Generalized Linear Mixed Models

Elizabeth A. Claassen, PhD
Senior Research Statistician Developer

JMP STATISTICAL DISCOVERY

Generalized Linear Mixed Models

Modeling non-Normal data with random effects

- Mixed Modeling has been the standard for analyzing data with more than one source of random variation (blocking, split-plots, etc.).
- The Linear Mixed Model (LMM) assumes the response is continuous with no bounds.
- What if your response is a discrete count? Or a binary response? Or a proportion in terms of y/n?

- Enter the Generalized Linear Mixed Model (GLMM)

Generalized Linear Model (GLM)

- Examples of GLMs
  - Logistic regression
  - Poisson regression
  - Normal regression
  - Analysis of variance models

GLMM – Defining Elements

- Distribution $$\mathbf{y} \mid \mathbf{b} \sim f(\boldsymbol{\mu}, \Sigma)$$
  - exponential family
- Linear Predictor $$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$$ $$\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$$
- Link $$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

- Linear predictor is the mixed model; the distribution and link function allow for non-Gaussian data

Motivating Example

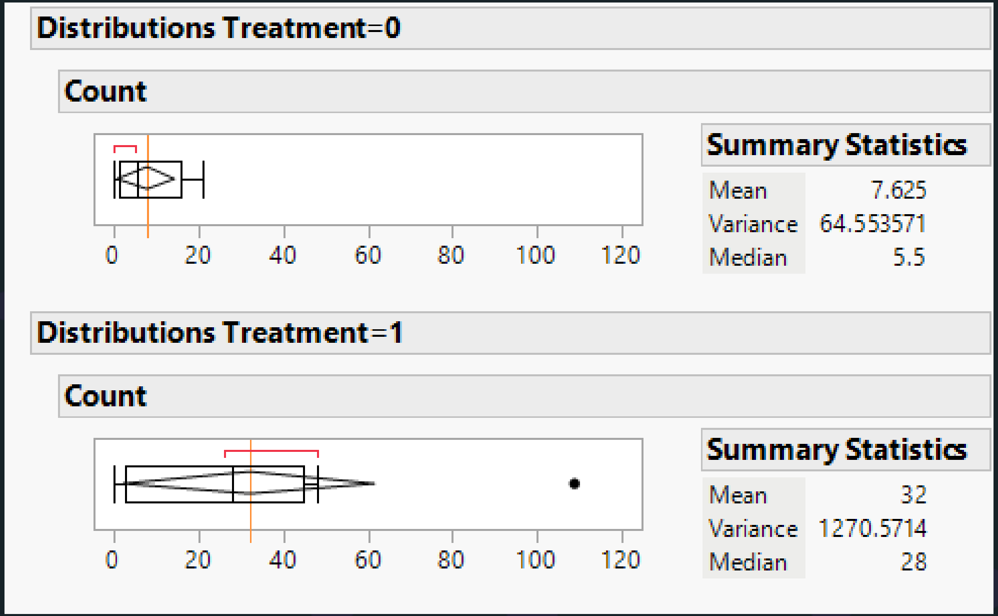
- Paired Comparison Experiment:
  - a.k.a. Randomized Complete Block Design
  - 8 Pairs / Blocks / Clinics
  - 2 Treatments – “Treatment 0” “Treatment 1”

    e.g. **“control”** & **“test”**
  - Response: **count**

    e.g. “**obs**” = **0, 1, 2, ...;** number of patients / claims / defects

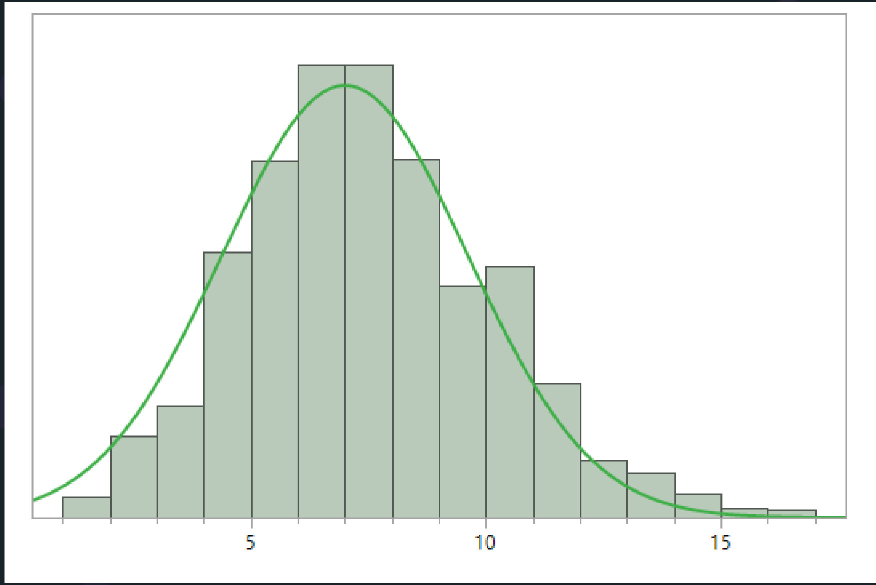
Example: The Data

Clinic	Treatment_0	Treatment_1
1	1	36
2	5	109
3	21	30
4	7	48
5	2	0
6	6	2
7	0	5
8	19	26



What Distribution?

Poisson λ=7



Count ~ Normal, ANOVA with count is okay, right?

Two Things a Model Must Do

- Plausibly describe the process that gives rise to the observed data
  - how explanatory variables affect response
  - probability distributions involved
- Allow / **facilitate** addressing the objective that motivated collecting the data
  - test a hypothesis
  - make a decision
  - estimate a parameter

Linear Model for RCBD Count Data

- ANOVA – linear model for RCBD
- Model: count = intercept + treatment + block + residual
  - $count_{ij} = \mu + \tau_i + b_j + e_{ij}$
- Implement in JMP with Standard Least Squares or JMP Pro with Mixed Model

ANOVA – selected results

Response Count

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Treatment	1	1	7	3.6387	0.0981

Multiple Comparisons for Treatment

Least Squares Means Estimates

Treatment	Estimate	Std Error	DF	Lower 95%	Upper 95%
0	7.625000	9.1348406	13.993	-11.96814	27.218143
1	32.000000	9.1348406	13.993	12.40686	51.593143

Problems with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}$

- Assumes $\mathbf{X}\boldsymbol{\beta}$ estimates $\mathrm{E}(\mathbf{y}) \propto \boldsymbol{\lambda}$
- $\hat{\lambda}$ must be > 0
- No guarantee $0 < \mathbf{X}\hat{\boldsymbol{\beta}}$
  - e.g. regression provides easy examples
- Logical issues
  - Poisson assumptions aren't the same as LMM

$$\begin{array}{ccc} E(y|b) = \lambda & & E(y|b) = X\beta \\ & \neq & \\ Var(y|b) = \lambda & & Var(y|b) = \sigma^2 \end{array}$$

- "Residual" has no meaning
- We need a better approach

Blocked Design: A Closer Look

"Experiment" (Study) Design		
Block	Unit	
Block 1		
Block 2		
Block 3		
Block 4		
Block 5		
Block 6		
Block 7		
Block 8		

Treatment Design	
0	1

Full Design		
Block	Unit	
Block 1	0	1
Block 2	1	0
Block 3	0	1
Block 4	0	1
Block 5	1	0
Block 6	1	0
Block 7	1	0
Block 8	0	1

Repurposed ANOVA Table

Experiment		Treatment		Combined	
Source	d.f.	Source	d.f.	Source	d.f.
block	7			block	7
		trt	1	trt	1
unit(block)	8*(2-1) =8	“parallels”	14	unit(block) | trt a.k.a. “residual” a.k.a “blk x trt”	8-1=7
Total	15	Total	15	Total	15

JMP STATISTICAL DISCOVERY

Repurposed ANOVA & Sensible Model

sensible model ➔ one-to-one ANOVA effect – model parameter match

combined		model			
Source	d.f.	LMM	naive GLM(M)	Poisson GLM(M) w/unit	Negative Binomial GLMM
block	7	b_j	b_j	b_j	b_j
treatment	1	τ_i	τ_i	τ_i	τ_i
unit(block) | trt block x trt "residual"	7	e_{ij} or σ^2	here's the problem ➔ overdispersion likely	bt_{ij}	ϕ
total	15				

Overdispersion: model fails to adequately account for variation in the data

Consequence: confidence intervals too narrow; inflated type I error rate

jmp STATISTICAL DISCOVERY

GLMM – Defining Elements

- Distribution $\mathbf{y} \mid \mathbf{b} \sim f(\boldsymbol{\mu}, \Sigma)$
  - exponential family
- Linear Predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$

$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$$

- Link $\boldsymbol{\eta} = g(\boldsymbol{\mu})$

- Linear predictor is the mixed model; the distribution and link function allow for non-Gaussian data

Repurposed ANOVA ➔ appropriate GLMM

combined	
Source	**d.f.**
block	7
treatment	1
unit(block) | trt block x trt "residual"	7
total	**15**

$$\Rightarrow b_j \text{ i.i.d. } N\left(0,\sigma_B^2\right); bt_{ij} \text{ i.i.d. } N\left(0,\sigma_{BT}^2\right)$$

$$\text{Linear predictor: } \eta+\tau_i+b_j+\left(bt\right)_{ij}$$

$$\text{Link: } \eta_{ij}=\log\left(\lambda_{ij}\right)$$

$$y_{ij} \mid b_j, bt_{ij} \sim \text{ind Poisson}\left(\lambda_{ij}\right)$$

$$\hat{\lambda}_i=\exp\left(\hat{\eta}+\hat{\tau}_i\right)$$

JMP Demo

Example 1

- *SAS for Mixed Models (2018),* Example 11.5; from Beitler & Landis (*Biometrics*, 1985)
- Multi-location clinical trial
- 8 clinics, two treatments: "CNTL" and "DRUG"
- n_{ij} patients assigned to treatment *i* at clinic *j*
- Response variable y_{ij} is number of patients with a favorable outcome
- Objective: does "DRUG" increase probability of favorable outcome & if so, how much?

Repurposed ANOVA Table		
SOURCE	**DF**	**MODEL EFFECT**
clinic	7	$c_j \sim N(0, \sigma_c^2)$
treatment	1	τ_i
group(clinic) | trt a.k.a. clinic × trt	7	$ct_{ij} \sim N(0, \sigma_{ct}^2)$
TOTAL	**15**	

Resulting GLMM

- distribution of observations: $y_{ij} | c_j, ct_{ij} \sim Binomial(n_{ij}, p_{ij})$
- logit link function: $\eta_{ij} = log\left(\frac{p_{ij}}{1-p_{ij}}\right)$
- linear predictor: $\eta_{ij} = \eta + \tau_i + c_j + ct_{ij}$

Example 2

- *SAS for Mixed Models (2018),* Example 12.3
- Multi-source, random coefficient regression
- 8 lots
- Amounts $X_1 = 0, X_2 = 2, X_3 = 4, \ldots X_6 = 10$ of finishing treatment applied to samples from each lot
- Response variable y_{ij} is number of aberrant micro-sites on finished product for amount i, lot *j* – discrete count
- Objectives:
    - estimate effect of increasing amount of finishing treatment on aberrant micro-site count
    - estimate above via linear regression
    - determine amount of finishing treatment required to assure expected aberrant micro-site count ≤ 10

Example 2, continued

Repurposed ANOVA Sources of Variation & Resulting Model Effects			
Study (Experiment) Design	**Treatment Design**	**Combined**	**Linear Regression Model Effect**
Lot		Lot	$B_{0j} + b_{1j}X_{1j} = \beta_0 + b_{0j} + b_{1j}X_{1j}$
	Amount	Amount	$\beta_1 X_j$
Sample (Lot)		sample(lot)|amount	$s_{ij} \sim N(0, \sigma_s^2)$

Terminology & Assumed Distributions

- b_{0j} called random intercept
- b_{1j} called random slope

together account for LOT

- random intercept & slope potentially correlated
- $\begin{bmatrix} b_{oj} \\ b_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}\right)$
- often assume $\sigma_{01} = 0$
- e.g. with only 8 lots, may not have enough replication to estimate σ_{01}

Resulting GLMM

- distribution of observations: $y_{ij} | b_{0j}, b_{1j}, s_{ij} \sim Poisson(\lambda_{ij})$
- log link function: $\eta_{ij} = log(\lambda_{ij})$
- linear predictor: $\eta_{ij} = \beta_0 + b_{0j} + (\beta_1 + b_{1j})X_i + s_{ij}$

Further Resources

SAS for Mixed Models: Introduction and Basic Applications (2018), Stroup, Milliken, Claassen and Wolfinger

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications (2012), Stroup

Statistically Speaking Webinar: *The "What, Why, and How" of Generalized Linear Mixed Models,* Stroup and Claassen