capsee

CENTER FOR ANALYSIS OF
POSTSECONDARY EDUCATION
AND EMPLOYMENT

Linking College and Labor Market Datasets for Research on the Returns to College

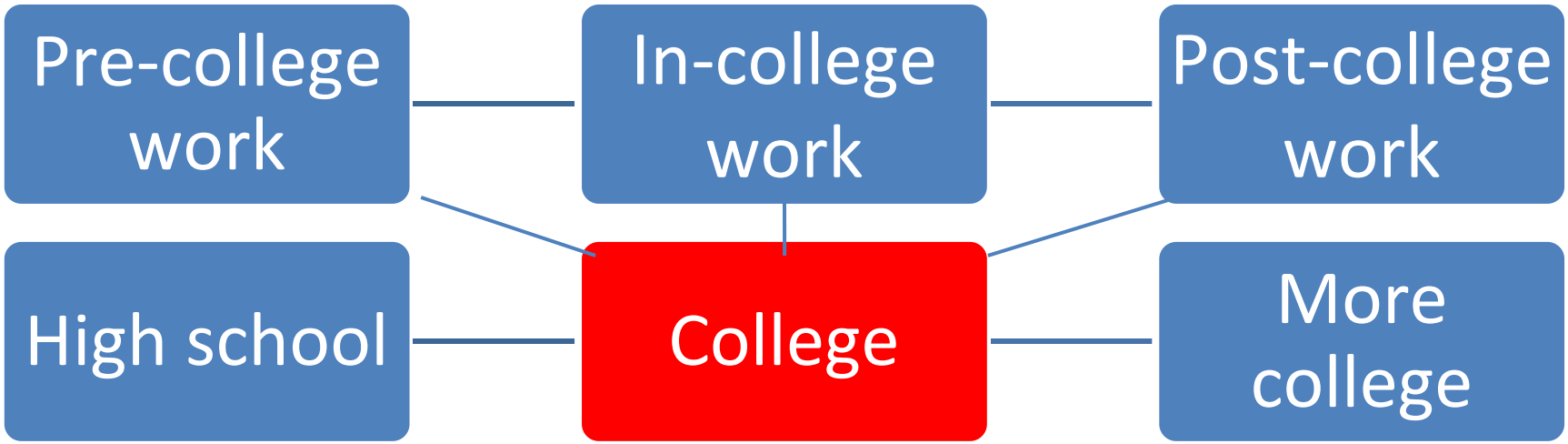
Di Xu
Thomas Bailey, Clive Belfield, Shanna Jaggars
Teachers College, Columbia University

March,2014 | AEFP Preconference Workshop

Outline

1. What data is linkable?
2. What does data typically look like?
3. Advantages of using linked data
4. Disadvantages of using linked data
5. Potential Problems with the analysis
6. Practicalities of obtaining and using linked data
7. Practice: cleaning and linking data sets

Student Progress



State Administrative College Data

- Begin with college data: link across, forward, back
- These data are different from longitudinal surveys:
  - Created for basic administration and compliance purposes
  - Variation by state in quality, comprehensiveness, history
- Coverage issues are very important:
  - Often limited to public sector within one state
  - University systems typically hold own data; community college districts or systems typically hold own data
  - Centralized states (collect and hold data across all publics) and decentralized states (data available college-by-college basis)

Datasets Linkable to College Data

- National Student Clearinghouse data on where students transfer to, how long they persist, award earned
  - Merge on name and birthday
  - High match rate: NSC coverage is very full (includes all Title IV colleges)
- State high school data with full transcript information
  - Merge on name/birthday/ID
  - Low match rate: student mobility and lagged/delayed college enrollment and enrollment out of publics or out of state
- College-level data from IPEDS or other sources; census data
  - Merge on geocode or college name

Linkable Labor Market Data

- Unemployment Insurance data for individual student earnings
  - Merge college and UI data using SSN
  - Moderate match rate: coverage of employment data may not be complete

- Labor market data may differ from national surveys
  - Self report vs. formal record
  - Different follow up vs. quarterly employment data
  - Total income vs. income data from all formal jobs
  - Sometimes hours worked and occupation

College Transcript Data: Course-Level

Obs	id	term	course	credits	grade
1	04_000000001	FA04	SPD100	3	B-Good
2	04_000000002	FA04	ART131	3	D-Poor
3	04_000000002	FA04	ENG01	5	U-Unsatisfactory
4	04_000000002	FA04	MTH02	5	W-Withdrawal
5	04_000000002	FA04	MUS121	3	W-Withdrawal
6	04_000000002	FA05	ART175	4	F-Failure
7	04_000000002	FA05	MTH01	4	U-Unsatisfactory
8	04_000000003	FA04	HIS121	3	C-Average
9	04_000000003	FA04	MAC131	2	W-Withdrawal
10	04_000000003	FA04	MTH271	3	A-Excellent

- Multiple rows per student
- One row per course
- Generally:
  - Semester course taken
  - Course name and number
  - Credits attempted
  - Grade
  - Typically not section number or information on instructor
- Can be used to derive semester-level and student-level variables

College Demographic Data: Student-Level

Obs	id	gender	race
1	04_000000001	1	1
2	04_000000002	1	1
3	04_000000003	1	1
4	04_000000004	1	2
5	04_000000005	1	1
6	04_000000006	1	4
7	04_000000007	2	1
8	04_000000008	2	1
9	04_000000009	2	1
10	04_000000010	1	1

- Looks just like survey data
- One row per student
- Generally:
  - Gender
  - Race
  - Birthdate
  - Zipcode sometimes

College Award Data: Student-Level

Obs	id	award_long	award_term	award_cip	award_major
1	2004_05_FA_000000008	Associate of Arts and Sciences	FA07	240101	Education
2	2004_05_FA_000000011	Associate of Applied Science	SU06	110101	Information Systems Technology
3	2004_05_FA_000000012	Associate of Applied Science	SU08	520399	Accounting

- Includes award, semester of award attainment, cip codes, major field of award
- Classification of Instructional Programs (CIP 2000): (https://nces.ed.gov/pubs2002/cip2000/)
- Variations across states in defining types of award
- Multiple Award
- Transfer students

Other College Administrative Data

- Placement test scores and assignment
  - Missing values
  - Multiple tests: reading, writing, math
  - Multiple scores
- Financial aid
  - Missing values: Only available for student who are eligible and applied for financial aid

NSC Data: Semester-Level

Obs	id	Enrollment_Begin	Public_Private	Type
1	04_000000001	20070827	Public	2
2	04_000000004	20060824	Public	4
3	04_000000008	20080114	Public	4
4	04_000000010	20060821	Public	4
5	04_000000010	20070108	Public	4
6	04_000000010	20070820	Public	4
7	04_000000010	20080114	Public	4
8	04_000000010	20080901	Public	4
9	04_000000010	20090120	Public	4
10	04_000000010		Public	4

- From National Student Clearinghouse – Enrollment begin and end dates
- Derive semester-level variables (e.g. co-enrollment; post-community college enrollment)

Levels of Measurement: Quarterly

Obs	id	quarter	wages	naics_code
1	04_000000001	20031	12003.89	722110
2	04_000000001	20032	12100.03	722110
3	04_000000001	20033	12060.12	722110
4	04_000000001	20034	12223.24	722110
5	04_000000001	20041	0.00	.
6	04_000000001	20042	3554.30	722211
7	04_000000001	20043	8500.66	722211
8	04_000000001	20044	8800.70	722211
9	04_000000001	20051	8322.68	722211
10	04_000000001	20052	8593.32	722211

- Example UI data
- Date of quarter won't match exactly with enrollment semesters
- Need to be adjusted for inflation
- Multiple entries in a quarter for one student
- North American Industry Classification System

(https://www.census.gov/eos/www/naics/)

Advantages with Linked Data (1)

- Longitudinal data
- Reduce bias from attrition
- Large sample sizes allow for subgroup analysis
  - Colleges, programs, courses
  - Demographic groups
- Address a lot of questions for education policy
- More precise, accurate, and various measures of educational attainment

Student Pathways: Transfers

According to NSC:

- One-third of all students transfer
- 14% of students who start at 4-year college transfer to 2-year college
- Transfer from 2-year to 4-year colleges
- Co-enrollment
- Implication for research?
  - Enrollment
  - Educational Award

Student Pathways: Course-taking

Students take many different courses:

- Below college-level courses
  - Remedial classes: reading, writing, math, biology, chemistry etc.
  - ESL classes
  - Basic skills
  - Student success courses
- College-level courses
  - Gatekeeper courses: course required for an award
  - Subject-specific courses

Advantages (2)

- Many pre-college controls
  - Ability measures
  - Proxies for non-cognitive attributes (e.g. credits accumulated in school for effort)
  - Time-varying controls
- Help reduce and test for omitted variable bias
- Opportunities to test for selection bias (variations in college practices, changes over time, compare students to themselves in other classes)

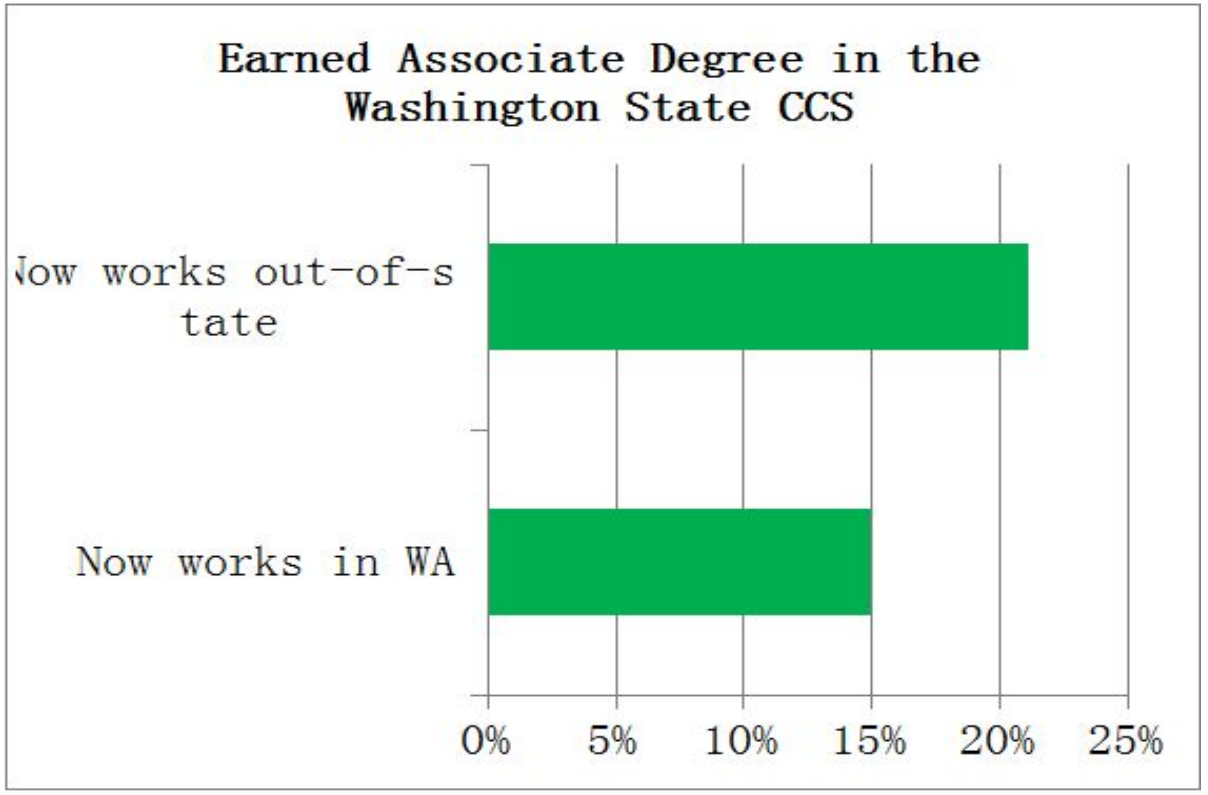
Advantages (3)

- More precise and accurate measures of earnings/income:
  - Self-reports less reliable at lower earnings (overstate low income): compress the education-earnings premium
  - Self-reports more measurement error for the less education (low education persons misstating their income): reduce precision
  - More educated persons have multiple jobs (bonuses/commissions)
  - No non-response missing data (CPS is 20-30%)
- Data on income over time, including before and during college, and quarterly (not annual)

Disadvantages with Linked Data (1)

- SES typically missing (use occupation, geocode, financial aid)
- Attitudinal data usually not available
- UI data does not cover everyone and sample truncation or censoring may be endogenous
  - Students who move across state lines, self-employed, military, some federal workers
  - Cannot be sure that missing earnings is zero

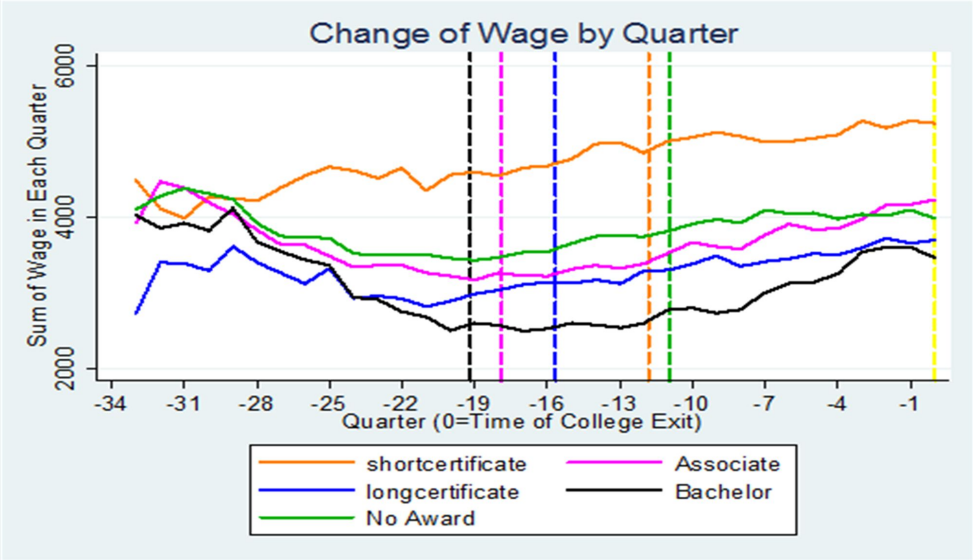
Endogenous Mobility



Disadvantages (2)

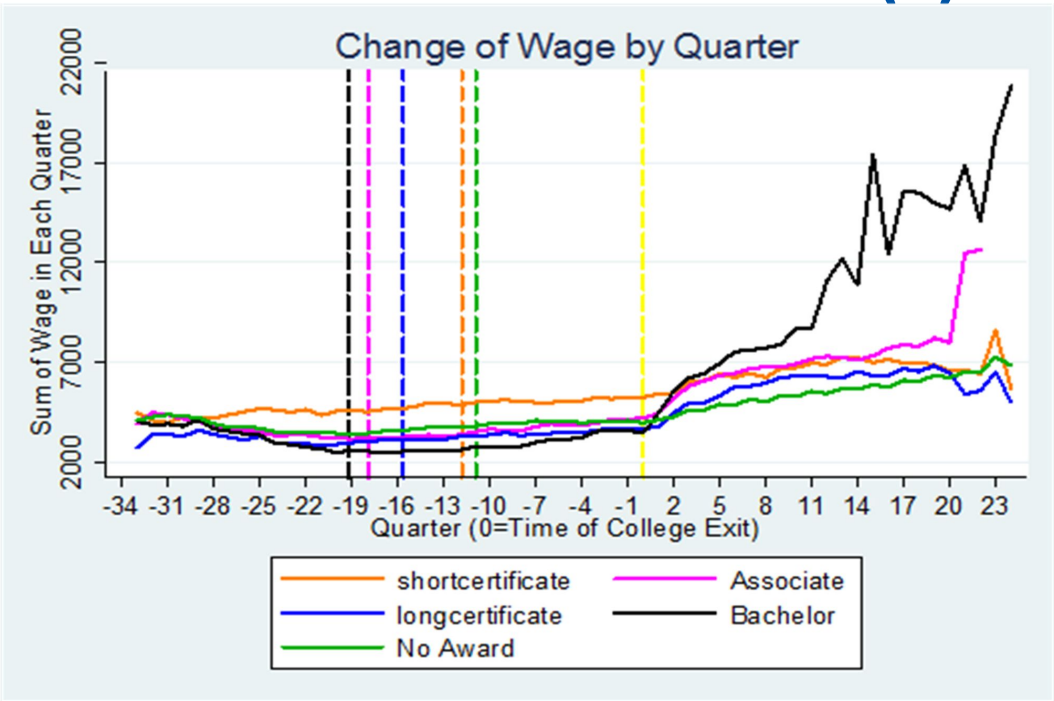
- Data cleaning and computation more complex – can require several months of work to complete
  - Information is recorded in different data structures; requires quite a bit of work to get them into the same structure so that you can analyze them together in the same model
  - Even basic variables require time to compute (e.g. number of college credits will need information on what is a college credit)

Examining Economics Returns to Education: Potential Problems (1)



- "Lock-in" effect
  - May exist even after controlling for opportunity cost
  - may vary across different award groups
  - Implication for Mincerian and Individual fixed effects models?

Potential Problems (2)



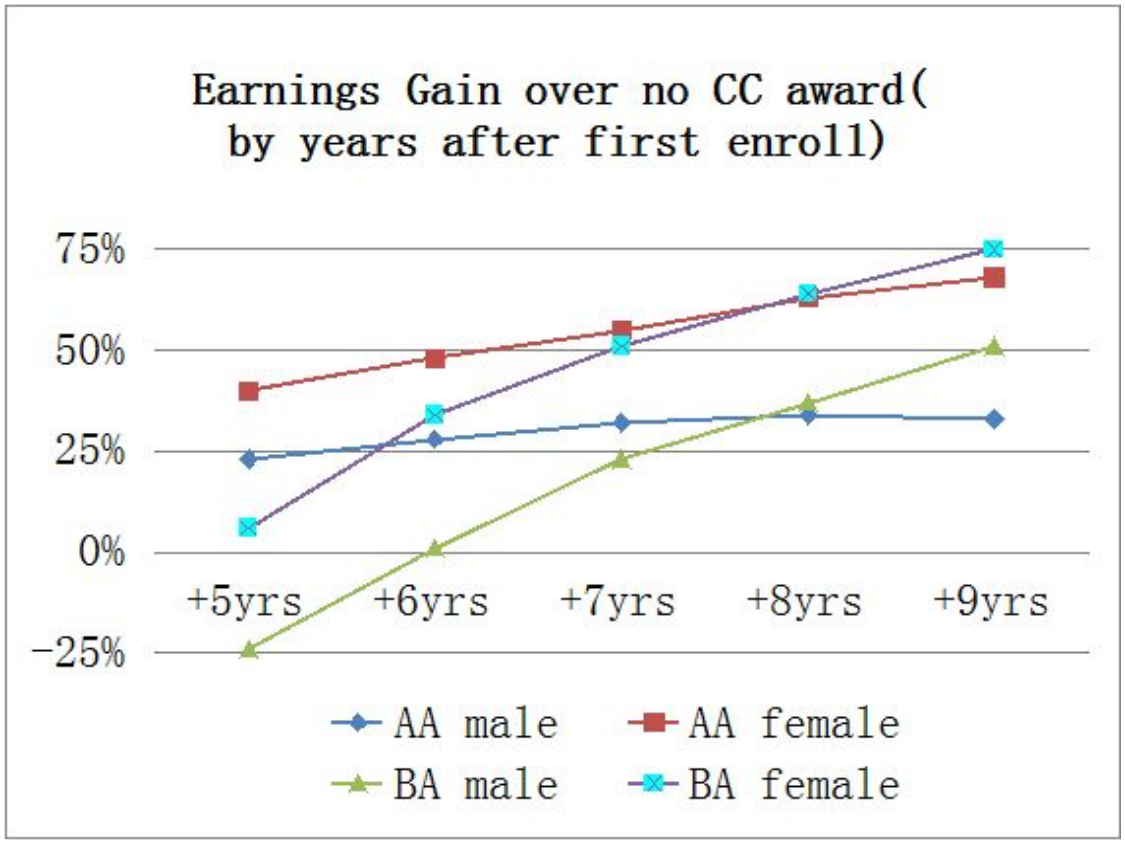
- Wage Growth
  - Higher post-college growth rate compared to pre-college period
  - May vary across different award groups: time out of college, growth rate
  - Implication for Mincerian and Individual fixed effects models?

Example: Mincerian Estimates Based on Different Model Specifications

	Dependent Variable: Quarterly Earnings				
	1	2	3	4	5
Highest degree: Bachelor	304.60*** (70.07)	603.14*** (70.05)	649.57 *** (71.83)	1569.79*** (130.52)	175.16** (80.79)
Highest degree: Associate	354.22*** (54.81)	364.75*** (53.93)	367.69*** (53.89)	1028.17*** (93.51)	51.73 (59.51)
Highest degree: Longcert	121.58 (101.38)	82.56 (100.18)	86.01 (100.08)	236.49 (161.89)	-71.45 (110.58)
Highest degree: Shortcert	459.38*** (131.991)	405.41*** (132.43)	412.91*** (132.42)	368.76** (168.62)	486.20 (171.65)
Still Enrolled by First Quarter of 2012		-1108.30*** (44.31)	-1019.03*** (57.82)	-493.63*** (68.82)	-699.12*** (62.74)
Number of Quarters Since College Exit			12.58 ** (5.38)	33.42*** (5.65)	21.13*** (5.48)
Enrolled*Bachelor				-1435.95*** (151.74)	
Enrolled*Associate				-1106.41*** (109.17)	
Enrolled*Longcert				-328.48 (202.99)	
Enrolled*Short Shortcert				190.30 (264.11)	
Quarters Since Exit *Bachelor					334.45*** (39.51)
Quarters Since Exit *Associate					112.93*** (14.07)
Quarters Since Exit *Longcert					33.48 (21.20)
Quarters Since Exit *Short Shortcert					-10.70 (20.85)
Observations	38,092	38,092	38,092	38,092	38,092
R-squared	0.1285	0.1424	0.1426	0.1466	0.1470

- Model 1: Traditional Mincerian
- Model 2: Control for whether still in college
- Model 3: Control for quarters out of college
- Model 4: Allow 2 to vary across award groups
- Model 5: Allow 3 to vary across award groups
- Model 6: Estimate returns by year

Rapid Early Growth in Earnings



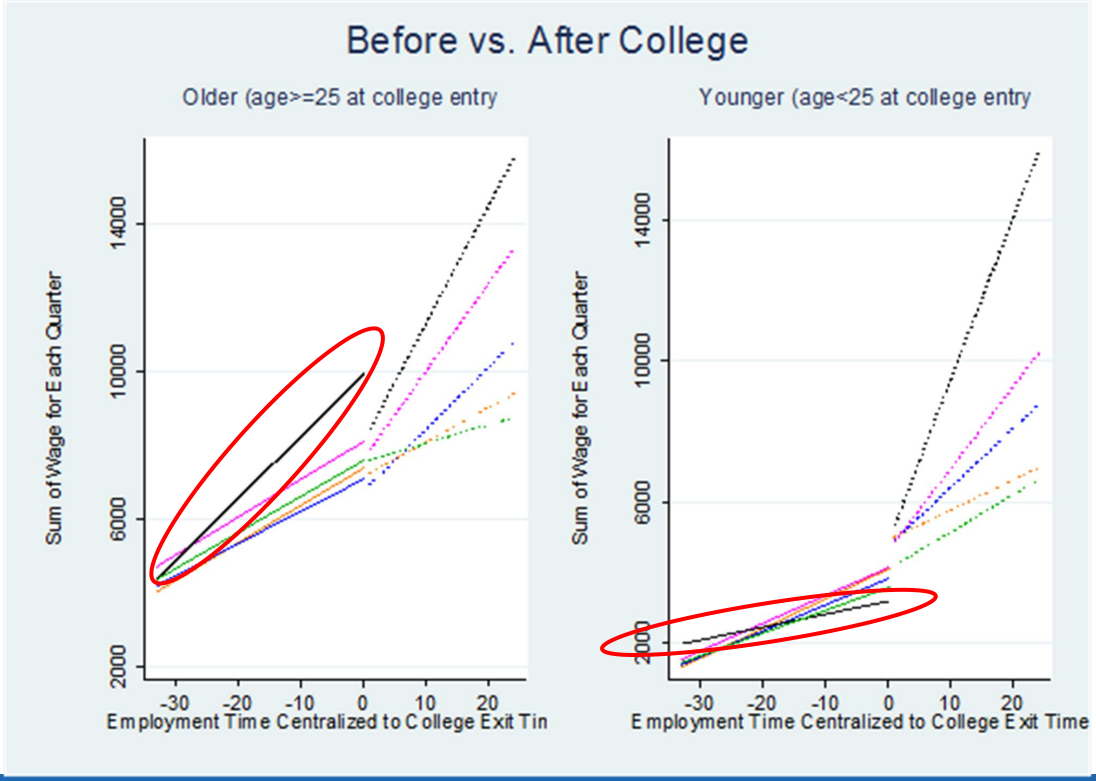
Potential Problems (3)

- Wage variations across industry

	Before College	After College	Average Quarterly Earnings
Admin & support & waste	3.97	3.33	4105.09
Construction	0.79	0	5582.46
Educational services	32.54	50.83	4023.51
Health care and social assistance	4.76	8.33	4895.49
Information & finance	5.56	2.5	5144.09
Manufacturing	10.32	0.83	8988.51
Public administration	2.38	2.5	6601.57
Retail and wholesale trade	14.29	8.33	3709.98
Services	25.4	22.5	2777.67
Others	0	0.83	6935.2
N	126	120	63,714

Potential Problems (4)

- Time-varying Factors that Influence both Degree Attainment and Wage



- For BA earners
- Totally different trajectories before college
- Similar trajectories after college
- Below average wage for young BA earners before college
- Why?

Potential Problems (5)

- Violation of "Strict Exogeneity Assumption" Underlying Fixed Effects Models
  - What is the assumption?
  - In what way could it be violoated?
  - How can we test it?
- Substantial variations in returns to different field of study

Practicalities of Using Linked Data (1)

- Need links with state system officials and UI data-holders
  - Personal relationships to persuade data-owners that research is useful
  - Work with many agencies; some have good mutual relationships
- This is not a priority for state officers; may take time
- Cannot ask repeatedly for more information
  - Need to know exactly how much data you need
  - Data-owners typically do not mind if ask for more years if data is all in same format

Practicalities (2)

- Cannot be a “lone wolf”
  - No carte blanche from data-owners
  - Must allow review of your work by data donors
- Cannot share
  - Data donors will likely not allow sharing of data: need to think about how this impacts on publication prospects
- Work may have direct policy implications:
  - States may ask for technical assistance or policy recommendations
  - States may not like results

Data Practice

- Fake data
- Created to resemble data structure in real data sets
- Six data files: course, student, award, nsc, cpi2010, wage
- Using STATA for data clean and merge
- Other software for data cleaning: e.g. SAS, R, SPSS, etc.

Some Useful Tips with STATA

- Use the "help" command: e.g. help reg
- Always create a "do" file instead of writing codes directly in the command window
- Difference between string and numeric values
  - if female==1 vs. if female=="1"
- Some useful command in data cleaning
  - use, save
  - generate, replace
  - keep, drop
  - rename, destring, tostring, substr
  - collapse, merge, append
  - tab, sum, scatter, hist, twoway

Example: Cleaning Data (1)

- Clean transcript data
  - flag college-level course

Coding Sheme: College-level course: course number>100 (e.g. ENG111)
Stata hint: substr, destring

	course	term	grade	program_cip	credits	id
1	BIO101	FA07	C-Average	240101	4	1
2	ITE115	FA07	A-Excellent	240101	3	1
3	MTH03	FA07	W-Withdrawal	240101	5	1
4	AST117	SP08	A-Excellent	240101	1	1

```
*find the course number;
use "C:\Users\DiX\Desktop\projects\AEFP\Data for AEFP\course", clear
gen csnum=substr(course,4,3)
destring csnum, replace

*flag college-level course;
gen crscl=0
replace crscl=1 if csnum>100
```

Example: Cleaning Data (1) continued

- Clean transcript data
  - create variable for the number of college-level credits earned for a course

  Coding Sheme: Pass a course: a letter grade D or above, P, S

  Stata hint: 1) whether the student passed the course

  2) credits*whether pass*whether college

```
*whether the student earned any credits from the course;
gen anycr=0
replace anycr=1 if grade=="A-Excellent" | grade=="B-Good" | grade=="C-Average" | grade=="D-Poor" |
      grade=="P-Pass" | grade=="S-Satisfactory"

*calculate number of college-level credits earned;
gen crsclcr=0
replace crsclcr=anycr*credits if crscl==1
```

Example: Cleaning Data (1) continued

- Clean transcript data
  - recode 'term' to indicate quarters elapsed since the third quarter of 2007 (summer 2007);

  Coding Sheme: term to quarter: spring (q1), summer (q3), fall (q4)

  Stata hint: jumps between spring and summer

```
gen time=0
replace time=1  if term=="FA07"
replace time=2  if term=="SP08"
replace time=4  if term=="SU08"
replace time=5 if term=="FA08"
replace time=6  if term=="SP09"
......
save "C:\DiX\Desktop\projects\AEFP\Data for AEFP\courseclean", replace
```

Example: Cleaning Data (2)

- Create student-level variables using cleaned transcript data
  - total number of college-level credits earned

  Stata hints: collapse

```
**total number of college-level credits earned
use "C:\DiX\Desktop\projects\AEFP\Data for AEFP\courseclean", clear
sort id
collapse (sum) crsclcr, by (id)
save "C:\DiX\Desktop\projects\AEFP\Data for AEFP\credits", replace
```

Example: Cleaning Data (3)

- Clean student-level demographic data
  - recode gender into female (1/0 dummy)

Coding Scheme: Gender: 1 -- Male; 2 -- Female

	gender	race	birthdate	id
1	2	1	08/02/1978	1
2	1	1	03/10/1986	2
3	1	1	03/16/1982	3
4	2	1	07/11/1986	4
5	2	1	05/23/1989	5
6	1	1	09/01/1981	6
7	1	1	05/03/1988	7

```
use "DiX\Desktop\projects\AEFP\Data for AEFP\student", clear

*recode gender;
gen female=0
replace female=1 if gender=="2"
```

Example: Cleaning Data (3) continued

- Clean student-level demographic data
  - recode race into a set of dummies

Coding Scheme: Race: 1 -- White; 2 -- Black; 3 -- American Indian;
4 -- Asian; 5 -- Hispanic; 6 -- Unknown

Stata hint: tab race

race	Freq.	Percent	Cum.
1	75	72.12	72.12
2	20	19.23	91.35
4	3	2.88	94.23
5	4	3.85	98.08
6	2	1.92	100.00
Total	104	100.00	

```
gen white=0
replace white=1 if race=="1"
gen black=0
replace black=1 if race=="2"
gen raceother=0
replace raceother=1 if race=="4" | race=="5" | race=="6"
```

Example: Cleaning Data (3) continued

- Clean student-level demographic data
  - calculate student age at the beginning of 2012

Stata hint: substr, destring, mdy: (date1-date2)/365.25

```
*calculate age at the beginning of 2012;
gen month_birth=substr(birthdate,1,2)
destring month_birth, replace
gen day_birth = substr(birthdate,4,2)
destring day_birth, replace
gen year_birth = substr(birthdate,7,4)
destring year_birth, replace

gen date_birth = mdy(month_birth,day_birth,year_birth)
gen date_2012 = mdy(1,1,2012)
gen agedays = date_2012 - date_birth
gen age2012 = agedays/365.25

keep id female white black raceother age2012
save "DiX\Desktop\projects\AEFP\Data for AEFP\studentclean", replace
```

Example: Cleaning Data (4)

- Create student-level variables using administrative and nsc data
  - highest degree received (BA or above, AA, Long-term Certificate, Short-term Certificate)

	type	public_pri~e	enrollment~n	enrollment~d	grad_date	degree_trans	id
22	4	Public	20090831	20091211			19
23	4	Public	20100119	20100506			19
24	4	Public			20100731	MASTER OF SOCIAL WORK	19
25	2	Public	20110119	20110510			22
26	4	Private	20120801	20121231			23
27	4	Private	20130101	20130430			23

```
use "DiX\Desktop\projects\AEFP\Data for AEFP\nsc", clear

*clean degree received;
tab degree_trans
gen award="BA"
replace award="AS" if degree_trans=="AAPSY"
replace award="" if degree_trans==""
```

degree_trans	Freq.	Percent	Cum.
AAPSY	1	16.67	16.67
BACHELOR OF ARTS	1	16.67	33.33
BACHELOR OF FINE ARTS	1	16.67	50.00
BACHELOR OF SCIENCE	2	33.33	83.33
MASTER OF SOCIAL WORK	1	16.67	100.00

Example: Cleaning Data (4) continued

```
keep  if award!=""
keep id award

*merge with award data;
append using
      "C:\Users\Fang\Desktop\projects\Capsee\works
      hop proposal\Data for AEFP\award"
*code degree ever earned;
gen ba=0
replace ba=1 if award=="BA"
gen aa=0
replace aa=1 if award=="AA" | award=="AA&S" |
      award=="AAA" | award=="AAS" | award=="AS"
gen lcert=0
replace aa=1 if award=="CERT" | award=="DIPL"
gen scert=0
replace scert=1 if award=="CSC"
```

```
**code the highest degree earned;
collapse (max) ba aa lcert scert, by (id)

gen bachelor=0
replace bachelor=1 if ba==1
gen associate=0
replace associate=1 if ba==0 & aa==1
gen longcertificate=0
replace longcertificate=1 if ba==0 &aa==0 & lcert==1
gen shortcertificate=0
replace shortcertificate=1 if ba==0 &aa==0 & lcert==0
      & scert==1

keep id bachelor associate longcertificate
      shortcertificate
save"DiX\Desktop\projects\AEFP\Data for
      AEFP\awardclean", replace
```

Example: Cleaning Data (5)

- Create quarter-level variables using wage data
  - adjust for CPI to 2010 dollars, formula? wagecpi=(100/CPI)*wage

Stata hint: destring, rename, merge

```
destring employment_year, generate(year)
rename employment_quarter quart

*merge with cip data;
sort year quart
merge m:1 year quart using "DiX\Desktop\projects\AEFP\Data for AEFP\cpi2010"

*adjust cpi;
generate wagecpi=100*(wage/cpi)
```

Example: Cleaning Data (5)

- Create quarter-level variables using wage data
  - calculate average quarterly earnings in 2012: Sum of wage/number of quarters worked

```
*only keep 2012 wage;
keep if year==2012
keep id wagecpi  quart

*collapse data so each student has only one entry for each quarter
collapse (sum)wagecpi, by (id time)

**calculate number of quarters observed and add wage together;
gen count=1
collapse (sum)wagecpi count, by (id)

*calculate average quarterly earnings;
gen wage2012=wagecpi/count
keep id wage2012
```

Merging Data

- Merge different data sets together:

```
drop _merge
merge 1:1 id using "DiX\Desktop\projects\AEFP\Data for AEFP\awardclean"
drop _merge
merge 1:1 id using "DiX\Desktop\projects\AEFP\Data for AEFP\credits"
drop _merge
merge 1:1 id using "DiX\Desktop\projects\AEFP\Data for AEFP\studentclean"
drop _merge
```

- Post-merging recode:

```
replace wage2012=0 if wage2012==.
replace bachelor=0 if bachelor==.
replace associate=0 if associate==.
replace longcertificate=0 if longcertificate==.
replace shortcertificate=0 if shortcertificate==.
```

Conclusions

- State administrative allows for exploration of the heterogeneity of pathways and course taking patterns
- Many different ways to test for how college influences student outcomes and earnings
- Many opportunities to perform validity checks
- Potential problems to watch out for
- Search for exogenous changes to identify causal influences of college choices on outcomes

Visit us on the web at capseecenter.org

We're also on Facebook and Twitter.

Center for Analysis of Postsecondary Education and Employment
Teachers College, Columbia University
525 West 120th Street, Box 174, New York, NY 10027
capsee@columbia.edu
212.678.3091

CAPSEE is funded through a grant (R305C110011) from the Institute of Education Sciences, U.S. Department of Education.