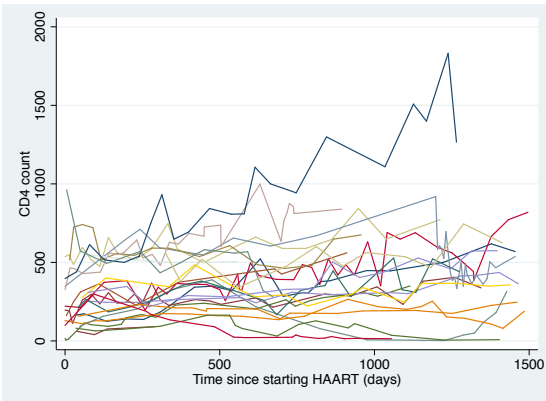# Longitudinal Data
Fall 2014

Naïve Analysis of Longitudinal Data
(Major Themes)

Instructors

Nick Jewell (*jewell@berkeley.edu*)

GSI

Robin Mejia (*mejia@nasw.org*)

1

Major Themes

1. Dependent data – impact on estimation and inference. Why can't I just use the tools I already know? MT1

2. Using longitudinal history in regression.
   – Avoiding, by default, treating the data like cross-sectional data. What am I missing by using the tools I already know? MT2

3. Estimating contribution of variance from different units. Can I better understand where my variation comes from MT3

4. Efficiency. Can I make my estimators more precise? MT4

2. Using longitudinal history in regression

- Make the covariates included in regression address the question of interest.

- In many cases, a cross-sectional study can be confounded in a way the longitudinal study is not.

- One way this can happen is how subjects are recruited over time.

- A hypothetical example: CD4 count and time since diagnosis of AIDS.

Longitudinal questions/parameters

- One might expect, particularly in an earlier era a true average decline in CD4 count with time since dx.

- However, if one recruits subjects and only measure once their CD4 counts and record the time since dx, a bias can result.

- Why?

  – (perhaps) subjects who live longer will have on average higher CD4 counts for their time since dx and,

  – because they live longer and they contribute more to the pool of people with longer times since diagnosis than subjects that have a steep decline, who tend to die earlier thus contributing less data.
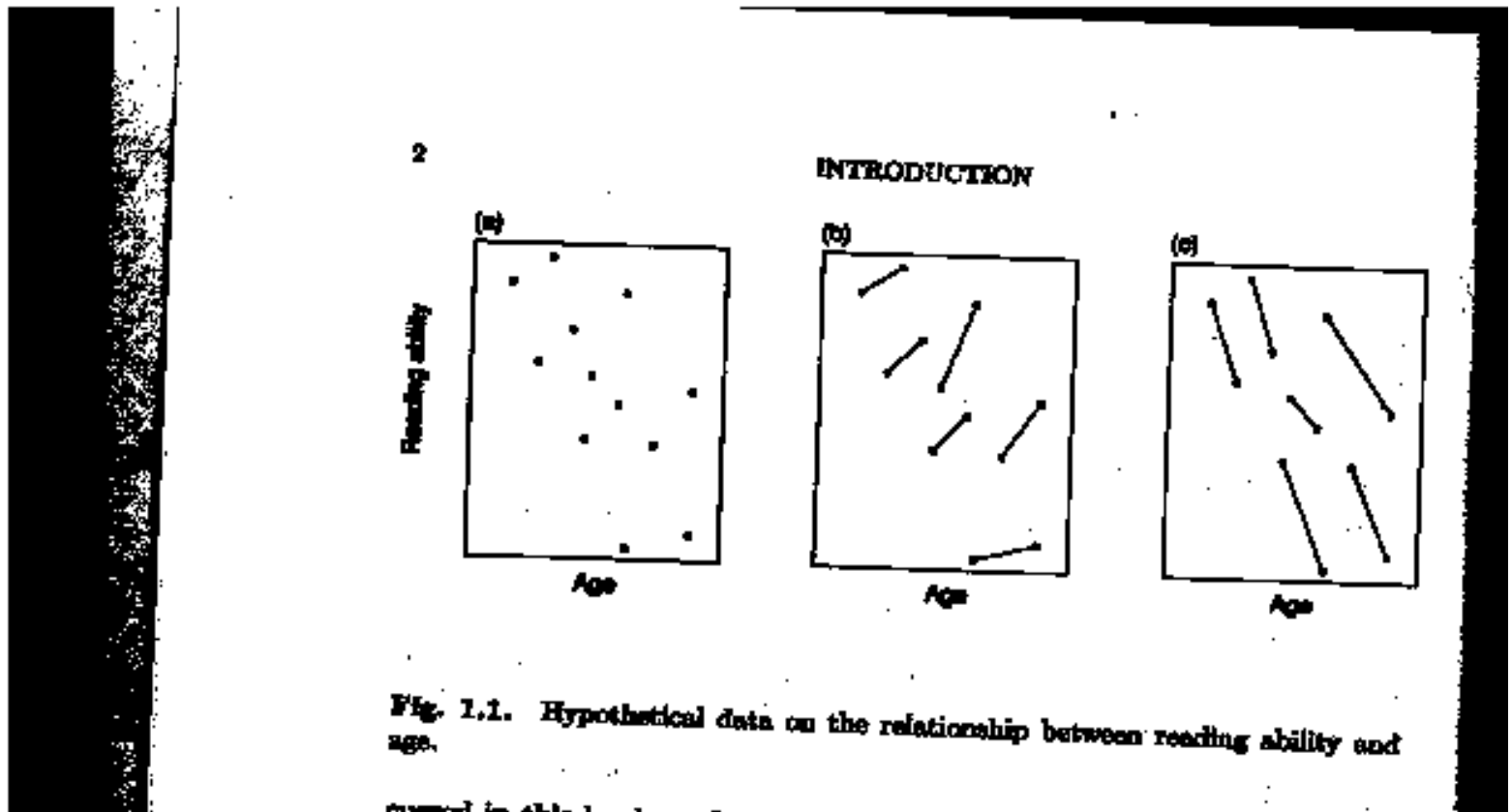
4

Cross-sectional vs Longitudinal Information
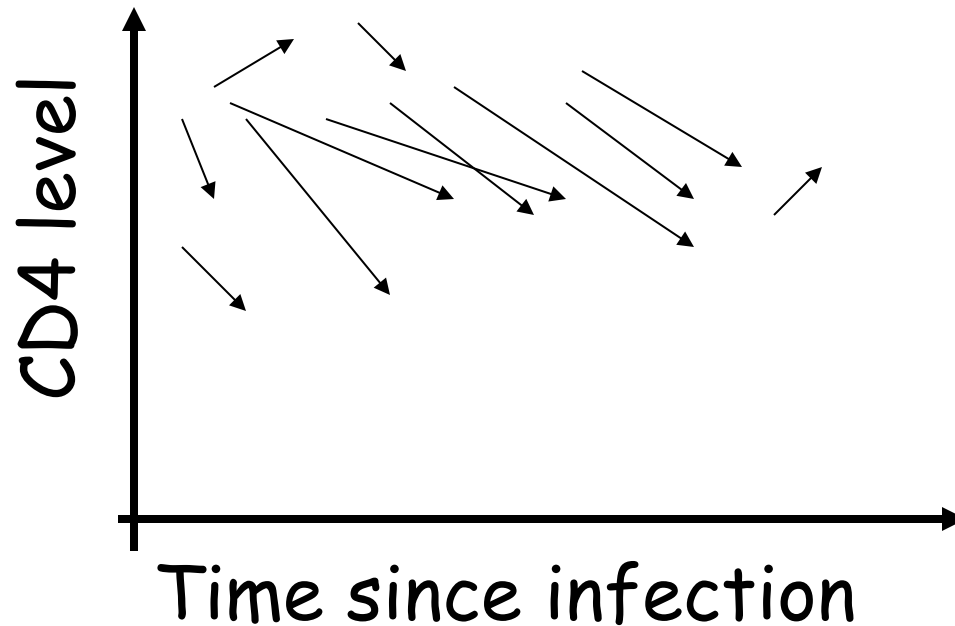
(a) cross-sectional: CD4 doesn't change much in time

(b) longitudinally: CD4 decreases (on average) in time

Example from Diggle, et al.

Fig. 1.1. Hypothetical data on the relationship between reading ability and age.

Separating out longitudinal (interesting) from cross-sectional (maybe less interesting) effects

- Consider the model:

$$E[Y_{ij} \mid X_{i1} = x_{i1}, X_{ij} = x_{ij}] =$$

$$\beta_0 + \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1})$$

- β_L represents the expected change in Y given a change in X_{ij} relative to the baseline value (X_{i1}) - longitudinal effect.

- β_C represents the expected difference in average Y across two sub-populations that differ by their baseline values, X_{i1} - cross-sectional effect.
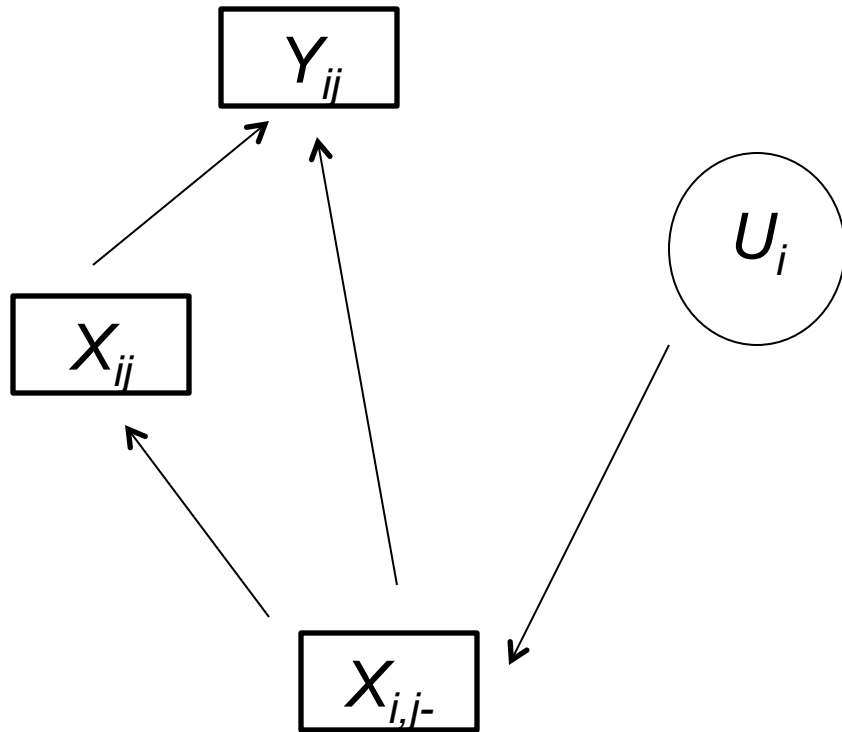
Only X-sectional Data
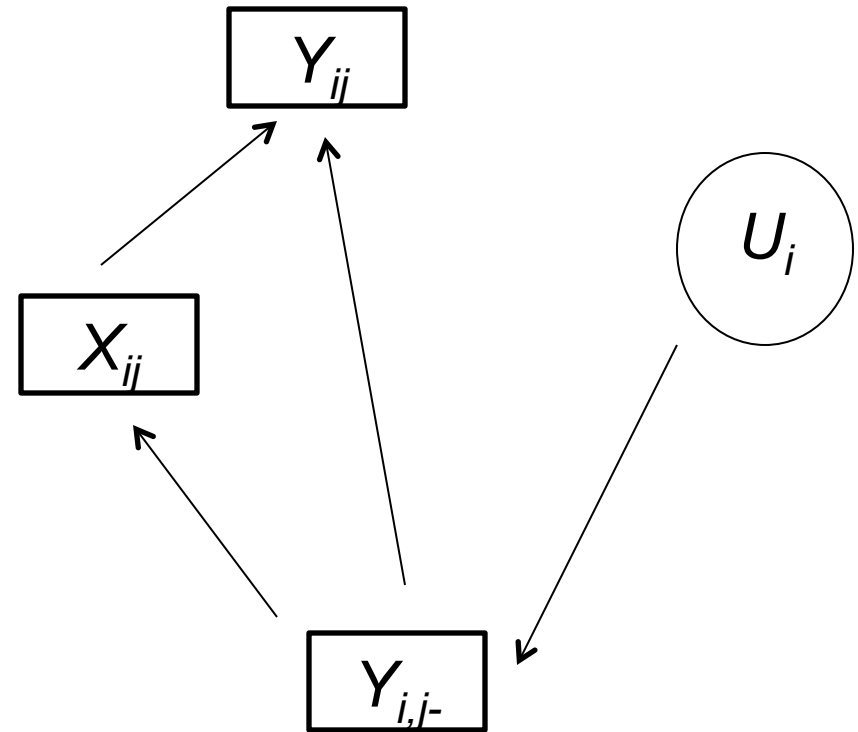
■ In a x-sectional study, only can estimate:

$$E[Y_{i1} \mid X_{i1} = x_{i1}] = \beta_0 + \beta_C x_{i1}$$

■ Can use cross-sectional data to estimate longitudinal effect only if $\beta_C = \beta_L$.

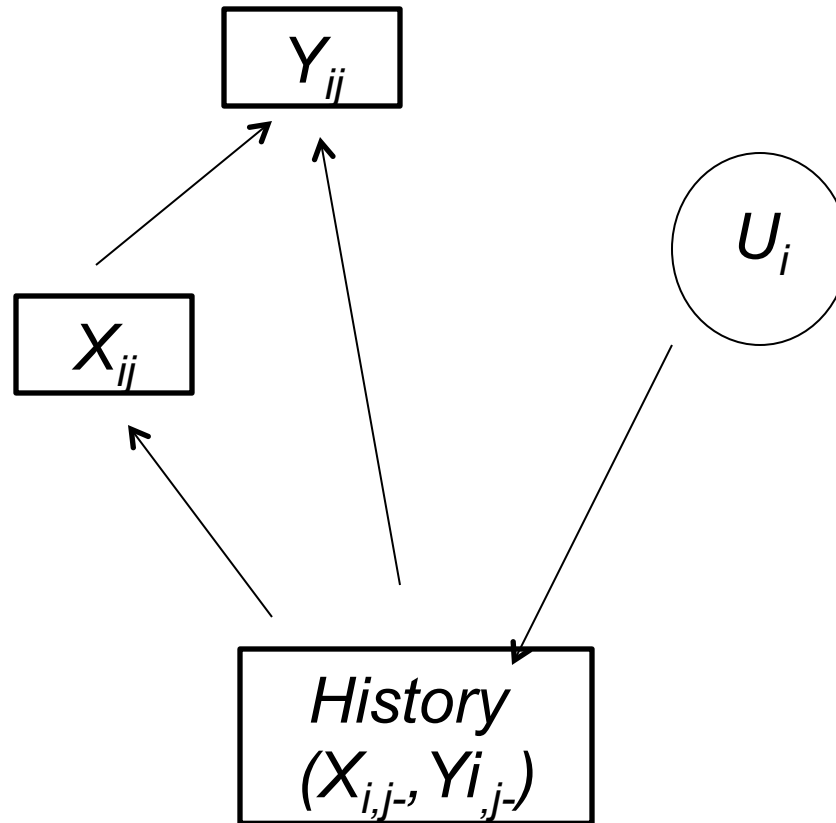# Longitudinal data gives more opportunity to adjust for unmeasured confounders

Unmeasured Confounder (U_i) blocked by past measure of covariates ($X_{i,j-}$).

Unmeasured Confounder (U_i) blocked by past measure of outcome ($Y_{i,j-}$).

Most Generally: Adjust for entire history

Y_{ij}

X_{ij}

U_i

History
$(X_{i,j\text{-}}, Yi_{,j\text{-}})$

Still begs the question of how to adjust for history– we will discuss more later in the term, but in general…..

Regression using More complicated functions of past

■ Parameterizing the model based on the measured past – i.e., whole past

$$E[Y_{ij} \mid \mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{i(j-1)}, \mathbf{X}_{ij}, Y_{i1}, Y_{i2}, ..., Y_{i(j-1)}]$$

3. Partition of Variance

- One can use the repeated measures to distinguish the degree of variation in Y across time for one person from the variation in Y among persons.

- E.g., subject *j* within family *i*, measured at time *k*:
$$Y_{ijk} = b_0 + b_{0i} + b_{0ij} + e_{ijk}$$

then under assumptions:

$$\text{var}(Y_{ijk}) = \text{var}(b_{0i}) + \text{var}(b_{0ij}) + \text{var}(e_{ijk})$$

4. Efficiency

■ If the within subject variability is high, can gain a lot of efficiency by taking repeated measurements on the same subject.

– Example - $$Y_{ij} = \mu + \alpha_i + e_{ij}$$

$$\bar{y}_{i.} \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \longrightarrow \mathrm{var}(\bar{y}_{i.}) = \sigma_\alpha^2 + \frac{\sigma_e^2}{n_i}$$

Re-Cap

- ■ Is ignoring correlation of measurements on same individual (unit) OK?
  - – For estimation – yes (usually) – although one can do better by not ignoring it.
  - – For inference – NO!

- ■ Advantages of Longitudinal Data
  - – Can distinguish x-sectional from longitudinal effects (can eliminate some of the confounding due to individual-level differences by looking at change in outcome vs. change in explanatory variable).
  - – Can model association of current outcome with entire history.
  - – Partition Variance
  - – Increased efficiency.