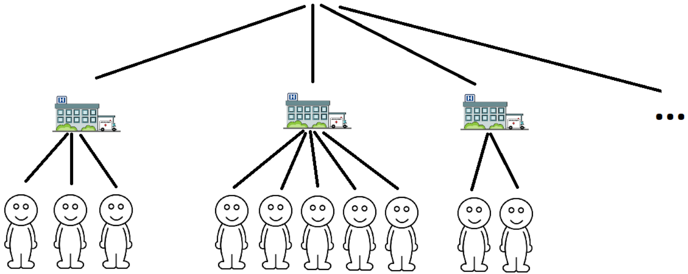
Lecture 13: Clustered Data, An Introduction to Applied Mixed Models

Author: Nick Reich / Transcribed by Gregory Guranich, edited by Bing Miu

Course: Categorical Data Analysis (BIOSTATS 743)

Clustered Data

- **Clustered Data:**
- Hierarchy, nested populations
- Longitudinal, Correlated observations or sets .eg repeated measure
- **Example: (Hierarchy of nested populations)**
- Patient $\in$ Hospital $\in$ Region



Clustered Data

- **Example: (Longitudinal)**
- Repeated measurements on the same unit of observation
  - Patients have multiple temperature measurements over time
  - Temperature $\in$ Patient $\in$ Clinic

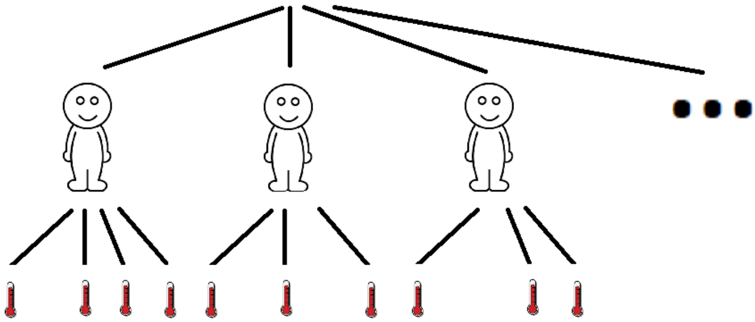
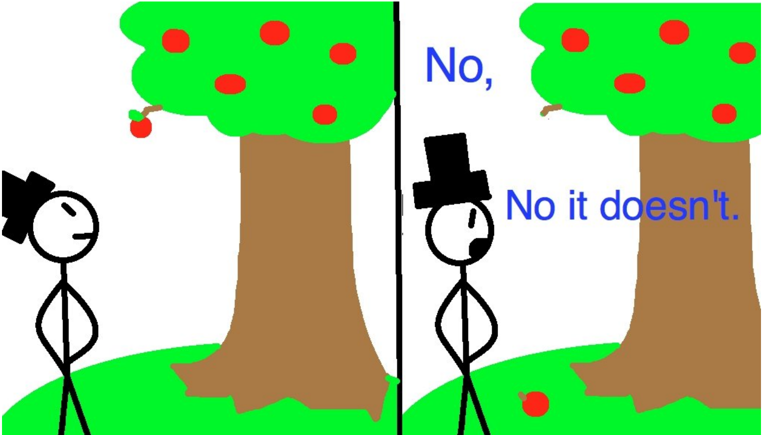


Figure 1: Alkema 2017

Implications of Clustered Data

- Observations in clusters violates our assumptions of independence
- Clustered data are less informative and less generalizable than independent data
- **Example:** Subjects within a household are more similar and will vary less. Similar genetics and behavior



Example Notation

Mixed models, also known as multilevel or hierarchical models, are used to model cluster data. We will use the following notation to model such data.

Y_{ij} = Response variable

Trt_{ij} = Treatment variable

Example Notation 2

$$j = 1,\ 2,\ \ldots,\ n$$

$$i = 1,\ 2,\ \ldots,\ n_i$$

$$N = \sum_i n_i = \text{total number of groups}$$

$$Treatment \quad Trt_{ij} = \begin{cases} 1, & \text{if } Trt_{ij} \text{ was treated} \\ 0, & \text{otherwise} \end{cases}$$

Treatment $_{ij}$	Region $_i$	Mortality Count	
1	1	72	
0	1	256	$\bar{y}_1$
1	1	91	
0	2	11	
1	2	12	$\bar{y}_2$
1	2	7	
1	2	13	

Example to begin with

- Let Y_{ij} denotes the counts of mortality

$$Y_{ij} \sim Poisson(\lambda_{ij}, P_i)$$

$$log(\lambda_{ij}) = \beta_0 + \beta_1 Trt_{ij} + \beta_2 X_i$$

- We want to make inference about β_1. What is the treatment effect?
- Observations are not iid, need "adjustment"
- How can we account for variance within groups?
  - 1. Marginal models with generalized estimating equations (GEE) for variance adjustment
  - 2. generalized linear mixed models (GLMM's)

GLMM's Models

- Account for modeling for individual and group level variation in estimating group-level coefficients
- Allow the proper measure of variation in individual level regression coefficients
- For point estimates, Shrinkage (relative to sample size) of parameters toward group means
- Rule of thumb
  - number of groups greater than 5
  - substainial variation among groups

GLMMS Model Notation

- where β is a fixed effect and μ_i are varying (random) effects

$$g(E[y_{ij}|\mu_i]) = X_{ij}^T \beta + Z_{ij}^T \mu_i$$
$$\mu_i \sim N(0, G_\theta)$$

- Agresti uses i to represent the group, Y_{ij} is the *jth* observation in *ith* group
- So here we have i different μ which are draws from a normal
- The μ_i have a common distribution
- Notice β has no subscript, β is fixed
- β is a global estimate and μ_i is a group specific estimate
- g is a glm link
- θ is a parameter that governs the distribution of the random effect

Estimation

$$
\begin{aligned}
L(\beta,\theta) &= f(\vec{y}|\beta,\theta) \\
&= \int f(\vec{y}|\vec{\mu}) f(\vec{\mu}) du
\end{aligned}
$$

- where we integrate across our marginal $f(\hat{\mu})$
- these problems usually do not have closed form solutions
- How to approximate? Use Bayesian MCMC or HMC algorithims to sample from posterior distribution of β and θ
- approximate with Laplace methods (some coefficients are subject to penalty terms)
- **Degrees of Freedom:** Approximated by estension of the Hat matrix
  - some closed for solutions (ex: beta-binomial conjugate in next lecture)

$$
\begin{aligned}
tr(H) &= p \\
p \leq tr(H) &\leq p + q
\end{aligned}
$$

Example

- Lets say we are looking at treatment(spinal implants) to relieve back pain. Rows represent visit 1 and columns represent visit 2

visit1/visit2	success	failure
success	63	16
failure	12	35

Figure 3:

- pain indicator Y_{ij} and group indicator X_{ij}

$$y_{ij} = \begin{cases} 1 & \text{patient j at visit i has no pain} \\ 0 & \text{patient j at visit i has pain} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{i=2; patient's second visit} \\ 0 & \text{i=1; pateint's first visit} \end{cases}$$

Example (possible models)

- logistic-normal model

$$logit(P(Y_{ij} = 1|\mu_i)) = \alpha + \beta X_{ij} + \mu_i$$
$$\mu_i \sim N(0, \sigma_\mu^2)$$

- Similar notation for bayesians

$$Y_i \sim N(\pi, \sigma_y^2)$$
$$logit(\pi) = \alpha_i + \beta X_{ij}$$
$$\alpha_i \sim N(\beta_0, \sigma_\alpha)$$
$$\beta_0 \sim N(0, 100)$$
$$\sigma_\alpha \sim U(0, 5)$$

Example (interpretation)

$$logit(P(Y_{ij} = 1|\mu_i)) = \alpha + \beta X_{ij} + \mu_i$$
$$\mu_i \sim N(0, \sigma_\mu^2)$$

- $\alpha \sim$ log odds of pain free at visit 1
- $\beta \sim$ change in log odds of being pain free comparing visit 2 to visit 1

Example (interpretation)

- With mixed effect μ_i our intercept can now vary