Estimating Treatment Effects in the Presence of Correlated Binary Outcomes and Contemporaneous Selection

Matthew P. Rabbitt*

Economic Research Service
U.S. Department of Agriculture

2017 Stata Conference
July 27-28, 2017

*The views expressed in this presentation are those of the author and do not necessarily reflect those of the Economic Research Service or the U.S. Department of Agriculture.

Outline

- Motivation and Background
- An Illustrative Model of Correlated Logistic Outcomes with Contemporaneous Selection
- Useful Average Treatment Effect (ATE) Forumations for Causal Inference with Correlated Logistic Outcomes
- ETXTLOGIT Command
- GSEM Reparameterization of Model for Estimation
- Monte Carlo Experiment
- Empircal Example: SNAP benefit receipt and children's food insecurity
- Next Steps

Motivation and Background

- Correlated binary outcomes are commonly encountered by researchers in the social sciences.
  - Longitudinal models (e.g., random effects logistic regression.)
  - Two-level or random-intercept models (e.g., random intercept logistic regression.)
  - Hazard and survival models (e.g., discrete-time logistic model.)
  - Seemingly unrelated regression (SUR) models (e.g., SUR logistic regression.)
  - Item Response Theory (IRT) models (e.g., 1-PL (Rasch) logistic IRT model.)
- Example applications of these models include health, demography, economics, and education topics among others.

Motivation and Background

- Causal inference with correlated binary outcomes is challenging because individual's often self select into the treatment group
- Methodological approaches to addressing self-selection bias with correlated binary outcomes
  - Longitudinal instrumental variables models (e.g, two-stage least square for longitudinal models.)
    - May lead to nonsensical predictions that affect inference because of unbounded probabilities (particularly important with behaviors that have probabilities close to 0 or 1)
  - IRT models (e.g., two-stage least squares or other methodolgy using summary measures of latent trait.)
    - Summary measures may lead to different analysis samples and are less efficient (Rabbitt,2017; Christensen,2006)

Illustrative Model of Correlated Logistic Outcomes

Item Reponse Theory (IRT) Measurement Model

- 1-PL Logistic (Rasch, 1960/1980) Model

$$Y_{ij}^{*} = \theta_i + \nu_{i_j}$$

- Key model assumptions
  1. Error in responses (ν_{ij}) is distributed according to a Extreme Value Type 1 (EV1) distribution
     $P\left(Y_{ij} = 1 \mid \theta_i, \delta_j\right) = \frac{\exp\left(\theta_i - \delta_j\right)}{1+\exp\left(\theta_i - \delta_j\right)}, j = 1, ..., J; i = 1, ..., N$
  2. Conditional independence
     $P\left(Y_{ij} = y_i \mid \theta_i, \delta_j\right) = \prod_{j=1}^{J} \frac{\exp\left(q_{ij}\left(\theta_i - \delta_j\right)\right)}{1+\exp\left(q_{ij}\left(\theta_i - \delta_j\right)\right)}$,where
     $q_{ij} = 2Y_{ij} - 1$

Illustrative Model of Correlated Logistic Outcomes

The Explanatory Model (De Boeck and Wilson, 2004)

- Explanatory variables (e.g., person-level characteristics) may be incorporated into the model by assuming

$$\theta_i = \beta_T T_i + \beta_X^{'} X_I + e_i,$$

where T_i is a treatment indicator, X_i is a matrix of control variables, and $e_i \sim N\left(0, \sigma^2\right)$.

- The probabiltiy of observing the response vector for person i is

$$P\left(Y_{ij} = y_i \mid \theta_i, \delta_j, e_i\right) = \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp(q_{ij}(\theta_i - \delta_j))}{1 + \exp(q_{ij}(\theta_i - \delta_j))} \frac{1}{\sigma} \phi\left(\frac{e_i}{\sigma}\right) de_i,$$

where ϕ is the standard normal pdf.

Illustrative Model of Correlated Logistic Outcomes

Explanatory 1-PL (Rasch) Selection Model (Rabbitt, 2014)

- Treatment participation decision

$$T_i = I\left(\alpha_X' X_i + \alpha_Z' Z_i + u_i > 0\right)$$

where $u_i \sim N\left(0, 1\right)$.

- Following Terza(2009), I assume the error component, e_i, may be respecified as $e_i = \lambda u_i + e_i^*$, so

$$\theta_i^* = \beta_T T_i + \beta_X' X_I + \lambda u_i + e_i,$$

where $e_i^* \sim N\left(0, \eta^2\right)$.

Illustrative Model of Correlated Logistic Outcomes

Explanatory 1-PL (Rasch) Selection Model (Rabbitt, 2014)

- Likelihood function

$$L = \prod_{i=1}^{N} T_i \int_{-\alpha'_X X_i - \alpha'_Z Z_i}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp(q_{ij}(\theta_i^* - \delta_j))}{1+\exp(q_{ij}(\theta_i^* - \delta_j))} \frac{1}{\eta} \phi\left(\frac{e_i^*}{\eta}\right) de_u^* \phi\left(u_i\right) du_i +$$

$$(1 - T_i) \int_{-\infty}^{-\alpha'_X X_i - \alpha'_Z Z_i} \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp(q_{ij}(\theta_i^* - \delta_j))}{1+\exp(q_{ij}(\theta_i^* - \delta_j))} \frac{1}{\eta} \phi\left(\frac{e_i^*}{\eta}\right) de_u^* \phi\left(u_i\right) du_i$$

Illustrative Model of Correlated Logistic Outcomes

Explanatory 1-PL (Rasch) Selection Model (Rabbitt, 2014)

- Reparmeterized Likelihood function

$$L = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi \left(q_{ij} \left(\alpha'_X X_i + \alpha'_Z Z_i + \lambda u_i \right) \right) \prod_{j=1}^{J} \frac{\exp(q_{ij}(\theta_i^* - \delta_j))}{1 + \exp(q_{ij}(\theta_i^* - \delta_j))} \frac{1}{\eta} \phi \left(\frac{e_i^*}{\eta} \right) de_u^* \phi$$

- For more details on the reparmeterization, see Skrondal and Rabe-Hesketh (2004).

Useful Average Treatment Effect Formulations

- The ATE will depend on the model and substantive knowledge of the behavior being analyzed. For example, when estimating an explantory IRT model the researcher may want to examine how a treatment affects the probabiltiy of an individual's latent ability falling in a specific range on the latent continuum.

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[P\left(Y_i > \tau \mid T_i = 1, X_i, u_i, e_i^*\right) - P\left(Y_i > \tau \mid T_i = 0, X_i, u_i, e_i^*\right)\right]\frac{1}{\eta}\phi\left(\frac{e_i^*}{\eta}\right) de_u^* \phi\left(u_i\right) du$$

- Alternatively, one may be interested in an ATE for each item, ATE_j.

ETXTLOGIT Command Syntax and Options

- Command syntax
  - etxtlogit *depvar*$_1$ *varlist*$_1$ (*depvar*$_2$= *varlist*$_2$) [*if*] [*in*] [*weight*], id(*varlist*) intpoints1(*integer 12*) intpoints2(*integer 12*)
- Options
  - noconstant suppresses the constant in the outcome equation.
  - from(*matname*) specifies starting values for estimation.
  - vce(*vcetype*) specifies the variance-covariance matrix is obtained by oim or opg.
  - lcon(*string*) constrains the selection parameter, λ, to a specific value.
  - gradient results in the display of the gradient.

ETXTLOGIT Command Output

```
Endog Treat. Random-Effects Logistic Regression Number of obs      =     15000
Group variable: id                              Number of groups   =      5000

Random effects e_i ~ Gaussian                   Obs per group: min =         3
Random effects u_i ~ Gaussian                                  avg =       3.0
                                                               max =         3

Integration method 1: mvghermite                Integration points =        15
Integration method 2: mvgsteen                  Integration points =        15

Log likelihood = -11846.208
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
s            |
           x |    1.01636   .0639408    15.90   0.000     .8910385    1.141682
           z |   1.134807   .0635548    17.86   0.000     1.010241    1.259372
       _cons |  -1.066662   .0500314   -21.32   0.000    -1.164722   -.9686027
-------------+----------------------------------------------------------------
y            |
           s |  -.6825051   .2652765    -2.57   0.010    -1.202437   -.1625728
           x |   .9411961   .1587848     5.93   0.000     .6299836    1.252408
         Th1 |   .6564859   .1120284     5.86   0.000     .4369142    .8760576
         Th2 |   1.246197   .1135879    10.97   0.000     1.023569    1.468825
         Th3 |   1.733079   .1154958    15.01   0.000     1.506712    1.959447
-------------+----------------------------------------------------------------
    /lnsig2u |   1.050815   .0689696    15.24   0.000     .9156372    1.185993
      lambda |   .7642504   .1690593     4.52   0.000     .4329003    1.095601
-------------+----------------------------------------------------------------
     sigma_u |   1.691148   .0583189                      1.580622    1.809402
         rho |   .2250801    .083162                      .0620856    .3880747
------------------------------------------------------------------------------
Likelihood-ratio test of lambda = 0:  chi2(1) =    20.56  Prob >= chi2 =  0.000

Instrumented:  s
Instruments:   x z
```

GSEM: An Alternative Estimation Approach for the Explanatory 1-PL (Rasch) Selection Model

- Command syntax
  - gsem (*depvar*$_{11}$ *depvar*$_{12}$ ... *depvar*$_{1J}$ <- *varlist*$_1$@myvarlist RE[*id*]@1 U@myU, logit) (*depvar*$_2$ <- *varlist*$_2$ U@myU, probit), var(U@1)
- Options
  - All command options are described in detail in the GSEM Stata documentation.

Monte Carlo Experiment

Data Generating Procedure

- Data for each experiment were generated according to the following assumptions.
  - Exogenous variables
    $X_i \sim U\left(0,1\right]$
    $Z_i \sim U\left(0,1\right]$
  - Endogenous variables
    $T_i^* = I\left(\alpha_X X_i + \alpha_Z Z_i + u_i > 0\right); u_i \sim N\left(0,1\right)$
    $Y_{ij} = \frac{\exp\left(\beta_T T_i + \beta_X X_i + \lambda u_i + e_i^* - \delta_j\right)}{1 + \exp\left(\beta_T T_i + \beta_X X_i + \lambda u_i + e_i^* - \delta_j\right)}; e_i^* \sim N\left(0, \eta^2\right)$

Monte Carlo Experiment

Table 1. Bias and RMSE for the person-level, variance, and selection parameters from the BRSM estimated using ETXTLOGIT and GSEM

		ETXTLOGIT		GSEM	
Parameter	True Value	Bias	RMSE	Bias	RMSE
β_T	-1.000	0.015	0.300	0.015	0.300
β_X	1.000	-0.009	0.175	-0.009	0.175
δ_1	0.500	0.003	0.123	0.003	0.123
δ_2	1.000	0.001	0.125	0.001	0.125
δ_3	1.500	-0.003	0.125	-0.002	0.125
λ	$1,000$	-0.007	0.191	0.265	0.319
η^2	2.718	-0.007	0.222	-0.615	0.671

Note: Calculations based on 1,000 replications of ETXTLOGIT and GSEM applied to simulated data of 5,000 individuals and 3 items.

Empirical Example

Table 2. Estimates of the effect of SNAP receipt on children's food insecurity

Variable	XTLOGIT	ETXTLOGIT
SNAP receipt, last 12 months	1.511^{***}	-1.186^{**}
	(0.184)	(0.597)
	[0.029]	[−0.038]
	[0.037]	[−0.037]
λ	–	1.613^{***}
	(–)	(0.352)
ρ	–	0.611
Log-likelihood	−6,427.548	−8,603.340
Time to convergence (min)	6.473	96.420

Note: Unweighted estimation was completed using a random sample of 5,000 low-income households with children from the 2001-2008 CPS-FSS.

Practical Considerations and Hints

- Exogenous models, estimated using **XTLOGIT**, may be more practical for initial model develpment
  - **XTLOGIT** may be utilized to determine the set of control variables
  - **quadchk** is useful for ensuring the numerical methods for this part of the full model have converged
- The the **lcon** option can be used to conduct a grid search over the most troublesome parameter, λ, to assess convergence
- **ETXTLOGIT** provides a likelihood-ratio (LR) test of the endogenous vs. exogenous models
- **GSEM** estimation approach may be preferred to **ETXTLOGIT** in some applications because of the computational burden; however, **ETXTLOGIT** appears to have an advantage in more complex model specifications

Next Steps

- Continue implementation of ETXTLOGIT options and certification tests
- Implement the analytic Hessian
- Implement postestimation options
  - predict $\left(e.g., P\left(Y_{ij} = 1 \mid \theta_i, \delta_j\right)\right)$
  - ATE estimation

Contact Information

Thank you!

For comments, questions, or suggestions:

Matthew P. Rabbitt

matthew.rabbitt@ers.usda.gov

(202)-694-5593

References

- Christensen, K.B. (2006). "From Rasch Scores to Regression." Journal of Applied Measurement, 7(2), 184-191.
- De Boeck, P., and Wilson, M. (2004). Descriptive and Explanatory Item Response Models. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach, 43-74.
- Rabbitt, M. P. (2017). Causal Inference with Latent Variables from the Rasch Model as Outcomes. Unpublished Manuscript.
- Rabbitt, M. P. (2014). Measuring the Effect of Supplemental Nutrition Assistance Program Participation on Food Insecurity Using a Behavioral Rasch Selection Model. Unpublished Manuscript. Greensboro: University of North Carolina.

References

- Rasch, G. (1960/1980). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).
- Skrondal, A., and Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. CRC Press.
- Terza, J. V. (2009). Parametric Nonlinear Regression with Endogenous Switching. Econometric Reviews, 28(6), 555-580.