Classmate Peer Effects in First-Year Core Courses

A Progress Report: Year 1 of 3

What Are Peer Effects?

Effects of peers on student learning

- Classmates
- Roommates
- Dormmates
- Fellow organization members
- Romantic partners
- Friends

Why Are Peer Effects Economics?

- Primary/secondary: Tracking vs. streaming
- Higher education
  - Rothschild and White: Peers help "produce" education services
  - Efficiency of resource allocation may be enhanced by compensating these peer providers
  - This argument can be used to support merit aid for good peers

Existing Studies: Primary/secondary

- Multitude of studies
- Peer effects seem to exist but are weak
- Some support for streaming: strong help weak more than weak hurt strong

Existing Studies: Higher Education

- Roommate studies at Dartmouth, Williams, C&B schools find weak and inconsistent roommate effects
- Hoel, Parker, and Rivenburg (2005) find nonlinear classmate quality effects, strong gender peer effects in Hum 110 at Reed.

Teagle Grant

Teagle Foundation grant to Reed, Whitman, and Lewis & Clark to examine peer effects in freshman core courses

- Quantitative study: evidence that core-course peers affect GPA and/or graduation outcomes
- Qualitative study: interviews with core instructors to get observations about peer effects in classroom

Measuring Peer Quality

- How to measure quality distribution of peers?
  - SAT
  - "Reader rating"
  - High school GPA
  - High school class rank
- All are important, not collinear
- We use predicted GPA based on all of the above plus ethnicity, gender, etc.

Missing Data Problems

- Can't have missing data because we need to measure quality of (nearly) all peers
- Numerous missing cases for some variables in data set (especially HS class rank)
- How to estimate models working around the missing variables?

Why Are Data Missing?

- ***Missing completely at random***: probability that a case is missing does not depend on observed or missing variables
- ***Missing at random***: probability that a case is missing does not depend on missing variables
- ***Missing not at random***: probability that a case is missing depends on missing and/or observed variables

Missing Data Solutions

- Complete Case Analysis
  - Can't use because we need prediction for all cases
- Dummy Variables for Missing
  - Used in 2004, but may have biased results
- Dummies with Interactions
  - Difficult to implement with complex patterns of missing data
- Multiple Imputation Models
  - Available in Stata via ICE do-file

Simple Illustrative Model

- Dependent variable is *G* (college GPA)
- Two regressors: *S* and *R* (SAT and HS class rank)
  - *S* has complete data
  - *R* has many cases missing completely at random
  - *M* is dummy variable that is one if *R* is missing and zero if it is observed (missing data indicator)

Simple Dummy Variable Model

$$R_{0,i} = \begin{cases} R_i \text{ if } M_i = 0, \\ 0 \text{ otherwise.} \end{cases}$$

$$G_i = \beta_0 + \beta_S S_i + \beta_R R_{0,i} + \beta_M M_i + \varepsilon_i$$

$$\text{For } M_i = 0 : G_i = \beta_0 + \beta_S S_i + \beta_R R_i + \varepsilon_i,$$

$$\text{For } M_i = 1 : G_i = \beta_0 + \beta_S S_i + \beta_M + \varepsilon_i.$$

Simple Dummy Variable Model

- OLS estimate of β_S is biased if *S* is correlated with *R*.
- Standard omitted-variable bias: *S* is picking up the effect of both *S* and *R* for those cases for which *R* is missing, but measures only (correct) effect of *S* for complete cases.

Dummy Model with Interaction

$$G_i = \beta_0 + \beta_M M_i + \beta_S S_i + \beta_{SM} S_i M_i + \beta_R R_{0,i} + \varepsilon_i$$

$$\text{For } M_i = 0 : G_i = \beta_0 + \beta_S S_i + \beta_R R_i + \varepsilon_i,$$

$$\text{For } M_i = 1 : G_i = \left(\beta_0 + \beta_M\right) + \left(\beta_S + \beta_{SM}\right) S_i + \varepsilon_i.$$

Equivalent to running separate regressions on complete and missing sub-samples.

Not feasible with complex pattern of missing data.

Multiple Imputation by Chained Equations (MICE)

- Imputation: Replace missing variables with random draws from their posterior (estimated) probability distribution conditional on observed variables
- Combination: Repeat the imputation process multiple times and combine the results to obtain a single estimator

MICE Example: Imputation

- For complete-data sub-sample, regress

$$R_i = \gamma_0 + \gamma_1 S_i + \eta_i$$

- Determine the distributions of the estimators for γ_0, γ_1, and the estimated distribution of η_i.
- For each missing case, draw *m* random samples from estimated distributions of γ_0, γ_1, and η_i and calculate *m* imputed values for missing R_i.

MICE Example: Estimation

- For each of the *m* imputed full samples, estimate the regression

$$G_i = \beta_0 + \beta_S S_i + \beta_R \hat{R}_i + \varepsilon_i$$

- Save the resulting estimates of β and related estimated variances and variances.

MICE Example: Combination

- Combine the *m* estimates from the imputed samples to form single estimator for β and its variance:

$$\bar{\beta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j$$

$$\hat{Var}\left(\bar{\beta}\right) = \frac{1}{m}\sum_{j=1}^{m}\hat{Var}\left(\hat{\beta}_j\right) + \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_j - \bar{\beta}\right)^2$$

MICE Complications: Imputation

- If missing data pattern is more complex, there may be few complete cases
  - Draw initial imputation for each missing value from univariate distribution of variable (estimated from available cases), then cycle through imputation of all missing variables *k* times (*k* = 20?).
- Missing regressors may not be continuous
  - Use logit, multinomial logit, or ordered logit as appropriate

MICE Complications: Post-estimation

- Still working on these issues:
  - Testing of composite hypotheses
  - Prediction

Preliminary Results

- Preliminary complete-case and MICE estimates for Reed and Whitman
- Still to do (next summer!):
  - Estimates for Lewis & Clark
  - Composite test statistics
  - Prediction of GPA based on imputation model
  - Calculation of classmate peer statistics
  - Estimation of models with peer measures

… and …

- Qualitative analysis
  - Interviews with ~10 core-course instructors at each school
  - Looking for evidence to complement, confirm, explain (or perhaps contradict) quantitative findings
  - Looking for ideas about variables to attempt to measure in future studies