DEVELOPMENT OF AN UNBIASED LONGITUDINAL MODEL ON FUNCTIONAL STATUS TRANSITIONS IN OLDER PERSONS

Xian Liu, Ph.D.[1,2]

Charles C. Engel, Jr., M.D., M.P.H.[1,2]

Han Kang, Dr.PH[3]

Kristie L. Gore, Ph.D.[1,2]

[1] Deployment Health Clinical Center
Walter Reed Army Medical Center
Washington, DC 20307-5001

[2] Department of Psychiatry
F. Edward Hebert School of Medicine
Uniformed Services University of the Health Sciences
Bethesda, Maryland 20814

[3] Environmental Epidemiology Service
Department of Veterans Affairs
Washington, DC 20420

ABSTRACT

In this study, we develop 2 two-stage mixed models seeking to reduce the selection biases in estimating longitudinal outcomes of the number of functional limitations, one parametric and one non-parametric. The parametric model is a two-step perspective, developed as an extension of Heckman's traditional two-step linear regression model, whereas the nonparametric mixed model uses a retransformation approach taking into account the prediction biases given skewed error distributions across time. Our empirical analysis found that the nonparametric adjusting method is the most appropriate approach for analyzing large-scale survey data for transitions in the number of functional limitations in older persons.

INTRODUCTION

During recent decades, the mean age of the population has sharply increased in the United States. Given the significant impact of such demographic changes on the demands for health services and on fundamental social aspects of life, studies analyzing the health of older persons have become topics of emerging interest. Much of this research relies on measures of health status as integral variables, and substantial attention has been placed upon transitions in an older person's functional limitations (Crimmins et al., xxxx; Liu et al., 1995). Yet demographers have devoted surprisingly little effort to the development of efficient statistical models to describe and analyze longitudinal outcomes of functional status in older persons.

Analyzing longitudinal data of functional status transitions poses special challenges to demographers and epidemiologists. Although most longitudinal surveys collect random and unbiased samples at baseline, a considerable proportion of the baseline respondents will not survive to the ensuing investigations, particularly among older persons. As a result, the longitudinal health outcomes would be based on several follow-up samples selected by values of the dependent health variable, since physically frailer and environmentally disadvantaged persons tend to be more likely to die. Direct application of conventional one-step linear mixed models would lead to inconsistent estimates of the effects on the number of functional limitations, in turn resulting in misleading conclusions on patterns of health transitions.

In this research, we develop two generalized mixed models to address this selection bias problem in estimating longitudinal outcomes of the number of functional limitations, one parametric and one non-parametric. The parametric model is a two-stage

perspective, developed as an extension of Heckman's traditional two-step linear regression model, whereas the nonparametric mixed model uses a retransformation approach taking into account the prediction biases given skewed error distributions. Lastly, we utilize empirical examples to demonstrate the new methods developed in this research and discuss merits and weaknesses in each of those models.

Impact of Selection Biases

We develop a two-stage regression model to demonstrate the selection bias problem in estimating the number of functional limitations, by adapting the traditional two-step linear models in econometrics. In terms of an original sample of I observations and J time points, we first assume the existence of the number of functional limitations at each of the subsequent time points for those who have perished between two adjacent observation time-points. We further assume that the number of functional limitations for the dead, denoted by Y^d, is greater than or equal to a constant C, and numbers of functional limitations for survivors, Y^s, are all smaller than this constant.

We begin with two longitudinal mixed models, one complete model that includes all members of the original sample and one truncated model that consists of survivors only, to clarify the impact of sample selection on estimates of the number of functional limitations, given by

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1 \qquad (1a)$$

$$\mathbf{Y}\big|\mathbf{Y} < \mathbf{C} = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}_2, \qquad (1b)$$

where $\mathbf{Y}$ represents the (n × 1) vector of observed data within the framework of a block design (n = I × J). The matrix $\mathbf{X}$ is a (n × p) matrix for p -1 independent variables and $\mathbf{Z}$ is a (n × r) design matrix for the random effects. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are matrices of parameters for $\mathbf{X}$

and **Z**, respectively. The random effects are assumed to be normally distributed with mean 0 and variance matrix **G**. The joint distribution of $\boldsymbol{\varepsilon_1}$ $\boldsymbol{\varepsilon_2}$ is assumed to be a singular distribution with covariance matrix $\boldsymbol{\sigma_{12}}$. While the random error term ε_1 can be assumed to be normally distributed with mean 0 and variance matrix σ_1^2, we cannot readily assume that $\boldsymbol{\varepsilon_2}$ be normally distributed with mean 0, since the error term in equation (1b) might not be independent of the independent variables, as shown below.

Because $\mathbf{Y^d}$ is not observable, we define a dichotomous factor δ_{it} that indicates the survival status for individual i between time t and t+1 (t = 0, 1, 2, J-1) and is used as a proxy for C, such that

$$\begin{cases} \delta_{it} = 0 \text{ if individual i dies between time t and t+1} \left(Y_{it} \geq C\right) \\ \delta_{it} = 1 \text{ if individual i survives from time t and time t+1} \left(Y_{it} < C\right) \end{cases}$$

Specifically, we view the number of functional limitations at time t+1 as a joint distribution of two sequential events – the likelihood of survival between time t and time t+1 (S_t; t = 0, 1,2, J-1) and the conditional density function on the number of functional limitations (Y_{t+1}) among those who have survived to t+1. Given the aforementioned assumptions, the expected number of functional limitations for individual i at time t+1 can be estimated by the following equation

$$\mathrm{E}\left(\mathbf{Y}_{\mathbf{i(t+1)}} \middle| \mathbf{X}_{2\mathbf{i}}, \boldsymbol{\delta}_{\mathbf{it}} = 1\right) = \Pr\left(\boldsymbol{\delta}_{\mathbf{it}} = 1 \middle| \mathbf{X}_{1\mathbf{i}}\right)\left\{\mathbf{X}_{2\mathbf{i}}\boldsymbol{\beta}_2 + \mathbf{Z}_{2\mathbf{i}}\boldsymbol{\gamma}_2 + \mathbf{E}\left[\boldsymbol{\varepsilon}_{2\mathbf{i}} < \mathbf{C} - \left(\mathbf{X}_{1\mathbf{i}}\boldsymbol{\beta}_1 + \mathbf{Z}_{1\mathbf{i}}\boldsymbol{\gamma}_1\right)\right]\right\}. \quad (2)$$

If ε_2 is independent of ε_1, the conditional mean of ε_2 is 0, and the sample selection process into the incomplete sample is random. However, in many circumstances the conditional mean of the disturbance in the incomplete sample is a function of $\mathbf{X}_{1i}$ and $\mathbf{Z}_{1i}$, as widely reported in mortality and health transitions literature. As a consequence, estimation of equation (2) without considering the covariance between ε_1 and ε_2 would

lead to inconsistent parameter estimates and serious prediction biases. Modeling transitions in the number of functional limitations thereby can be far more complex than applying a one-step mixed model.

Model Specification

We attempt to develop two two-step mixed regression models to overcome the aforementioned selection biases in describing and estimating transitions in the number of functional limitations, one parametric and one nonparametric.

(A) The two-step parametric mixed model.

There are a variety of statistical approaches to estimate survival rates, both continuous and discrete. Since the outcome data in this analysis are discrete in nature, we develop a discrete survival model. As the probit function is well behaved in describing a binomial distribution, we construct a probit mixed model to estimate survival rates between time t and time t+1 (t =0, 1, …., J – 1). For individual i at time t, his or her chance of survival up to time t+1 can be modeled as

$$\Pr\left(\mathbf{Y}_{it} \middle| \delta_{it} = 1\right) = \Phi\left(\mathbf{X}_{it}\beta_{p} + \mathbf{Z}_{it}\gamma_{p}\right)$$
$$t = 1, 2, 3, ..., t - 1, \qquad (3)$$

where Φ(.) represents the cumulative normal distribution function (probit function). From this equation, we can obtain survival rates for each individual at J – 1 time observation intervals. Then we save the estimate of $\Phi(\mathbf{X\beta} + \mathbf{Z\gamma})$ for each individual in each observation period as an unbiased estimate of the survival rate.

Given the assumption on the latent number of functional limitations at time t+1 (t = 0, 1, 2, …., J-1) for those who have died between time t and time t+1 ($\delta_{it} = 0$), a

survivor's number of functional limitations is truncated on the right. The inverse Mills ratio for individual i at time t+1 is therefore given by

$$\lambda_{it} = \frac{\varphi\left(\mathbf{X}_{it}\boldsymbol{\beta}_p + \mathbf{Z}_{it}\boldsymbol{\gamma}_p\right)}{\Phi\left(\mathbf{X}_{it}\boldsymbol{\beta}_p + \mathbf{Z}_{it}\boldsymbol{\gamma}_p\right)} \quad \text{if } \delta_{it} = 1\left(Y < C\right), \qquad (4a)$$

$$\lambda_{it} = -\frac{\varphi\left(\mathbf{X}_{it}\boldsymbol{\beta}_p + \mathbf{Z}_{it}\boldsymbol{\gamma}_p\right)}{1-\Phi\left(\mathbf{X}_{it}\boldsymbol{\beta}_p + \mathbf{Z}_{it}\boldsymbol{\gamma}_p\right)} \quad \text{if } \delta_{it} = 0\left(Y \geq C\right), \qquad (4b)$$

Where φ(.) represents the standard normal density function. Values of λ's at time 0 (first wave) are all zero assuming there is no selection bias at the outset of the longitudinal investigation.

Given the variable **λ** created, we develop a conditionally unbiased truncated mixed model on the number of functional limitations at time t. For individual i, the longitudinal mixed model can be written as

$$\mathbf{Y}\left(\mathbf{Y}|\boldsymbol{\delta}=1\right) = \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{Z}_3\boldsymbol{\gamma}_3 + \boldsymbol{\sigma}_{12}\boldsymbol{\lambda} + \boldsymbol{\varepsilon}_3, \qquad (5)$$

where $\boldsymbol{\sigma}_{12}$ is a vector of covariance between $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$, specified in the estimation process as the regression coefficients of **λ**. The error term $\boldsymbol{\varepsilon}_3$ has mean 0 and variance σ_t^2 and is assumed to be uncorrelated with $\mathbf{X}_3$, **γ** and **λ**.

Notice that in Equation (5), the inclusion of **λ** and **σ** takes into account the covariance between two error terms, $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$, thereby indicating that the joint distribution of two sequential equations, represented by Equation (2), is actually embedded in Equation (5).

(B) The two-stage nonparametric mixed model.

When the assumption of normality for **ε** cannot be satisfied in estimating a longitudinal event, as is often the case in health transitions (Liu, 2000; Manning, Duan

and Rogers, 1987), Equation (5) cannot derive correct estimates of the number of functional limitations. In this case, we extend Duan's (1983) and Liu's (2000) retransformation method in the context of transitions in the number of functional limitations. The two-stage nonparametric mixed model is given by

$$\Pr(Y > 0) = \Phi(\mathbf{X}_1\boldsymbol{\beta}_4 + \mathbf{Z}_2\boldsymbol{\gamma}_4) \quad (6a)$$

$$\mathbf{Y}(\mathbf{Y} > 0) = \log(\mathbf{X}_1\boldsymbol{\beta}_5 + \mathbf{Z}_1\boldsymbol{\gamma}_5 + \boldsymbol{\varepsilon}_5)\boldsymbol{\xi}, \quad (6b)$$

where ξ serves as a nonparametric adjustment factor for selection bias from mortality. We fit the natural logarithm of the number of functional limitations to address the possible non-linearity of its distribution among those with any functional limitations. The expected number of functional limitations at various time points can be expressed by the following joint distribution:

$$\mathbf{E}(\hat{\mathbf{Y}}|\mathbf{S} = 1) = \Phi(\mathbf{X}_1\hat{\boldsymbol{\beta}}_4 + \mathbf{Z}_2\hat{\boldsymbol{\gamma}}_4)\log(\mathbf{X}_1\hat{\boldsymbol{\beta}}_5 + \mathbf{Z}_1\hat{\boldsymbol{\gamma}}_5)\hat{\boldsymbol{\xi}} \quad (7)$$

We use empirical data to estimate values in ξ. First, assuming **X** to have full rank, we have

$$\begin{aligned} \mathrm{E}(\mathbf{Y}|\mathbf{Y} > 0) &= \mathbf{E}[\log(\mathbf{X}_1\boldsymbol{\beta}_5 + \mathbf{Z}_1\boldsymbol{\gamma}_5 + \boldsymbol{\varepsilon}_5)] \\ &= \int [\log(\mathbf{X}_1\boldsymbol{\beta}_5 + \mathbf{Z}_1\boldsymbol{\gamma}_5 + \boldsymbol{\varepsilon}_5)]\mathbf{dF}(\boldsymbol{\varepsilon}_5). \end{aligned} \quad (8)$$

When the error distributional function F is unknown, we replace this cumulative density function F by its empirical estimate $\hat{\mathbf{F}}_j$ at time-point j, which is referred to as the

$$\begin{aligned} \mathbf{E}(\hat{\mathbf{Y}}_j|\mathbf{Y}_j > 0) &= \mathbf{E}\int [\log(\mathbf{X}_{1j}\boldsymbol{\beta}_5 + \mathbf{Z}_{1j}\boldsymbol{\gamma}_5 + \boldsymbol{\varepsilon}_{5j})\mathbf{d}\hat{\mathbf{F}}_{n_j}(\boldsymbol{\varepsilon}_{5j})] \\ &= \frac{1}{\mathbf{n}_j}\sum_{i=1}^{n_j}\log(\mathbf{X}_{1ij}\boldsymbol{\beta}_5 + \mathbf{Z}_{1ij}\boldsymbol{\gamma}_5 + \hat{\boldsymbol{\varepsilon}}_{5ij}) \\ &= \log(\mathbf{X}_{1j}\hat{\boldsymbol{\beta}}_5 + \mathbf{Z}_{1j}\hat{\boldsymbol{\gamma}}_5)\mathbf{n}_j^{-1}\sum_{i=1}^{n_j}\exp(\hat{\boldsymbol{\varepsilon}}_{5ij}), \end{aligned} \quad (9)$$

Where n_j is the number of observations at time j with the number of functional limitations greater than zero, and $\hat{\beta}_5$ and $\hat{\gamma}_5$ can be estimated by employing the maximum likelihood procedure without considering random disturbances (Liu, 2000). When the sample size for a longitudinal analysis is large enough, such a smearing estimate for the retransformation in log-linear equations is consistent, robust and efficient (Duan, 1983; Liu, 2000; Manning et al., 1987).

The estimate of ξ at time-point j is thus given by

$$\xi_j = \frac{\sum_{i=1}^{n_j} \exp\left[\log\left(Y_{ij} \middle| Y_{ij} > 0\right) - \left(X_{1ij}\beta_5 + Z_{1ij}\gamma_5\right)\right]}{n_j}. \qquad (10)$$

Empirical Examples

Data used for empirical demonstrations come from the Survey of Asset and Health Dynamics among the Oldest Old (AHEAD), a nationally representative investigation of older Americans. This survey, conducted by Institute of Social Research (ISR), University of Michigan, is funded by National Institute on Aging as a supplement to the Health and Retirement Study (HRS). To date, the Survey consists of six waves of investigation. The Wave I survey was conducted between October 1993 and April 1994. Specifically, a sample of individuals aged 70 or older (born in 1923 or earlier) was identified throughout the HRS screening of an area probability sample of households in the nation. This procedure identified 9,473 households and 11,965 individuals in the target area range. AHEAD obtains detailed information on a number of domains, including demographics, health status, health care use, housing structure, disability, retirement plans, and health and life insurance. Survival information throughout the six

waves has been obtained by a link to the data of National Death Index (NDI). The present study uses data of all six waves (1993, 1995, 1998, 2000, 2002 and 2004) for analyzing transitions in the number of functional limitations in older Americans.

We measure functional status by a score of activities of daily living (ADL), instrumental activities of daily living (IADL), and other types of functional limitations (Liu, Engel, Kang, & Cowan 2005). A score of one is given to an individual who has any difficulty with a specific physical or social activity, and the number of items for which difficulties are reported is then summed. As a result, the score ranges from 0 (functional independence) to 15 (maximum disability). Covariates include time (0 to 5), veterans status (1 = veteran, 0 = not veteran), age, gender (female = 1), education (years in school), ethnicity (1 = white, 0 = others), marital status (1 = currently married, 0 = other), smoking cigarettes and drinking alcohol, the number of serious health conditions, and self-rated health (5 scales: 1 = poor, 5 = excellent). Because the interval between two adjacent time points is not equally spaced in the AHEAD longitudinal dataset, we use REPEATED/TYPE = SP in executing the SAS PROC.MIXED procedure to represent the autoregressive error structure of the data (Littell et al., 2006).

Table 1 shows four sets of number of functional limitations in older Americans at six time points, 1993, 1995, 1998, 2000, 2002 and 2004, derived from, respectively, observed values and the three types of mixed models. The conventional linear mixed model systematically overestimates the number of functional limitations at every subsequent time point and this overestimation increases with time, compared to the observed values. The parametric two-step mixed model somewhat reduces such overestimation, but the biases still appear considerable and systematic. Lastly, the

nonparametric approach derives the most accurate estimates to describe transitions in the number of functional limitations in older Americans.

<Table 1 about here>

Figure 1 further demonstrates differences in the predicted number of functional limitations among results derived from the three mixed models. In Panel A, there are distinct and systematic separations between the two growth curves. At each time point, the predicted number of functional limitations derived from the conventional mixed model is considerably higher than the corresponding observed value. The predicted growth curve in Panel B, derived from the parametric approach, displays narrowed separations from the observed line; however, the biases remain sizable. In Panel C, separations of the two curves almost disappear, thereby demonstrating the validity and reliability of the nonparametric approach.

<Figure 1 about here>

Discussion

Our analysis demonstrates that direct application of one-step linear mixed models on longitudinal data can be associated with serious prediction biases when the impact of selection bias is strong. We introduce two refined approaches to overcome such biases when applying the mixed model to analyze large-scale longitudinal data of an older person's number of functional limitations. The parametric adjusting approach is an extension of Heckman's traditional two-step model, based on several assumptions about the parametric distribution of selection errors. We show that the parametric approach on the mixed model reduces some of the biases incurred by the misspecification of

disturbances in a one-step mixed model; however, the correction is limited and the biases are still substantial, evidenced by the clear separation between the growth curve generated from this method and the curve from the observed data (see Table 1).

Our nonparametric adjusting method offers a new way to manage selection biases in applying the mixed model to analyze large-scale survey data with high attrition. This approach takes into account the selection biases with appropriate empirical adjustment, minimizing such biases considerably. The convergence of the two growth curves perhaps provides the strongest evidence that because health transitions among older persons are highly selective, the assumption of normality in both mortality and health status is not appropriate in analyzing longitudinal data with high attrition rates.

References

Duan, N. 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* 78, 605-610.

Fu, Vincent Kang. 2004. Sample selection bias models. In Melissa Hardy and Alan Bryman (eds.), *Handbook of Data Analysis*, Pp. 409-430. London: Sage.

Heckman, James J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5:475-492.

Heckman, James J. 1979. Sample selection bias as a specification error. *Annals of Econometrica* 47:153-161.

Liang, Kung-Yee and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrica* 73:13-22.

Littell Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. 2006. *SAS for Mixed Models*. Second edition. Gary, NC: SAS Institute, Inc.

Liu, Xian. 2000. Development of a structural hazard rate model in Sociological research. *Sociological Methods & Research* 29:77-117.

Liu, X., Charles C. Engel, Han Kang, and David Cowan. 2005. The effect of veteran status on mortality among older Americans and its pathways. *Population Research and Policy Review* 24:573-592

Manning, W., Duan, N., & Rogers, W. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 35:59-82.

Winship, Christopher and Robert D. Mare. 1992. Models for sample selection bias. *Annual Reviews of Sociology* 18:327-350.

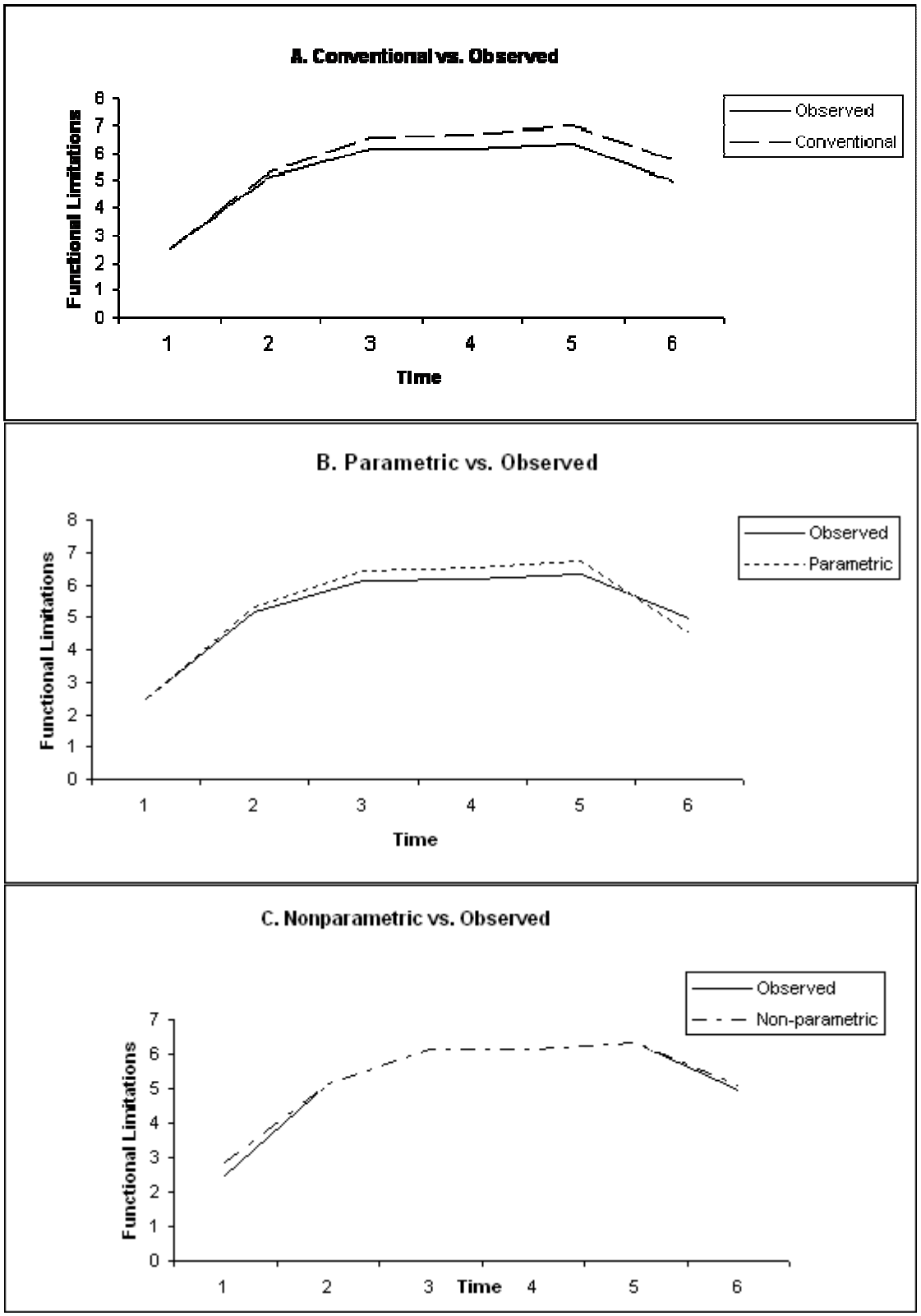
Zeger, Scott L. and Kung-Yee Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121-130.

Predicted Number of Functional Limitations in Older Americans

Derived from three Prediction Approaches (n = 8443)

Time Points and Parameters	Observed and Predicted Number of Functional Limitations			
	Observed	Conventional	Parametric	Nonparametric
1993	2.4887	2.4931	2.4649	2.8638
1995	5.1514	5.3390	5.2996	5.1622
1998	6.1378	6.5476	6.4242	6.1396
2000	6.1602	6.7048	6.5147	6.1609
2002	6.3348	6.9821	6.7546	6.3378
2002	4.9608	5.7510	4.4911	5.1012
Covariance between ε_1 and ε_2			6.5998**	
ξ_1				1.4354**
ξ_2				1.2562**
ξ_3				1.1350**
ξ_4				1.1422**
ξ_5				1.1367**
ξ_6				1.2910**
Model Chi-squares		9800.07**	6450.57**	4790.20**

** $P < 0.01$



Transitions in the Number of Functional Limitations in Older Americans:
Growth Curves derived from three Mixed Models