Modeling Semi-Continuous Outcomes in Longitudinal Studies: Methods and Applications

Jichuan Wang, Ph.D.
Center for Clinic and Community Research (CCCR)
Children's National Medical Center (CNMC)
The George Washington University School of Medicine

Abstract

Very often researchers may encounter continuous outcome measures with a large number of observed values clustered at zero, called a semi-continuous measure. Such an outcome has a mixture of two distributions representing occurrence (non-zero vs. zero values) and intensity (amount of non-zero values) of the outcome measure. As the likelihood of occurrence and the intensity of non-zero values are often correlated with each other, the traditional econometric approach of modeling the two parts of outcome measures separately is inappropriate. The recently developed analytical methods, such as the Mixed-Effect Mixed Distribution Model and the two-part latent growth model (LGM) are often used to model semi-continuous outcome measures. This study demonstrates application of the two-part LGM. The data used for model demonstration were collected in a natural

history study (N=248) of rural stimulant users in rural counties in Western Ohio. Frequency of crack-cocaine use in the past 30 days measured at the baseline and every 6 months in the first two years of study period were used for the model. Both unconditional and conditional two-part LGMs are demonstrated.

1) Introduction

In real research continuous outcome measures are often not normally distributed, but with a large number of observed values clustered at zero (a.k.a., a semi-continuous measure). This problem is traditionally handled by

- Log-transformation: it cannot solve the problem of excess zeros in the measure.

- Recoding data into a dichotomous categorical variable (0 vs. 1): would discard important information.

- Econometric two-part model: it implements a logistic or probit regression to model the probability of having a non-zero value, and a linear regression to model the non-zero values, assuming two separate or unconnected models (Manning, Duan, & Rogers, 1987; Wanning, Duan, & Rogers, 1987; Duan, Manning, Morris, & Newhouse, 1983; Olsen & Schafer, 2001).

Appropriate methods have been developed to deal with the extra zeros in the semi-continuous outcome:

- The mixed-effect mixed distribution model : Tooze et al. (2002) have

proposed the *Mixed-Effect Mixed Distribution Model* and developed a SAS program to fit semi-continuous outcome in longitudinal data.

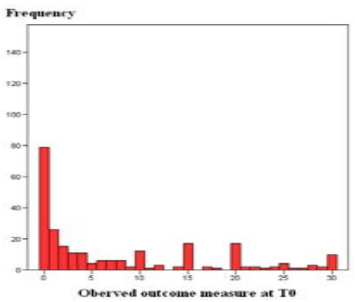
- Two-part latent growth model (LGM): The original distribution of outcome measure is decomposed into two parts (*likelihood* and *amount*) that are considered as two associated growth processes and are modeled simultaneously (Brown et al., 2005; Olsen & Schafer, 2001).

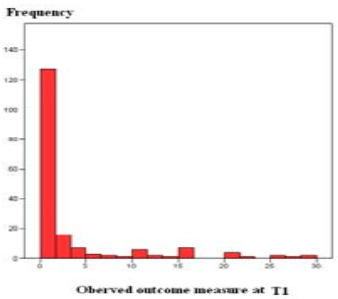
2) Application of the two-part LGM

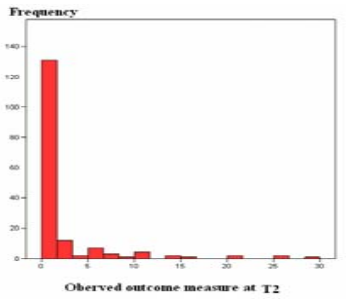
Data: The data used for model demonstration were collected in a natural history study ($N = 249$) of stimulant users in rural counties in Western Ohio between October 2002 and September 2004 (Siegal et al., 2006). Frequency of crack cocaine use in the past 30 days measured at the baseline and every

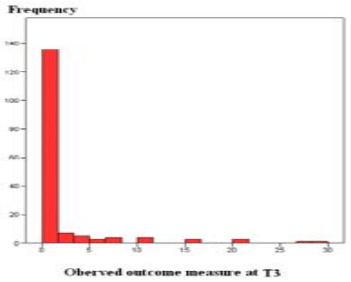
6 months in the first two years of study period were used as five repeated measures of the outcome.

The frequency distributions of the outcome measures are shown in Figure 1. The zero frequency in the measure is very large at each time point, indicating the outcome measures are semi-continuous variables. As such, the two-part LGM is appropriate for modeling such longitudinal data.









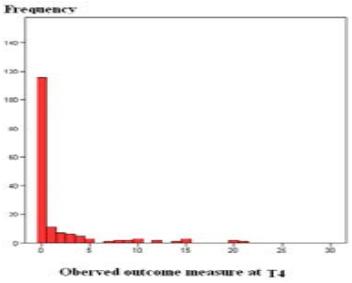


Figure 1. Frequency of crack-cocaine use in the past 30 days over time.

Unconditional Two-part LGM:

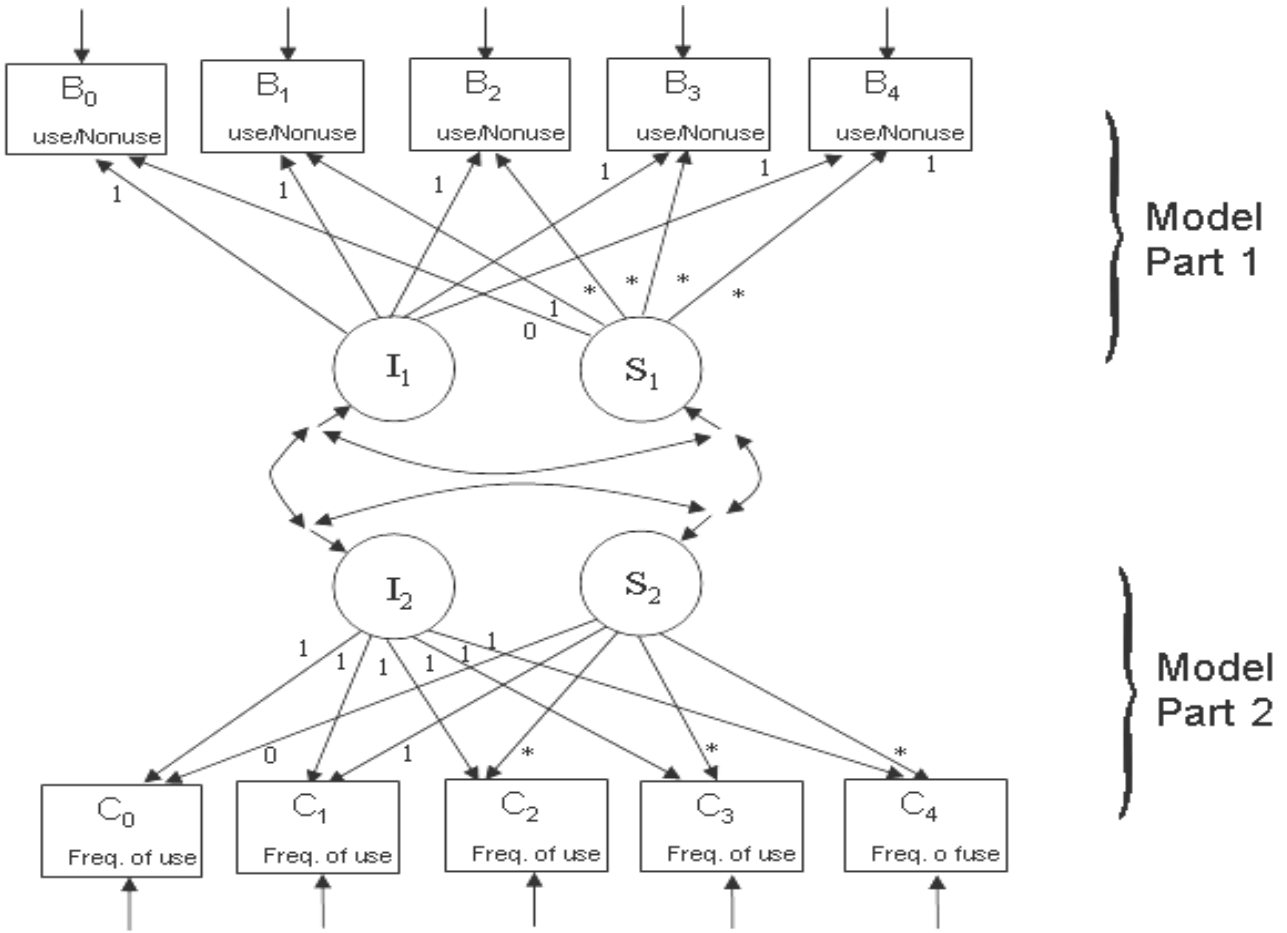


Figure 2. Two-part LGM

In Part 1 of the model:

"*No use*" of crack cocaine was separated from the distribution of the observed continuous outcome measure (i.e., number of days used crack cocaine in the past 30 days), and new binary outcome variables B_0, B_1, B_2, B_3, and B_4 are created to represent ever used crack cocaine in the past 30 days prior to each interview (1− Used crack; 0 − No use).

In Part 2 of the model:

New continuous outcome variables C_0, C_1, C_2, C_3, and C_4 are created to represent frequency of crack cocaine use only among those who had ever used crack in the past 30 days prior to each interview. Those who did not use crack cocaine in the past 30 days were treated as missing cases in the *C*

variables. The LGM with the *B* variables in Part 1 model and the LGM with the *C* variables in Part 2 model are estimated simultaneously. Associations or causal relationships between the latent growth factors in the two models can be specified.

For Part 1 model: Significant slope growth factor *ETA1B* (−0.790, p = 0.000) indicates that the likelihood of reporting ever used crack in the past 30 days significantly declined over time.

Part 2 model: to study the frequency of crack use among those who used crack in the past 30 days. Significant slope growth factor *ETA1B* (−0.790, p = 0.000) indicates that the likelihood of reporting ever used crack in the past 30 days significantly declined over time. In addition, the frequency of crack use among.

For two-part LGM, Mplus does not provide the familiar model fit indices like comparative fit index (*CFI*), Tucker-Lewis fit index (*TLI*), root mean square error of approximation (*RMSEA*), and standardized root mean square residual (*SRMR*). In stead, Mplus Output only shows the loglikelihood and information criteria for the overall model that can be used for model comparisons. *Pearson Chi-Square* and the *Likelihood Ratio* (*LR*) *Chi-Square* statistics are also provided for the binary outcome measures. These two *Chi-square* statistics are supposed to agree with each other. Otherwise neither of them is trustable. The *Pearson* and the *LR Chi-Squares* in this example are not close to each other (the model results are not reported here) likely due to a large number of zero cells in the contingency table of the binary outcome measures.

Selected results of the unconditional two-part LGM (Table 1):

Table 1. Selected Model Results of Unconditional Two-part LGM

ETA0B WITH				
ETA1B	0.000	0.000	999.000	999.000
ETA0C WITH				
ETA0B	0.774	0.308	2.514	0.012
ETA1B	-0.085	0.116	-0.735	0.463
ETA1C WITH				
ETA0B	0.004	0.114	0.038	0.970
ETA1B	0.185	0.066	2.823	0.005
ETA0C	-0.078	0.068	-1.153	0.249
Means				
ETA0B	0.000	0.000	999.000	999.000
ETA1B	-0.790	0.117	-6.747	0.000
ETA0C	1.597	0.119	13.463	0.000
ETA1C	-0.397	0.072	-5.501	0.000

Conditional Two-part LGM:

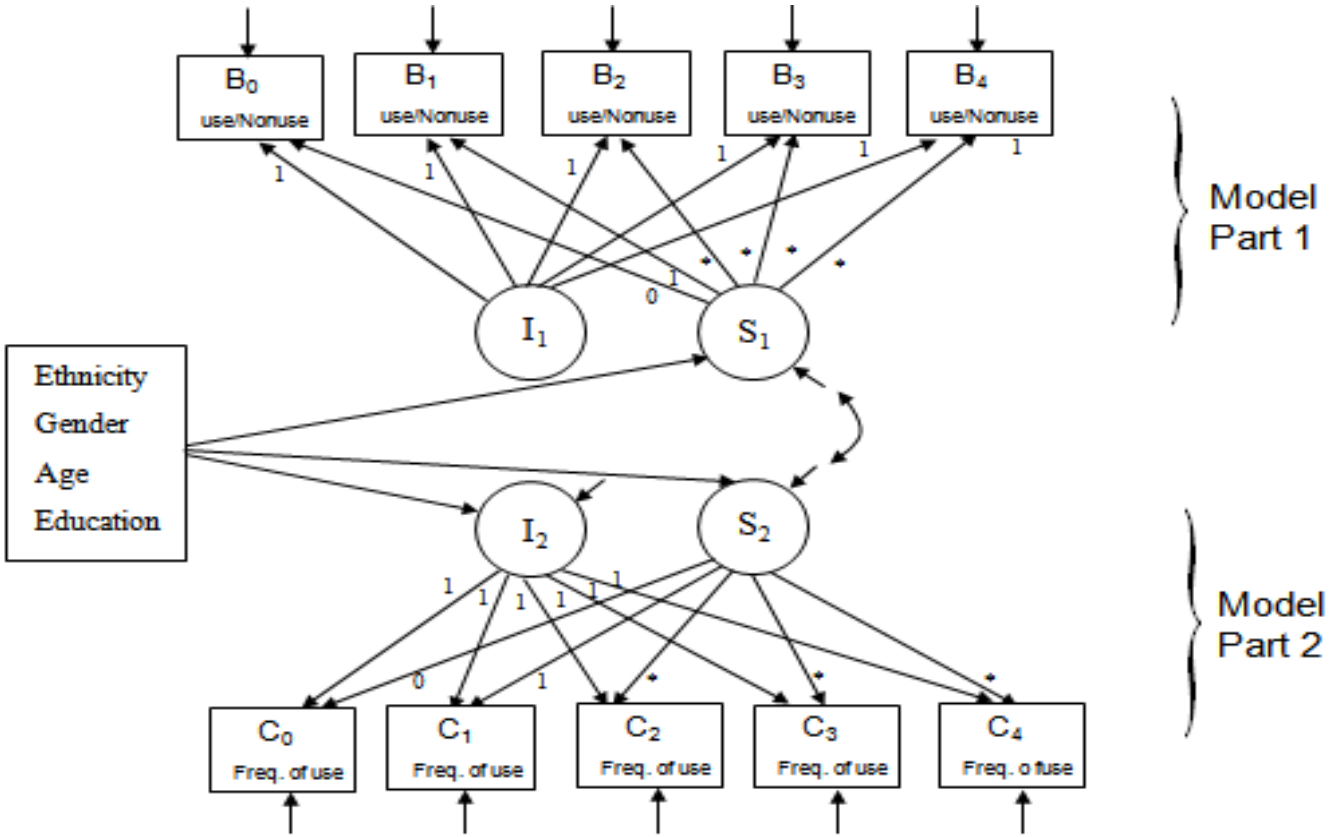


Figure 3. Conditional Two-Part LGM

Table 2. Selected Model Results of Conditional Two-part LGM

ETA1B	ON			
GENDER	-0.124	0.161	-0.770	0.441
WHITE	-0.712	0.268	-2.656	0.008
AGE	0.014	0.009	1.490	0.136
EDUC	-0.131	0.124	-1.057	0.290
ETA0C	ON			
GENDER	-0.349	0.179	-1.951	0.051
WHITE	0.188	0.220	0.852	0.394
AGE	0.011	0.010	1.166	0.244
EDUC	-0.073	0.142	-0.514	0.607
ETA1C	ON			
GENDER	0.068	0.084	0.806	0.420
WHITE	-0.233	0.109	-2.147	0.032
AGE	0.004	0.005	0.822	0.411
EDUC	0.050	0.066	0.752	0.452
ETA0C	WITH			
ETA1B	0.104	0.091	1.144	0.253
ETA1C	WITH			
ETA1B	0.153	0.050	3.057	0.002
ETA0C	-0.051	0.073	-0.704	0.481

The model results show that gender, ethnicity, age, and education did not have significant effects on initial level of crack cocaine use frequency at baseline (i.e., no effects on the intercept growth factor *ETA0C*). However, Whites had a significant negative effect (-0.233, p=0.032) on the slope growth factor *ETA1C*), indicating that Whites reported lower frequency of crack cocaine use over time. In addition, it seems that Whites were also less likely to use crack cocaine over time (the effect White on the latent slope growth factor ETA1B is -0.712 (p=0.008)).

References

Brown, E. C., Catalano, C. B., Fleming, C. B., Haggerty, K. P. & Abbot, R. D. 2005. Adolescent substance use outcomes in the Raising Healthy Children Project: A two-part latent growth curve analysis. *Journal of Consulting and Clinical Psychology*, *73*, 699-710.

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. 1983. A comparison of alternative models for the demand for medical care. *Journal of Economic and Business Statistics*, *1*, 115–126.

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

Muthén, L. & Muthén, B. 1998-2008. *Mplus User's Guide*. Los Angeles, CA: Muthén & & Muthén.

Olsen, M. K., & Schafer, J. L. 2001. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association, 96,* 730-745.

Siegal, H. A., Draus, P. J., Carlson, R. G., Falck, R. S., & Wang, J. 2006. Perspectives on health among adult users of illicit stimulant drugs in rural Ohio. *Journal of Rural Health, 22*, 169-173.

Tooze, A. J., G. K. Grunwald & R.H. Jones. 2002. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research.* Vol. 11: 341-355.

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