M3 conference, June 27-28, 2023: Storrs, CT

The Ohio State University

# Model Typology and Demonstration for Over-dispersion and Extra-zeros in Multilevel Models for Discrete Count Data

Ann A. O'Connell[a], Nivedita Bhaktha[b], Krisann Stephany[a], Winifred Wilberforce[a] & Abena Anyidoho[a]
[a]The Ohio State University
[b]GESIS Leibniz-Institut für Sozialwissenschäften
Contact: Oconnell.87@osu.edu

M3 presentation, data and R code are at:
**https://go.osu.edu/aoconnell_website**

http://www.barbaragrygutis.com/garden-of-constants

1

## Today's presentation

- Typology of models for multilevel discrete count data
  - Event counts: 0, 1, 2, etc.
  - Adaptations for rates or densities (counts per area, time, etc.)
- Tutorial based on several models used most often in educational and social sciences
  - Topics include equidispersion, over-dispersion, zero-inflation, sampling vs. structural zeros
  - Illustrations: Generalized Linear Mixed Models (GLMM) for Poisson (MLP), Negative Binomial (MLNB) and Generalized Poisson (MLGP), and Multilevel ZIP, ZINB, and Hurdle models
- Guidance for researchers on selection and use of these models

Additional details on multilevel counts in general: O'Connell, Bhaktha & Zhang (2022). Chapter 8: Single and multilevel models for count outcomes. In *Multilevel Modeling Methods with Introductory and Advanced Applications (*O'Connell, McCoach & Bell (Eds.)) IAP 2

2

# Approach

M3 presentation, data and R code are at:
https://go.osu.edu/aoconnell_website

- Intro: Working with count data and count regression
- Typology
- Data sources for our examples
- Generalized Linear Mixed-Models (GLMMs) for Counts
- Impact of clustering on count data
- Interpreting model and parameter estimates
- Comparing multilevel models and model-fit
  - Information criteria or likelihood ratio tests
- Summary and Guidance for Researchers

GLMMs
We're focusing on 5 models in the Typology:
Multilevel P, NB, GP, ZI and H

3

3

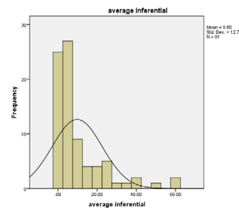
# Intro: Working with Count Data and Count Regression

4

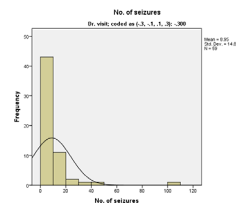
4

# Special nature of count data

- Numbers or rates of events typically vary across individuals or cases in a sample
  - Non-negative integers: 0, 1, 2, …
- Distribution of counts is often skewed/reverse J-shaped
  - Lots of people with no occurrence of the event (zeros)
  - Several people with one or more occurrences
  - A few people with a very high number of occurrences
- Events may occur within a specific range, time span, or geographic area (i.e., "exposure" opportunities, which also may vary across cases)
  - Rates of events




Number of teachers' inferential-type questions




Number of patient seizures at baseline in an RCT

5



# Poisson Distribution

- Most common *initial* distribution used to describe count data, $Y$
  - Poisson process counts the number of times that an event occurs
  - Probability mass function describes probability of observing specific count responses:

$$P(Y = y) = \frac{\lambda^{y} e^{-\lambda}}{y!} \quad \text{for } y = 0, 1, 2, \ldots$$

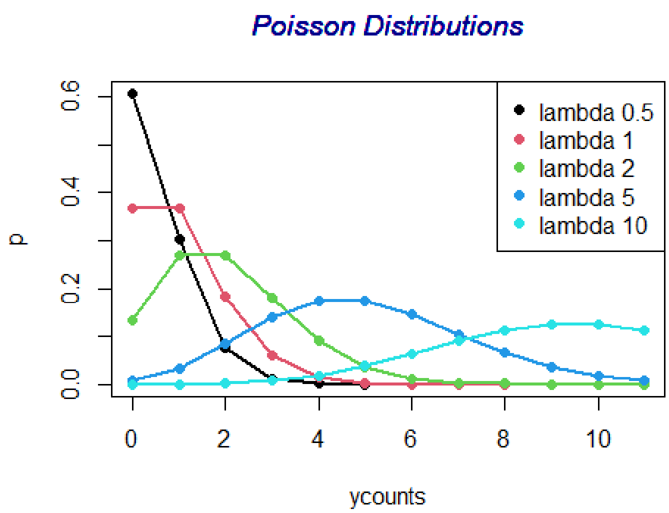
  - Assumption: Data following a true Poisson distribution have a variance that is equal to their mean
    - "***equidispersion***" E(Y) = Var($Y$) = $\lambda$

One parameter, the mean $\lambda$ or $\mu$

As the mean increases, so does the variability!

6

# Probability functions for selected values of lambda




- As $\lambda$ increases, the probability of zeros decreases, and variability in the counts also increases
- $\lambda$ can be any positive value (not necessarily an integer, since it's an average)
  - It is the mean, or the "rate parameter"
  - Average number of events (counts) that occurs in the identified observation period
    - If "exposure" time varies for each case – can be adjusted for as an "offset" in prediction models

7

7

# Poisson Regression Model (single-level)

- In the regression model, the number of events is assumed to follow a Poisson distribution with a *conditional mean* based on each individual's characteristics:

$$P(y_i \mid \mathbf{x}_i) = \frac{\lambda_i^{\,y_i} e^{-\lambda_i}}{y_i!}$$

No random error term in the regression model, since we already assume that the variance is equal to the mean

Conditional mean →

$$\lambda_i = E(y_i \mid \mathbf{x}_i) = \exp(\mathbf{x}_i \boldsymbol{\beta})$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots \beta_p x_{ip}$$

A generalized linear model (GLM), for which the parameter estimates are found through maximum likelihood estimation

8

8

# GLM for Counts

- Poisson Regression is a Generalized Linear Model
  - Random component: $Y_i \sim \text{Poisson}(\lambda_i)$
  - Systematic component: regression function is linear (additive) in the parameters
  - Uses log link for the linear predictor

$$\eta_i = \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots \beta_p x_{ip}$$

$$\lambda_i = E(y_i \mid \mathbf{x}_i) = \exp\left(\mathbf{x}_i \beta\right)$$

Transformation for the expected value of the counts

Log transformation used to model the non-linear form between $\lambda_i$ and X, and assures that $\lambda_i$ will be > 0, while regr coeffs can still range +/- ∞





# Example: Relationship between Y and X

Simple simulation:

Imagine counting the number of child-related outbursts in classrooms within a year or month, etc., across a sample n = 100 principals

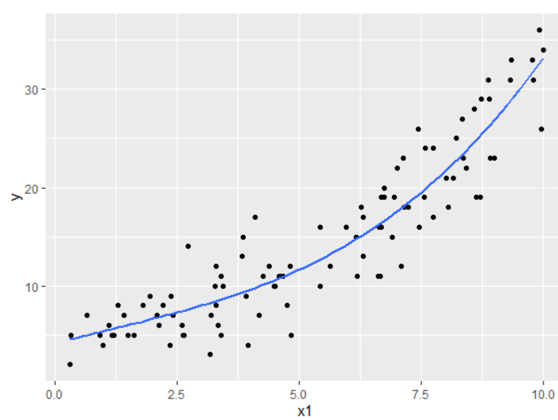
Assume X is rating on discipline severity of the principal, we'll assume Uniform ranging from 0 to 10 on this scale.

$$Y \sim P\left(n = 100, \lambda_i = \exp(\beta_0 + \beta_1 X_1)\right)$$

$$\beta_0 = 1.5$$

$$\beta_1 = .2$$

$$X_1 \sim Unif\left(n = 100, \min = 0, \max = 10\right)$$



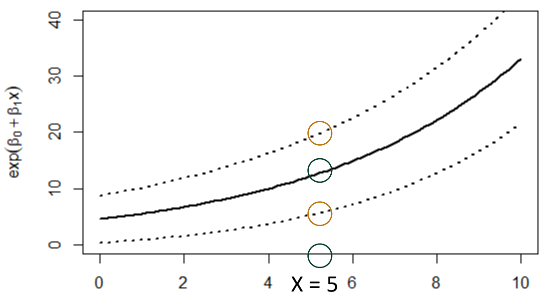

Curve is $\lambda_i$ | X

# Relationship between E(Y|X) and X

$$\lambda_i = \exp(b_0 + b_1 x_i)$$

$\beta_0 = 1.5$
$\beta_1 = .20$

- Mean function ($\lambda_i$) by X, with plausible range for Y




Dotted lines indicate the range of plausible values (95%) for Y

Regression is heteroscedastic

- As X increases, lambda increases, and so does the variability in the counts!

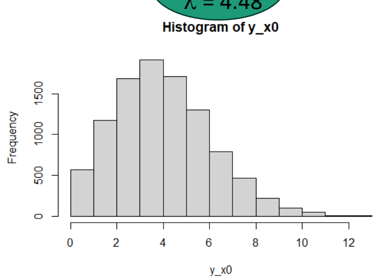
| $x$ | $\lambda_i = \exp(b_0 + b_1 X)$ | E(Y\|X= $x$) | Var(Y\|X=$x$) | 95% Plausible Range for Y $\lambda_i \pm 2 \cdot \sqrt{Var(Y)}$ | Count Range |
|---|---|---|---|---|---|
| 0 | 4.48 | 4.48 | 4.48 | (0.25, 8.71) | 8.46 pts |
| 5 | 12.18 | 4.48 | 4.48 | (5.20, 19.16) | 13.96 |
| 10 | 33.12 | 4.48 | 4.48 | (21.61, 44.62) | 23.01 |





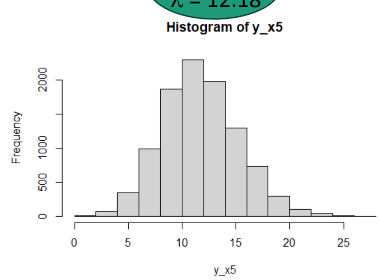
# Distributions at selected values for X

X = 0
$\lambda$ = 4.48



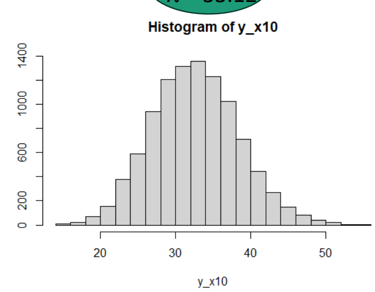

$\bar{Y} = 4.50$ and $\hat{\sigma}^2 = 4.60$

X = 5
$\lambda$ = 12.18




$\bar{Y} = 12.22$ and $\hat{\sigma}^2 = 12.24$

X = 5
$\lambda$ = 33.12




$\bar{Y} = 33.12$ and $\hat{\sigma}^2 = 32.99$

- 10,000 random values generated at each X
- Each distribution exhibits equi-dispersion
- But across the regression line (as X increases), we see the heteroscedasticity

Q: How to investigate the assumption of <u>equi-dispersion</u> for the Poisson regression?

# Model residuals

- One way to investigate equidispersion when using Poisson Regression is to consider the distribution of model *residuals* (adjusted for variability of the counts)

- Pearson Residuals

$$r_m = \frac{y_m - \hat{\lambda}_m}{\sqrt{\hat{\lambda}_m}}, \quad m = 1 \ldots M$$

$M$ = number of unique covariate patterns

$$X^2 = \sum r_i^2$$

- Deviance Residuals ($s_i$ = sign of $y_i - \hat{\lambda}_i$ )

$$d_i = s_i \sqrt{2\left\{ y_i \cdot \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - \left(y_i - \hat{\lambda}_i\right)\right\}}$$

$$G^2 = \sum d_i^2 = \text{Residual Deviance}$$
$$= -2\left[LL_{\text{model}} - LL_{\text{saturated}}\right]$$

13

13

# Simple code (snippets) in R - Poisson

"M3_clean_sim2023.r"

```
##########################################################################
####  Simple (single-level) Poisson Regression - using sim1dat
##########################################################################

m1_poisson <- glm(y ~ x1,
                  family="poisson", data = sim1dat)

m1_poisson
summary(m1_poisson)                              ## for coeffs se and z-test
```

```
##################################
### Raw Residuals
##################################

ctpreds <- fitted.values(m1_poisson)
rawresids <- y - ctpreds
```

```
##################################
### Pearson Residuals
##################################

presids <- residuals(m1_poisson, type="pearson")  ## model
```

```
##################################
### Deviance Residuals
##################################

dresids <- residuals(m1_poisson, type="deviance")              ## model
```

14

14

# From our generated data… $\hat{\eta}_i = 1.45 + .20x_{i1}$

```
>
> m1_poisson

Call:  glm(formula = y ~ x1, family = "poisson", data = sim1dat)

Coefficients:
(Intercept)           x1
     1.4584       0.2029

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      476.9
Residual Deviance: 73.85        AIC: 513.5
> summary(m1_poisson)  ## for coeffs se and z-test

Call:
glm(formula = y ~ x1, family = "poisson", data = sim1dat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.45839    0.07591   19.21   <2e-16 ***
x1           0.20289    0.01060   19.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 476.898  on 99  degrees of freedom
Residual deviance:  73.849  on 98  degrees of freedom
AIC: 513.47

Number of Fisher Scoring iterations: 4

> presids <- residuals(m1_poisson, type="pearson")  ## model
> sum(presids^2)       ##  This is statistic called X-squared
[1] 72.19203
>
```

- Residuals information

```
>
> sum(presids^2)
[1] 72.19203
>
>
> sum(dresids^2)
[1] 73.84911
```

n – 2 df

- Since 72.19/98 = .74, there is no suggestion of overdispersion
- Poisson seems to be a reasonable model for the data





# Interpretation – coeffs in the simple model

$$\hat{\eta}_i = 1.45 + .20x_{i1}$$

- We said Y = number of classroom outbursts, X = discipline severity of principal, range from 0 to 10
- For a non-harsh principal (0 on this scale) 1.45 is expected log(number of classroom outbursts)
  - This is prediction on the link or log scale
- Exp(1.45) = 4.26 = expected count on the Y scale, for principal of very low discipline severity
- As ratings for principals increase by 1 unit, predicted log(number of outbursts) tends to increase by .20 points
- Exp(.20) is referred to as a Rate Ratio

# More on coefficients and Rate Ratios

- In our example exp(.20) = 1.22

$$\hat{\eta}_i = 1.45 + .20x_{i1}$$

- Expected increase in counts as X increase by 1 unit
  - Log(prediction | X = 2) = 1.85
  - Log(prediction | X = 3) = 2.05

Difference in log(predictions) = 2.05 – 1.85 = .20

- As a Rate Ratio:

$$RR_X = \frac{\hat{Y}_{X=3}}{\hat{Y}_{X=2}} = \frac{\exp(2.05)}{\exp(1.85)} = \frac{7.77}{6.36} = 1.22$$

- As percent change: Use 100*(exp(b) – 1)
  - 100*(exp(.20) – 1) = 100*(1.22 – 1) = 22% increase in number of outbursts, as X increases by 1 unit

17

17

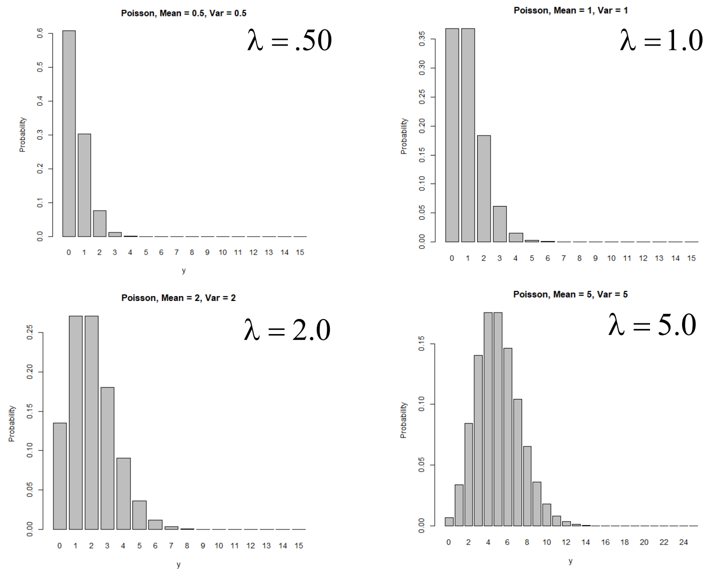
# Coefficient Interpretation (general)

- Each estimated coefficient, $\hat{\beta}_j = b_j$ , represents the expected change in the natural log of counts given a one-unit change in the predictor, holding other predictors constant
- Coefficients can be exponentiated to represent change in actual counts (rather than change in the log-counts) to yield the *rate ratio* (RR) or *incidence rate ratio* (IRR)
  - exp($b_j$) = RR
- Each RR explicitly compares the *rates* of two values of the predictor as it increases by one unit
  - For each one-unit increase in the predictor, the expected count will multiplicatively increase (or decrease, if the parameter is negative) by the value of the RR, holding other predictors constant

18

18

# More on equi- versus over-dispersion




Poisson Distribution will <u>always</u> have this property of equi-dispersion

Actual data for counts often don't meet this assumption

<u>Poisson</u> Density:

$$f(y \mid \lambda) = \frac{\lambda^{y} e^{-\lambda}}{y!}$$

$$\mu = \sigma^2 = \lambda$$

One-parameter distribution

19

19

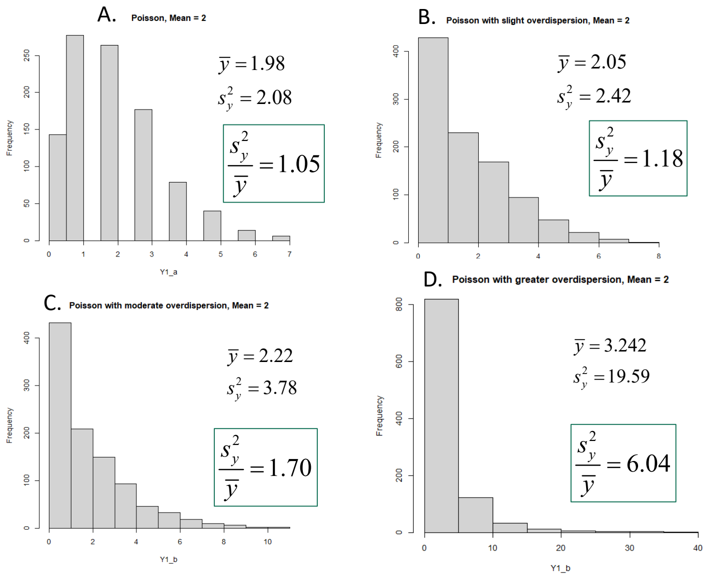
# Contributions to Overdispersion

- Overdispersion:
  - A property of the data *relative to* the model/distribution we choose
  - Occurs when the conditional variances of the observed counts are <u>greater</u> than the conditional means
  - *Standard errors for parameter estimates are underestimated* – predictors may appear to be statistically significant when they are not!
- Many factors can contribute to overdispersion
  - Unaccounted for clustering in the data (multilevel designs)
  - Missing covariates or interaction terms (poor specification)
  - Unreliable data collection
- We often don't know the distribution of the extra dispersion
  - Need to make assumptions about how it's distributed
- The data aren't wrong, but our model choice or methods might be!

20

20

# Add some additional variability to the Poisson mean
# Example: λ = 2 and n = 1000




Sampled n = 1000 observations

A: data sampled from a Poisson, with λ = 2

B, C & D are Poisson with λ = 2 but with random error added

- B exp(N(1000, 0, .25))
- C exp(N(1000, 0, .50))
- D exp(N(1000, 0, 1.0))

Ratio of Var to Mean for each sample?

- 1st : 1.05 variance 5% larger than mean
- 2nd: 1.18 variance 18% larger
- 3rd: 1.70 variance 70% larger
- 4th: 6.04 variance 504% larger!

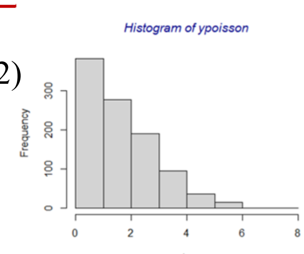
21

21

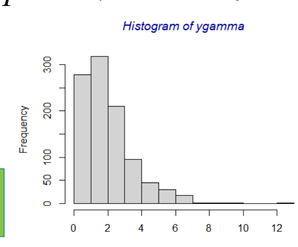
# Gamma: A more reasonable distribution for extra-dispersion

- Has some skew, and is not symmetrical
- A common Poisson-Gamma mixture for overdispersed counts is the *Negative Binomial* distribution
- Gamma has two parameters: <u>shape</u> and <u>scale</u>
  - Gamma mean: shape*scale
  - Gamma variance: shape*scale$^2$
- I generated Gamma dispersion to have a mean of 2

$$Poisson(n = 1000, \lambda = 2)$$
$$\bar{y} = 2.07$$
$$s_y^2 = 1.93$$




$$Gamma(n = 1000, shape = 2, scale = 1)$$
$$\bar{y} = 1.99$$
$$s_y^2 = 2.19$$

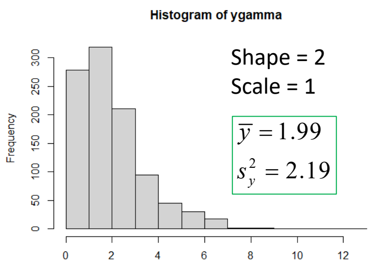


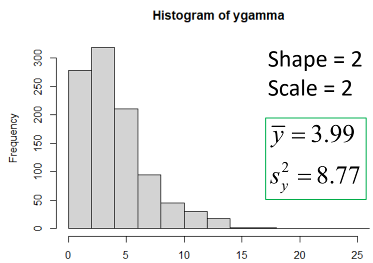

Mean and variance are equal when scale = 1

22

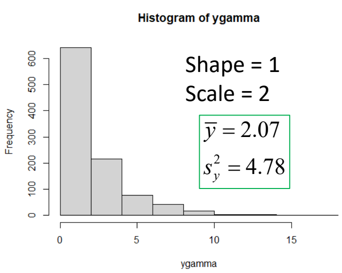
# Other Gamma Examples

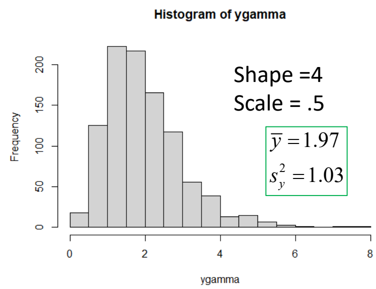
Gamma mean: shape*scale
Gamma variance: shape*scale²

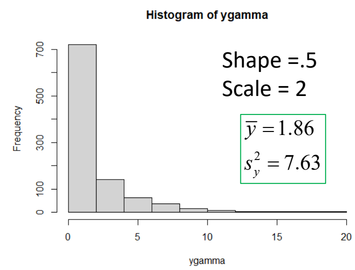



In general, the shape parameter affects the distribution of counts; the scale (or location) increases the possible range or spread, and thus the mean



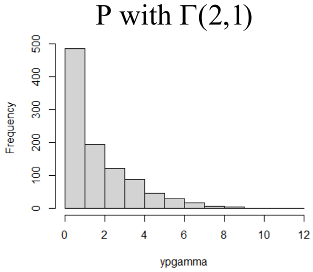




23

# Now adding some extra Gamma dispersion to the Poisson counts (Poisson λ = 2) and (n = 1000)

using $\Gamma(shape, scale)$

P with $\Gamma(2,1)$

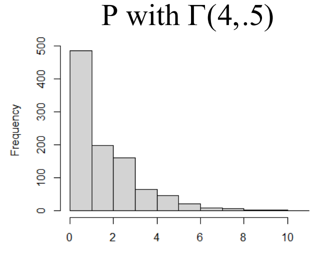



$\bar{y} = 2.01$

$s_y^2 = 3.77$

$\frac{s_y^2}{\bar{y}} = 1.88$

P with $\Gamma(4,.5)$

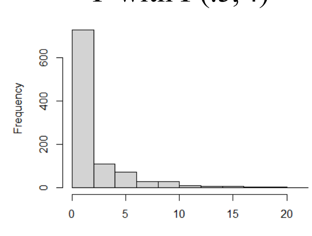



$\bar{y} = 1.96$

$s_y^2 = 3.08$

$\frac{s_y^2}{\bar{y}} = 1.57$

P with $\Gamma(.5,4)$




$\bar{y} = 2.12$

$s_y^2 = 11.21$

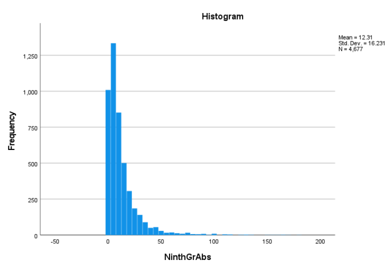
$\frac{s_y^2}{\bar{y}} = 5.29$

24

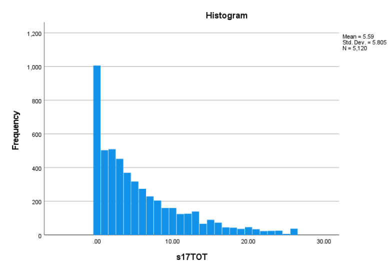
24

# Fundamental questions

- What count model may be most appropriate for the data
  - Equi-dispersion? Overdispersion?
- Are all zeros coming from the same population or process?
  - Zero-inflation versus Hurdle models?
- Can variation in counts or rates be explained by one or more predictors, across one or more levels of the data? (multilevel)




NLSY79 n=4677 9th grade absences




SSP_2017 n=5171 students, aggression past 30 days





# Approaches

- Overdispersion can also contribute to "extra" zeros in a count variable
  - i.e., beyond the expected number of zeros (or other counts) governed by a Poisson (P) distribution
  - Some accommodations include models that estimate an additional dispersion parameter, to account for the "extra" variability
- Zero-inflation (ZI) models
  - Some counts may have more 0's than expected through P or NB
  - Mixture models: ZIP or ZINB – overall distribution captured by a mix of 2 processes or distributions
- Hurdle (H) models
  - Two-part models rather than a mixture (HP, HNB)
  - All zeros treated as structural zeros; positive counts modeled through a truncated P or NB model

*Sampling zeros*: occurring based on the distribution for the counts
*Structural zeros*: perhaps some 0's are not "at risk" for the event!

## Structural versus Sampling zeros

- Example: Y = number of delinquent behaviors in 30-day period
  - Zeros occur because students with a history of delinquent behaviors ("at risk") did not engage in (or report) any of the behaviors in prior 30 days (sampling zeros)
  - OR
  - A student has never engaged in the behavior (not "at risk") (excess or structural zeros)

**Sampling zeros**

- Students who are "at risk"
- history of behavior but did not report or experience the event during the study period
- Some zero counts are expected based on study design, variable being measured, etc.

**Structural zeros**

- students with 0 are not "at risk"
- also called excess zeros
- subpopulation of subjects who cannot or won't experience the event
  - Distinguishable from students with positive counts



27

## Why are these distinctions important?

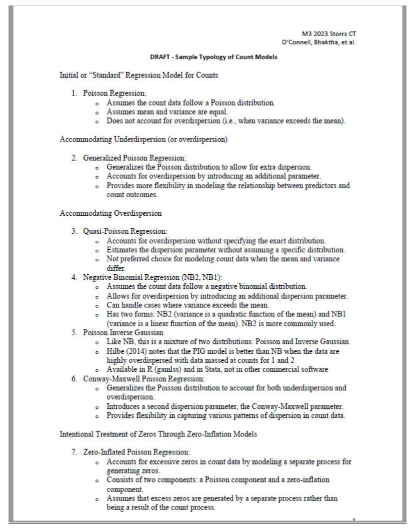
- Rarely do counts "in the wild" follow a true Poisson distribution
  - Typically we see over- or extra-dispersion in the data (more variability than the Poisson would suggest)
- There are many kinds of models to accommodate count distributions that deviate from a Poisson distribution
  - Negative Binomial; over-dispersed NB; overdispersed or quasi-Poisson; generalized Poisson
  - ZIP, ZINB, HP, HNB
  - And others
- The ZI and H models are approaches for intentionally focusing on the zero counts when estimating the model, even if there may be extra-dispersion in the counts



28

# Typology

- https://go.osu.edu/aoconnell_website
- Brief description of 13 different models that
- Our goal for today is to help you make sense of when/why to use these different models
- Focus on clustered data/multilevel models for counts: GLMMs for P, NB,GP, ZIP, ZINB, HP, HNB
- Provide code and resources for fitting these basic multilevel models in the typology
- Code at my web page above



M3 2023 Storrs CT
O'Connell, Bhaktha, et al.

**DRAFT - Sample Typology of Count Models**

Initial or "Standard" Regression Model for Counts

1. Poisson Regression:
   - Assumes the count data follow a Poisson distribution.
   - Assumes mean and variance are equal.
   - Does not account for overdispersion (i.e., when variance exceeds the mean).

Accommodating Underdispersion (or overdispersion)

2. Generalized Poisson Regression:
   - Generalizes the Poisson distribution to allow for extra dispersion.
   - Accounts for overdispersion by introducing an additional parameter.
   - Provides more flexibility in modeling the relationship between predictors and count outcomes.

Accommodating Overdispersion

3. Quasi-Poisson Regression:
   - Accounts for overdispersion without specifying the exact distribution.
   - Estimates the dispersion parameter without assuming a specific distribution.
   - Not preferred choice for modeling count data when the mean and variance differ.
4. Negative Binomial Regression (NB2, NB1):
   - Assumes the count data follow a negative binomial distribution.
   - Allows for overdispersion by introducing an additional dispersion parameter.
   - Can handle cases where variance exceeds the mean.
   - Has two forms: NB2 (variance is a quadratic function of the mean) and NB1 (variance is a linear function of the mean). NB2 is more commonly used.
5. Poisson Inverse Gaussian
   - Like NB, this is a mixture of two distributions: Poisson and Inverse Gaussian
   - Hilbe (2014) notes that the PIG model is better than NB when the data are highly overdispersed with data massed at counts for 1 and 2
   - Available in R (gamlss) and in Stata, not in other commercial software
6. Conway-Maxwell Poisson Regression:
   - Generalizes the Poisson distribution to account for both underdispersion and overdispersion.
   - Introduces a second dispersion parameter, the Conway-Maxwell parameter.
   - Provides flexibility in capturing various patterns of dispersion in count data.

Intentional Treatment of Zeros Through Zero-Inflation Models

7. Zero-Inflated Poisson Regression:
   - Accounts for excessive zeros in count data by modeling a separate process for generating zeros.
   - Consists of two components: a Poisson component and a zero-inflation component.
   - Assumes that excess zeros are generated by a separate process rather than being a result of the count process.

1

29

29

# Data – available on my website

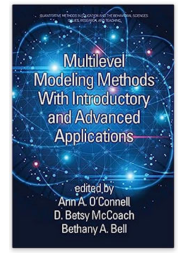
**https://go.osu.edu/aoconnell_website**

30

30

# Data

- O'Connell, Bhaktha & Zhang (2022) for Multilevel P, NB, GP
  - Modeled after ECLS-K data
  - J = 720 schools, n = 11,301 Kindergarten students
  - DV = counts representing proficiency on early numeracy/literacy items
  - COVs = both child/family- and school-level predictors of proficiency
- Bowen et al., (2017, 2022) for Zero-inflation and Hurdle
  - SSP = School Success Profile data
  - J = 17 schools, n = 5171 students, grades 6 - 9
  - DV = count of aggression/microaggression issues experienced by students at school
  - COVs = school/teacher/adult(parent) items, student association with deviant peers gender, etc. I chose 6 preds. 2 are school level
  - Contact: Dr. Natasha Bowen, OSU, Bowen.355@osu.edu




School Success Profile

31

31

# GLMMs for Counts

32

32

## Some recent (multilevel) examples in the literature


- Ames, et al., (2016). Food insecurity and educational achievement: A multi-level generalization of Poisson regression. International Journal of Food and Agricultural Economics, 4(1), 21-34
- Barth & Schmitz (2021). Interviewers' and respondents' joint production of response quality in open-ended questions: A multilevel negative-binomial regression approach. *methods, data, analysis, 15*(1), 43-76.
- Chambers & Erausquin (2018). Race, sex, and discrimination in school settings: A multilevel analysis of associations with delinquency. *Journal of School Health, 88*(2), 159-166.
- Fenta, S. M., & Fenta, H. M. (2020). Risk factors of child mortality in Ethiopia: application of multilevel two-part model. *PloS one*, *15*(8), e0237640.
- Fenta, et al. (2020). The best statistical model to estimate predictors of under-five mortality in Ethiopia. *Journal of Big Data, 7:63*.
- Geremew, et al. (2020). Factors affecting under-five mortality in Ethiopia: A multilevel negative binomial model. *Pediatric Health, Medicine and Therapeutics, 11*, 525-534.
- LARRC, Lo & Xu (2022). Impacts of the Let's Know curriculum on the language and comprehension related skills of PK and K children. *Journal of Educational Psychology*.






# Multilevel P, NB, GP

# Demonstration Example 1 – Multilevel P, NB, GP

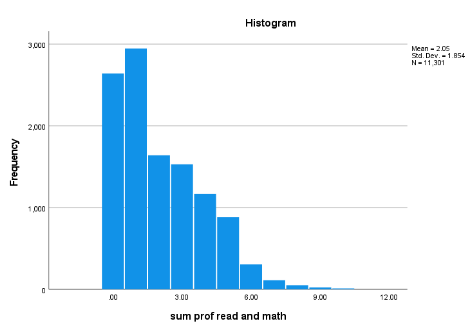
Datafile is "M3_n11301_profcounts.sav"

- Data modeled after proficiency counts from the ECLS-K
- Baseline data from 1998 – 99, followed up through 1st grade, 3rd grade, 5th grade, 8th grade
- Students sampled from both public and private schools
- Modified/simulated for demonstration purposes (not meant to be representative of the ECLS-K original sample)
- $Y_{ij}$ = proficiency counts (*profcount*) were collected for a sample of $n$ = 11,301 Kindergarten children sampled from within $J$ = 720 early-grade schools.
  - Scored based on the number of correct responses to a set of twelve early reading and early math skills items, with the total possible count ranging from 0 to 12 ($\bar{X}$ = 2.05 items, $S$ = 1.85, $S^2$ = 3.44).

35

35

# Proficiency Count Distribution – Kindergarten




(Not adjusted for clustering by school)

**Statistics**

profcount sum prof read and ma

| | | |
|---|---|---|
| N | Valid | 11301 |
| | Missing | 0 |
| Mean | | 2.0526 |
| Median | | 2.0000 |
| Mode | | 1.00 |
| Std. Deviation | | 1.85431 |
| Variance | | 3.438 |
| Range | | 10.00 |
| Minimum | | .00 |
| Maximum | | 10.00 |

*var > mean*

$$\frac{var}{mean} = 1.67$$

36

36

# Descriptive Statistics – Student Level (n = 11301)

**Statistics**

| | | male child sex | momed mom's highest level of education | c_momed centered mom's highest level of education | foodinsec 0 'none' 1 'food insecure' |
|---|---|---|---|---|---|
| N | Valid | 11301 | 11301 | 11301 | 11301 |
| | Missing | 0 | 0 | 0 | 0 |
| Mean | | .5017 | 4.45 | .0000 | .0760 |
| Median | | 1.0000 | 5.00 | .5506 | .0000 |
| Mode | | 1.00 | 3 | -1.45 | .00 |
| Std. Deviation | | .50002 | 1.714 | 1.71383 | .26503 |
| Range | | 1.00 | 8 | 8.00 | 1.00 |
| Minimum | | .00 | 1 | -3.45 | .00 |
| Maximum | | 1.00 | 9 | 4.55 | 1.00 |

- 50% male
- 8% food insecure
- Median Mom's Ed is 5 – at least some college
- Mean Mom's Ed is 4.45, so vocational/tech, 2 yr college or at least some college

37

37

# Descriptive Statistics – School Level (J = 720)

**Statistics**

| | | public | nbhoodprobs |
|---|---|---|---|
| N | Valid | 720 | 720 |
| | Missing | 0 | 0 |
| Mean | | .7750 | 6.2944 |
| Median | | 1.0000 | 5.0000 |
| Mode | | 1.00 | 4.00 |
| Std. Deviation | | .41787 | 3.38564 |
| Range | | 1.00 | 18.00 |
| Minimum | | .00 | 3.00 |
| Maximum | | 1.00 | 21.00 |

- 78% public schools
- Nbhoodprobs represents degree of crime/conflict issues in vicinity of school (crime, trash, vacant buildings, drug-use, etc.), possible range from 0 to 21:
  - Mean = 6.29
  - Std dev = 3.39

38

38

# Poisson or Negative Binomial or Generalized Poisson?

- Poisson Regression models the *conditional* distribution of Y, that is, the (count) distribution that occurs when Y is *conditional* on the values of the predictors
- However, the Poisson (P) model makes a very strong assumption that the variance of the distribution is equal to the mean:
  - E(Y) = Var(Y) = $\lambda$ known as "*Equidispersion*"
- The Negative Binomial (NB) model adjusts for and includes an additional parameter that captures the extra variation present in the data
  - Mixture of Poisson and Gamma distributions
- The Generalized Poisson (GP) estimates a scale parameter that adjusts both the variance and the mean
  - Mixture of Poisson and Poisson distributions

39

39

# Negative Binomial Regression Model (single-level)

- Includes a distinct source of dispersion over the assumed equidispersion Poisson variance

$$\log(\tilde{\lambda}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots \beta_p x_{ip} + \varepsilon_i$$

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i \boldsymbol{\beta}) \cdot \exp(\varepsilon_i) = \lambda_i \cdot \exp(\varepsilon_i) = \lambda_i \cdot \delta_i$$

But: How to characterize the $\delta_i = \exp(\varepsilon_i)$

- Allows for count estimation models where the variance can be different from (greater than) the mean
- Rates may vary across individuals with the same collection of predictors
  - Randomly perturbed or vary by $\delta_i$
- Interpretation of fixed model parameters and RR same as in the P regression
- To *identify* the NB regression model, we make an assumption, $E(\delta_i) = 1.0$
- Expected counts then are the same as they were in the P regression:

$$E(\tilde{\lambda}_i) = E(\lambda_i \delta_i) = \lambda_i E(\delta_i) = \lambda_i$$

40

40

# NB Dispersion Parameter

- The distribution of $P(y_i \mid x_i, \delta_i)$ is still Poisson, but we can't solve this equation without specifying the form of pdf for the $\delta_i$. (Cameron & Trivedi, 2013, p. 115-119; Long, 1997,p. 231-232; Long & Freese, 2014, p. 507-508; Stroup, 2013, p. 352-354)
  - (note these refs do not always use the same symbols OR parameterization when presenting their equations)
- The most commonly used distribution for the extra dispersion represented by $\delta_i$ is the *Gamma* distribution.
- This form of NB model is a mixed Gamma-Poisson model (NB2 or quadratic form)
- The Gamma distribution in general has two parameters: one governing scale or location and one for shape ($\theta$).
  - With $E(\delta_i) = 1.0$ for identification, we are interested in the shape parameter, $\theta$
- Degree of overdispersion in the conditional means is governed by $\theta$, and thus the variance of the NB distribution is a function of both the Poisson and Gamma variances

41

41

# Variance of NB (NB2) as the Poisson-Gamma Mixture

- A distribution that is Gamma($\lambda$, $\theta$) has variance $\lambda^2/\theta$
- The variance for the mixture is: $V(Y) = \lambda + \lambda^2/\theta$
- This represents an *indirect* relationship between $\theta$ and dispersion
  - As $\theta$ tends to infinity, the extra dispersion component tends to 0
- To capture a *direct* association, the NB model is often parameterized so that $\alpha = 1/\theta$ represents the extra dispersion: $V(Y) = \lambda + \alpha\lambda^2$
  - As $\alpha$ increases, so does the extra variance component
  - As $\alpha$ tends to 0, less extra dispersion is added, and the variance of the mixture converges to that of the Poisson
- Under this parameterization, $\alpha$ is referred to as the heterogeneity or NB dispersion parameter.
- There are other ways to express an NB model, but this representation is referred to as the *NB2* or simply the *NB* model

Gamma(scale, shape)

Poisson is a reduced form of the P-G mixture, thus allowing for nested comparisons of the models

42

42

# Generalized Poisson

NOTE: dispersion parameter is defined for *each family* of distribution; the same symbols do not necessarily refer to same quantity or parameter

- Alternative to the NB: Mixture of Poisson with Poisson = Generalized Poisson (Consul, 1998; Joe & Zhu, 2005; Harris, Yang & Hardin, 2012; SAS, 2022)

$$f(y_i, \lambda_i, \theta) = \lambda_i(\lambda_i + \theta y_i)^{y_i - 1} e^{-\lambda_i - \theta y_i} / y!$$
$$\text{for } \lambda_i > 0, \text{ and } [\max(-1, -\lambda_i / 4)] \leq \theta < 1$$

(NOTE: $\theta < 1$)

- The mean and variance of Y under GP distribution are:

$$E(Y_i) = \mu_i = \frac{\lambda_i}{1-\theta}$$

$$Var(Y_i) = \frac{\lambda_i}{(1-\theta)^3} = \frac{1}{(1-\theta)^2} E(Y_i) = \phi^2 E(Y_i)$$

When $\theta = 0$ there is equidispersion, and GP ➔ Poisson
When $\theta > 0$ we have over-dispersion
When $\theta < 0$ we have under-dispersion

- GP Regression Model: $\log\left(\frac{\lambda_i}{1-\theta}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots \beta_p x_{ip}$

43

43

# Clustering, and some additional issues

- Presence of clustering in the data can contribute to overdispersion
  - Random intercepts and slopes can be added, similar to standard HLM models
  - NB and GP include random effects at the lower-level (dispersion parameters) even after adjusting for clustering
- Some notes between *R* (glmmTMB) and *SAS (*Proc Glimmix)
- NB (slide 41)
  - *R* estimates what we've called $\theta$
  - SAS estimates what we've called $\alpha$, which is $1/\theta$
- GP (slide 42)
  - R estimates what we've called $\phi^2$ (and calls it "sigma"), where $\phi^2 = 1/(1-\theta)^2$
  - SAS estimates $\theta$, and labels as the "scale" parameter

44

44

# Assessing the ICC for Example Data

- ICC estimation for mixed-effects count models established by Leckie et al. (2020) and Austin et al. (2018)
- For our example data (Y = *profcounts* of K children assessed within schools), we have:
- Normal approximation (for crude comparison)
  - ICC = .24 (24%)
- Multilevel Poisson
  - ICC = .37 (37%)
- Multilevel NB
  - ICC = .28 (28%)

<u>Summary</u>: Once clustering is accounted for, the NB model indicates somewhat smaller estimated correlation between students from the same school, <u>but</u> 28% of the (marginal) variance in *profcounts* can be attributed to systematic difference between schools.





# Comparison of Competing Models

- Several ways to choose between competing models
  - Deviance statistics represent badness of fit or discrepancy between a model and the actual data
- Most common is through the *Likelihood Ratio Test*
  - As with other generalized and mixed linear models, a likelihood ratio test can be constructed to compare the *difference* -2LL between two nested models
  - Difference will have *df* equal to the difference in number of parameters between the models being compared
- Information criteria (AIC or BIC)
- Others: observed to predicted counts, graphical comparison, etc.
- Goal for today is to illustrate how these models work, clarify interpretation, provide for general comparison between them

# "Full" Model – Random Intercepts Poisson

$$\eta_{ij} = \beta_{0j} + \gamma_1(male) + \gamma_2(c\_momed) + \gamma_3(foodinsec) + \gamma_4(public) + \gamma_5(nbhoodprobs)$$

$$\beta_{0j} = \gamma_0 + u_{0j}$$

NB and GP are modifications of the P model

```
mlm_P2_tmb <- glmmTMB(profcount ~ male + c_momed +foodinsec
                        + public + nbhoodprobs
                        + (1 | schoolid),
                        family = poisson(link = "log"),
                        data=myctdata)
```

47

47

# glmmTMB code for the NB and GP models

```
mlm_NB2_tmb <- glmmTMB(profcount ~ male + c_momed + foodinsec
                        + public + nbhoodprobs
                        + (1 | schoolid),
                        family = nbinom2(link = "log"),
                        data=myctdata)
```

GLMM - NB

```
mlm_GP1_tmb <- glmmTMB(profcount ~ male + c_momed + foodinsec
                        + public + nbhoodprobs + (1 | schoolid),
                        family = genpois(link = "log"),
                        data=myctdata)
```

GLMM - GP

48

48

Table 1. Parameter estimates (standard errors) for proficiency count data using two-level random-intercept Poisson, Negative Binomial (NB2), and Generalized Poisson Models: Results in R using glmmTMB.

| | Random Intercepts Models | | |
|---|---|---|---|
| Parameter/Statistic | Random-Intercepts Poisson | Random-Intercepts Negative Binomial | Random-Intercepts Generalized Poisson |
| $\gamma_{00}$= *Intercept* | 1.12 (.04) [a] | 1.12 (.04) [a] | 1.14 (.04) [a] |
| $\gamma_{10}$ = *male* | −0.09 (.01) [a] | −0.09 (.02) [a] | −0.10 (.02) [a] |
| $\gamma_{20}$ = *c_momed* | 0.12 (.004) [a] | 0.12 (.01) [a] | 0.12 (.005) [a] |
| $\gamma_{30}$ = *foodinsec* | −0.20 (.03) [a] | −0.21 (.04) [a] | −0.21 (.04) [a] |
| $\gamma_{01}$ = *public* | −0.29 (.04) [a] | −0.28 (.04) [a] | −0.29 (.04) [a] |
| $\gamma_{02}$ = *nbhoodprobs* | −0.04 (.005) [a] | −0.04 (.005) [a] | −0.04 (.005) [a] |
| $\tau_{00}$ = school intercept variance | 0.13 (.01) [a,b] | 0.12 (.01) [a,b] | 0.12 (.01) [a,b] |
| Disp Param (R)<br>Disp Param (SAS) | | 8.7 (.009) [a,c]<br>.12 (.009) [a,c] | |
| GP Param (R)<br>GP Param (SAS | | | $\zeta$ = 1.31 (.008) [a,c]<br>$\zeta$ = 0.13 (.008) [a,c] |
| -2LL | 40182.89 | 39955.28 | 39801.72 |
| Pearson $\chi^2$/df | 1.19 | 0.98 | 0.92 |
| AIC | 40196.89 | 39971.28 | 39817.72 |

[a] $p < .0001$ [b] pvalues for random effects from SAS or in R used confint(model)
[c] pval in SAS or in R used confint(model, parm="sigma"); GP param in R rel to SAS = 1/(1-.13)

GP:

As Mom's education increases by 1 unit, ln(counts) increases by 0.12 (sig), holding all else constant.

The ln(counts) decreases by .29 for students in public school, holding all else equal.

For random effects, I used confint(model) to look at 95% CI for the standard dev of the variance component.

49

49

# Rate Ratios (RR) = exp(fixed effects)

**Table 2.**

Rate Ratios for the GP Model

| Fixed Effect | GP Rate Ratios (se) |
|---|---|
| $\gamma_{00}$= *Intercept* | 3.11 (.13) |
| $\gamma_{10}$ = *male* | 0.91 (.01) |
| $\gamma_{20}$ = *c_momed* | 1.13 (.006) |
| $\gamma_{30}$ = *foodinsec* | 0.81 (.03) |
| $\gamma_{01}$ = *public* | 0.75 (.03) |
| $\gamma_{02}$ = *nbhoodprobs* | 0.96 (.01) |

- Intercept: $\gamma_{00}$ = 1.14 (*sig*)
  - Exp(1.14) = 3.11
  - Expected proficiency count when all predictors are 0
    - Given random effects of 0
    - Dispersion estimate doesn't affect parameter interpretation, just tells us about unaccounted for variation around our estimates

Slope: *c_momed,* $\gamma_{20}$ = 0.12 (sig): Exp(0.12) = IRR = 1.13

As momed increases by 1 unit (i.e., from 5 = some college to 6 = Bach. Degree), the expected proficiency counts increase 1.13 times

- Controlling for other predictors in the model, and for a "typical" participant (random effects of 0)
- Represents a percentage change of 100(1.13 – 1) = 13% increase

50

50

# AIC comparison

- We used the bbmle package to compare AICs for the set of Level-1 random intercept models and Level-1 and Level-2 random intercept models
- Results show the lower AIC for the GP models, but it is not a very strong difference (biggest jump is between the NB Level-1 only and Level-1 and 2 models, compared to GP2)

```
> #install.packages("bbmle")
> library(bbmle)
>
> AICtab(mlm_P1_tmb, mlm_P2_tmb, mlm_NB1_tmb, mlm_NB2_tmb, mlm_GP1_tmb, mlm_GP2_tmb)
            dAIC  df
mlm_GP2_tmb   0.0 8
mlm_GP1_tmb 140.9 6
mlm_NB2_tmb 153.6 8
mlm_NB1_tmb 285.7 6
mlm_P2_tmb  379.2 7
mlm_P1_tmb  513.8 5
```

GP1, NB1, P1 are level-1 only random-intercept models

GP2, NB2, P2 are level-1 and level-2 random-intercept models

51

51

# Summary for Multilevel P, NB, GP

- Multilevel GP is a reasonable alternative to the NB
  - Easy to fit and interpret
  - Also appropriate for under-dispersion (while the NB is not)
- However, in our example the difference wasn't too dramatic
- Software parameterizations are not consistent across platforms, so some caution is needed when interpreting
- Next steps: Can we fit a better model by examining possible zero-inflation in the data?

52

52

## Next consideration: Zero-Inflation

- In some situations, count data may exhibit extra numbers of zeros, even beyond what might be expected through overdispersion.
- In ZI models, a mixture of two distributions is used to model two kinds of zeros as well as the counts
- The excess zeros may arise if the population consists of two subgroups of individuals
- Sampling zeros – Usually based on Poisson or NB for the count distribution, and these zeros are believed to occur by chance (according to that distribution
- Structural zeros – These arise from the other part of the mixture, for which subjects always produce zero counts.

53

53

# Zero Inflation Models

54

54

## General Form of the ZI model

- The general probability function for the ZI mixture is:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)p(y_i = 0 \mid \mu_i) & \text{for } y_i = 0 \\ (1-\pi_i)p(y_i \mid \mu_i) & \text{for } y_i > 0 \end{cases}$$

- The parameter $\pi$ is the proportion of individuals who would always return a zero (***structural*** zeros or "extra" zeros)
- Then 1 - $\pi$ represents the proportion who could be called "at risk" for the event. The count distribution selected for the ZI model governs the distribution of all these counts 0, 1, 2, … (*sampling* zeros)

55

55

## Zero-inflated Poisson (ZIP)

- We may have a collection of predictors governing either component of the mixture
- For example, if Poisson is used to govern distribution of the counts (0,1, 2, …) then:
  - Y = 0 with probability $\pi(z)$
  - Y ~ Poisson($\mu(x)$) with probability $1 - \pi(z)$
- In the ZIP model:
  - Assume a logistic model for $\pi(z)$

$$\text{logit}(\pi(z)) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \ldots \gamma_r z_r$$

  - Assume the Poisson log link is used for counts including 0

Predictors do not have to be the same for either component

$$\log(\mu(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$$

56

56

# ZINB

- A similar structure is obtained when the NB model is used to govern the counts
  - Replace the log(μ(x)) with the usual NB regression model
- A challenge with ZI models is that there are many different ways to include predictors within each component
- For GLMMs, we could also include or exclude random components
- MANY different variations
  - A simplifying approach is to fit the ZI model allowing for a single zero-inflation parameter that is independent of predictors (no "z" predictors as on previous slide!)
- Overall goal is to find a well-fitting model that explains the "extra" zeros as well as the distribution of all counts (including zeros on the count side)

57

57

# Example 2 Data: School Success Profiles

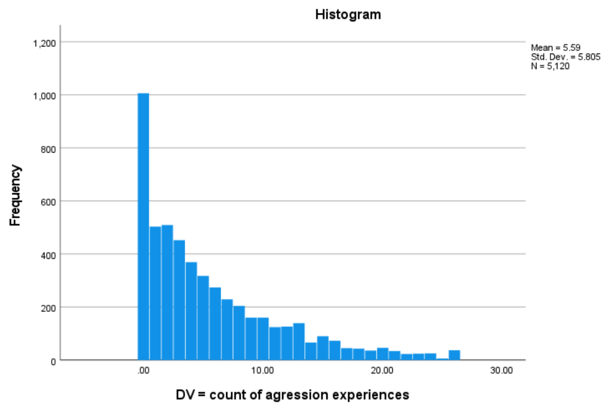
- The SSP was developed by a team of researchers at UNC as part of a district-wide effort for data-informed approaches to address student needs.
- For these data, n = 5171 students from grades 6 – 9 in 17 schools completed the SSP.
- DV = counts of aggression/microaggression events reported by students at the school in the past 30 days

Data file: SSP2017_M3_subset.sav

58

58

# SSP – Counts of Aggression/Microaggression Events




(Not adjusted for clustering by school)
Proportion of zeros is 19.6%

**Statistics**

s17TOT

| | | |
|---|---|---|
| N | Valid | 5120 |
| | Missing | 51 |
| Mean | | 5.5877 |
| Median | | 4.0000 |
| Mode | | .00 |
| Std. Deviation | | 5.80514 |
| Variance | | 33.700 |
| Range | | 26.00 |
| Minimum | | .00 |
| Maximum | | 26.00 |

(s17TOT)

*var > mean*

$$\frac{var}{mean} = 6.03$$

59

59

# Student-level Predictors (n = 5171)

**Statistics**

| | | lgdevpeers | Lang01oth | famsupp | boy1grl0 |
|---|---|---|---|---|---|
| N | Valid | 5073 | 5149 | 5064 | 5148 |
| | Missing | 98 | 22 | 107 | 23 |
| Mean | | .2621 | .0893 | 2.5606 | .5002 |
| Median | | .2007 | .0000 | 2.7143 | 1.0000 |
| Mode | | .00 | .00 | 3.00 | 1.00 |
| Std. Deviation | | .27407 | .28526 | .51637 | .50005 |
| Variance | | .075 | .081 | .267 | .250 |
| Range | | 1.10 | 1.00 | 2.00 | 1.00 |
| Minimum | | .00 | .00 | 1.00 | .00 |
| Maximum | | 1.10 | 1.00 | 3.00 | 1.00 |

- For **lgdevpeers**, we used log(number of friends engaging in deviant/aggressive behaviors)
- Almost 9% of students spoke language **other** than English at home
- **Family support** is a multi-item scale with ratings on strength of support (3 = high perceived support)
- 50% male (**boy1grl0**)

60

60

# School-level Predictors (J = 17)

**Statistics**

| | | L2_schlsupp | MS0HS1 |
|---|---|---|---|
| N | Valid | 17 | 17 |
| | Missing | 0 | 0 |
| Mean | | 3.1451 | .3529 |
| Median | | 3.1296 | .0000 |
| Mode | | 2.83[a] | .00 |
| Std. Deviation | | .18394 | .49259 |
| Variance | | .034 | .243 |
| Range | | .87 | 1.00 |
| Minimum | | 2.83 | .00 |
| Maximum | | 3.69 | 1.00 |

a. Multiple modes exist. The smallest value is shown

- **L2_schlsupp**: Average student ratings on perceived support from school/teachers
  - Items were rated from 1 to 4; 4 indicates greater perceived support
- **MS0HS1**: About 35% of the schools were High Schools (grades 9 – 12); 65% were Middle Schools (grades 6 – 8)

61

61

# R Code for the ZI models

Equation 1:

```
### random intercept Poisson ZIP (Poisson1)
eq1 <- s17TOT ~ (lgdevpeers + Lang01oth + famsupp + boy1grl0 + (1 | SchID2))
```

Equation 2:

```
### with two level-2 predictors (Poisson2)
eq2 <- s17TOT ~ (L2_schlsupp + MS0HS1 + lgdevpeers + Lang01oth + famsupp + boy1grl0 + (1 | SchID2))
```

Full Model GLMM ZIP

```
zpoisson4 <- glmmTMB(eq2,
                     data=myctdata,
                     na.action=na.exclude,
                     ziformula=~1,
                     family = poisson)
summary(zpoisson3)
```

Full Model GLMM ZINB

```
znbmod4 <- glmmTMB(eq2,
                   data=myctdata,
                   na.action=na.exclude,
                   ziformula=~1,
                   family = nbinom2)
summary(znbmod4)
```

62

62

# Results – Multilevel ZIP and ZINB

Table 2. Parameter estimates (standard errors) for SSP count data using two-level random-intercept ZIP and ZINB: Results in R using glmmTMB.

| | Random Intercepts Models | | Random Intercepts Models | |
|---|---|---|---|---|
| Parameter/Statistic | Poisson | ZIP | NB | ZINB |
| Conditional Model | | | | |
| *Intercept* | 3.06 (.47)*** | 3.06 (.31)*** | 3.07 (.53)*** | 3.14 (.04)*** |
| *L2_schlsupp* | -0.31 (.15)* | -0.28 (.10)** | -0.32(.16) | -0.32(.14)* |
| *MS0HS1* | -0.25 (.04)*** | -0.20 (.04)*** | -0.23 (.06)*** | -0.22 (.05)*** |
| *lgdevpeers* | 0.86 (.02)*** | 0.63 (.02)*** | 0.93 (.06)*** | 0.83 (.06)*** |
| *Lang01oth* | -0.05 (.02)* | -0.02 (.02) | -0.08 (.06) | -0.06 (.05) |
| *famsupp* | -0.26 (.01)*** | -0.19(.01)*** | -0.27(.03)*** | -0.25(.03)*** |
| *boy1grl0* | 0.14 (.01)*** | 0.16 (.01)*** | 0.15 (.03)*** | 0.17 (.03)*** |
| $\tau_{00}$ = school intercept variance | 0.01 | 0.003 | 0.006 | 0.003 |
| Zero-Inflation Model | | | | |
| *Intercept* | | -1.44 (.03)*** | | -2.22 (.10)*** |
| Model Info | | | | |
| NB disp Param | | | 1.02 | 1.48 |
| Pearson $X^2$/df | 26841.33/4991 = 5.38 | | 4548.116/4990 = .91 | |
| -2LL | 39845.0 | 34170.5 | 27552.5 | 27450.0 |
| AIC | 39861.0 | 34188.50 | 27570.5 | 27476.1 |

For ZI Interpretation

Estimated mean function is in the "conditional model" section

Estimated P(excess 0) is in "ZI model" section

Prob: Odds/(1+Odds)

For ZIPoisson:
P(excess 0) = 19.26%

For ZINB:
P(excess 0) = 9.78%

Actual P(0) = 19.6%

63

# Interpretation/Summary

- Estimated mean function for the counts is in the "Conditional Model" section
- Patterns are similar among all 4 models although variable significance is affected by choice of model
- Poisson does a poor job of predicting counts, given the large proportion of "excess" zeros predicted by the model
- ZINB has smallest deviance of these 4 models
- Remember – there are MANY ways to modify either/both parts of this model!
  - Ziformula =~ 0
  - Ziformula =~1
  - Ziformula =~.
  - Or can specify directly in this statement (or through setting previous equation)

64

64

# Hurdle Models





## Hurdle Models

- Assumes that all zeros are structural zeros
- One model estimates proportion of structural zeros
  - Typically a logit model as with ZI
- Second model estimates counts starting with 1, 2, 3, etc (no zeros!)
  - Truncated count distribution, typically P or NB
- General Structure:

Counts 1, 2, 3, etc., are predicted only if the case "crosses the hurdle"

$$P(Y_i = y_i) = \begin{cases} \pi_i & \text{for } y_i = 0 \\ (1-\pi_i)\dfrac{P(y_i \mid \mu_i)}{1 - P(y_i = 0 \mid \mu_i} & \text{for } y_i > 0 \end{cases}$$

$\pi_i$ is the probability of a case belonging to the zero component

$P(y_i \mid m_i)$ corresponds to the probability function for either P or NB counts

# Results – Hurdle Models

Table 3. Parameter estimates (standard errors) for SSP count data using two-level random-intercept Hurdle models. Results in R using glmmTMB, using **ziformula** = ~**1**

| | Random Intercepts Models | | Random Intercepts Models | |
|---|---|---|---|---|
| Parameter/Statistic | Poisson | Hurdle Poisson | NB | Hurdle NB |
| Conditional Model | | | | |
| *Intercept* | 3.06 (.47)*** | 3.03 (.31)*** | 3.07 (.53)*** | 2.90 (.04)*** |
| *L2_schlsupp* | -0.31 (.15)* | -0.27 (.10)** | -0.32(.16) | -0.25 (.11)* |
| *MS0HS1* | -0.25 (.04)*** | -0.19 (.03)*** | -0.23 (.06)*** | -0.19 (.04)*** |
| *lgdevpeers* | 0.86 (.02)*** | 0.62 (.02)*** | 0.93 (.06)*** | 0.70 (.05)*** |
| *Lang01oth* | -0.05 (.02)* | -0.02 (.02) | -0.08 (.06) | -0.03 (.05) |
| *famsupp* | -0.26 (.01)*** | -0.20 (.01)*** | -0.27(.03)*** | -0.22(.03)*** |
| *boy1grl0* | 0.14 (.01)*** | 0.16 (.01)*** | 0.15 (.03)*** | 0.19 (.03)*** |
| $\tau_{00}$ = school intercept variance | 0.01 | 0.003 | 0.006 | 0.0005 |
| Zero-Inflation Model | | | | |
| *Intercept* | | -1.42 (.04)*** | | -1.42 (.10)*** |
| Model Info | | | | |
| NB disp Param | | | 1.02 | 1.57 |
| Pearson $X^2$/df | 26841.33/4991 = 5.38 | | 4548.116/4990 = .91 | |
| -2LL | 39845.0 | 34183.8 | 27552.5 | 27565.5 |
| AIC | 39861.0 | 34201.8 | 27570.5 | 27585.5 |

- Note zero-component (logit) are the same, as all zeros are treated as structural zeros
- The Hurdle NB does not have the lowest AIC here

67

67

# Comparing AIC across ZIP and Hurdle models

```
> AICtab(poisson4, zpoisson4, hpoisson4, nbmod4, znbmod4, hnbmod4, hnbmod5)
          dAIC    df
hnbmod5      0.0 17
znbmod4    133.3 10
nbmod4     227.8 9
hnbmod4    242.8 10
zpoisson4 6845.8 9
hpoisson4 6859.0 9
poisson4 12518.2 8
```

- Here I added a model that used the same equation (eq2) for the zero-inflation component (see results next page)
- Hurdle for this new model has best fit criteria based on the AIC
- There are so many options for how to fit both components!

68

68

# Hurdle with ziformula = ~.   (eq2)

Section 1

```
> summary(hnbmod5)
 Family: truncated_nbinom2  ( log )
Formula:          s17TOT ~ (L2_schlsupp + MSOHS1 + lgdevpeers + Lang01oth + famsupp +
    boy1gr10 + (1 | SchID2))
Zero inflation:          ~.
Data: myctdata

     AIC      BIC   logLik deviance df.resid
 27342.8  27453.5 -13654.4  27308.8     4982

Random effects:

Conditional model:
 Groups Name        Variance  Std.Dev.
 SchID2 (Intercept) 0.0005259 0.02293
Number of obs: 4999, groups:  SchID2, 17

Zero-inflation model:
 Groups Name        Variance Std.Dev.
 SchID2 (Intercept) 0.06853  0.2618
Number of obs: 4999, groups:  SchID2, 17

Dispersion parameter for truncated_nbinom2 family (): 1.57
```

69

69

# Hurdle with ziformula = ~.   (eq2)

Section 2

```
Conditional model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.90017    0.36250   8.000 1.24e-15 ***
L2_schlsupp -0.25040    0.11262  -2.223   0.0262 *
MSOHS1      -0.19302    0.03900  -4.950 7.44e-07 ***
lgdevpeers   0.70381    0.05428  12.967  < 2e-16 ***
Lang01oth   -0.02600    0.05313  -0.489   0.6245
famsupp     -0.21842    0.02805  -7.788 6.81e-15 ***
boy1gr10     0.18663    0.02877   6.488 8.73e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zero-inflation model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.86218    1.49112  -1.919   0.0549 .
L2_schlsupp  0.16675    0.47386   0.352   0.7249
MSOHS1       0.26312    0.17259   1.525   0.1274
lgdevpeers  -1.91513    0.17538 -10.920  < 2e-16 ***
Lang01oth    0.24342    0.12577   1.935   0.0529 .
famsupp      0.44734    0.08333   5.368 7.95e-08 ***
boy1gr10     0.10716    0.07352   1.458   0.1449
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Same pattern of effect for the mean structure (conditional model)

**NOTE: Zero-inflation component is predicting P(0)**

Students' exposure to peers engaging in deviant behavior yields lower probability of count = 0

Higher family support yields higher probability of count = 0

70

70

# Guidance for Researchers

- Distinction between zeros – important
- Typology – there are other possibilities (too many) and MANY variations on the models presented here
- Residuals – complex form for GLMM for ZI and Hurdle
  - More research needed in this area
- Graphical methods not easily adapted for GLMM because of random effects, may want to focus on population-averaged versions (marginal effects) for these models
- Need large sample size for mixed models – particularly true for counts or limited dependent variables
- GLMMTMB and GLMMAdaptive – both work well, steeper learning curve than GLM packages or other mixed model packages
- Challenges with convergence (not for us but we know of many – especially with smaller sample sizes or complex models, interactions terms, random slopes, etc. )

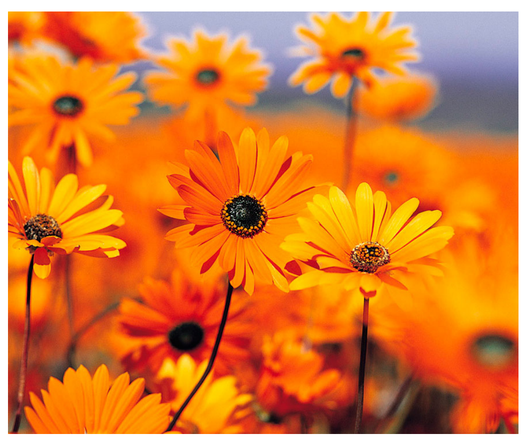
71

71

# Thank you!

- Please visit my website for data and code

M3 presentation, data and R code are at:
**https://go.osu.edu/aoconnell_website**



Ann A. O'Connell

Oconnell.87@osu.edu

References follow …

72

72

# References for M3 (1)


- Austin, P. C., Stryhn, H., Leckie, G., & Merlo, J. (2018). Measures of clustering and heterogeneity in multilevel Poisson regression analyses of rates/count data. *Statistics in Medicine*, *37*(4), 572-589.
- Bilder, C. R., & Loughin, T. M. (2014). *Analysis of categorical data with R*. CRC Press.
- Bolker, B., et al. (2021, June 28). GLMM FAQ. https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html
- Bowen, N.K., Lucio, R., Patak-Pietrafesa, M., & Bowen, G. (2020) The SSP 2020: The Revised School Success Profile. Children & Schools, 42(1).
- Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.
- Consul, PC & Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics, 21(1*), 89-109
- Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R* (Vol. 53). New York: Springer.
- Fox, J. (2017). *Applied regression analysis and generalized linear models* (3rd ed.). Sage.
- Harris, T, Yang, Z, & Hardin, JW. Modeling underdispersed count data with generalized Poisson regression. *The Stata Journal, 12*(4), 736-747






# References for M3 (2)


- Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.
- Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.
- Joe, H & Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with Negative Binomial Distribution. *Biometrical Journal, 2*, 219-229.
- Leckie, G., Browne, W. J., Goldstein, H., Merlo, J., & Austin, P. C. (2020). Partitioning variation in multilevel models for count data. *Psychological methods*.
- Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Sage.
- Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed). Stata Press.
- McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models, vol. 37 CRC Press.
- O'Connell, A. A., McCoach, D. B., & Bell, B. A. (Eds.). (2022). *Multilevel Modeling Methods with Introductory and Advanced Applications*. IAP.
- Snijders, T.A.B. & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd Ed.). Sage.
- Stroup, W.W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. CRC Press