Generalized Latent Trait Models for Multiple Correlated Health Endpoints

Xuefeng Liu* Kesheng Wang†

Abstract

Latent trait models have an abroad application in education, health science, psychology and other areas. There are two common assumptions in latent trait models: local independence of manifest outcomes and normal distribution of latent traits. In practice, these assumptions may not be satisfied, especially for the normality of latent traits. In this study, a class of generalized latent trait models and modified Gauss−Newton algorithms for multiple outcomes are proposed. Instead of assuming latent traits to be normal, we specify a skew normal distribution for latent traits of which a normal distribution is a special case, and then model the conditional probability of each outcome as a nonlinear quadratic function of latent traits, which has properties similar to the logistic function. The estimated generalized nonlinear least-square method is used to solve equations for parameters of interest. The models are applied to an infant morbidity study to develop a new single variable, called infant morbidity index (IMI) that functions as a summary of four infant morbidity outcomes and represents propensity for infant morbidity, is developed. The validity of this index as a measure of propensity for infant morbidity needs to be further investigated in future research.

Key Words: Latent variable models, Latent traits, Modified Gauss-Newton algorithms, Non-linear least square methods, Infant morbidity index

1. Introduction

Infants have varying propensities for morbidity. The propensity for morbidity (denoted by S) is unobserved and manifest in morbidity outcomes. There are several different morbidity outcomes to consider, including birth defect, abnormal conditions in born, developmental delay and/or disability, and abnormal birth weight. Most current studies have focused on assessment of effects of risk factors on individual outcomes or of the relationship between these outcomes or both. It is not clinically useful to identify risk factors and quantify the strength of the relationship between risk factors and infant morbidity only by analyzing effects of risk factors on individual morbidity outcomes. A single comprehensive measurement of morbidity analogous to, say, blood pressure as a measure of cardiac output is needed. Since these outcome variables are major causes of infant morbidity, the effects of risk factors on these outcome variables and their relationship have been analyzed in many studies. Yet little attention has been paid to the development of a composite index which is a summary construct of infant morbidity. Such an index, nevertheless, could provide a single comprehensive measure that might be useful in studies of infant health. Compared with the subjective face-valid index, the new index not only identifies all the patterns of morbidity outcomes, but also allows early intervention programs that aim to improve the health of infants with morbidity. The objective of this study is to develop latent trait models for an index of infant morbidity by combining four pregnancy outcomes and to assess the validity of the index.

Latent variable models are a class of models that link a set of manifest variables to a set of unobserved latent variables. When manifest variables are categorical and latent variables are continuous, they become latent trait models. Latent trait models have an abroad application in education, health science, psychology and other areas. There are two common

*Department of Systems, Population, and Leadership, University of Michigan, 400 North Ingalls Street, MI 48109, Email: liuxf@umich.edu

†Department of Biostatistics and Epidemiology, East Tennessee Sate University, Johnson City, TN 37614

assumptions in latent trait models: local independence of manifest outcomes and normal distribution of latent traits. In practice, these assumptions may not be satisfied, especially for the normality of latent traits. Liu $etal$ (2008) proposed lognormal distribution to account for the skewness of the latent trait while the conditional independence of manifest outcomes is still assumed. Dunson (2006) developed a Bayesian semiparametric approach to the assumption problem, which is an alternative to dependent Dirichlet process used as a class of priors for a collection of unknown distributions. Typically, the manifest outcomes are related to a latent response variable through a factor analytic model, with a scale mixture of underlying normals used to characterize flexibly the measurement error distributions. Their primary focus was on developing a hierarchical Dirichlet process for assessing dynamic changes in the latent response distribution. Miyazaki $etal$ (2009) proposed a new semiparametric model using a Dirichlet process mixture logistic distribution. The method does not rely on assumptions of the local independence of manifest outcomes. However, the latent trait models that involve a Dirichlet process can be computationally challenging and are difficult to understand by health science researchers.

In this paper, we develop a latent trait model such that the conditional probability of manifest outcomes are related to the latent trait through a quadratic model with similar properties of logistic regression models while the conditional probability could be anchored as 0 when the value of latent trait is 0. Instead of assuming a sysmmetric normal distribution, we use a skewed normal distribution with non-zero mean to explain the potential asysmmetry of the latent trait distribution. These models are utilized to combine manifest outcomes into a single infant morbidity index (IMI) that reflects the propensity for morbidity (latent trait). An overall assessment of validity of the IMI is developed by correlating it with each outcome, with infant mortality and with a face valid index of morbidity outcomes. In our analysis, we assume that the likelihood of each of these four outcomes is affected by an underlying propensity for morbidity that has a common influence on all four morbidity outcomes.

This paper is structured as follows. In Section 2, we describe the source and creation of infant morbidity data used in this study. Section 3 presents generalized latent trait models for multiple multinomial outcomes, and quadratic functions for modelling conditional probabilities of manifest outcomes given latent traits. In Sections 4 and 5, we introduce modified Gauss-Newton Algorithms for estimating the parameters of interest in latent trait models and test the goodness of fit of the models. We develop an infant morbidity index in Section 6. Several issues are discussed in the final section.

2. Data Sample and Definitions of Manifest Outcomes

The infant morbidity data were derived from the merger of four data sources. The base data set was drawn from Florida's Birth Vital Statistics, 1997-2010. This data set contained sociodemographic and perinatal health factors, as well as a measure of tobacco use, for pregnant women who had children born in the state of Florida in 1997 and 2010. It was augmented by three other data sources: 1) Florida Birth Defects Registry which contained information on children who were diagnosed as having birth defect; 2) Children's Medical Services Early Intervention Program (CMS-EIP) which contained information on children who were assessed for developmental delay or disability and received evaluation or intervention services, 1997-2011; 3) Medicaid status from the Agency for Health Care Administration (AHCA) Medicaid eligibility files. Two exclusion criteria were applied to this merged population-based data set: 1) multiple births; 2) missing values of birth weight (BW), demographic, behavioral or perinatal health factors. The first criterion was used to satisfy the independency of individuals in the study population. After applying

the above criteria and deleting BW<350 and BW≥6,000, 385,485 records were available for analysis. Birth defect (BD), abnormal condition (AC) of the newborn, developmental delay or disability (DDD) and low birth weight (LBW) are four major contributors to infant morbidity. In this data set, dichotomous outcomes were BD diagnosed in the first year of life, AC of the newborn, and DDD diagnosed under the age of 1 year. Birth weight (BW) was classified into four categories: extremely low birth weight (ELBW, 350-999g), very low birth weight (VLBW, 1,000-1,499g), low birth weight (LBW, 1,500-2,499g) and normal birth weight (NBW, 2,500-5,999g).

BD is defined as abnormal development of the fetus resulting in death, malformation, growth retardation, or functional disorders. The major risk factors of BD are environmental exposure, maternal alcohol consumption, and maternal cigarette smoking during pregnancy. AC of the newborn denotes infants who have anemia (HCT<39/HGB<13), birth injury, fetal alcohol syndrome, hyaline membrane disease/RDS, meconium aspiration syndrome, seizures or assisted ventilation. Developmental delay is the slowed or impaired development of a child under 5 years old. Development disability when applied to infants and young children means individuals from birth to age 5 years, inclusive, who have substantial developmental delay or specific congenital or acquired conditions with a high probability of a mental or physical impairment or combination of mental and physical impairments. Many children show problems of DDD with time, e.g., 6-7% by 1 year of age and 12-14% by school age. An outcome variable, which is related with DDD is LBW. LBW is a strong predictor of DDD in early childhood. LBW refers to infants born less than 2,500 grams.

3. STATISTICAL MODELS

3.1 Generalized Latent Trait Models

To develop generalized latent trait models for multiple multinomial morbidity outcomes with complete responses, we define S to be a univariate unobservable latent variable of interest with values $s \in [0, \infty)$ and Y_m $(m = 1, \cdots, M)$ to be the m-th manifestation of S with potential values $y_m \in \{1, \cdots, C_m\}$. For simplicity, here we omit the subscript denoting individuals when giving the model for a single individual. Denote $\pi(S)$ to be the marginal density of S and $\pi(Y_m = y_m|s)$ to be the conditional probability that given s, the individual will have a response y_m to outcome m $(m = 1, \cdots, M;\ y_m = 1, \cdots, C_m)$. Under latent trait models, all these outcome variables are associated because the population under study is a scaled mixture of subpopulations. As applied to our infant morbidity data, four manifestations of S are: Y_1, BD with two categories (Yes, No); Y_2, AC with two categories (Yes, No); Y_3, DDD which is dichotomous (Yes, No); Y_4, BW, which has four categories (extremely low birth weight, very low birth weight, low birth weight and normal birth weight). In this case, it is natural to associate the latent variable S with underlying infant morbidity and consider observed outcomes to be surrogates for S, which can be thought as propensity for infant morbidity. Here $\pi(S)$ denotes the marginal distribution of propensity for infant morbidity, and $\pi(Y_m = y_m|s)$ is the probability of individuals who will have the response y_m to infant morbidity outcome m (m=1,2,3,4) given latent value s. Since S is the propensity for infant morbidity, it is reasonable to treat it as a continuous variable. Then the joint distribution of morbidity outcomes can be written as

$$\pi(Y_1 = y_1, \cdots, Y_M = y_M) = \int_0^\infty \pi(S) \prod_{m=1}^{M} \prod_{y_m=1}^{C_m} (\pi(Y_m = y_m|S))^{a_{y_m}} \mathrm{d}S, \qquad (1)$$

where $a_{y_m} = 1$ if the individual falls into the y_mth category of outcome m and 0 otherwise. The expression in (1) will be referred to as the cell probability in Section 4, which is the function of the vector of $\pi(Y_m = y_m|s)$ and $\pi(S)$. The term $\prod_{y_m=1}^{C_m}(\pi(Y_m = y_m|S))^{a_{y_m}}$ is a multinomial process associated with outcome m, where $\pi(Y_m = y_m|S)$ is the conditional probability for the y_mth category of outcome m subject to $\sum_{y_m=1}^{C_m}\pi(Y_m = y_m|S) = 1$ $(m = 1, \cdots, M)$. For BD, AC and DDD, it is a Bernoulli, but for BW, it is a multinomial distribution with n equal to 1. Note that in (1), the latent variable is continuous and observed outcomes are categorical. In fact, the distribution (1) is a scaled mixture of product multinomial processes with mixing weights $\pi(S)$. One basic assumption implicit in (1) is that given latent value s, responses of morbidity outcomes are independent, i.e.,

$$\begin{aligned}\pi(Y_1 = y_1, \cdots, Y_M = y_M|S) &= \prod_{m=1}^{M}\pi(Y_m = y_m|S) \\ &= \prod_{m=1}^{M}\prod_{y_m=1}^{C_m}(\pi(Y_m = y_m|S))^{a_{y_m}},\end{aligned} \tag{2}$$

This conditional independence is equivalent to the axiom of local independence. In our analysis, as discussed by Bandeen-Roche *et al* (1997), not only is it convenient, but also it defines the sense in which the S serves as a summary construct. Information about S is available from (1) through posterior distributions of the latent variable

$$f(S|Y_1 = y_1, \cdots, Y_M = y_M) = \frac{\pi(S)\prod_{m=1}^{M}\prod_{y_m=1}^{C_m}(\pi(Y_m = y_m|S))^{a_{y_m}}}{\int_0^\infty \pi(T)\prod_{m=1}^{M}\prod_{y_m=1}^{C_m}(\pi(Y_m = y_m|T))^{a_{y_m}}\,\mathrm{d}T}, \tag{3}$$

These posterior distributions capture information about the unknown latent variable given observed indicators and hence are useful in development of a composite morbidity index for our example.

3.2 Quadratic Models for Conditional Distributions

Suppose that data of M distinct morbidity outcomes with C_m $(m = 1, \cdots, M)$ categories in outcome m are collected on an infant. Let $\pi(Y_m = y_m|s)$ for $y_m = 1, \cdots, C_m$ and $s \in [0, \infty)$ be the conditional probability that the infant will have a response y_m to outcome m. We capture information to develop a composite morbidity index through modelling the conditional probability of each morbidity outcome as a function of latent variable S. Since S is the continuous latent variable associated with manifest morbidity outcomes with domain $[0, \infty)$, it is reasonable to assume that the conditional probability of an infant having an adverse morbidity outcome will be zero when S is equal to zero. For this reason, the following models for conditional probability are considered

$$\pi(Y_m = y_m|S) = \frac{(\beta_{my_m}S)^{2\alpha_{my_m}}}{1 + \sum_{j=1}^{C_m-1}(\beta_{mj}S)^{2\alpha_{mj}}}, \tag{4}$$

where β_{my_m} and α_{my_m} $(m = 1, \cdots, M; y_m = 1, \cdots, C_m - 1)$ are parameters linking the latent variable to the conditional probability $\pi(Y_m = y_m|S)$ subject to $\sum_{y_m=1}^{C_m}\pi(Y_m = y_m|S) = 1$. They are the parameters of interest we need to estimate in the conditional distribution of S given the observed outcomes.

For every morbidity outcome, we treat the normal category as the reference category. That is to say, we specify $\alpha_{mC_m} = 0$ for any $m = 1, \cdots, M$. For example, BD is the first

dichotomous outcome (Yes, No) we introduced in our study. So the category BD=No is the reference category ($\alpha_{12} = 0$) and the model for the conditional probability of BD=Yes is

$$\pi(Y_1 = 1|S) = \frac{(\beta_{11}S)^{2\alpha_{11}}}{1 + (\beta_{11}S)^{2\alpha_{11}}}.$$

BW is the fourth outcome under study with four categories (ELBW, VLBW, LBW, NBW). Then the model for the conditional probability of ELBW is

$$\pi(Y_4 = 1|S) = \frac{(\beta_{41}S)^{2\alpha_{41}}}{1 + \sum_{j=1}^{3}(\beta_{4j}S)^{2\alpha_{4j}}}.$$

Similarly, we can easily write out the models for VLBW and LBW. Here we treat NBW as the reference category ($\alpha_{44} = 0$).

3.3 Latent Trait Distribution

In our study, the latent variable S represents propensity for infant morbidity which is manifest in four outcomes: BD, AC, DDD, and LBW. Our prior belief is that most infants are exposed to low-level risks such that the latent variable may have an asymmetric distribution with a long right tail. We assume that S has the following distribution

$$\pi(S) = \frac{2}{\omega}\phi\left(\frac{S-\theta}{\omega}\right)\Phi\left(\gamma \cdot \left(\frac{S-\theta}{\omega}\right)\right), \tag{5}$$

where $\phi(\cdot)$ denotes a standard normal distribution with a cumulative distribution of $\Phi(\cdot)$. θ and ω are location and scale parameters, and γ is the shape parameter. It is easy to verify that the normal distribution is a special case of this distribution when $\gamma = 0$, and that the absolute value of the skewness increases as the absolute value of γ increases. The distribution is right skewed if $\gamma > 0$ and is left skewed if $\gamma < 0$. For simplicity, we refer to (4) combined with (1) and (5) as s-models.

Model (4) has good properties similar to those of logistic models with the difference that the conditional probabilities in model (4) will be zero if S is anchored zero. This is generally consistent with the hypotheses in studies of infant medical and health care that prevalence for some disease is zero in absence of risk factors. When the shape parameter γ=0, the relationship between (4) and the logistic model is reflected in the following reparameterization

$$\pi(Y_m = y_m|Z) = \frac{\exp(\beta^*_{0my_m} + \beta^*_{my_m}Z)}{1 + \sum_{j=1}^{C_m-1}\exp(\beta^*_{0jy_m} + \beta^*_{jy_m}Z)}, \tag{6}$$

where $\beta^*_{0my_m} = 2\alpha_{my_m}(\mu + \log\beta_{my_m})$, $\beta^*_{my_m} = 2\alpha_{my_m}/\omega$ and $Z = (S-\mu)/\omega$ ($m = 1, \cdots, M; y_m = 1, \cdots, C_m$). Here Z denotes the variable which has a standard normal distribution. The relationship of (4) to the logistic model, together with the medical hypothesis, provides us a good reason to use these models in our analysis. Note that we require $S > 0$ and $\beta_{my_m} > 0$ in model (4) to complete this translation.

4. ESTIMATION METHOD

Consider M infant morbidity outcomes with each having C_m categories. Define β to be the vector of β_{my_m} and α_{my_m} ($m = 1, \cdots, M; y_m = 1, \cdots, C_m - 1$) which are parameters of interest in s-models defined in (4). Suppose that observations on all subjects can be arranged in $N = \prod_{m=1}^{M} C_m$ cell counts. Let $n = (n_1, \cdots, n_N)$ be the

Table 1: Estimates of parameters in s-models

BD		AC		DDD		BW					
α_{11}	β_{11}	α_{21}	β_{21}	α_{31}	β_{31}	α_{41}	β_{41}	α_{42}	β_{42}	α_{43}	β_{43}
1.02	1.08	1.12	2.06	2.37	2.08	3.51	2.16	3.14	2.15	1.25	2.40

vector of cells, where $n_i's$ $(i = 1, \cdots, N)$ have a multinomial distribution with vector of cell probabilities $\pi(\beta) = (\pi_1(\beta), \cdots, \pi_N(\beta))'$. Here $\pi(\beta)$ is defined by expressing (1) and (4) as functions of $\beta = (\alpha_{11}, \beta_{11}, \cdots, \alpha_{M(C_M-1)}, \beta_{M(C_M-1)}, \theta, \omega, \gamma)$. Considering model/paraemter identifiability, we fix the scale parameter ω to be 1. Let $p = (p_1, \cdots, p_N)'$ be the vector of sample proportions ($p_i = n_i/n$). Then the sample proportions of p are unbiased and consistent estimates of cell probabilities of $\pi(\beta)$. That is,

$$p = \pi(\beta) + \epsilon$$

where the expectation of ϵ is $\mathrm{E}(\epsilon) = 0$ and the variance is $\mathrm{var}(\epsilon) = n^{-1}V$ and $V = n(\mathrm{diag}(\pi) - \pi\pi')$. From (1) and (4), it is easy to see that $\pi(\beta)$ is a nonlinear function of β. Following the nonlinear least square theory, the estimated generalized non-linear least square (EGNLS) estimates of β can be obtained by minimizing the quadratic form:

$$Q(\beta, \hat{V}) = n(p - \pi(\beta))'\hat{V}^{-1}(p - \pi(\beta)) \tag{7}$$

where $\hat{V}$ represents a consistent estimator of V and hence does not depend on β. Considering the algorithm convergence, we choose $\hat{V} = n(\mathrm{diag}(p) - pp')$, which only depends on the sample proportion p. By taking the derivative on both sides of (7), we obtain the estimating equations

$$\frac{\partial Q(\beta, \hat{V})}{\partial \beta} = \left(\frac{\partial \pi(\beta)}{\partial \beta}\right)' \hat{V}^{-1}(p - \pi(\beta)) = 0 \tag{8}$$

Note that if we substitute V for $\hat{V}$, (8) becomes the optimal estimating equations which are used to compute the maximum likelihood estimates (MLEs). Because of the dependence of V on β, the MLEs often attain local maxima. The idea of using the simplification (8) obtained by substituting $\hat{V}$ for V in the optimal estimating equations is to avoid the local maxima problem that we face in computing the MLEs. Let $\beta^{(k)}$ denote the estimate at the kth iteration. By Hartley's modified Gauss-Newton method, the estimate $\beta^{(k+1)}$ at iteration $k + 1$ is:

$$\begin{aligned}\beta^{(k+1)} =& \beta^{(k)} + \lambda^{(k)} \left[\left(\frac{\partial \pi(\beta^{(k)})}{\partial \beta}\right)' \hat{V}^{-1} \left(\frac{\partial \pi(\beta^{(k)})}{\partial \beta}\right)\right]^{-1} \\ & \left(\frac{\partial \pi(\beta^{(k)})}{\partial \beta}\right)' \hat{V}^{-1}(p - \pi(\beta^{(k)}))\end{aligned} \tag{9}$$

where $\lambda^{(k)} \in [0, 1]$ is the stepping coefficient for the k-th iteration. We adjust $\lambda^{(k)}$ to guarantee that $\beta^{(k+1)}$ is a better approximation to the least square estimator $\hat{\beta}$ than $\beta^{(k)}$ in the sense that $Q(\beta^{(k+1)}, \hat{V}) \leq Q(\beta^{(k)}, \hat{V})$. Although $\pi(\beta)$ and the derivatives of $\pi(\beta)$ with respect to β have no closed forms, we can calculate these integrals using common numerical integration methods since the integral over S is only one-dimensional. In our study, there are four morbidity outcomes: BD, AC, DDD, and BW. The first three are

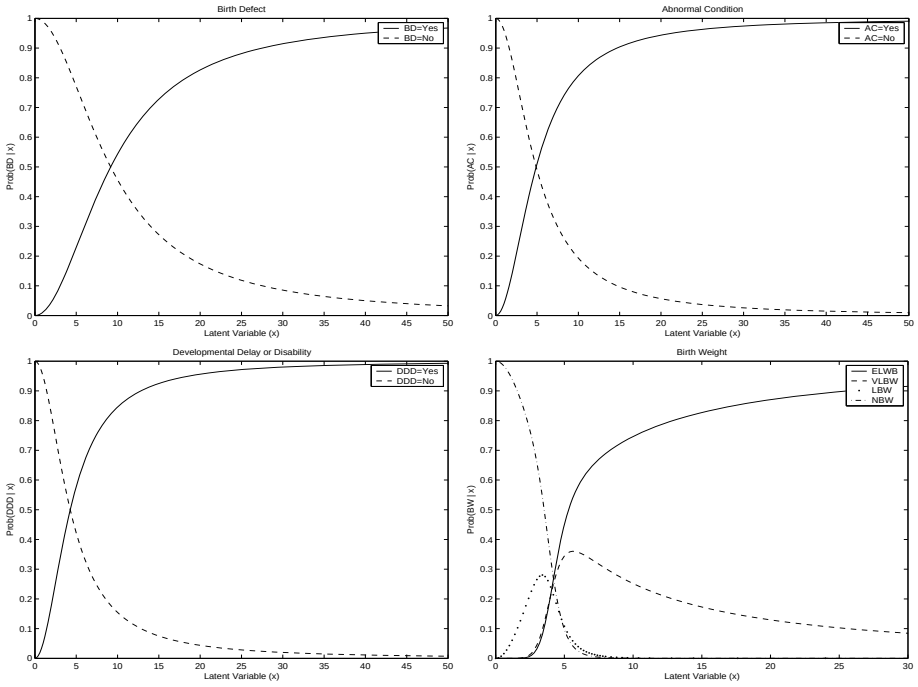

Figure 1: Associations of the latent trait with the risk of individual manifest outcomes

dichotomous and the last one is polytomous with four categories. Based on s-models, we have 13 parameters to be estimated. Note that considering the identifiability of parameters in s-models, we fix the variance/scale parameter ω in (5) at 1 in our estimation.

5. ESTIMATION RESULTS

Using modified Gauss-Newton algorithm described in Section 4, we obtain estimates of parameters in s-models (Table 1). Substituting these estimates into model (4), we calculate conditional probabilities given s for each outcome.

Figure 1 shows conditional probabilities for each outcome variable. Change patterns are similar for three dichotomous outcomes: BD, AC, and DDD. The conditional probabilities for adverse outcomes (BD: Yes, AC: Yes, DDD: Yes) decrease, and those for normal outcomes (BD: No, AC: No, DDD: No), however, increase with s increasing. In the graph for the multinomial outcome BW, there are four lines which correspond to NBW, LBW, VLBW, and ELBW, respectively. Note that trends of curves associated with LBW and VLBW are different from those of NBW and ELBW. They increase first, attain maximum, and then decrease. The changes for curves of NBW and ELBW are similar to those of BD, AC, and DDD. The order in which curve peaks for BW appear is NBW, LBW, VLBW, and ELBW.

Based on (1) and (4), we calculate expected probabilities and counts for each combination of four morbidity outcomes. Since the sample size in our study is very large ($n = 385485$), it will not be appropriate to use χ^2 statistics to test the goodness of fit of s-models. We suggest using the weighted regression method to do it. To perform the test, assume ϵ is the random vector with mean $E(\epsilon) = 0$ and variance $V(\epsilon) = n\,[\text{diag}(\pi) - \pi\pi']$. Let the sample count be the dependent variable, denoted by Y and the expected count be the covariate, denoted by X. Then we can use SAS PROC REG to fit the model

$$Y^* = X^*\beta + \epsilon^*$$

where $Y^* = \hat{V}^{-\frac{1}{2}}Y$, $X^* = \left(\hat{V}^{-\frac{1}{2}}1, \hat{V}^{-\frac{1}{2}}X\right)$ and $\epsilon^* = \hat{V}^{-\frac{1}{2}}\epsilon$. In the above transformation, $\hat{V} = n\left[\text{diag}(p) - pp'\right]$ is the consistent estimator of V and $\hat{V}^{-\frac{1}{2}}$ is equal to $ED^{-\frac{1}{2}}E'$, where E is the matrix of eigenvectors of $\hat{V}$ and D is the diagonal matrix with eigenvalues of $\hat{V}$ down the diagonal. The test result ($R^2 = 0.98$) shows that our s-models fit the data set well.

6. DEVELOPMENT OF IMI

In section 5, we derived estimates of parameters related to s-models. Substituting parameter estimates into (3), we obtained posterior distributions of latent variable S (propensity for infant morbidity). Let $\hat{s} = \hat{E}(S|Y = y)$, then

$$\hat{s} = \int Sf(S|Y_1 = y_1, \cdots, Y_M = y_M)dS \tag{10}$$

where $f(S|Y_1 = y_1, \cdots, Y_M = y_M)$ is given by (3).

To adapt for use in infant helath, we re-scale $\hat{s}$ on the 1-100 scale and arrive at the proposed IMI. Thus, the IMI is estimated as a function of its manifestations: BD, AC, DDD, and BW. It is a composite index of propensity for infant morbidity which is developed from s-models. Table 2 lists results for $\hat{s}$ and IMI, given the pattern of four morbidity outcomes. We see that all infants who have the same pattern of outcomes have equal IMI values. The IMI can identify patterns of four outcomes. The range of the IMI is from 1 to 100. The maximal value is generated by infants who have the outcome pattern of BD=Yes, AC=Yes, DDD=Yes, and ELBW=Yes (the first row), and the minimal value is associated with infants whose pattern is BD=No, AC=No, DDD=No, and NBW=Yes (the last row). In addition, given the pattern of any 3 outcomes, there is an obvious trend of the IMI with the last outcome, i.e., values of the IMI increase with the last outcome getting worse. Further study will be needed to determine whether the IMI can exactly identify the pattern of outcomes in general situations.

7. DISCUSSION

Latent variable models have been applied in many clinical setting to describe disease in populations. Recent applications include gerontology, genetics, medical care, ophthalmology, and cytometry. Specifically in psychiatric research, latent variabl models have been used in studies of alcoholism, autism, social phobias, schizophrenia, psychiatric syndromes, and psychiatric disorder. In this study, latent variable models were applied to the study of infant morbidity, in which we developed a composite infant morbidity index based on four major outcomes BD, AC, DDD, and BW. This method for developing a composite single variable may be applied and extended to other areas.

Our study demonstrates an approach to development and validation of a composite index where the latent variable is continuous and observed outcomes are dichotomous or multinomial using latent variable modelling and modified Gauss-Newton estimation. In fact, this approach would have been latent trait setting since the latent variable S under study is continuous (we assumed that S was distributed as skew normal). This approach can be generalized to other latent variable models, such as latent profile models, where the latent variable is categorical and the observed indicators are continuous, and latent structure models, where the latent variable and the observed indicators are all continuous. In all these latent variable models, the key is to associate general health status such as infant morbidity,

Table 2: Estimates of posterior means of the latent variable $\hat{s}$ and infant morbidity index (IMI) across groups of manifest outcomes

BD	AC	DDD	BW	$\hat{s}$	IMI
Yes	Yes	Yes	350-999	0.5599	100.00
Yes	Yes	Yes	1000-1499	0.5461	97.04
Yes	Yes	Yes	1500-2499	0.4581	78.19
Yes	Yes	Yes	2500-5999	0.3864	62.83
Yes	Yes	No	350-999	0.4524	76.97
Yes	Yes	No	1000-1499	0.4310	72.36
Yes	Yes	No	1500-2499	0.3208	48.78
Yes	Yes	No	2500-5999	0.2391	31.27
Yes	No	Yes	350-999	0.5155	90.49
Yes	No	Yes	1000-1499	0.4972	86.57
Yes	No	Yes	1500-2499	0.3948	64.65
Yes	No	Yes	2500-5999	0.3196	48.52
Yes	No	No	350-999	0.3888	63.34
Yes	No	No	1000-1499	0.3661	58.48
Yes	No	No	1500-2499	0.2488	33.37
Yes	No	No	2500-5999	0.1614	14.63
No	Yes	Yes	350-999	0.5207	91.62
No	Yes	Yes	1000-1499	0.5030	87.81
No	Yes	Yes	1500-2499	0.4016	66.11
No	Yes	Yes	2500-5999	0.3267	50.06
No	Yes	No	350-999	0.3956	64.80
No	Yes	No	1000-1499	0.3730	59.98
No	Yes	No	1500-2499	0.2568	35.06
No	Yes	No	2500-5999	0.1695	16.36
No	No	Yes	350-999	0.4656	79.80
No	No	Yes	1000-1499	0.4445	75.28
No	No	Yes	1500-2499	0.3356	51.95
No	No	Yes	2500-5999	0.2555	34.78
No	No	No	350-999	0.3292	50.57
No	No	No	1000-1499	0.3054	45.48
No	No	No	1500-2499	0.1796	18.52
No	No	No	2500-5999	0.0978	1.00

physical disability, and mental illness status, of which there is no single measure analogous to, say, blood pressure as a measure of cardiac output, with the unobservable latent variable.

REFERENCES

Agresti, A. (2002), "Categorical Data Analysis" (2nd edtion), Wiley: New York.

Agustines, L.A., Kub, Y.G., Rumney, P.J., Lu, M.C., Bonebrake, R., Asrat, T., Nageotte, M. (2000), "Outcomes of extremely low-birth-weight infants between 500 and 750g," *American Journal of Obstetrics and Gynecology*,182,1113–1116.

Bandeen-Roche, K., Munoz, B., Tielsch, J.M., West, S.K., Schein, O.D. (1997), "Self-reported assessment of dry eye in a population-based setting," *Investigative Opthalmology and Visual Science*, 38, 2469–2475.

Benning, S.D., Patrick, C.J., Salekin, R.T., Leistico, A.M.R. (2005), "Convergent and discriminant validity of psychopathy factors assessed via self-report," *Assessment*, 12, 270–289.

Bucher, H.U., Killer, C., Ochsner, Y., Vaihinger, S., Fauchere, J.(2002), "Growth, developmental milestones and health problems in the first 2 years in very preterm infants compared with term infants: population based study," *European Journal of Pediatrics*, 161, 151–156.

Bucholz, K.K., Heath, A.C., Reich, T., Hesselbrook, V.M., Kramer, J.R., Nurnberger, J.I. Jr., Schuckit, M.A. (1996), "Can we subtype alcoholism?," *Aocoholism: Clinical and Experimental Research*, 20, 1462–1471.

Collin, M.F., Jalsey, C.L., Anderson, C.L. (1991), "Emerging developmental sequelae in the 'normal' extremely low birth weight infants," *Pediatrics*, 88, 115–120.

Dunson, D.B. (2006), Bayesian dynamic modeling of latent trait distributions, *Biostatistics*, 7, 551–568.

Eaton, W.W., Dryman, A., Sorenson, A., McCutcheon, A.(1989), "DSM-III major depressive disorder in community: a latent class analysis of data from the NIMH Epidemiologic Catchment Area Program," *British Journal of Psychiatry*, 155, 48–54.

Eaton, W.W., McCutcheon, A., Dryman, A., Sorenson, A. (1989), "Latent class analysis of anxiety and depression," *Sociological Methods and Research*, 18, 104–125.

Eaves, L.J., Silberg, J.L., Hewitt, J.K., Rutter, M., Meyer, J.M., Neale, M.C., Pickles, A. (1993), "Analyzing twin resemblance in multisymptom data: genetic applications of a latent class model for symptoms of conduct disorder in juvenile boys," *Behaviroal Genetics*, 23, 5–19.

Fergusson, D.M., Horwood, L.J., Lynskey, M.T. (1995), "The prevalence and risk factors associated with abusive or hazardous alcohol consumption in 16-year-olds," *Addiction*, 90, 935–946.

Forbes, J.F., Pickering, R.M. (1988), "Development of a neonatal case-mix classification," *Medical Care*, 26, 1033–1045.

Gallant, R.A. (1987), Nonlinear Statistical Model. Wiley: New York.

Goldenberg, R.L., Hauth, J.C., Andrews, W.W. (2000), "Intrauterine infection and preterm delivery," *New England Journal of Medicine*, 342, 1500–1507.

Griffin, J.E., Steel, M.F.J., (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 473, 179-194.

Hack M., Fanaroff, A.A. (1999), "Outcomes of children of extremely low birth weight and gestational age in the 1990's," *Early Human Development*, 53, 193–218.

Hartley, HO. (1961), "The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares," *Technometrics*, 3, 269–280.

Hogan, D.P., Park, J.M. (2000), "Family factors and social support in the developmental outcomes of very low-birth weight children," *Clinical Perinatology*, 27, 433–459.

Kendler, K.S., Karkowski, L.M., Prescott, C.A., Pedersen, N.L. (1998), "Latent class analysis of temperance board registrations in Swedish male-male twin pairs born 1902 to 1949: searching for subtype of alcoholism," *Psychological Medicine*, 28, 803–813.

Kendler, K.S., Karkowski, L.M., Walsh, D. (1998), "The structure of psychosis: latent class analysis of probands from the Roscommon Family Study," *Archives of General Psychiatry*, 55, 492–499.

Kessler, R.C., Stein, M.B., Berglund, P. (1998), "Social phobia subtypes in the National Comorbidity Survey," *American Journal of Psychiatry*, 155, 613–619.

Lieff, S., Olshan, A.F., Werler, M., Strauss, R.P., Smith, J., Mitchell, A. (1999), "Maternal cigarette smoking during pregnancy and risk of oral clefts in newborns," *American Journal of Epidemiology*, 150, 683–694.

Liu, X.F., Roth, J.R., (2008), "Development and validation of an infant morbidity index using latent variable models," *Statistics in Medicine*, 27, 971–989.

Lorente, C., Cordier, S., Goujard, J., Ayme, S., Bianchi, F., Calzolari, E., De Walle, H.E., Knill-Jones, R. (2000), "Tobacco and alcohol use during pregnancy and risk of oral clefts, Occupational Exposure and Congenital Malformation Working Group," *American Journal of Public Health*, 90, 420–423.

MacEachern, S.N. (2000), Dependent dirichlet processes, unpublished paper.

Mathews, T.J., Curtin, S.C., MacDorman, M.F. (2000), "Infant mortality statistics from the 1998 period linked birth/infant death data set," *National Vital Statistics Reports*, 48, 1–25.

Mathews, T.J., MacDorman, M.F., Menacker, F. (2002), "Infant mortality statistics from the 1999 period linked birth/infant death data set," *National Vital Statistics Reports*, 50, 1–28.

McKeith, I.G., Fairbairn, A.F., Bothwell, R.A., Moore, P.B., Ferrier, I.N., Thompson, P., Perry, R.H. (1994), "An evaluation of the predictive validity and inter-rater reliability of clinical diagnostic criteria for senile dementia of Lewy body type," *Neurology*, 44, 872–877.

McCutcheon, A.L. (1987), "Latent Class Analysis," Sage: Newbury Park, CA.

Laz Lazarsfeld, P.F., Henry, N.W. (1968), "Latent Structure Analysis," Houghton-Mifflin: New York.

Melton, B., Liang, K.Y., Pulver, A.E. (1994), "Extended latent class approach of the study of familial/sporadic forms of disease: its application of the study of heterogeneity of schizophrenia," *Genetic Epidemiology*, 11, 311–327.

Miyazaki, K., Hoshino, T. (2009), "A Bayesian Semiparametric Item Response Model with Dirichlet Process Priors," *Psychometrika*, 74, 375–393.

Nestadt, G., Hanfelt, J., Liang, K.Y., Lamacz, M., Wolyniec, A., Pulver, A.E. (1994), "An evaluation of the structure of schizophrenia spectrum personality disorders," *Journal of Personality Disorders*, 8, 288–298.

Pickles, A., Bolton, P., MacDonald, H., Bailey, A., LeCouteur, A., Sim, C.H., Rutter, M. (1995), "Latent class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history stydy of autism," *American Journal of Human Genetics*, 57, 717–726.

Roche, K.B., Miglioretti, D.L., Zeger, S.L., Rathouz, P.J. (1997), "Latent variable regression for multiple discrete outcomes," *Journal of the American Statistical Association*, 92, 1375–1386.

Saigal, S., Rosenbaum, P., Stoskopf, B., Hoult, L., Furlong, W., Feeny, D., Hagan, R. (2005), "Development, reliability and validity of a new measure of overall health for pre-school children," *Quality Life Research*, 14, 243–257.

Saigal, S., Stoskopf, B.L., Streiner, D.L., Burrows, E. (2001), "Physical growth and current health status of infants who were of extremely low birth weight and controls at adolescence," *Pediatrics*, 108, 407–415.

Schendel, D.E., Stockbauer, J.W., Hoffman, H.J., Herman, A.A., Berg, C.J., Schramm, W.F. (1997), "Relation between very low birth weight and developmental delay among preschool children without disabilities," *American Journal of Epidemiology*, 146, 740–749.

Shaw, G.M., Lammer, E.J. (1999), "Maternal periconceptional alcohol consumption and risk for orofacial clefts," *Journal of Pediatrics*, 134, 298–303.

Sham, P.C., Castle, D.J., Wessely, S., Farmer, A.E., Murray, R.M. (1996), "Further exploration of latent class typology of schizophrenia," *Schizophrenia Research*, 20, 105–115.

Szatmari, P., Volkmar, F., Walter, S. (1995), "Evaluation of diagnostic criteria for autism using latent class models," *Journal of the Academy of Child and Adolescent Psychiatry*, 34, 216–222.

Sullivan, P.F., Kendler, K.S. (1998), "Typology of common psychiatric syndromes: an empirical study," *British Journal of Psychiatry*, 173, 312–319.

Thompson, J.R., Carter, R.L., Edwards, A.R., Roth, J., Ariet, M., Ross, N.L., Resnick, M.B. (2003), "A population based study of the effects of birth weight on early developmental delay and disability in children," *American Journal of Perinatology*, 20, 321–332.

van Putten, W.L., de Vries, W., Reinders, P., Levering, W., van der Linden, R., Tanke, H.J., Bolhuis, R.L., Gratama, J.W. (1993), "Quantification of fluorescence properties of lymphocytes in peripheral blood mononuclear cell suspensions using a latent class model," *Cytometry* , 14, 86–96.

Vinceti, M., Rovesti, S., Bergomi, M., Calzolari, E., Candela, S., Campagna, A., Milan, M., Vivoli, G. (2001), "Risk of birth defects in a population exposed to environmental lead pollution," *The Science of The Total Environment*, 278, 23–30.

Voigt, R.G., Brown, F.R., Fraley, J.K., Liorente, A.M., Rozelle, J., Jensen, C.L., Heird, W.C. (2003), "Concurrent and predictive validity of the cognitive adaptive test/clinical linguistic and auditory milestone scale (cat/clams) and the mental developmental index of the bayley scales of infant development," *Clinical Pediatrics*, 42, 427–432.