-	.53	.19
GEE							
	M	1.28	.92	.53	-	-.67	-2.38
	S	.34	.28	.33	-	.36	.51
	Q	.50	.26	.33	-	.37	.50
	R	.28	.24	.54	-	.54	.19
RL							
	M	1.78	.99	.57	.25	-.69	-2.44
	S	.37	.29	.35	.53	.39	.56
	Q	.69	.28	.35	.58	.41	.56
	R	.37	.24	.48	-	.58	.20
RL							
	M	1.83	.98	.56	-	-.71	-2.46
	S	.38	.29	.35	-	.39	.53
	Q	.67	.29	.35	-	.41	.54
	R	.36	.25	.48	-	.50	.19

Table 5: Estimates obtained from Brazil-DHS

Hazard Models.

Model	Variable	Estimate	Standard Error
CC	Pint1	1.560	.181
	Pint2	.875	.147
	Pint3	.495	.172
	Pdied	.574	.154
SC	Pint1	.723	.267
	Pint2	.735	.217
	Pint3	.627	.239
	Pdied	-1.215	.205

Binary Response Models

Model	Variable	Estimate	Standard Error
CL	Pint1	1.693	.201
	Pint2	.913	.156
	Pint3	.511	.184
	Pdied	.593	.171
Rl	Pint1	1.697	.232
	Pint2	.929	.172
	Pint3	.542	.199
	Pdied	.161	.218
GEE	Pint1	1.644	.196
	Pint2	.895	.159
	Pint3	.509	.174
	Pdied	.426	.196

Table 6: Numerical evaluation of error in estimate of population attributable risk

Case I : Normal distribution of unmeasured characteristic (mean=-2.5; st.dev=1)			
	β=.50		
	Values of w		
	.10	.50	.90
Correct PAR	.42	.20	.034
Incorrect PAR	.49	.22	.037
	β=1.50		
	Values of w		
	.10	.50	.90
correct PAR	.49	.22	.038
Incorrect PAR	1.83	.56	.077
Case II: Non parametric distribution of unmeasured characteristic (mean=-2.15; st.dev=1.65)			
	β=.50		
	Values of w		
	.10	.50	.90
Correct PAR	.33	.16	.030
Incorrect PAR	.46	.21	.041
	β=1.50		
	Values of w		
	.10	.50	.90
correct PAR	.97	.38	.058
Incorrect PAR	1.83	.56	.077

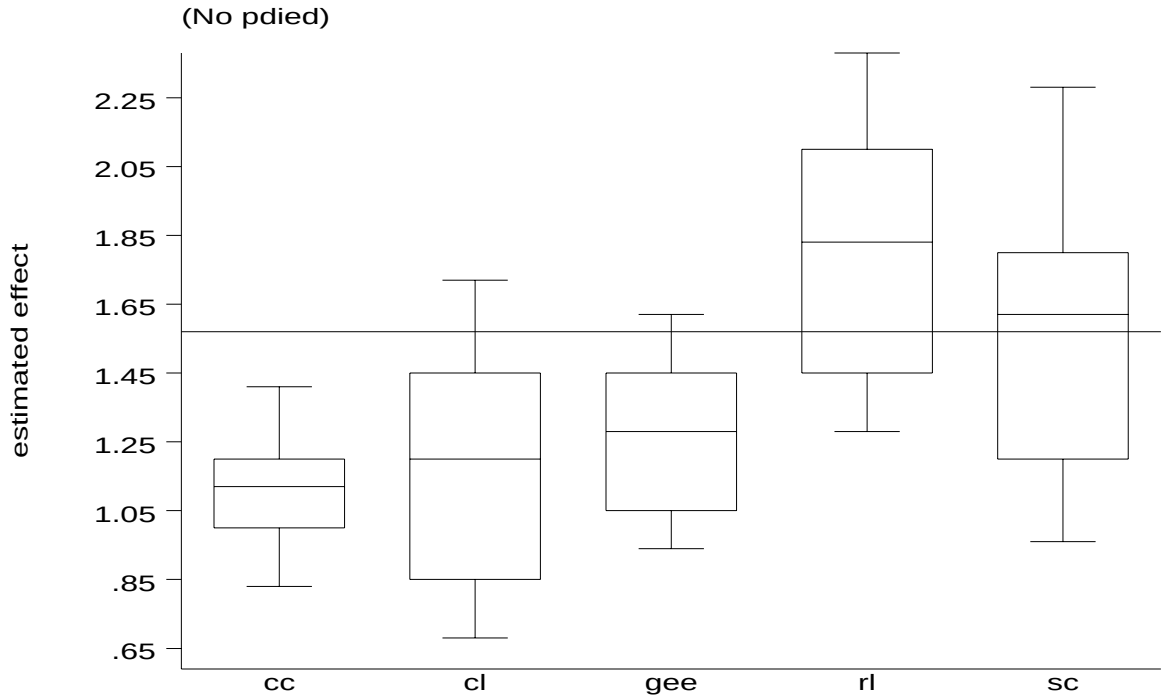


Figure 1a: Estimated effects of pint1 by estimation procedure

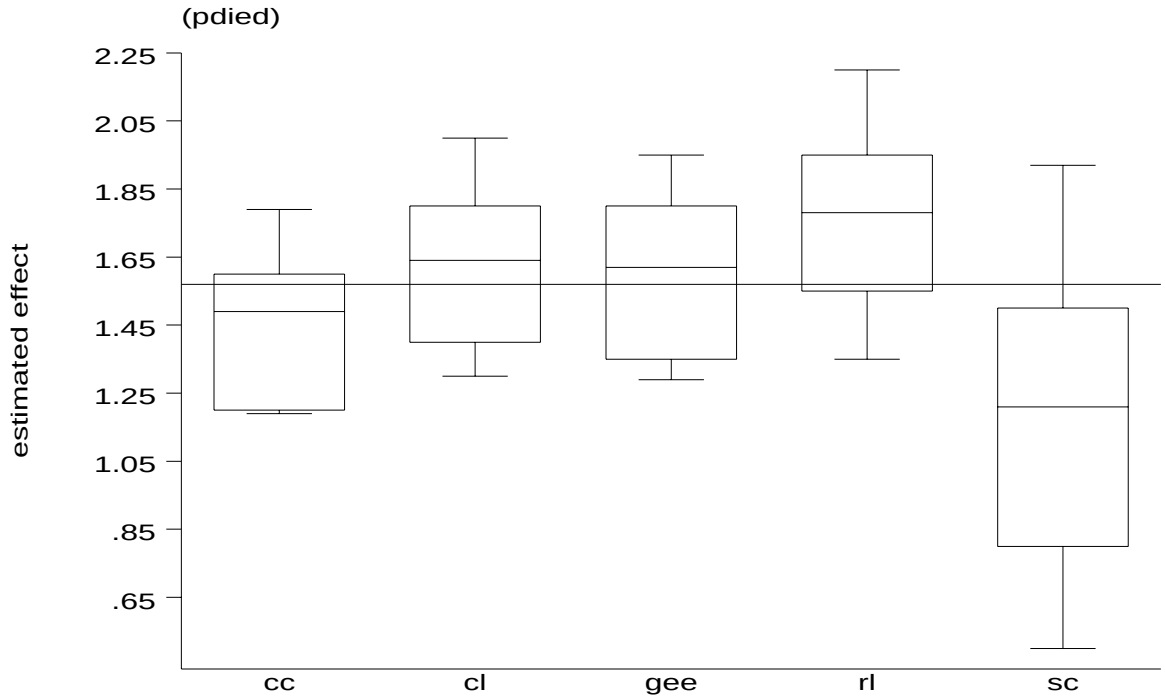


Figure 1b: Estimated effects of pint1 by estimation procedure

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
email: palloni@ssc.wisc.edu