Modeling Longitudinal Count Data with Excess Zeros and Time-Dependent Covariates: Application to Drug Use

Trent L. Lalonde

University of Northern Colorado

November 17, 2014

UNIVERSITY of NORTHERN COLORADO

Presentation Outline

I EMA Example and Data Issues

II Correlated Count Regression Models

III Time-Dependent Covariate Estimation

IV Models for Correlated Counts with Excess Zeros

V Hurdle Generalized Method of Moments

VI Example Data Analysis: EMA

EMA Example and Data Issues

Motivating Example

EMA: Ecological Momentary Assessment

- Interest: Honest Reporting of Marijuana Usage
- Connection to Craving, Motivation, Social Context

Motivating Example

EMA: Ecological Momentary Assessment

- Interest: Honest Reporting of Marijuana Usage
- Connection to Craving, Motivation, Social Context
- Subjects recruited based on usage history, physiological testing
- Respond to text messages 3 times per day for 14 consecutive days.

Properties of the Data

EMA Data:

- Count Response Variable
- Longitudinal Responses
- Excess Zeros Expected (And Observed)
- Predictors Change Over Time

Correlated Count Regression Models

Ordinary Count Regression Models

Poisson regression:

Random Component: Poisson Distribution

$$Y_i \sim \mathrm{Poi}(\lambda(\mathbf{x}_i))$$

Systematic and Link Components: Log Link

$$\ln(\lambda(\mathbf{x}_i)) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Ordinary Count Regression Models

Parameter Estimation typically proceeds using Maximum Likelihood (implemented using Iterative Re-Weighted Least Squares)

$$l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N} \left[y_i \ln(\lambda(\boldsymbol{\beta}; \mathbf{x}_i)) - \lambda(\boldsymbol{\beta}; \mathbf{x}_i) - \ln(y_i!) \right]$$

Hypothesis Testing performed using Wald Statistics

Implicitly assumes $\mathrm{Var}(Y_i) = \mathrm{E}[Y_i]$

Overdispersion

When $\mathrm{Var}(Y_i) > \mathrm{E}[Y_i]$ the data are overdispersed

Positive autocorrelation in the data leads to overdispersion

Consequence: Inflation of Type I Error Rate

Correlated Count Regression Models

Accounting for autocorrelation in count responses:

Conditional Models: Include random subjects effects

Subject-Specific Interpretations

Correlated Count Regression Models

Accounting for autocorrelation in count responses:

Conditional Models: Include random subjects effects

Subject-Specific Interpretations

Marginal Models: Determine marginal moments directly

Population-Averaged Interpretations

Conditional Correlated Count Regression

Mixed Poisson Count Regression:

Random Component: Poisson Distribution / Gamma Random Effect

$$Y_{it} | u_i \sim \mathrm{Poi}(\lambda(\mathbf{x}_{it}, \mathbf{z}_{it}))$$

$$u_i \sim Gamma(\alpha, \beta)$$

Systematic and Link Components: Log Link, Random Effects Design

$$\ln(\lambda(\mathbf{x}_{it}, \mathbf{z}_{it})) = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{v}(\mathbf{u})$$

Conditional Correlated Count Regression

Parameter Estimation:

Maximum Likelihood, h-Likelihood, Markov Chain Monte Carlo, EM Algorithm

The above model is often referred to as a **Random Intercept** model. Sometimes **Random Slopes** models are applied, adding columns from $\mathbf{X}$ to $\mathbf{Z}$.

Marginal Correlated Count Regression

Marginal Correlated Poisson Count Regression:

Random Component: Mean and Variance Specified

$$Y_{it} \sim \mathcal{D}\left(\lambda(\mathbf{x}_{it}), \phi V(\lambda(\mathbf{x}_{it}))\right)$$

The marginal model is specified through the mean and variance structure, as defining a quasi-likelihood.

Marginal Correlated Count Regression

The mean is specified through the link and systematic components:

$$\ln(\lambda(\mathbf{x}_{it})) = \mathbf{x}_{it}^T \boldsymbol{\beta}$$

The variance-covariance structure is specified directly:

$$V(\lambda(\mathbf{x}_{it})) = \mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2}$$

Marginal Correlated Count Regression

Estimate parameters by solving estimating equations:

$$\sum_{i=1}^{N} \left(\frac{\partial \lambda(\boldsymbol{\beta}; \mathbf{x}_i)}{\partial \boldsymbol{\beta}} \right)^T \left[\phi \mathbf{V}_i(\lambda(\boldsymbol{\beta}; \mathbf{x}_i)) \right]^{-1} \left(\mathbf{Y}_i - \lambda(\boldsymbol{\beta}; \mathbf{x}_i) \right) = \mathbf{0}$$

- Dispersion parameters estimated similarly, using GEE2
- Test hypotheses using Sandwich Wald Tests / Generalized Score Tests

Time-Dependent Covariate Estimation

Time-Dependent Covariates

Predictors that include variation both between and within subjects

- Exogenous versus Endogenous
- External versus Internal
- Also "Time-Varying Covariates" or "Within-Subjects Covariates"

Time-Dependent Covariate Models

Consider three approaches:

1. **Conditional Models**: Mixed Correlated Count Regression
2. **Marginal Models**: GEE for Count Regression
3. **Marginal Models**: GMM for Count Regression

Conditional Time-Dependent Covariate Models

Directly split coefficients into “within” and “between” components (Neuhaus and Kalbfleisch (1998)).

Traditional Mixed Poisson:

$$\ln(\lambda(x_{it}, \mathbf{z}_{it})) = \beta_0 + \beta_1 x_{it} + \mathbf{z}_{it}^T \mathbf{v}(\mathbf{u})$$

Conditional Time-Dependent Covariate Models

Directly split coefficients into "within" and "between" components (Neuhaus and Kalbfleisch (1998)).

Traditional Mixed Poisson:

$$\ln(\lambda(x_{it}, \mathbf{z}_{it})) = \beta_0 + \beta_1 x_{it} + \mathbf{z}_{it}^T \mathbf{v}(\mathbf{u})$$

Mixed Poisson with TDC Decomposition:

$$\ln(\lambda(x_{it}, \mathbf{z}_{it})) = \beta_0 + \beta_{1B}\bar{x}_{i.} + \beta_{1W}(x_{it} - \bar{x}_{i.}) + \mathbf{z}_{it}^T \mathbf{v}(\mathbf{u})$$

Conditional Time-Dependent Covariate Models

Coefficient interpretations:

β_B represents the expected effect on the (transformed) response mean from changes across individuals

β_W represents the expected effect on the (transformed) response mean from changes within individuals

Marginal Time-Dependent Covariate Models: GEE

GEE Approach: For longitudinal data, Pepe and Anderson (1994) argued

- Use a **diagonal working correlation** structure

  or

- Verify the sufficient condition:

$$\mathrm{E}[Y_{it}|X_{it}] = \mathrm{E}[Y_{it}|X_{ij}, j = 1, \dots, T]$$

Either will guarantee the expectation of the GEE is the **zero vector**

Marginal Time-Dependent Covariate Models: GEE

GEE Approach:

- Use of “Independent” working correlation structure recommended
- Fitzmaurice (1995) noted losses in efficiency with this approach
- Efficiency depends on **strength of autocorrelation**

Marginal Time-Dependent Covariate Models: GMM

Generalized Method of Moments Approach:

- Lai and Small (2007) proposed a method of avoiding diagonal working correlation structures
- Idea: Select combinations of derivative and residual terms **without** a working correlation structure
- Ensure that expectation is zero, depending on nature of time-dependent covariate

Marginal Time-Dependent Covariate Models

GMM Process: Minimum Quadratic Form estimation

Minimize

$$Q(\beta) = \mathbf{G}^T(\beta; \mathbf{Y}, \mathbf{X})\mathbf{W}^{-1}\mathbf{G}(\beta; \mathbf{Y}, \mathbf{X})$$

Where $\mathbf{G}(\beta; \mathbf{Y}, \mathbf{X})$ is an average vector of **valid moment conditions** constructed according to the type of TDC such that

$$\mathrm{E}[\mathbf{G}(\beta; \mathbf{Y}, \mathbf{X})] = \mathbf{0}$$

Models for Correlated Counts with Excess Zeros

Excess-Zero Count Model Options

Hurdle Poisson:

- "Certain Zero" comes from one process
- Once "hurdle" is cleared, responses are positive

Zero-Inflated Poisson:

- "Zero" comes from two processes
- Either "Certain Zero" or part of Poisson process

Excess-Zero Correlated Count Model Options

Correlated Count Model Options:

1. **Conditional Models**: Mixed Hurdle
2. **Conditional Models**: Mixed ZIP
3. **Marginal Models**: Hurdle GEE

Conditional Model: Mixed Hurdle Poisson Model

Mixed Hurdle Poisson Distributional Component

$$Y_{ij}|u_i \sim HurP(\pi(\mathbf{x}_{it}, \mathbf{z}_{it}; u_i), \lambda(\mathbf{x}_{it}, \mathbf{z}_{it}; u_i))$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2)$$

$$f_{ij}(y_{ij}|u_i; \pi_{it}, \lambda_{it}) = \begin{cases} \pi_{it} & y_{ij} = 0 \\ (1-\pi_{it})\frac{f(y_{ij}|u_i;\lambda_{it})}{1-f(0;\lambda_{it})} & y_{ij} > 0 \end{cases}$$

Conditional Model: Mixed ZIP Model

Mixed Zero-Inflated Poisson Distributional Component

$$Y_{ij}|u_i \sim ZIP(\pi(\mathbf{x}_{it}, \mathbf{z}_{it}; u_i), \lambda(\mathbf{x}_{it}, \mathbf{z}_{it}; u_i))$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2)$$

$$f_{ij}(y_{ij}|u_i; \pi_{it}, \lambda_{it}) = \begin{cases} \pi_{it} + (1 - \pi_{it})f(0; \lambda_{it}) & y_{ij} = 0 \\ (1 - \pi_{it})f(y_{ij}|u_i; \lambda_{it}) & y_{ij} > 0 \end{cases}$$

Conditional Model: Mixed ZIP Model

Mixed Hurdle / ZIP Systematic Components

$$\text{logit}(\pi_{it}) = \mathbf{x}_{l,it}\boldsymbol{\alpha} + \mathbf{z}_{it}\mathbf{u}$$

$$\ln(\lambda_{it}) = \mathbf{x}_{c,it}\boldsymbol{\beta} + \mathbf{z}_{it}\mathbf{u}$$

Conditional Modeling

Mixed Hurdle and ZIP Models:

- Estimation proceeds using likelihood methods (MCMC)
- Time-dependent covariates can again be split into “within” and “between” effects
- Models show high Type I Error rates in the presence of time-dependent covariates

Marginal Modeling: Hurdle GEE

GEE for "zero-inflation" presented by Dobbie and Welsch (2001)

Construct two response vectors:

- Binary: "Certain Zero" Indicator:

$$Y_{bin,it} \sim \mathcal{D}(\pi_{it}, \pi_{it}(1-\pi_{it}))$$

- Count: "Positive" counts with Positive Poisson Moments:

$$Y_{it}|(y_{it} > 0) \sim \mathcal{D}\left(\mu(\lambda_{it}), V(\lambda_{it})\right)$$

Marginal Modeling: Hurdle GEE

Hurdle GEE: Construct two models:

- Binary Response:

$$\text{logit}\left(\pi(\mathbf{z}_{it})\right) = \mathbf{z}_{it}^T \boldsymbol{\alpha}$$

- Positive Count Response:

$$\ln\left(\lambda(\mathbf{x}_{it})\right) = \mathbf{x}_{it}^T \boldsymbol{\beta}$$

Marginal Modeling: Hurdle GEE

Hurdle GEE: Solve two estimating equations:

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\alpha}} \right) \mathbf{V}_{l,i}^{-1} (\mathbf{y}_{bin} - \boldsymbol{\pi}_i) = \mathbf{0}$$

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right) \mathbf{V}_{c,i}^{-1} \left(\mathbf{I}_{(y_i > 0)} \right) (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

(Use Independent Working Correlation Structure for Time-Dependent Covariates)

Hurdle Generalized Method of Moments

Hurdle GMM

Why not use existing methods?

- Need to account for autocorrelation, excess zeros, time-dependent covariates
- Other methods have a single treatment for all Time-Dependent Covariates
- Marginal method (Independent GEE) imposes independence assumption, with consequences of lost efficiency

Hurdle GMM: Model

Joint Quasi-Generalized Linear Model: Random Components

- Certain Zero:

$$Y_{bin,it} = I(Y_{it} = 0) \sim \mathcal{D}(\pi_{it}, \pi_{it}(1 - \pi_{it}))$$

- Positive Count:

$$Y_{it} | (y_{it} > 0) \sim \mathcal{D}\left(\mu(\lambda_{it}), V(\lambda_{it})\right)$$

Hurdle GMM: Model

Joint Quasi-Generalized Linear Model: Random Components

- Positive Count:

$$\mu(\lambda_{it}) = \frac{\lambda_{it}}{1 - e^{-\lambda_{it}}}$$

$$V(\lambda_{it}) = \mu(\lambda_{it})\left[1 - \lambda_{it} + \mu(\lambda_{it})\right]$$

Hurdle GMM: Model

Joint Quasi-Generalized Linear Model: Systematic Components

- Certain Zero:

$$\mathrm{logit}\left(\pi(\mathbf{z}_{it})\right) = \mathbf{z}_{it}^{T}\boldsymbol{\alpha}$$

- Positive Count:

$$\ln\left(\lambda(\mathbf{x}_{it})\right) = \mathbf{x}_{it}^{T}\boldsymbol{\beta}$$

Hurdle GMM: General Process

Hurdle GMM Process: Independently Minimize Quadratic Forms:

$$Q_I(\alpha) = \left(\mathbf{G}_I(\alpha; \mathbf{Y}, \mathbf{Z})\right)^T \mathbf{W}_I^{-1} \left(\mathbf{G}_I(\alpha; \mathbf{Y}, \mathbf{Z})\right)$$

$$Q_c(\beta) = \left(\mathbf{G}_c(\beta; \mathbf{Y}, \mathbf{X})\right)^T \mathbf{W}_c^{-1} \left(\mathbf{G}_c(\beta; \mathbf{Y}, \mathbf{X})\right)$$

Hurdle GMM: General Process

Hurdle GMM Process: Select **Valid** Moment Conditions:

$$\mathbf{G}_l(\alpha; \mathbf{Y}, \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_{l,i}(\alpha; \mathbf{Y}_i, \mathbf{Z}_i)$$

$$\mathbf{G}_c(\beta; \mathbf{Y}, \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_{c,i}(\beta; \mathbf{Y}_i, \mathbf{X}_i)$$

$$\mathrm{E}[g_{l,ij}] = \mathrm{E}[g_{c,ij}] = 0$$

Hurdle GMM: General Process

Hurdle GMM Process: Structure of Valid Moment Conditions:

$$g_{l,ij}(\boldsymbol{\alpha}; \mathbf{Y}_i, \mathbf{Z}_i) = \frac{\partial \pi(\mathbf{z}_{is})}{\partial \alpha_k} \left(I_{(Y_{it}=0)} - \pi(\mathbf{z}_{it}) \right)$$

$$g_{c,ij}(\boldsymbol{\beta}; \mathbf{Y}_i, \mathbf{X}_i) = \frac{\partial \mu(\lambda(\mathbf{x}_{is}))}{\partial \beta_k} \left(I_{(Y_{it}>0)}[Y_{it} - \mu(\lambda(\mathbf{x}_{is}))] \right)$$

It remains to determine how to construct these Valid Moment Conditions

Hurdle GMM: Valid Moment Conditions

Determining Valid Moment Conditions:

1. "Types" as selected by researcher
2. "Extended Classification" using data

Hurdle GMM: Types

Validity of Moment Conditions depends on the expectation:

$$\mathrm{E}|_{\boldsymbol{\beta}} \left[\frac{\partial \mu_{is}}{\partial \beta_j} (Y_{it} - \mu_{it}) \right] = 0$$

Lai and Small (2007) proposed using expected characteristics of individual Time-Dependent Covariates to make decisions on combinations of s and t that would lead to independent components.

Hurdle GMM: Types

- Type I TDC: Expectation holds **for all** s **and** t
- Type II TDC: Expectation holds **for** $s \geq t$
- Type III TDC: Expectation holds **for** $s = t$
- Type IV TDC: Expectation holds **for** $s \leq t$

Hurdle GMM: Types

- Type I TDC: The response and time-dependent covariate are associated only at the **same time**.
- Type II TDC: The response is associated with **prior values** of the time-dependent covariate.
- Type III TDC: A **feedback loop** exists between the time-dependent covariate and the response.
- Type IV TDC: The time-dependent covariate is associated with **prior values** of the response.

Hurdle GMM: Types

Construct subject vectors of Valid Moment Conditions using values of s and t that satisfy the chosen "type" of TDC:

$$g_{l,ij}(\boldsymbol{\alpha}; \mathbf{Y}_i, \mathbf{Z}_i) = \frac{\partial \pi(\mathbf{z}_{is})}{\partial \alpha_k} \left(I_{(Y_{it}=0)} - \pi(\mathbf{z}_{it}) \right)$$

$$g_{c,ij}(\boldsymbol{\beta}; \mathbf{Y}_i, \mathbf{X}_i) = \frac{\partial \mu(\lambda(\mathbf{x}_{is}))}{\partial \beta_k} \left(I_{(Y_{it}>0)} [Y_{it} - \mu(\lambda(\mathbf{x}_{it}))] \right)$$

Hurdle GMM: Extended Classification

Extended Classification Process:

1. Estimate derivative, residual terms of expectation using initial estimates (Independent GEE)
2. Select Valid Moment Conditions individually based on empirical independence of standardized derivative, residual terms (using all subjects)
3. Construct vectors of Valid Moment Conditions using empirically supported combinations of s and t

Hurdle GMM: Extended Classification

Using initial parameter estimates, calculate component-wise independent vectors:

$$\hat{d}_{sj} = \frac{\partial \hat{\boldsymbol{\mu}}_s}{\partial \beta_j}$$

$$\hat{r}_t = \mathbf{y}_t - \hat{\boldsymbol{\mu}}_t$$

Standardized Values:

$$\tilde{d}_{sji} \text{ and } \tilde{r}_{ti}$$

Hurdle GMM: Extended Classification

Calculate Correlation:

$$\hat{\rho}_{sjt} = \frac{\sum(\tilde{d}_{sji} - \bar{\tilde{d}}_{sj})(\tilde{r}_{ti} - \bar{\tilde{r}}_t)}{\sqrt{\sum(\tilde{d}_{sji} - \bar{\tilde{d}}_{sj})^2 \sum(\tilde{r}_{ti} - \bar{\tilde{r}}_t)^2}}$$

Assuming all fourth moments exist and are finite,

$$\rho^*_{sjt} = \frac{\hat{\rho}_{sjt}}{\sqrt{\hat{\mu}_{22}/N}} \sim \mathcal{N}(0,1)$$

$$\left(\hat{\mu}_{22} = (1/N)\sum_i (\tilde{d}_{sji})^2(\tilde{r}_{ti})^2\right)$$

Omit potential moment conditions with **significant** association

Hurdle GMM: Estimation Process

0 (Based on Hurdle GEE (Independent), evaluate associations in potential moment conditions)

1 Construct separate vectors of Valid Moment Conditions for two components of Joint Quasi-GLM

2 Using initial parameter estimates, estimate optimal weight matrices for each component

3 Separately minimize two Quadratic Forms for two components of Joint Quasi-GLM

Hurdle GMM: Estimation Options

Implementation of GMM:

- **Two-Step GMM:** Estimate weight matrix $\hat{\mathbf{W}}$ using initial parameter estimates, minimize $Q(\beta)$
- **Iterated GMM:** Iterate between estimation of $\hat{\mathbf{W}}$ and minimization of $Q(\beta)$
- **Continuously Updating GMM:** Minimize $Q(\beta)$, where $\mathbf{W}(\beta)$ is a function of unknown parameters

Hurdle GMM: Two-Step Estimation

Implementation of GMM:

- **Two-Step GMM:** Estimate weight matrix $\hat{\mathbf{W}}$ using initial parameter estimates, minimize $Q(\beta)$
- **Iterated GMM:** Iterate between estimation of $\hat{\mathbf{W}}$ and minimization of $Q(\beta)$
- **Continuously Updating GMM:** Minimize $Q(\beta)$, where $\mathbf{W}(\beta)$ is a function of unknown parameters

Hurdle GMM: Two-Step Estimation

$$\hat{\boldsymbol{\alpha}} = \arg\ \min\left[Q_l(\boldsymbol{\alpha})\right] \qquad , \qquad \hat{\boldsymbol{\beta}} = \arg\ \min\left[Q_c(\boldsymbol{\beta})\right]$$

$$\mathrm{Cov}(\hat{\boldsymbol{\alpha}}) = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{g}_l}{\partial \boldsymbol{\alpha}}\right)^T \hat{\mathbf{V}}_{g_l}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{g}_l}{\partial \boldsymbol{\alpha}}\right)$$

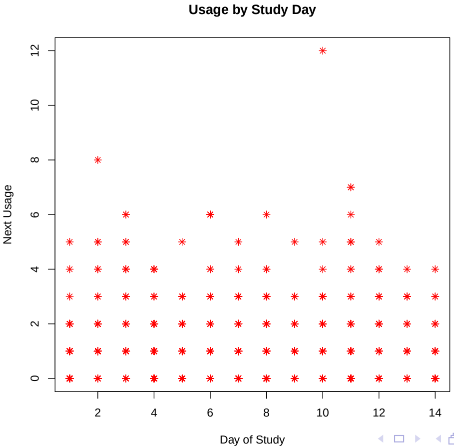
$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{g}_c}{\partial \boldsymbol{\beta}}\right)^T \hat{\mathbf{V}}_{g_c}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{g}_c}{\partial \boldsymbol{\beta}}\right)$$

Example Data Analysis: EMA

EMA Data Analysis

Predict next usage using craving, controls (day of the week, academics)

Usage over Time

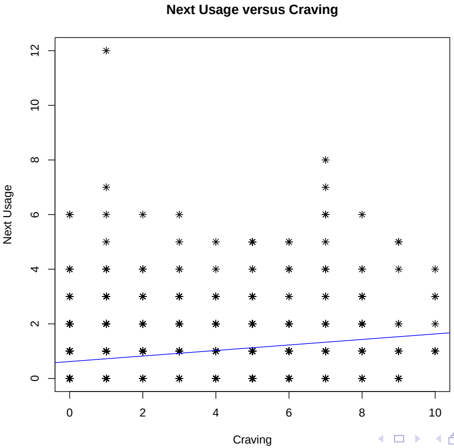


Usage Reports

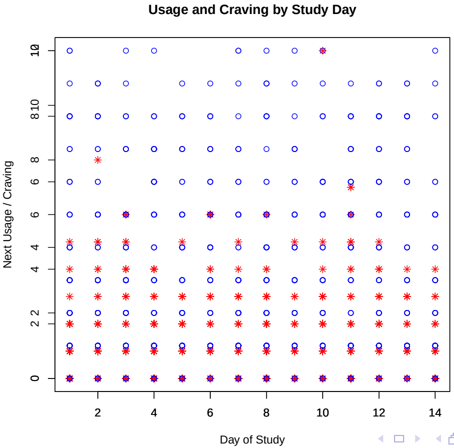
Times Used	0	1	2	3	4	5	6	7	8	12
Frequency	513	318	120	55	25	13	8	2	1	1

48.58% Zeros

Usage and Craving



Usage and Craving



Models Fit

Three models fit:

1. Mixed Hurdle Poisson, with Between / Within Decomposition of “craving”
2. GEE Hurdle, with Independent Working Correlation Structure
3. GMM Hurdle, with “craving” as Type II TDC, Day as Type I TDC

Model Results

	Mixed Hurdle	**Hurdle IGEE**	**Hurdle GMM**
Logistic			
craving	0.223^{***} (W) $-0.219^{\cdot}$ (B)	-0.170^{***}	-0.169^{***}
controls	*Not Significant*	.	**
Count			
craving	-0.077^{***} (W) 0.160 (B)	0.049^{**}	0.048^{***}
cum GPA	0.056	-0.056	-0.056^{***}
controls	*Not Significant*	.	**

Model Results

- Populations with higher craving:
  - Lower probability of certain zero
  - Higher expected positive count, once hurdle is cleared
- Higher within-subject variation:
  - Higher probability of certain zero
  - Lower expected positive count, once hurdle is cleared

Concluding Remarks

- GMM: Initial Values, Estimation Method
- Simulating “Types” of TDC’s
- GMM Fit Statistics
- ARE versus IGEE

Modeling Longitudinal Count Data with Excess Zeros and Time-Dependent Covariates: Application to Drug Use

Trent L. Lalonde

University of Northern Colorado

November 17, 2014