Analysis of Scale Error Data Using Hurdle Random Effects Models

Yunje Han · Keunbaik Lee

Department of Statistics, Sungkyunkwan University, Seoul, 03063, Korea

Abstract

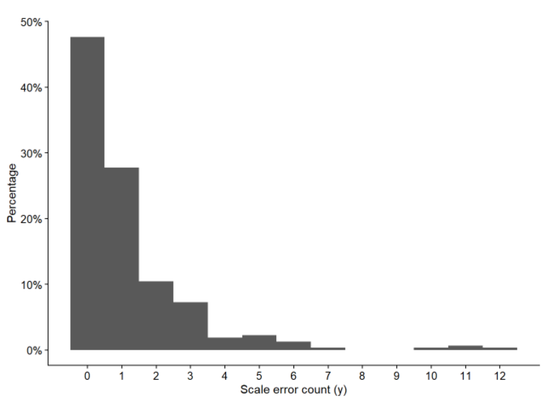
Zero-inflated longitudinal count data are commonly encountered in fields such as developmental psychology where repeated measurements often exhibit an excess number of zeros. Overall marginalized hurdle random effects models (OMHREMs) provide a flexible framework for analyzing such data by jointly modeling the zero and positive count processes, accounting for within-subject correlation through random effects, and yielding marginal interpretations of covariate effects. In this study, we evaluate the performance of the OMHREM under the missing at random (MAR) assumption through extensive simulation studies across a range of missing rates and sample sizes. To demonstrate its practical applicability, the model is applied to longitudinal scale error data from developmental psychology.

Introduction

- The Zero-inflated Poisson model (Lambert, 1992) assumes two sources of zeros: structural and sampling zeros.
- Hurdle model (Cragg, 1971; Mullahy, 1986) accommodates two separate latent processes when zero observations are generated by a single mechanism and positive counts only occur after an initial event.
- Unlike ZIP models, hurdle models remain applicable when zero-deflation is present.

Motivating Data

- Scale error refers to perception–action dissociation, where children act as if objects differ in size from their actual dimensions (DeLoache et al., 2004).
- Iconic examples include attempting to sit on a miniature chair or enter a small toy car.
- In classroom environments, children are typically observed repeatedly over multiple sessions, yielding longitudinal count data with repeated measurements.
- Advanced statistical methods are increasingly needed to analyze scale error data (Hagihara et al., 2024).





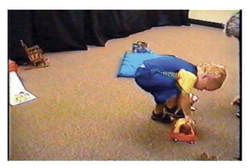


Table 1. Summary statistics of variables

Variable	Mean	SD
Scale error count (Y)	1.13	1.76
Age (months)	22.07	4.67
Session	3.54	2.60
Number of target objects	3.24	0.43

Models for Longitudinal Zero-Inflated Count Data

Hurdle Random Effects Model (Min & Agresti, 2005)

- y_{it} is the count response for subject i at time t.

$$\Pr\left(Y_{it}=y_{it}\mid \mathrm{b}_{i1},\mathrm{b}_{i2}\right)=\begin{cases}\mathrm{p_{it}}(\boldsymbol{b}_{i1}), & y_{it}=0,\\ \{1-\mathrm{p_{it}}(\boldsymbol{b}_{i1})\}\dfrac{g\left(y_{it};\lambda_{\mathrm{it}}(\boldsymbol{b}_{i2})\right)}{1-g\left(0;\lambda_{\mathrm{it}}(\boldsymbol{b}_{i2})\right)}, & y_{it}>0\end{cases}$$

$$\mathrm{logit}\{\mathrm{p_{it}}(\boldsymbol{b}_{i1})\}=\boldsymbol{x}_{it1}^{\mathrm{T}}\boldsymbol{\alpha}+\mathrm{z_{it1}}\boldsymbol{b}_{i1},\qquad \log\{\lambda_{\mathrm{it}}(\boldsymbol{b}_{i2})\}=\boldsymbol{x}_{it2}^{\mathrm{T}}\boldsymbol{\beta}+\mathrm{z_{it2}}\boldsymbol{b}_{i2}$$

- p_{it}^{M} is the mixing probability and $g\left(y_{it};\lambda_{it}^{M}\right)$ is Poisson probability mass function with mean λ_{it}^{M}.

Overall Marginalized Hurdle Random Effects Model (Lee et al., 2026)

Marginal model:

The marginal Poisson hurdle random effects model (OMHREM) directly specifies the population-averaged probability of zero outcomes alongside the marginal mean of positive counts.

$$\Pr\left(Y_{it}=y_{it}\mid \boldsymbol{x}_{it}\right)=\begin{cases}p_{it}^{\mathrm{M}}, & y_{it}=0,\\ \left(1-p_{it}^{\mathrm{M}}\right)\dfrac{g\left(y_{it};\lambda_{it}^{\mathrm{M}}\right)}{1-\exp\left(-\lambda_{it}^{\mathrm{M}}\right)}, & y_{it}=1,2,\ldots.\end{cases}$$

$$\mathrm{logit}(\mathrm{p_{it}^{M}})=\mathrm{x_{it1}^{T}}\boldsymbol{\gamma},\quad \log(\lambda_{\mathrm{it}}^{\mathrm{M}})=\mathrm{x_{it1}^{T}}\boldsymbol{\beta}+\log(\mathrm{off_{it}}).$$

- $\boldsymbol{\gamma}$: p_1 x 1 regression parameter vector for the binary component
- $\boldsymbol{\beta}$: p_2 x 1 regression parameter vector for the count component

Dependence model:

The dependence model accounts for within-subject correlation through random effects.

$$\Pr\left(Y_{it}=y_{it}\mid \boldsymbol{x}_{it},\boldsymbol{b}_{i}\right)=\begin{cases}p_{it}^{\mathrm{c}}(\boldsymbol{b}_{i}), & y_{it}=0,\\ \{1-p_{it}^{\mathrm{c}}(\boldsymbol{b}_{i})\}\dfrac{g\left(y_{it};\boldsymbol{b}_{i},\lambda_{it}^{\mathrm{c}}(\boldsymbol{b}_{i})\right)}{1-\exp\left\{-\lambda_{it}^{\mathrm{c}}(\boldsymbol{b}_{i})\right\}}, & y_{it}=1,2,\ldots.\end{cases}$$

- $g\left(y_{it};\boldsymbol{b}_{i},\lambda_{it}^{c}(\boldsymbol{b}_{i})\right)$: conditional Poisson probability mass function

The conditional model components are specified using generalized linear models:

$$\mathrm{logit}(\mathrm{p_{it}^{c}}(b_i))=\Delta_{it1}+z_{\mathrm{it1}}^{\mathrm{T}}b_{i1},\ \log(\lambda_{\mathrm{it}}^{c}(b_i))=\Delta_{it2}+z_{\mathrm{it2}}^{\mathrm{T}}b_{i2}$$

$$\boldsymbol{b}_i=(\boldsymbol{b}_{i1}^{\mathrm{T}},\boldsymbol{b}_{i2}^{\mathrm{T}})^{\mathrm{T}}\sim iid\ N(0,\Sigma_i)\ where\ \Sigma_i=\begin{pmatrix}\sigma_{i1}^2 & \rho\sigma_{i1}\sigma_{i2}\\ \rho\sigma_{i1}\sigma_{i2} & \sigma_{i2}^2\end{pmatrix}$$

- $\boldsymbol{b}_i=(\boldsymbol{b}_{i1}^{\mathrm{T}},\boldsymbol{b}_{i2}^{\mathrm{T}})^{\mathrm{T}}$: vector of subject-specific random effects that captures the within-subject correlation.

We assume generalized linear models for the standard deviations σ_{i1} and σ_{i2}, and the correlation ρ_i as follows:

$$\log(\sigma_{i1})=h_{\mathrm{i}}^{\mathrm{T}}\zeta_1,\qquad \log(\sigma_{i2})=h_{\mathrm{i}}^{\mathrm{T}}\zeta_2,\qquad \frac{1}{2}log\left(\frac{1+\rho_i}{1-\rho_i}\right)=w_{\mathrm{i}}^{\mathrm{T}}\delta$$

- $\boldsymbol{h}_i$ and $\boldsymbol{w}_i$: subject-specific covariate vectors, $\zeta_1(a\times 1),\zeta_2(b\times 1),\delta\ (c\times 1)$: unknown parameter vectors

To preserve consistency between the marginal and conditional formulations, we assume $\mathrm{E}(p_{it}^{c}(b_i))=p_{it}^{\mathrm{M}}$ and $\mathrm{E}(\lambda_{it}^{c}(b_i))=\lambda_{it}^{\mathrm{M}}$, which yield deterministic, closed-form expressions for Δ_{it1} and Δ_{it2}:

$$\Delta_{it1}=\sqrt{1+\sigma_{i1}^2}\,logit\left(p_{it}^{\mathrm{M}}\right),\ \Delta_{it2}=\log\left(\lambda_{it}^{M}\right)-\frac{1}{2}\sigma_{i2}^2$$

The marginal and dependence models are interconnected through the overall marginal mean $\mu_{\mathrm{it}}=\mathrm{E}(Y_{it})$.

Simulation Study

We assume a missing at random (MAR) mechanism, where the probability of missingness depends on the observed outcomes but not on the unobserved ones. (Diggle 2002)

$$\mathrm{logit}\left(p_{ij}\right)=\phi_0+\phi_1 y_{i,j-1}$$

Table 2. Simulation results of the OMHREM under MAR mechanisms with 5\% and 15\% dropout rates for sample sizes N=500, 700, and 1000.

	N = 500		N = 700		N = 1000	
Missing Rate	5%	15%	5%	15%	5%	15%
10^3 FROB(Σ)	0.977	0.014	0.169	1.031	1.420	0.313
APRB	0.965	1.208	0.647	1.395	0.705	0.834
ASE(ASD)	0.082 (0.092)	0.088 (0.098)	0.070 (0.080)	0.075 (0.084)	0.059 (0.067)	0.063 (0.071)

- Frobenius norm(FROB): $\mathrm{FROB}=\mathrm{tr}\left[\left(\hat{\Sigma}_{\mathrm{i}}\Sigma_{\mathrm{i}}^{-1}-\mathrm{I}\right)^2\right]$
- Average absolute percent relative bias (APRB): $\mathrm{APRB}=\frac{1}{\mathrm{L}}\sum_{\mathrm{l}=1}^{\mathrm{L}}|\mathrm{PRB}(\hat{\varepsilon}_{\mathrm{l}})|$
- Average standard error (ASE): $\mathrm{ASE}=\frac{1}{\mathrm{L}}\sum_{\mathrm{l}=1}^{\mathrm{L}}\mathrm{SE}(\hat{\varepsilon}_{\mathrm{l}})$
- Average standard deviation (ASD): $\mathrm{ASD}=\frac{1}{\mathrm{L}}\sum_{\mathrm{l}=1}^{\mathrm{L}}\mathrm{SD}(\hat{\varepsilon}_{\mathrm{l}})$

❖ As the sample size increases, both the ASE and ASD consistently decrease.
❖ The APRB expectedly increases as the missingness severity rises from 5% to 15% for all sample sizes.
❖ ASE and ASD tend to expand slightly under the 15% missing rate condition compared to the 5% baseline.
❖ The overall biases remain negligible, and the parameter estimates exhibit robust and superior performance even under a 15% missing rate.

Analysis of Scale Error Data

Table 3. Models with heterogeneous covariance structures

Model	Description	Variance Structure	Correlation Structure
Model 1	OMHREM	λ_{j0}	ν_0
Model 2	OMHREM	$\lambda_{j0}+\lambda_{j1}n_{obj,i}$	ν_0
Model 3	OMHREM	$\lambda_{j0}+\lambda_{j1}n_{obj,i}$	$\nu_0+\nu_1\log(age_i)$
Model 4	OMHREM	λ_{j0}	$\nu_0+\nu_1\log(age_i)$

Table 4. Maximum likelihood estimates, maximized log-likelihoods, AIC, and BIC for Models 1–5 fitted to the scale error data

	Model 1	Model 2	Model 3	Model 4
Zero model				
γ_0 (Intercept)	−0.380 (0.222)	−0.287 (0.238)	−0.226 (0.240)	−0.317 (0.224)
γ_1 ($\mathrm{age_z}$)	−0.034 (0.172)	−0.041 (0.173)	−0.073 (0.174)	−0.052 (0.174)
γ_2 ($\mathrm{session_z}$)	0.134 (0.134)	0.142 (0.135)	0.165 (0.132)	0.168 (0.133)
Count model				
β_0 (Intercept)	0.603* (0.176)	0.638* (0.132)	0.630* (0.123)	0.570* (0.176)
β_1 ($\mathrm{age_z}$)	0.103 (0.166)	0.079 (0.148)	−0.007 (0.130)	−0.008 (0.158)
β_2 ($\mathrm{age_z}^2$)	−0.093 (0.101)	−0.083 (0.089)	−0.064 (0.079)	−0.063 (0.092)
β_3 ($\mathrm{session_z}$)	−0.159* (0.070)	−0.168 (0.089)	−0.189* (0.086)	−0.180* (0.069)
β_4 ($n_{obj,c}$)	−0.536 (0.287)	−0.670 (0.398)	−0.685 (0.448)	−0.475 (0.284)
Random-effects covariance model				
λ_{10} (Intercept)	−0.291 (0.209)	−0.351 (0.251)	−0.383 (0.255)	−0.319 (0.216)
λ_{11} (n_{obj})	–	0.350 (0.532)	0.551 (0.498)	–
λ_{20} (Intercept)	−0.873* (0.284)	−1.269* (0.423)	−1.290* (0.409)	−0.831* (0.268)
λ_{21} (n_{obj})	–	1.039 (0.556)	1.206* (0.542)	–
ν_0 (Intercept)	−0.265 (0.347)	−0.164 (0.395)	11.622 (7.912)	7.717 (7.252)
ν_1 (log(age))	–	–	−3.815 (2.573)	−2.567 (2.408)
Criterion				
Maximized log-likelihood	-450.608	-447.132	-446.474	-450.095
AIC	923.216	920.264	920.948	924.190
BIC	950.591	952.616	955.789	954.054

$$\mathrm{logit}(\hat{\mathrm{p}}_{\mathrm{it}}^{\mathrm{M}})=\text{-}0.287\text{-}0.041\mathrm{age_{z,it}}+0.142\mathrm{session_{z,it}}$$

$$\log(\hat{\mu}_{\mathrm{it}}^{\mathrm{M}})=0.638+0.079\ \mathrm{age_{z,it}}\text{ - }0.083\mathrm{age_{z,it}}^2\text{-}0.168\ \mathrm{session_{z,it}}\text{ -}0.670\ \mathrm{n_{obj,it.}}$$

$\hat{\sigma}_{\mathrm{i1}}=\exp(-0.351+0.350)=0.999$: four target objects | $\hat{\sigma}_{\mathrm{i1}}=\exp(-0.351)=0.704$: three target objects
$\hat{\sigma}_{\mathrm{i2}}=\exp(-1.269+1.039)=0.794$: four target objects | $\hat{\sigma}_{\mathrm{i2}}=\exp(-1.269)=0.281$: three target objects

$$\hat{\rho}_{\mathrm{i}}=\frac{\exp\{2(-0.164)-1\}}{\exp\{2(-0.164)+1\}}\approx-0.163\qquad \hat{\Sigma}_{four}=\begin{pmatrix}0.998 & -0.130\\ -0.130 & 0.632\end{pmatrix}\qquad \hat{\Sigma}_{three}=\begin{pmatrix}0.496 & -0.032\\ -0.032 & 0.079\end{pmatrix}$$

Conclusion

- This study evaluates the performance of OMHREM under a wide range of missing rates and varying sample sizes.
- The simulation results indicate that OMHREM provides an improved model fit relative to competing models while offering interpretable marginal covariate effects.
- OMHREM is applied to longitudinal scale error data from the field of developmental psychology.
- One limitation of OMHREM is that it assumes a Poisson distribution, which can lead to underestimated standard errors in the presence of overdispersion. Hence, marginalized negative binomial hurdle model can be considered.