Multi-level Analysis I
Recognizing the Problem

Maureen Smith, MD PhD
Dept. of Population Health Sciences
University of Wisconsin-Madison

June 5, 2004

A day in the life of a researcher

- We have data
  - ID (observation #)
  - X (variable 1)
  - Y (variable 2)

- We want to use the value of X to explain the value of Y

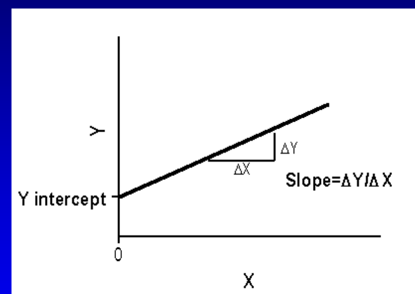
ID	X	Y
1	60	3
2	75	6
3	81	10
4	70	7
5	65	5

Welcome to the fantasy world of linear regression

- A simple model

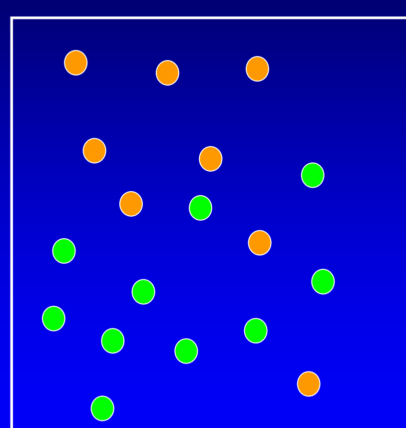
  y_i = intercept + slope(x_i) + error
  i indicates observations (1…N)

- Assumptions
  - Linearity
  - Independence
  - Normality
  - Constant variance



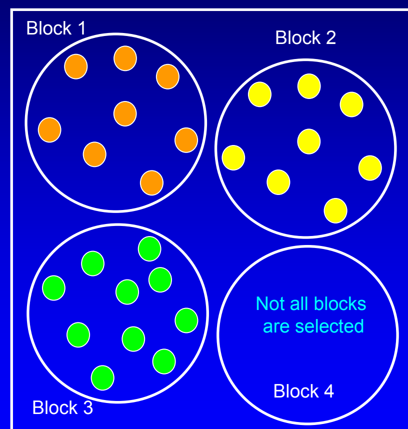
Reality check

- How often are observations truly independent from one another?
  - Dot indicates geographic location of teenager
  - Orange or green indicates hair color
- Do these teenagers look independent?



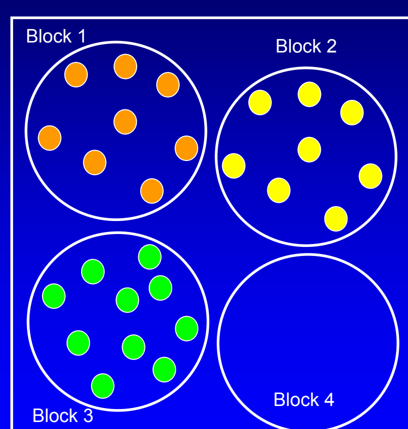
1) Clustering introduced in sampling

- Multistage sampling
  - Circles represent city blocks
  - Blocks randomly sampled
  - All persons in block surveyed to determine attitudes
- Persons in one block are more like their neighbors than persons who live in another block
- **Nesting** or **clustering** of data
  - Persons within blocks



Effect of sample design on errors

- Errors in linear regression
  - Assume independence
  - Each person => info
  - Each person worth "1"
- If clustering occurs
  - Obs not independent
  - Each person => less info
  - Each person worth < "1"



Simple linear regression won't work!

- Violates assumption of independence
- If don't account for it
  - Standard errors are too small
  - Makes coefficients look more significant
  - "You think there is more information in the data than actually exists"

How much information is lost?

"Design Effect"

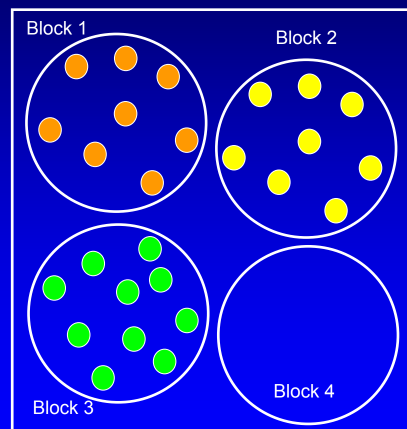
- If designing a study using multistage sampling, need to increase sample size to account for loss of information
- Design effect
  - Each observation is "worth less"
  - Need to estimate your "effective" sample size
  - Used for sample size calculations in multi-stage sampling

$$N_{effective} = \frac{N_n}{\text{Design effect}}$$

Questions – Pair up!

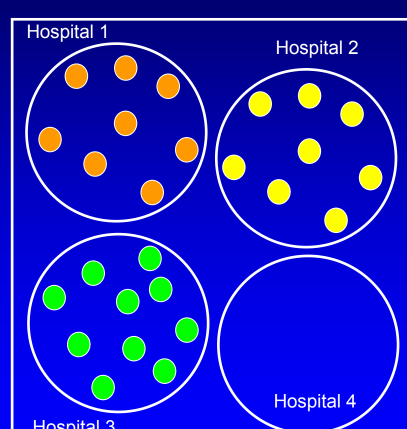
- Multi-stage sample design
  - City blocks N= 3
  - Persons N=26
- Design effect = 2

1. What is the effective sample size?
2. What sample size would you use in your power calculations?



2) Clustering introduced naturally

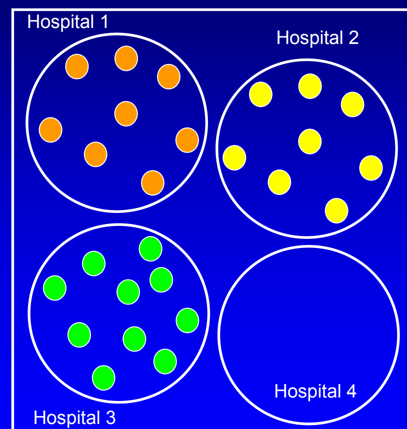
- Analyze costs of care for hospitalized patients
- Patients in one hospital are more alike than patients in another hospital
- **Nesting** or **clustering** of data
  - Patients within hospitals



Effect of natural clusters on errors

- **<u>Same effect</u>** on errors
  - Obs not independent
  - Each person => less info
  - Each person worth < "1"

- Simple linear regression won't work!



What do we do?

- First question - do we care?
  - Is clustering a nuisance?

    OR
  - Is clustering an interesting phenomenon?

- Leads to different analytic strategies

If clustering is a nuisance

- Example - Multi-stage sampling
  - Don’t care how people vary within city blocks versus between city blocks
  - Artificially imposed by the sampling design
  - Not interested in measuring it
  - Just want to correct for it
- Use analytic strategies that correct for clustering

How to correct errors for clustering

- Robust estimates of variance
  - Stata “, robust cluster (____)”
  - SAS empirical estimates of variance
- Programs that account for complex survey design (weights, strata, clusters)
  - Stata “svy” commands
  - SAS “survey___” commands
- Other strategies

If clustering is interesting

- Example - examine costs for hospitalized patients
- Split out the variation in costs
  - How much variation due to differences in patients?
  - How much variation due to differences in hospitals?
- Examine factors that explain variation in costs
  - Characteristics of patients
  - Characteristics of hospitals
- Analytic strategy = Multi-level modeling!

Questions

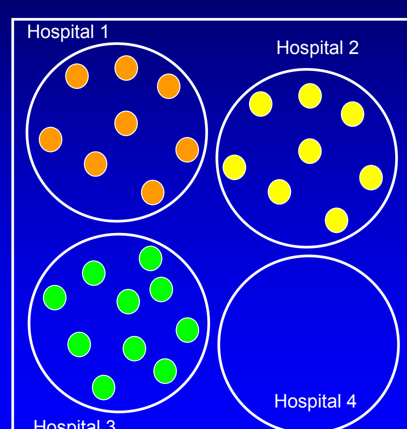
1. Identify 3 patient characteristics that might explain variation in costs
2. Identify 3 hospital characteristics that might explain variation in costs
3. Do you think more of the variation in costs is explained by the patient or the hospital?

Multi-Level Models

(Hierarchical linear models)
(Random effects models)

The concept of "levels"

- Our example – 2 levels
  - Micro = patients (N=26)
    - Micro-level = "units"
  - Macro = hospitals (N=3)
    - Macro-level = "groups"
- At each level
  - Patient characteristics
  - Hospital characteristics



Data Structure - Patient

Patient-level data (= "unit-level data")

Patient ID	Hospital ID	Age (X)	Cost (Y)
1	1	60	3
2	1	75	6
3	2	81	10
4	2	70	7
5	2	65	5

- **Y** represents a patient characteristic
  - Cost (thousands of $)
- **X** represents a patient characteristic
  - Age
  - Note – understand process at each step
  - "Older patients are sicker and tend to cost more"

Simple Linear Regression

$$y_i = a + bx_i + e_i$$

- i indexes patients (i=1 to N)
- Relates x to y
- Both variables are patient characteristics
- Remember the assumptions

Questions

$$cost_i = a + b(age_i) + e_i$$

1. Is there a problem with this model when applied to these data?
2. If so, what?

Patient ID	Hospital ID	Age (X)	Cost (Y)
1	1	60	3
2	1	75	6
3	2	81	10
4	2	70	7
5	2	65	5

The Problem

- Does not account for the clustering of patients within hospitals
  - Data have a structure that is not represented
  - e_i - Assumption of independence is not met
- Do we care?
  - If clustering is nuisance => Stata robust option
  - If clustering is interesting => Multilevel model

Data Structure - Hospital

Hospital-level data (= “group-level data”)

Hospital ID	Beds (W)
1	10
2	65

- **W** represents a hospital characteristic
  - # of beds in the hospital
- Bigger hospitals are more expensive
  - More technology
  - More high-cost specialists
  - “A built bed is a filled bed”

Combined Data Structure

Patient-level data

Hospital ID	Patient ID	Age (X)	Cost (Y)
1	1	60	3
1	2	75	6
2	3	81	10
2	4	70	7
2	5	65	5

\+

Hospital-level data

Hospital ID	Beds (W)
1	10
2	65

= ?

Combined Data Structure

Patient- and hospital-level data

Patient ID	Hospital ID	Age (X)	Cost (Y)	Beds (W)
1	1	60	3	10
2	1	75	6	10
3	2	81	10	65
4	2	70	7	65
5	2	65	5	65

- Age **(X)** and Cost **(Y)**
  - Variation between patients
- Beds **(W)**
  - Only variation between hospitals
  - No variation within hospitals

WARNING – Equations coming up!

Remember - In multi-level modeling …

SUBSCRIPTS ARE YOUR FRIENDS!

Simple Linear Regression

(one approach to modeling this data structure)

$$y_{ij} = a + bx_{ij} + dw_j + e_{ij}$$

- j indexes hospitals (j=1 to N)
- i indexes patients within hospitals (i=1 to n_j)

$$\mathbf{cost}_{ij} = a + b(\mathrm{age}_{ij}) + d(\mathrm{beds}_j) + e_{ij}$$

- Frequently used

Questions

$$\mathbf{cost}_{ij} = a + b(\mathrm{age}_{ij}) + d(\mathrm{beds}_j) + e_{ij}$$

1. Is there a problem with this model when applied to these data?
2. If so, what?

Patient ID	Hospital ID	Age (X)	Cost (Y)	Beds (W)
1	1	60	3	10
2	1	75	6	10
3	2	81	10	65
4	2	70	7	65
5	2	65	5	65

The Problem, Part 2

- You must assume that all of the data structure is represented by the explanatory variables
- Unlikely this will account for the clustering of patients within hospitals
  - Assumes that all clustering within hospitals is explained by the number of beds in the hospital (**W**)
  - If "beds" does not explain all clustering, then assumption of independence is not met for e_{ij}

How do we represent the clustering?

- Let the regression coefficients vary from group to group

$$y_{ij} = a_j + b_j x_{ij} + dw_j + e_{ij}$$

- Groups j can have higher or lower values of $\mathbf{a}_j$ and $\mathbf{b}_j$
- Why not create $\mathbf{d}_j$?

Starting simple – random intercept

- Model the clustering between groups
  - Let the intercept only (**a_j**) vary from group to group
  - Take out all group-level variables (**W**)

$$y_{ij} = a_j + bx_{ij} + e_{ij}$$

- Groups j - higher or lower values of **a_j** only
- Assumes some groups tend to have, on average, higher or lower values of **Y**

Question

$$y_{ij} = a_j + bx_{ij} + e_{ij}$$

1. Why take the group-level variable (**W**) out of this model?
2. Must **W** be taken out of the model?

How do we want to model variation between groups?

- **W** – a "partial" way to model variation between groups
  - If included, it will pick up part of the variation between groups
  - "Part of the variation in costs between hospitals will be explained by the number of beds in the hospital"

- Goal of a random intercept model
  - Model the actual structure of the data
  - Let groups vary, on average, in Y
  - "Let the hospitals vary, on average, in cost"

How do we actually do it?

$$\mathbf{y_{ij} = a_j + bx_{ij} + e_{ij}}$$

- Split $\mathbf{a_j}$ into ($\mathbf{a_0}$ + $\mathbf{u_j}$)

$$\mathbf{y_{ij} = a_0 + u_j + bx_{ij} + e_{ij}}$$

- $\mathbf{a_0}$ = average intercept (constant)
- $\mathbf{u_j}$ = deviation from the average intercept for group j
  = conditional on **X**, individuals in group j have **Y** values that are $\mathbf{u_j}$ higher than in the average group

- "Conditional on patient age, patients in Hospital j have costs that are $\mathbf{u_j}$ higher than the average costs for all patients"

What do we do with $\mathbf{u_j}$?

Part 1 – Fixed effects

- Are groups j regarded as unique?
  - Do you want to draw conclusions about each group?

TREAT AS "FIXED EFFECTS"

- Create j – 1 indicator variables (0/1)
- Leads to j – 1 regression parameters

Questions

$$\mathbf{cost_{ij} = a_0 + b(age_{ij}) + u_j + e_{ij}}$$

1. For our data, what does this equation look like if $\mathbf{u_j}$ is modeled as a fixed effect?
2. Are all indicator variables in a model also fixed effects?

Patient ID	Hospital ID	Age (X)	Cost (Y)
1	1	60	3
2	1	75	6
3	2	81	10
4	2	70	7
5	2	65	5

Modeling u_j as a fixed effect

(u_j = “differences between hospitals”)

$$cost_{ij} = a_0 + b(age_{ij}) + c(hosp2_{ij}) + e_{ij}$$

- **hosp2** = 0/1
  - 1 = patient i in hospital 2, 0 = patient i in hospital 1
- Do we need index j? No – why?

$$cost_i = a_0 + b(age_i) + c(hosp2_i) + e_i$$

- What assumptions does this model make?

What do we do with u_j?

Part 2 – Random effects

- Three issues
  - Are groups regarded as sample from pop.?
  - Do you want to test the effect of group level variables (remember **W** = # beds)?
  - Do you have small group sizes (2-50 or 100)?

TREAT AS “RANDOM EFFECTS”

- Model u_j explicitly
- Additional assumption that u_j is i.i.d.
  - Groups (hospitals) considered exchangeable
- Can include group-level explanatory variables (**W**)

Questions

$$y_{ij} = a_0 + b(x_{ij}) + u_j + e_{ij}$$

1. For our data, what does this equation look like if $\mathbf{u_j}$ is modeled as a random effect?
2. How would we include our hospital-level explanatory variable?

Patient ID	Hospital ID	Age (X)	Cost (Y)	Beds (W)
1	1	60	3	10
2	1	75	6	10
3	2	81	10	65
4	2	70	7	65
5	2	65	5	65

Modeling $\mathbf{u_j}$ as a random effect

($\mathbf{u_j}$ = "differences between hospitals")

$$cost_{ij} = a_0 + b(age_{ij}) + u_j + e_{ij}$$

- $\mathbf{u_j}$ = deviation from the average cost for hospital j
  = estimated using HLM, SAS, Stata (get a number!)

$$cost_{ij} = a_0 + b(age_{ij}) + d(beds_j) + u_j + e_{ij}$$

- Uses the number of beds in the hospital to explain some of the variation in $\mathbf{u_j}$
- Last question - what happens to $\mathbf{u_j}$ if the number of beds explains all of the differences between hospitals?

What we did and didn't do today

- We discussed
  - Clustering (artificial and natural)
  - Accounting for clustering
    - Nuisance = robust estimates of variance
    - Interesting = multilevel models
  - Representing clustering in simple model
    - Fixed effects
    - Random effects with group-level explanatory variables
- We didn't discuss
  - Random coefficients other than the intercept
  - Interaction terms (cross-level effects)
  - Many other things

Follow-up

maureensmith@wisc.edu