Using Selection on Unobserved Variables to Address Selection on Unobserved Variables

Christopher Taber

University of Wisconsin

October 18, 2016

Papers

I am basically going to talk about three different papers here:

- “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools, with J. Altonji and T. Elder,*Journal of Political Economy*, Vol. 113, February 2005.
- “An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," with J. Altonji and T. Elder, *Journal of Human Resources*, Fall 2005.
- “Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables,” with J. Altonji, T. Conley, and T. Elder, 2012.

In giving these lectures I will start from the beginning and talk about the basic framework, (which was done early), but I will use our current notation (which was done later)

I will discuss the Catholic school work and describe where it fits in and then discuss what is new.

The IV Model

To start lets think about a standard instrumental variables model.

$$Y_i = \alpha T_i + W_i' \Gamma + u_i.$$

The key assumption is that we have some instrument Z_i which is correlated with T_i, but

$$cov(Z_i, u_i) = 0.$$

One can never verify this assumption but must take it on face value

Of course a special case of this model is OLS in which $Z_i = T_i$.

Virtually all causal empirical work in economics makes some assumption analogous to this in some place.

The best justification for the instrument is random assignment

However, if Z_i were truly randomly assigned, it should not be correlated with the observable covariates either

Researchers have recognized this for a long time

- Balancing tests are standard in randomized control trials
- It is common to run a regression of Z_i on W_i and test whether these are related

For the standard reasons, testing is not the right way to guide empirical researchers. The problem manifests itself in two ways:

- Just because we don't reject the null does not mean that the assumption is right (perhaps just low power)
- If we do reject the null that doesn't mean the assumption is not approximately true

In other words what we really care about is the magnitude of the relationship not just the F-statistic

In order to judge the magnitude one needs a framework for thinking about it

Basic Model

$$Y_i = \alpha T_i + X_i' \Gamma_X + W_i^{*\prime} \Gamma^*$$

where W_i^* contains all possible covariates-those we get to see and those we might not get to see

We can write this as

$$\begin{aligned} W_i^{*\prime}\Gamma^* &= \sum_{j=1}^{K^*} W_{ij}\Gamma_j \\ &= \sum_{j=1}^{K^*} S_j W_{ij}\Gamma_j + \sum_{j=1}^{K^*} \left(1 - S_j\right) W_{ij}\Gamma_j \\ &= W_i'\Gamma + u_i \end{aligned}$$

where S_j is an indicator for whether W_{ij} is contained in the data set.

We need some way to characterize what it means for **"The Observables to be like the Unobservables"**

The most natural is to think of S_j as i.i.d so that the observables are just a random set of stuff that I could have observed.

This motivates the main idea: **If Selection on the unobservables is the same as selection on the observables how large would the bias be?**

Think about running a regression of Z_i on the observable index and unobservable:

$$\begin{aligned}&\text{Proj}\left(Z_i \mid X_i, W_i'\Gamma, u_i\right) \\ &\quad = \phi_0 + \phi_x' X_i + \phi\left(W_i'\Gamma\right) + \phi_u u_i.\end{aligned}$$

It turns out that when S_j is i.i.d.,

$$\phi_u \approx \phi$$

We actually want to think about this as an extreme case. Imagine two types of data collectors:

- An incompetent data collector would have no idea what he was doing and choose S_j at random. We show that yields the condition that (asymptotically)

$$\phi_u = \phi$$

- By contrast suppose we had a perfect data collector. That person would collect all of the variables that were correlated with Z_i so that the only unobservables left would be uncorrelated with Z_i. In that case

$$\phi_u = 0$$

The truth is probably somewhere in between.

We formalize this idea in two different ways.

The first is by adding the possibility of another unobservable ξ_i so that

$$Y_i = \alpha T_i + \sum_{j=1}^{K^*} S_j W_{ij} \Gamma_j + \sum_{j=1}^{K^*} \left(1 - S_j\right) W_{ij} \Gamma_j + \xi_i$$

Without this variable we will get "observables like unobservables"

It is there to pick up the fact that we think this is an extreme assumption and that selection on observables is likely greater than selection on unobservables.

Structurally it can represent measurement error or unanticipated events that occur between the data collection on W_{ij} and when the outcome Y_i is realized. That is if X_i, W_i, Z_i are all determined at time 0,

$$E(Y_i - \alpha T_i \mid \mathcal{I}_0) = X_i' \beta + W_i' \Gamma$$

Then

The second way we will formalize it is to allow the distribution of (W_{ij}, Z_i, Γ_j) conditional on $S_j = 1$ to differ from the distribution of (W_{ij}, Z_i, Γ_j) conditional on $S_j = 0$

This is the part of the approach I in progress and I won't focus on it

I will come back to that later, but forget about it for a little while

The Econometric Model

Lets formalize the model:

Since S_j does not vary across people, to get bight from its iidness we need our set of potential covariates to be growing large, so this will be thought of as a sequence of models

$$Y_i = \alpha T_i + X_i' \Gamma_x + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij} \Gamma_j + \xi_i$$

It embodies the idea that a large number of factors are important in determining outcomes in social science data and that none dominate.

It will turn out that X_i plays no important role in this going forward, so we can use the same trick we used in the IV lecture notes and regress everything else in the model on X_i and taking residuals and then just working with that.

Thus for generic variable M_{ij} define

$$\widetilde{M}_{ij} \equiv M_{ij} - Proj(M_{ij} \mid X_i; \mathcal{G}^K)$$

The $\mathcal{G}^K$ represents a two stage process:

1. First the micro data generation process is determined
2. Given the Data generation process $\mathcal{G}^K$, the data is generated

Conditional on $\mathcal{G}^{K^*}$ (which is what we would observe in a particular data set) the variance of elements of W_{ij} will differ.

Assumptions

$$\widetilde{Y}_i = \alpha \widetilde{T}_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \widetilde{W}_{ij} \Gamma_j + \xi_i$$

We need 4 basic assumptions which are essentially

1. variance of $\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \widetilde{W}_{ij} \Gamma_j$ doesn't blow up as K^* gets large
2. $cov(\widetilde{Z}_i, \widetilde{Y}_i)$ is well behaved as K^* gets large
3. S_j is i.i.d.
4. ξ_i is independent of everything else

Theorem

Define ϕ and ϕ_u such that

$$Proj\left(\widetilde{Z}_i \mid \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\tilde{W}_{ij}\Gamma_j, \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\tilde{W}_{ij}\Gamma_j+\xi_i; \mathcal{G}^{K^*}\right)$$
$$=\phi\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\tilde{W}_{ij}\Gamma_j\right)+\phi_u\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\tilde{W}_{ij}\Gamma_j+\xi_i\right).$$

Then under our assumptions , if the probability limit of ϕ is nonzero, then

$$\frac{\phi_u}{\phi}\ \underset{K^*\to\infty}{\xrightarrow{p}}\ \frac{\left(1-P_s\right)A}{\left(1-P_s\right)A+\sigma_\xi^2}$$

where

$$A\equiv\lim_{K^*\to\infty} E\left(\frac{1}{K^*}\sum \sigma_{j,j}^{K^*}\left(\Gamma_j\right)^2\right).$$

We can write this as:

Corollary

When $0 < P_s < 1$ *and* $\sigma_\xi^2 > 0$,

either

$$0 < \; plim(\phi_u) < plim(\phi),$$

$$plim(\phi) < plim(\phi_u) < 0,$$

or

$$0 = plim(\phi_u) = plim(\phi).$$

Lets not lose the forest for the trees-ϕ tells us the relationship between the instrument and the observables so the closer is ϕ to zero the smaller is the possible range of ϕ_u (with $\phi_u = 0$ when $\phi = 0$ which corresponds to

Intuition:

Think about the linear projection:

$$\widetilde{Z}_i = \phi_0 + \phi^* \sum_{j=1}^{K^*} \widetilde{W}_{ij} \Gamma_j + \varepsilon_i$$

Then

$$\widetilde{Z}_i = \phi_0 + \phi^* \sum_{j=1}^{K^*} S_j \widetilde{W}_{ij} \Gamma_j + \phi^* \sum_{j=1}^{K^*} (1 - S_j) \widetilde{W}_{ij} \Gamma_j + \varepsilon_i$$

Since $\sum_{j=1}^{K^*} \widetilde{W}_{ij} \Gamma_j$ is orthogonal to ε_i, $\sum_{j=1}^{K^*} S_j \widetilde{W}_{ij} \Gamma_j$ and $\sum_{j=1}^{K^*} (1 - S_j) \widetilde{W}_{ij} \Gamma_j$ will be as well.

Estimation of the Effects of Catholic Schools

- Goal: Measure Average Effect of Catholic High Schools on Test Scores, HS Graduation, College Attendance
- Why?
    - Assess merits of private schooling
    - Lessons for public schools?
    - Consequences of expansion of school choice, vouchers

Previous Literature:Very large.

Coleman, Hoffer and Kilgore (1982) and Coleman and Hoffer (1987) find positive effects on HS GRAD, College, test scores

They essentially just regress these variables on a dummy variable for attendance of a Catholic school

Result is highly controversial. Selection problem

Cain and Goldberger point out that Catholic school is not randomly assigned.

In particular, parents who send their children to Catholic school have shown an interest in their children by not picking standard option

They may be different in a lot of other important ways.

Is positive relationship causal? A couple papers have tried IV approaches.

Evans and Scwab and Neal

These guys want to estimate the effect of Catholic school on outcomes

Ideally one would like to have instrumental variables

They use two:

- Catholic religion
- Proximity of a Catholic school

Both are presumably closely related to whether a students attends a Catholic school or not

Neither should obviously be correlated with outcomes otherwise

(probably hard to believe that either are randomly assigned)

Both Evans and Schwab and Neal focus on high school graduation and college attendance because that is where effects seem strongest (not much with test scores)

Since dependent variable is binary they don't want to use 2SLS

Problem with linear probability model is that

$$E(Y_i \mid X_i) = X_i'\beta$$

but if Y_i is binary

$$\begin{aligned} E(Y_i \mid X_i) &= 1 \Pr(Y_i = 1 \mid X_i) + 0 \Pr(Y_i = 0 \mid X_i) \\ &= \Pr(Y_i = 1 \mid X_i) \end{aligned}$$

so

$$\Pr(Y_i = 1 \mid X_i) = X_i'\beta$$

But this is kind of weird

The probability has to be between 0 and 1, but this isn't guaranteed for a linear model

Instead use bivariate model

$$
\begin{aligned}
CH_i &= 1(g(X_i) + u_i > 0) \\
Y_i &= 1(\alpha CH_i + f(Z_i) + e_i > 0)
\end{aligned}
$$

This nonlinear/nonparametric model is identified if there is an exclusion restriction: i.e. something in X_i that isn't in Z_i

This is analogous to instrumental variables/selection

Evans and Schwab

Evans and Schwab focus on “Catholic” as an instrument

It is certainly correlated with whether one goes to Catholic school

No obvious reason for it to be correlated with outcomes

They play around with some other instruments as well.

TABLE II

EDUCATIONAL OUTCOMES OF HIGH SCHOOL STUDENTS BY SCHOOL TYPE

Sample	*HIGH SCHOOL GRADUATE*		*COLLEGE ENTRANT*[a]	
	Public schools	Catholic schools	Public schools	Catholic schools
Full sample	0.79	0.97	0.32	0.55
SOPHOMORE TEST SCORE MISSING	0.71	0.98	0.22	0.50
SOPHOMORE TEST FIRST QUARTILE	0.63	0.91	0.11	0.25
SOPHOMORE TEST SECOND QUARTILE	0.80	0.96	0.19	0.40
SOPHOMORE TEST THIRD QUARTILE	0.89	0.98	0.37	0.56
SOPHOMORE TEST FOURTH QUARTILE	0.95	0.99	0.62	0.78
PARENT EDUCATION MISSING	[illegible]	[illegible]	[illegible]	[illegible]

TABLE III
PROBIT ESTIMATES OF *HIGH SCHOOL GRADUATE* AND *COLLEGE ENTRANT* MODELS

	HIGH SCHOOL GRADUATE		*COLLEGE ENTRANT*	
Independent variable[a]	Probit coefficient	Marginal effect[b]	Probit coefficient	Marginal effect[b]
CATHOLIC SCHOOL	0.777	0.117	0.384	0.144
	(0.056)	(0.014)	(0.032)	(0.012)
FEMALE	0.041	0.006	0.021	0.008
	(0.029)	(0.004)	(0.026)	(0.010)
BLACK	0.132	0.020	0.170	0.064
	(0.045)	(0.007)	(0.042)	(0.014)
HISPANIC	0.080	0.012	−0.160	−0.060
	(0.037)	(0.006)	(0.036)	(0.014)
OTHER RACE	0.346	0.052	0.316	0.118
	(0.067)	(0.011)	(0.060)	(0.022)
FAMILY INCOME MISSING	−0.111	−0.017	−0.382	−0.143
	(0.068)	(0.010)	(0.055)	(0.021)

TABLE VI

MAXIMUM LIKELIHOOD ESTIMATES OF *HIGH SCHOOL GRADUATE* AND *COLLEGE ENTRANT* BIVARIATE PROBIT MODEL USING *CATHOLIC RELIGION* AS AN INSTRUMENT

Model	Other variables in X_i[b]	MLE estimates of bivariate probit model: Coefficient on *CATHOLIC SCHOOL*	Marginal effect[c]	Average treatment effect	ρ	2SLS estimate of coefficient on *CATHOLIC SCHOOL*
HIGH SCHOOL GRADUATE[a]						
(1)		0.777	0.117	0.130		0.096[d]
		(0.056)	(0.014)	(0.007)		(0.008)
(2)		0.859	0.133	0.141	−0.053	0.127
		(0.115)	(0.022)	(0.014)	(0.067)	(0.024)
(3)	*10TH GRADE TEST SCORE AND TEST MISSING*	0.678	0.078	0.114	0.028	0.103
		(0.126)	(0.018)	(0.017)	(0.072)	(0.024)
(4)	*STATE EFFECTS*	0.911	0.142	0.144	−0.050	0.114
		(0.121)	(0.027)	(0.015)	(0.072)	(0.024)
(5)	*10TH GRADE TEST SCORE, TEST MISSING, AND STATE EFFECTS*	0.746	0.124	0.121	0.025	0.134
		(0.132)	(0.028)	(0.016)	(0.077)	(0.030)
COLLEGE ENTRANT[a]						
(6)		0.384	0.144	0.132		0.137[d]
		(0.032)	(0.012)	(0.011)		(0.011)
(7)		0.288	0.109	0.098	0.067	0.148
		(0.079)	(0.033)	(0.028)	(0.049)	(0.030
(8)	*10TH GRADE TEST SCORE AND TEST MISSING*	0.211	0.078	0.064	0.124	0.098
		(0.083)	(0.034)	(0.026)	(0.052)	(0.024)
(9)	*STATE EFFECTS*	0.341	0.110	0.115	0.056	0.092
		(0.084)	(0.032)	(0.029)	(0.053)	(0.024)
(10)	*10TH GRADE TEST SCORE, TEST MISSING, AND STATE EFFECTS*	0.277	0.071	0.082	0.113	0.098
		(0.090)	(0.026)	(0.027)	(0.046)	(0.028)

Neal

Neal noticed that effects on high school graduation are larger for urban students

He focuses on number of Catholics or density of Catholic schools in county

Should be closely related to Catholic school attendance

No reason to expect it to be related to outcome

Table 4
Probit Analysis of High School Graduation

	Urban Counties		Nonurban Counties	
	Whites	Blacks and Hispanics	Whites	Blacks and Hispanics
Black	...	.211	...	.236
		(.059)		(.079)
Female	.107	.281	.202	.206
	(.058)	(.055)	(.053)	(.070)
Mom—high school graduate	.364	.257	.547	.361
	(.070)	(.066)	(.062)	(.097)
Dad—high school graduate	.342	.145	.277	.366
	(.072)	(.069)	(.063)	(.102)
Mom—college graduate	.252	.306	.230	.360
	(.131)	(.197)	(.166)	(.312)
Dad—college graduate	.113	.265	.148	.411
	(.100)	(.155)	(.124)	(.328)
Mom—professional	.149	.090	.120	−.127
	(.126)	(.131)	(.124)	(.213)
Dad—professional	.234	.099	.175	−.051
	(.082)	(.129)	(.086)	(.213)
Two-parent family	.506	.334	.403	.115
	(.068)	(.059)	(.061)	(.079)
Numerous family reading materials	.294	.199	.207	.197
	(.062)	(.067)	(.060)	(.101)
No family reading materials	−.600	−.141	−.539	−.367
	(.148)	(.083)	(.097)	(.082)
County population, 1980: 500,000–1,000,000	−.014	−.240	...	...
	(.070)	(.080)		
County population, 1980: >1,000,000	−.041	−.370	...	...
	(.075)	(.078)		
Percentage of families on welfare—county, 1980	−1.524	−.786	−1.287	.607
	(.577)	(.418)	(.604)	(.547)
Catholic school	.361	.854	.255	.511
	(.120)	(.177)	(.202)	(.431)
Sample graduation rate	.76	.64	.74	.70
Attending catholic schools	.09	.05	.03	.01
Sample size	2,626	2,434	3,110	1,597

Table 6
Bivariate Probit Analysis of High School Graduation Students from Urban Counties

	Catholic School Attendance		High School Graduation	
	White	Black and Hispanic	White	Black and Hispanic
Black	. . .	.179	. . .	.220
		(.157)		(.062)
Female	.124	.233	.100	.277
	(.081)	(.109)	(.060)	(.056)
Mom—high school graduate	.076	.303	.368	.249
	(.121)	(.156)	(.073)	(.069)
Dad—high school graduate	.180	.238	.333	.132
	(.120)	(.153)	(.073)	(.072)
Mom—college graduate	.147	.439	.251	.335
	(.133)	(.231)	(.139)	(.205)
Dad—college graduate	.194	.048	.105	.263
	(.107)	(.191)	(.106)	(.160)
Mom—professional	.032	−.017	.163	.083
	(.151)	(.185)	(.120)	(.132)
Dad—professional	.106	.417	.249	.097
	(.095)	(.162)	(.087)	(.144)
Two-parent family	.070	.284	.474	.321
	(.113)	(.129)	(.069)	(.062)
Numerous family reading materials	.128	.096	.273	.200
	(.096)	(.121)	(.064)	(.066)
No family reading materials	.215	−.602	−.572	−.134
	(.267)	(.301)	(.147)	(.086)
County population, 1980: 500,000–1,000,000	.104	.386	−.053	−.246
	(.105)	(.189)	(.071)	(.082)
County population, 1980: >1,000,000	.093	.446	−.097	−.371
	(.109)	(.188)	(.078)	(.080)
Percentage of families on welfare—county, 1980	1.113	1.148	−2.183	.053
	(.961)	(1.105)	(.617)	(.566)
Catholic	1.034	.831	. . .	. . .
	(.092)	(.140)		
Catholics/county population—1980	.199	.956	.608	. . .
	(.274)	(.485)	(.210)	
Catholic schools/square mile—county	1.739	.479	. . .	−.595
	(.534)	(.400)		(.243)
Catholic school	. . .	. . .	.724	1.122
			(.321)	(.686)
Error covariance	. . .	. . .	−.237	−.125
			(.179)	(.330)

Evaluation of Instrumental Variable Strategies

by Altonji, Elder, and Taber

- Explore the validity of Catholic, Proximity, and the interaction in three ways
  - Face plausibility of validity of exclusion restrictions based on observable factors
  - Reduced-form estimates in the sample of those who attend public grammar schools (8th graders)
  - Altonji, Elder, and Taber methodology applied to instrumental variables

Outline

- Data and differences in means of outcomes and observables by Catholic religion.
- Single-equation and IV estimates of Catholic schooling effects
- The effect of Catholic religion for students from public eighth grades
- Using the observables to assess the bias from unobservables
- Repeat for distance and the interaction $C_i \times D_i$.

Data

- NELS:88, restricted use file (eighth graders in 1988)
- NLS-72 public use file (seniors in HS in 1972)
- Dependent Variables:
  - High School graduation by 1994 (GED's considered dropouts)
  - College Attendance: Enrolled in 4-year college at 1994 survey
  - College Attendance: Enrolled in 4-year college by 1976 survey (NLS-72)
  - 12th grade math and reading test scores in both datasets
- Catholic high school (CH_i): 1 if current or last high school attended was Catholic, 0 otherwise
- Instruments:
  - Catholic religion (C_i): 1 if parents report being Catholic church members, 0 otherwise
  - Distance (D_i): Set of indicators for distance to closest Catholic high school
  - $C_i \times D_i$ [Can put in main effects and use interactions as instrument]

Using Religion

Model:

$$Y_i = \alpha CH_i + X_i' \gamma + \varepsilon_i,$$

where X_i is uncorrelated with ε_i.

The problem is that CH_i and potentially C_i may be correlated with the error term.

Table 1

Probit, Bivariate Probit, OLS, and 2SLS Estimates of Catholic Schooling Effects
NELS:88 and NLS-72
Weighted, Marginal Effects of Nonlinear Models Reported, (Huber-White Standard Errors in Parentheses)

	Excluded Instruments		
	(1)	(2)	(3)
	Catholic (C_i)	*Distance* (D_i)	*Catholic*×*Distance* ($C_i \times D_i$)
HS Graduation (NELS:88)			
Probit (controls exclude "instrument")	0.065 (0.025)	0.047 (0.025)	0.052 (0.026)
Bivariate Probit	0.128 (0.032)	-0.007 (0.085)	-0.022 (0.119)
OLS	0.041 (0.014)	0.021 (0.014)	0.023 (0.015)
2SLS	0.34 (0.08)	-0.04 (0.10)	0.09 (0.11)
College in 1994 (NELS:88)			
Probit (controls exclude "instrument")	0.094 (0.022)	0.085 (0.022)	0.077 (0.022)
Bivariate Probit	0.170 (0.055)	0.103 (0.062)	-0.043 (0.070)
OLS	0.128 (0.026)	0.119 (0.026)	0.111 (0.026)
2SLS	0.40 (0.10)	0.31 (0.11)	-0.11 (0.12)
College in 1976 (NLS-72)			
Probit (controls exclude "instrument")	0.068 (0.016)	0.070 (0.016)	0.067 (0.016)
Bivariate Probit	-0.002 (0.028)	-0.052 (0.035)	-0.080 (0.035)
OLS	0.071 (0.015)	0.075 (0.016)	0.072 (0.016)
2SLS	0.06 (0.04)	0.44 (0.20)	-0.25 (0.11)

Notes:

Table 2

OLS and 2SLS estimates of Catholic Schooling Effects
NELS:88 and NLS-72
Weighted, (Huber-White Standard Errors in Parentheses)

	Excluded Instruments		
	(1)	(2)	(3)
	Catholic (C_i)	*Distance* (D_i)	*Catholic*$\times$*Distance* $(C_i \times D_i)$
12th Grade Reading Score (NELS:88)			
OLS	1.16 (0.37)	1.03 (0.37)	1.14 (0.38)
2SLS	1.40 (1.54)	-1.09 (1.84)	1.24 (1.82)
12th Grade Math Score (NELS:88)			
OLS	1.03 (0.31)	1.00 (0.31)	0.92 (0.32)
2SLS	2.64 (1.21)	2.43 (1.45)	-2.63 (1.57)
12th Grade Reading Score (NLS-72)			
OLS	2.06 (0.34)	2.54 (0.37)	2.50 (0.36)
2SLS	-1.34 (0.99)	8.69 (4.53)	0.50 (2.32)
12th Grade Math Score (NLS-72)			
OLS	1.52 (0.33)	1.77 (0.35)	1.71 (0.36)
2SLS	-0.07 (0.96)	11.05 (4.47)	-3.94 (2.27)

Notes:

- OLS: Find positive effect of CH_i on HS graduation (0.04), college attendance (0.13 and 0.07) and test scores
- In NELS:88, IV estimates for HS grad and college are implausibly large. Test scores also show “negative selection”
- In NLS-72, IV estimates for college are roughly the same as OLS, and test scores are negative
- Key question: Do IV estimates bolster probit and OLS evidence? Is the true effect substantial?

Table 3a

Comparison of Means of Key Variables by Value of Distance, Catholic, and their Interaction NELS:88

	(1) Overall Mean	(2) Difference by C_i	(3) Difference by D_i	(4) Difference by $C_i \times D_i$
Demographics				
Female	0.50	0.01	0.00	0.00
Asian	0.04	0.01	0.04	-0.02
Hispanic	0.10	0.19	0.08	0.03
Black	0.13	-0.15	0.08	-0.13
White	0.73	-0.05	-0.20	0.12
Family Background				
Mother's education	13.14	-0.26	0.17	-0.36
Father's education	13.42	-0.07	0.17	-0.31
Log of family income	10.20	0.11	0.12	-0.02
Mother only in house	0.15	-0.04	0.02	-0.03
Parent married	0.78	0.06	-0.02	0.03
Geography				
Rural	0.32	-0.15	-0.44	0.05
Suburban	0.44	0.06	0.08	0.00
Urban	0.24	0.09	0.36	-0.05
Expectations				
Schooling expectation	15.17	0.15	0.31	-0.06
Very sure to graduate high school	0.83	-0.01	0.00	-0.01
Parents expect some college	0.88	0.04	0.05	-0.02
Parents expect college grad	0.78	0.03	0.06	-0.04
Expect white collar job	0.46	0.03	0.06	-0.01
8th Grade Variables				
Delinquency Index	0.69	-0.05	0.03	-0.04
Got into fight	0.27	-0.01	0.01	0.05
Rarely completes homework	0.21	-0.05	0.00	0.00
Frequently disruptive	0.13	-0.02	-0.01	0.00
Repeated grade 4-8	0.08	-0.03	0.01	-0.03
Risk Index	0.72	-0.07	-0.01	0.01
Grades Composite	2.89	0.04	0.00	0.07
Unpreparedness Index	10.82	0.00	0.08	-0.09
8th Grade reading score	50.32	0.40	0.03	1.15
8th Grade math score	50.33	0.55	0.45	0.06
Outcomes				

Table 3b

Comparison of Means of Key Variables
by Value of Distance, Catholic, and their Interaction
NLS-72

	(1)	(2)	(3)	(4)
	Overall Mean	Difference by C_i	Difference by D_i	Difference by $C_i \times D_i$
Demographics				
Female	0.50	-0.01	0.03	0.03
Hispanic	0.04	0.11	0.01	-0.07
Black	0.15	-0.15	0.04	-0.08
Family Background				
Mother's education	12.19	-0.13	0.16	-0.33
Father's education	12.43	0.06	0.40	-0.32
Log of family income	8.93	0.07	0.11	-0.03
Father Blue Collar	0.24	0.01	-0.03	-0.01
Low SES Indicator	0.29	-0.05	-0.06	0.00
English Primary Language	0.92	-0.06	-0.02	0.03
Family Receives Daily Newspaper	0.88	0.04	0.06	0.01
Mother Works	0.50	-0.06	0.03	0.01
Geography				
Rural	0.23	-0.14	-0.30	0.05
Suburban	0.48	0.06	0.02	-0.04
Urban	0.29	0.08	0.28	-0.01
Expectations				
Decided to go to college pre-HS	0.41	-0.01	0.04	-0.06
Outcomes				
Enrolled in college by 1976	0.38	0.01	0.05	-0.06
Reading Score	50.01	0.30	0.46	0.55
Math Score	49.98	0.58	0.40	-0.10
Years of Academic PSE, 1979	1.61	0.03	0.22	-0.23
Attended Catholic HS	0.06	0.19	0.07	0.15

Means by Catholic religious affiliation

- Differences by C_i appear in variables measured prior to 8th grade enrollment
- Overall picture: use of C_i as an IV will likely positively bias estimates in NELS:88, and perhaps in NLS-72

Effect of Catholic Religion for Public Eighth Graders

- Starting point: identify a sample of persons for whom Catholic high school is not a serious option.
- Only 0.3% of public school 8th graders attend Catholic high school
- Interpret the coefficient on C_i in a single equation model as an estimate of the direct effect of Catholic religion on the outcome

Bias Formula

To figure out bias you can use partitioned regression.

That is if

$$Y = X_1'\beta_1 + X_2'\beta_2 + v$$

then running Y on X is equivalent to running X_1 on X_2 taking residuals and then running Y on those residuals (that is $\widehat{\beta}_1$ is identical)

Suppose

$$\begin{aligned}
\text{Proj}\left(CH_i \mid X_i, C_i\right) &= X_i'\beta + \lambda C_i \\
\widetilde{CH}_i &= \text{Proj}\left(CH_i \mid X_i, C_i\right) - X_i'\beta - \lambda C_i \\
\text{Proj}\left(C_i \mid X_i\right) &= X_i'\pi \\
\widetilde{C}_i &= C_i - X_i'\pi
\end{aligned}$$

Then think about regression Y_i on $X_i'\beta + \lambda C_i$ and X_i.

First think of regressing $X_i'\beta + \lambda C_i$ on X_i and taking residual

$$\begin{aligned} X_i'\beta + \lambda C_i - Proj\left(X_i'\beta + \lambda C_i \mid X_i\right) &= X_i'\beta + \lambda C_i - X_i'\beta - \lambda X_i'\pi \\ &= \lambda \widetilde{C}_i \end{aligned}$$

Now use the fact that for a simple regression model, the coefficient on the slope coefficient is $Cov(Y, X)/Var(X)$.

So regression of Y_i on $\lambda\widetilde{C}_i$ can be written as

$$\begin{aligned} \widehat{\alpha} &\to \frac{Cov(\lambda\widetilde{C}_i, Y_i)}{Var(\lambda\widetilde{C}_i)} \\ &= \frac{Cov(\widetilde{C}_i, \alpha\left[X_i'\beta + \lambda X_i'\pi + \lambda\widetilde{C}_i + \widetilde{CH}_i\right] + X_i'\gamma + \varepsilon_i)}{\lambda Var(\widetilde{C}_i)} \\ &= \alpha + \frac{Cov(\widetilde{C}_i, \varepsilon_i)}{\lambda Var(\widetilde{C}_i)} \end{aligned}$$

How do we estimate the bias?

- Suppose there is an event p_i for which $\Pr(CH_i = 1 \mid p_i) = 0$.
  - In our application this event is attendance of a public eighth grade by individual i.
- Consider a regression of Y_i on X_i and C_i conditional on p_i. Coefficient on C_i in this regression will converge to $\frac{Cov(\widetilde{C}_i, \varepsilon_i)}{Var(\widetilde{C}_i)}$.
- Obtain a consistent estimate of the bias ψ by taking the ratio $\frac{Cov(\widetilde{C}_i, \varepsilon_i)}{Var(\widetilde{C}_i)} / \lambda$ or by estimating the parameter ψ in the regression model

$$Y_i = X_i'\gamma + [C_i\widehat{\lambda}]\psi + \omega_i$$

  on the public eighth grade sample.
- This isn't perfect. There is still a problem of Selection bias, public school eigth graders are a selected sample– however, positive selection into Catholic 8th grades among Catholic students Will bias ψ downward

Table 4

Comparison of 2SLS Estimates[1] and Bias Implied by OLS Estimation of $Y_i = X_i'\gamma + [Z_i'\widehat{\lambda}]\psi + \omega_i$ on the Public Eighth Grade Subsample[2]; Various Outcomes and instruments; NELS:88 Sample Weighted, (Huber-White Standard Errors in Parentheses)

OUTCOME (Y)	**INSTRUMENTS** (Z_i)		
	(1)	(2)	(3)
	Catholic	*Distance*	*Catholic*×*Distance*
High School Graduation			
Implied Bias in 2SLS (ψ)	0.34 (0.08)	-0.05 (0.12)	0.15 (0.12)
2SLS Coefficient	0.34 (0.08)	-0.04 (0.10)	0.09 (0.11)
College Attendance			
Implied Bias in 2SLS (ψ)	0.29 (0.11)	0.37 (0.12)	-0.23 (0.13)
2SLS Coefficient	0.40 (0.10)	0.31 (0.11)	-0.11 (0.12)
12th Grade Reading Score			
Implied Bias in 2SLS (ψ)	0.54 (1.68)	-0.51 (2.08)	-0.50 (1.99)
2SLS Coefficient	1.40 (1.54)	-1.09 (1.84)	1.24 (1.82)
12th Grade Math Score			
Implied Bias in 2SLS (ψ)	1.85 (1.41)	1.83 (1.69)	-4.37 (2.06)
2SLS Coefficient	2.64 (1.21)	2.43 (1.45)	-2.63 (1.57)

Using observables to assess bias due to unobservables:

As I talked about before we derived an alternative to the assumption that

$cov(\widetilde{C}_i, \varepsilon_i) = 0.$

In this case we can write it as

$$\frac{cov(\widetilde{C}_i, \varepsilon_i)}{var(\varepsilon_i)} = \frac{cov(X_i'\pi, X_i'\gamma)}{var(X_i'\gamma)}$$

For an indicator variable such as C_i, This condition can be rewritten as

$$\frac{E(\varepsilon_i \mid C_i = 1) - E(\varepsilon_i \mid C_i = 0)}{Var(\varepsilon_i)}$$
$$= \frac{E(X_i'\gamma \mid C_i = 1) - E(X_i'\gamma \mid C_i = 0)}{Var(X_i'\gamma)}.$$

Use this assumption to approximate bias in 2SLS estimates:

plim$(\widehat{\alpha} - \alpha)$

$$
\begin{aligned}
&= \frac{cov(\widetilde{C}_i, \varepsilon_i)}{\lambda var\left(\widetilde{C}_i\right)} \\
&= \frac{var\left(C_i\right)}{\lambda var\left(\widetilde{C}_i\right)} \left[E(\varepsilon_i \mid C_i = 1) - E(\varepsilon_i \mid C_i = 0)\right] \\
&= \frac{var\left(C_i\right)}{\lambda var\left(\widetilde{C}_i\right)} \frac{Var(\varepsilon_i)}{Var(X_i'\gamma)} \left[E(X_i'\gamma \mid C_i = 1) - E(X_i'\gamma \mid C_i = 0)\right].
\end{aligned}
$$

Using *AET* Methodology, NELS:88
Weighted, (Huber-White Standard Errors in Parentheses)

	Excluded Instruments		
	(1)	(2)	(3)
	Catholic	*Distance*	*Catholic×Distance*
HS Graduation			
2SLS Coefficient	0.34 (0.08)	-0.04 (0.10)	0.09 (0.11)
Bias 1	0.52 (0.23)	0.15 (0.16)	0.14 (0.24)
Bias 2	0.84 (0.26)	0.06 (0.14)	...
College in 1994			
2SLS Coefficient	0.40 (0.10)	0.31 (0.11)	-0.11 (0.12)
Bias 1	0.45 (0.21)	0.46 (0.22)	0.15 (0.26)
Bias 2	0.45 (0.21)	0.40 (0.20)	...
12th Reading Score			
2SLS Coefficient	1.40 (1.54)	-1.09 (1.84)	1.24 (1.82)
Bias 1	1.18 (1.06)	2.49 (1.59)	2.59 (1.14)
Bias 2	1.42 (1.07)	2.11 (1.40)	...
12th Math Score			
2SLS Coefficient	2.64 (1.21)	2.43 (1.45)	-2.63 (1.57)
Bias 1	2.02 (0.75)	1.76 (1.03)	1.42 (0.88)
Bias 2	1.87 (0.74)	1.72 (0.98)	...

Comparison between Bivariate Probits and 2SLS

Evans and Schwab (1995) and Neal (1997) use bivariate probits with exclusion restrictions, with sensible results.

Identification in BP model:

$$\begin{aligned} CH_i &= 1(g(X_i) + u_i > 0) \\ Y_i &= 1(\alpha CH_i + f(Z_i) + e_i > 0) \end{aligned}$$

Identification of α requires two assumptions:

1. Either parametric assumptions on the distribution of u_i and e_i, or support conditions on $g(\cdot)$
2. Either an exclusion restriction or parametric restrictions on $f(\cdot)$ and $g(\cdot)$

- So, an exclusion restriction is not necessary to identify BP models in practice.

- Procedure: loose replication on urban minority subsamples
  - Coefficients are not very sensitive to exclusion restrictions in BP models
  - More importantly, standard errors also insensitive

Comparison of Linear and Non-Linear Models of College Attendance in NLS-72
(Standard Errors in Parentheses)
[Marginal Effects of Non-Linear Models in Brackets]

	Sample					
	Non-whites in cities (N=1532)			Whites in cities (N=5326)		
	Nonlinear Models (Probits) (1)	Linear Models (OLS/2SLS) (2)	Nonlinear Models Holding X_i Constant[4] (3)	Nonlinear Models (Probits) (4)	Linear Models (OLS/2SLS) (5)	Nonlinear Models Holding X_i Constant[4] (6)
Single Equation Model (OLS/Probit)	0.640	0.239		0.253	0.093	
	(0.198)	(0.070)		(0.062)	(0.022)	
	[0.239]			[0.093]		
Two Equation Models:						
Excluded Instruments:						
$\%CCH_i$ and CH/P_i	1.471	1.375	5.541	0.048	0.115	0.084
	(0.442)	(0.583)	(2.082)	(0.250)	(0.158)	(0.783)
	[0.517]		[0.706]	[0.018]		[0.031]
C_i and $\%CCH_i$	0.879	0.054	0.012	-0.090	-0.036	-0.084
	(0.523)	(0.309)	(1.443)	(0.121)	(0.050)	(0.148)
	[0.329]		[0.004]	[-0.033]		[-0.031]
C_i, $\%CCH_i$, and CH/P_i	1.106	0.331	1.302	-0.085	-0.034	-0.069
	(0.460)	(0.254)	(0.706)	(0.118)	(0.048)	(0.125)
	[0.409]		[0.471]	[-0.031]		[-0.025]
C_i only	0.761	-0.093	-0.505	-0.133	-0.056	-0.149
	(0.543)	(0.324)	(1.638)	(0.130)	(0.054)	(0.151)
	[0.285]		[-0.148]	[-0.049]		[-0.054]
$C_i \times D_i$	1.333	2.572	1.409	-0.121	-0.395	2.624
	(0.516)	(2.442)	(1.276)	(0.262)	(0.169)	(5.173)
	[0.478]		[0.497]	[-0.044]		[0.559]
None	1.224	...		-0.094	...	
	(0.542)	...		(0.301)	...	

Conclusions

- Unfortunately, none of the candidate instruments appears to be valid in this situation
- It appears important to isolate effects of exclusion restrictions from effects of functional form in identification
- We are forced to try to use some other approach

Altonji, Elder, Taber Approach

- Few students from Public 8th grade attend Catholic high schools
- Many students from Catholic 8th grade go on to Public High School
- Among Catholic 8th Grade students:
    - On the basis of observables doesn't look like much selection (at least in comparison with the full sample).
    - However, there is a huge difference in high school dropouts: 2% versus 10%
    - Selection on unobservables would have to be huge to explain this finding.
- We formalized the idea that using degree of selection on observables can provide a guide to bias from selection on unobservable variables in this paper

Outline

- Data and Comparison of Catholic 8th graders and full sample by High School Sector
- Probit and regression results for Catholic 8th graders
- What do observables say about selection bias for Catholic Schools?
- Conclusions.

Table 1
Comparison of Means of Key Variables by Sector

Variable	Full Sample			Catholic 8th Grade		
	Public 10th	Cath 10th	Difference	Public 10th	Cath 10th	Difference
Demographics	(N=11,167)	(N=672)		(N=366)	(N=640)	
FEMALE	0.52	0.45	-0.07	0.61	0.50	-0.11
ASIAN	0.03	0.04	0.01	0.05	0.05	0.00
HISPANIC	0.09	0.09	0.00	0.08	0.09	0.01
BLACK	0.10	0.09	-0.01	0.07	0.11	0.04
WHITE	0.78	0.78	0.00	0.80	0.74	-0.06
Family Background						
MOTHER'S EDUCATION IN YEARS	13.21	13.96	0.75	13.34	13.88	0.54
FATHER'S EDUCATION IN YEARS	13.49	14.51	1.01	13.39	14.38	0.99
LOG OF FAMILY INCOME	10.23	10.72	0.49	10.47	10.66	0.19
MOTHER ONLY IN HOUSE	0.14	0.09	-0.05	0.07	0.09	0.02
PARENT MARRIED	0.79	0.89	0.10	0.90	0.88	-0.02
PARENTS CATHOLIC	0.28	0.82	0.54	0.84	0.84	0.00
Geography						
RURAL	0.36	0.03	-0.33	0.13	0.01	-0.12
SUBURBAN	0.45	0.51	0.06	0.40	0.48	0.08
URBAN	0.19	0.46	0.27	0.47	0.51	0.04
DISTANCE TO CLOSEST CATHOLIC HS, MILES	22.16	2.97	-19.19	6.91	2.37	-4.53
Expectations [1]						
SCHOOLING EXPECTATIONS IN YEARS	15.25	15.97	0.72	15.52	15.92	0.40
VERY SURE TO GRADUATE HS	0.84	0.89	0.05	0.84	0.90	0.06
PARENTS EXPECT AT LEAST SOME COLLEGE	0.89	0.98	0.09	0.94	0.98	0.04
PARENTS EXPECT AT LEAST COLLEGE GRAD	0.79	0.92	0.13	0.88	0.91	0.03
STUDENT EXPECTS WHITE-COLLAR JOB	0.47	0.61	0.14	0.55	0.59	0.04
8th Grade Variables						
DELINQUENCY INDEX, RANGE FROM 0 TO 4	0.64	0.53	-0.11	0.54	0.46	-0.08
STUDENT GOT INTO FIGHT	0.24	0.23	-0.02	0.20	0.19	-0.01
STUDENT RARELY COMPLETES HOMEWORK	0.19	0.08	-0.11	0.08	0.06	-0.01
STUDENT FREQUENTLY DISRUPTIVE	0.12	0.08	-0.05	0.08	0.08	0.00
STUDENT REPEATED GRADE 4-8	0.06	0.02	-0.05	0.03	0.02	-0.01
RISK INDEX, RANGE FROM 0 TO 4	0.69	0.35	-0.34	0.39	0.39	0.00
GRADES COMPOSITE	2.94	3.16	0.22	3.09	3.20	0.11
UNPREPAREDNESS INDEX, FROM 0 TO 25	10.77	11.08	0.31	10.84	11.02	0.17
8TH GRADE READING SCORE	51.19	55.05	3.86	54.12	55.59	1.47
8TH GRADE MATHEMATICS SCORE	51.13	54.57	3.44	52.89	53.98	1.09
Outcomes						
10TH GRADE READING STANDARDIZED SCORE	51.02	54.69	3.66	54.63	54.62	-0.01
10TH GRADE MATH STANDARDIZED SCORE	51.12	55.03	3.91	53.40	54.52	1.12
12TH GRADE READING STANDARDIZED SCORE	51.20	54.60	3.40	53.25	54.70	1.45
12TH GRADE MATH STANDARDIZED SCORE	51.20	55.54	4.34	53.13	55.63	2.49
ENROLLED IN 4 YEAR COLLEGE IN 1994	0.31	0.59	0.28	0.38	0.61	0.23
HS GRADUATE	0.85	0.98	0.13	0.88	0.98	0.10

Notes:
(1) The Expectations variables are not included in our empirical models

Means of Controls and Outcomes by 8th Grade

- Huge difference in HS grad rates, college attendance
- smaller differences in test scores
- 8th grade outcomes more favorable for kids in Catholic high schools.
- Difference in observables is much smaller for Catholic 8th grade sample.

Table 3

OLS and Probit Estimates of Catholic High School Effects[1,2]
in Subsamples of NELS:88
Weighted, (Huber-White Standard Errors in Parentheses)
[Marginal Effects in Brackets[3]]

	Full Sample				*Catholic 8th Grade Attendees*			
	Controls							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	None	Fam. BG, city size, and region.[4]	(2) plus 8th grade tests	(3) plus other 8th grade measures[5]	None	Fam. BG, city size, and region.[4]	(2) plus 8th grade tests	(3) plus other 8th grade measures[5]
HS Graduation								
Probit	0.97	0.57	0.48	0.41	0.99	0.88	0.95	1.27
	(0.17)	(0.19)	(0.22)	(0.21)	(0.24)	(0.25)	(0.27)	(0.29)
	[0.123]	[0.081]	[0.068]	[0.052]	[0.105]	[0.084]	[0.081]	[0.088]
Pseudo R^2	0.01	0.16	0.21	0.34	0.11	0.35	0.44	0.58
College in 1994								
Probit	0.73	0.37	0.33	0.32	0.60	0.48	0.56	0.60
	(0.08)	(0.09)	(0.09)	(0.09)	(0.13)	(0.15)	(0.15)	(0.15)
	[0.283]	[0.106]	[0.084]	[0.074]	[0.236]	[0.154]	[0.154]	[0.149]
Pseudo R^2	0.02	0.19	0.29	0.34	0.04	0.18	0.29	0.36
12th Grade Reading Score								
OLS	4.28	2.08	1.18	1.14	1.92	0.17	0.37	0.33
	(0.47)	(0.54)	(0.38)	(0.38)	(0.82)	(0.98)	(0.63)	(0.62)
R^2	0.01	0.19	0.60	0.60	0.01	0.19	0.59	0.62
12th Grade Math Score								
OLS	4.86	1.98	1.07	0.92	2.79	1.10	1.46	1.14
	(0.44)	(0.54)	(0.34)	(0.32)	(0.77)	(1.00)	(0.53)	(0.46)
R^2	0.01	0.26	0.72	0.74	0.02	0.26	0.73	0.77

Notes:
(1) NELS:88 third follow-up and 2nd follow-up panel weights used for the educational attainment and 12th grade models, respectively.
(2) Sample sizes for Full sample: N=8560 (HS Graduation), N=8315 (College Attendance), N=8116 (12th Reading), N=8119 (12th Math).
For Catholic 8th Grade sample, N=859 (HS Graduation), N=834 (College Attendance), N=739 (12th Reading), N=739 (12th Math).
(3) Marginal effects of probit models are computed as average derivatives of the probability of an outcome with respect to catholic high school attendance.
(4) Control sets (2)-(4) include race (white/nonwhite), hispanic origin, gender, urbanicity (3 categories), region (8 categories), and distance to the nearest Catholic high school (5 categories). Family background variables used as controls include log family income, mother's and father's education, 5 dummy variables for marital status of the parents, and 8 dummy variables for household composition.
(5) "Other 8th grade measures" include measures of attendance, attitudes toward school, academic track, achievement, and behavioral problems (from teacher, parent, and student surveys). The NELS:88 variables used are bys55a, bys55e, bys55f, byt1_2, bys56e, byp50, byp57e bylep, bys55b, bys55d byrisk, bygrads, byp51, and bys78a-c, and also teacher survey variables regarding whether a student performs below ability, completes homework, is attentive or disruptive in class, or is frequently absent or tardy. See Appendix A for more details.
(6) The pseudo-R^2 for probit models is defined as $\frac{var(X'\gamma)}{1+var(X'\gamma)}$.

Table 4

OLS, Fixed Effect, and Probit Estimates of Catholic High School Effects by Race and Urban Residence. Full Set of Controls[1,2]
(Huber-White Standard Errors in Parentheses)
[Marginal Effects in Brackets[3]]

	Sample			
	(1)	(2)	(3)	(4)
	Urban and Suburban White Only	Urban and Suburban Minorities Only	Urban White Only	Urban Minorities Only
HS Graduate	(N=3799)	(N=1308)	(N=1002)	(N=697)
Sample Mean	0.88	0.80	0.88	0.80
Probit	0.443	0.524	1.176	1.592
	(0.279)	(0.338)	(0.417)	(0.673)
	[0.046]	[0.085]	[0.091]	[0.191]
College in 1994	(N=3695)	(N=1258)	(N=981)	(N=666)
Sample Mean	0.37	0.26	0.32	0.26
Probit	0.354	0.697	0.506	0.677
	(0.107)	(0.201)	(0.167)	(0.303)
	[0.087]	[0.158]	[0.110]	[0.144]
12th Grade Reading Score	(N=3638)	(N=1051)	(N=978)	(N=561)
Sample Mean	52.94	47.72	53.33	47.61
OLS	1.30	-0.72	1.59	-0.19
	(0.44)	(0.98)	(0.67)	(1.39)
12th Grade Math Score	(N=3638)	(N=1053)	(N=979)	(N=563)
Sample Mean	53.09	47.33	53.90	48.88
OLS	1.07	1.17	1.69	1.25
	(0.35)	(0.76)	(0.52)	(1.09)

Notes:
(1) All models include controls for hispanic origin, gender, region, citysize, distance to the nearest Catholic school (5 categories), family background, 8th grade tests, and other 8th grade measures. (from teacher, parent, and student surveys). See Table 3 notes 1 and 2.
(2) NELS:88 third follow-up and 2nd follow-up panel weights used for the educational attainment and 12th grade models, respectively.
(3) Marginal effects of probit models are computed as average derivatives of the probability of an outcome with respect to Catholic school attendance.

Probit and Regression Estimates for Catholic 8th graders

- Find a strong positive effect of Catholic High School on high school graduation (.08) and college attendance (.15)
- The effect on 10th grade test scores is small (standard deviation of test score is around 10)
- Key question: How much of the estimated high school effect on educational attainment is real, and how much is due to selection bias?

Consider the following “treatment effect” model without exclusion restrictions,

$$
\begin{aligned}
CH_i^* &= g(X_i) + u_i \\
CH_i &= 1(CH_i^* \geq 0) \\
Y_i &= \alpha CH_i + h(X_i) + \varepsilon_i,
\end{aligned}
$$

The econometrician observes (X_i, CH_i, Y_i), but not the unobservables (u_i, ε_i), or the latent variable CH_i^*.

Assume the unobservables (u_i, ε_i) are independent of the observables X_i and consider identification of the parameter α.

- We are essentially only one parameter (or one equation) short of identification. In particular if α were known, the system of equations would be identified.
- Under normality and linear indices, model is identified, but semiparametric identification requires such an excluded variable.
- We treat this model as if it were underidentified by one parameter. In particular, we act as if ρ is not identified.
- Relationship between observables can solve identification problem.

Table 5

Sensitivity Analysis: Estimates of Catholic High School Effects Given Different Assumptions on The Correlation of Disturbances in Bivariate Probit Models in Subsamples of NELS:88[1]. Modified Control Set[2].
(Huber-White Standard Errors in Parentheses) [Marginal Effects in Brackets]

	Correlation of Disturbances[3]					
	$\rho = 0$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$
HS Graduation:						
Full Sample	0.459	0.271	0.074	-0.132	-0.349	-0.581
(Raw difference=0.12)	(0.150)	(0.150)	(0.150)	(0.148)	(0.145)	(0.140)
	[0.058]	[0.037]	[0.011]	[-0.021]	[-0.060]	[-0.109]
Catholic 8th Graders	1.036	0.869	0.697	0.520	0.335	0.142
(Raw difference=0.08)	(0.314)	(0.313)	(0.310)	(0.306)	(0.299)	(0.290)
	[0.078]	[0.064]	[0.050]	[0.038]	[0.025]	[0.011]
Urban	1.095	0.905	0.706	0.499	0.282	0.053
Minorities	(0.526)	(0.538)	(0.549)	(0.560)	(0.570)	(0.578)
(Raw difference=0.22)	[0.176]	[0.157]	[0.132]	[0.101]	[0.062]	[0.013]
College Attendance:						
Full Sample	0.331	0.157	-0.019	-0.196	-0.376	-0.558
(Raw difference=0.31)	(0.070)	(0.070)	(0.070)	(0.068)	(0.067)	(0.064)
	[0.084]	[0.039]	[-0.005]	[-0.047]	[-0.087]	[-0.125]
Catholic 8th Graders	0.505	0.336	0.165	-0.008	-0.184	-0.362
(Raw difference=0.23)	(0.121)	(0.120)	(0.119)	(0.117)	(0.114)	(0.110)
	[0.140]	[0.093]	[0.045]	[-0.002]	[-0.050]	[-0.099]
Urban	0.447	0.269	0.090	-0.091	-0.272	-0.455
Minorities	(0.282)	(0.282)	(0.280)	(0.276)	(0.269)	(0.259)
(Raw difference=0.30)	[0.116]	[0.062]	[0.020]	[-0.020]	[-0.057]	[-0.091]

Observed and Unobserved Variables

Now we will use the assumption that selection on observables is similar to selection on unobservables

Does this make sense with Catholic Schools?

- Data on a broad set of family background measures, teacher evaluations, test scores, grades, and behavioral outcomes in eighth grade
- Measures have substantial explanatory power for the outcomes that we examine, and a large number of the variables play a role, particularly in the case of high school graduation and college attendance.
- The relatively large number and wide variety of observables that enter into our problem suggests that observables may provide a useful guide to the unobservables.

- Relationship among the unobservables likely to be weaker than the relationship among the observables because shocks that occur after eighth grade are excluded from X. These will influence high school outcomes but not the probability of starting a Catholic high school. Consequently,

$$\frac{cov(g(X_i), h(X_i))}{var(h(X_i))} > \frac{cov(u_i, \varepsilon_i)}{var(\varepsilon_i)}$$

- We think of our estimates of α that impose the conditions as an informal lower bound for α.

Using the Condition to identify Model

Estimate

$$\begin{aligned} CH_i &= 1(X_i'\beta + u_i > 0) \\ Y_i &= 1(X_i'\gamma + \alpha CH_i + \varepsilon > 0) \\ (u_i, \varepsilon_i) &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \end{aligned}$$

subject to the restriction

$$\rho = \frac{cov(u_i, \varepsilon_i)}{var(\varepsilon_i)} = \frac{Cov(X_i'\beta, X_i'\gamma)}{Var(X_i'\gamma)}$$

Table 6

Sensitivity of Estimates of Catholic Schooling Effects on College Attendance and HS Graduation to Assumptions about Selection Bias in NELS:88, Catholic 8th Grade Subsample[1,2], Modified Control Set[3] (Huber-White Standard Errors in Parentheses) [Marginal Effects in Brackets]

Model:

$$CH = 1(X'\beta + u > 0)$$

$$Y = 1(X'\gamma + \alpha CH + \epsilon > 0)$$

Estimation Method 1: β, γ, and α estimated simultaneously as a constrained bivariate probit model:

Model	Constraint on ρ	HS Graduation Coefficients		College Attendance Coefficients	
		$\widehat{\rho}$	$\widehat{\alpha}$	$\widehat{\rho}$	$\widehat{\alpha}$
(1)	$\rho = \frac{cov(X\beta, X\gamma)}{var(X\gamma)}$	0.24	0.59	0.24	0.11
		(0.13)	(0.33)	(0.06)	(0.16)
			[0.05]		[0.03]
(2)	$\rho = 0$	0	1.04	0	0.51
			(0.31)		(0.12)
			[0.08]		[0.14]

Estimation Method 2: 2-step, with β obtained from a univariate probit, γ from a univariate probit using the public 8th grade subsample. Next, α is computed from a bivariate probit with β fixed at this initial value and γ fixed up to 6 proportionality factors.[4]

Model	Constraint on ρ	HS Graduation Coefficients		College Attendance Coefficients	
		$\widehat{\rho}$	$\widehat{\alpha}$	$\widehat{\rho}$	$\widehat{\alpha}$
(3)	$\rho = \frac{cov(X\beta, X\gamma)}{var(X\gamma)}$	0.09	0.94	0.27	0.06
		(0.08)	(0.30)	(0.05)	(0.10)
			[0.07]		[0.02]

Estimation Method 3: β, γ, and α estimated simultaneously as a constrained semiparametric model[5]:

$$CH = 1(X'\beta + \theta + u > 0)$$

$$Y = 1(X'\gamma + \alpha CH + \theta + \epsilon > 0)$$

Model	Constraint on ρ, where $\rho = \frac{var(\theta)}{1+var(\theta)}$	HS Graduation Coefficients		College Attendance Coefficients	
		$\widehat{\rho}$	$\widehat{\alpha}$	$\widehat{\rho}$	$\widehat{\alpha}$
(4)	$\rho = \frac{cov(X\beta, X\gamma)}{var(X\gamma)}$	0.25	0.80	0.25	0.15
		(0.16)	(0.37)	(0.09)	(0.22)
			[0.05]		[0.04]

Notes:

(1) Estimation performed on a sample of Catholic 8th grade attendees from NELS:88. N=859 for the HS graduation sample, and N=834 for the college attendance sample.

(2) NELS:88 3rd follow-up sampling weights used in the computations.

(3) Due to computational difficulties, several variables were excluded from the control sets in the bivariate probit models. See Table 5, note 2.

(4) The categories of proportionality factors are demographics/family background, test scores, behavioral problems, school attendance and attitudes toward school, grades and achievement, and distance measures. The coefficients and (standard errors) of the proportionality factors for these categories are 0.82 (0.19), 0.87 (0.22), 0.92 (0.03), 1.07 (0.04), 0.59 (0.08), and 0.90 (6.08) respectively, in the high school graduation case. For college attendance, the coefficients and (standard errors) are 0.80 (0.01) , 1.01 (0.04), 0.95 (0.15), 0.43 (0.17), 1.44 (0.03), and 1.04 (1.59).

(5) Models estimated as univariate probits conditional on θ, the distribution of which is estimated nonparametrically.

Results:

- We use two alternative methods to estimate γ.
- For Method 1, in the case of High School graduation,

The estimate of

$$\rho = \frac{cov(u, \varepsilon)}{var(u)} = \frac{Cov(X_i'\beta, X_i'\gamma)}{Var(X'\gamma)} = 0.24$$

and the estimate of α falls to 0.59 (0.33)[0.05].

- For method 2, ρ is only 0.09, and α is 0.94 (0.30)[0.07].
- Consequently, even with the extreme assumption of equal selection on observables and unobservables imposed, there is evidence for a substantial positive effect of attending Catholic high school on high school graduation.
- The results for college attendance follow a similar pattern, but with the extreme assumption imposed most of the effect of Catholic High School is gone.

Summary of Empirical Work

- **Catholic High Schools:**
  - Substantially raise the graduation rate
  - Probably increase college attendance.
    Have little effect on math or reading scores. Perhaps a small positive effect on 12th grade math.
  - We don’t provide precise point estimates of effect sizes. “Lower bound” estimate is large.
- Correlation between indices of the observed control variables is useful in assessing the importance of selection bias in both single equation

Problems Addressed in Work in Progress

There are two major problems with what we did before

(we were aware of both of them when we wrote the paper, but we didn't think they were driving the results and we didn't want to do everything in one paper)

First Issue

A key assumption of the model that K^* is increasing with sample size

The standard errors did not account for this complication

In order to do that we must take the model more seriously

Second Issue

The argument is that the observables are like the unobservables.

However in our empirical work we assumed that u_i (the W_is we don't observe) is uncorrelated with the W_is we do observe

However, the W_is are pretty clearly correlated with each other, so this is a really goofy assumption

Note that it is not the theorem that is wrong-that allowed for the observables and the unobservables to be correlated

The problem is that the theorem applies to the actual Γ which you will not be able to estimate without further assumptions

There is a natural solution to this

- Write down a model for the relationship between covariates
- Estimate the model using the observables
- Use the model to get the relationship between the observables and unobservables

This is what we do here, the most natural is the factor model

The factor model

We make use of a factor model:

$$\begin{aligned} \widetilde{W}_{ij} &= \frac{1}{\sqrt{K^*}} \widetilde{F}_i' \Lambda_j + v_{ij} \\ \sigma_j^2 &\equiv Var(v_{ij}) \end{aligned}$$

where all of these error terms are iid

Dividing by $\sqrt{K^*}$ guarantees that the variance of $\widetilde{Y}_i$ that is due to the factor, $\widetilde{F}_i$ is stable as K^* rises

This model satisfies the two technical assumptions above that keep the variance and covariance finite.

The rest of the model

$$\begin{aligned}
\widetilde{Y}_i &= \alpha_0 \widetilde{T}_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \widetilde{W}_{ij} \Gamma_j + \left[\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \left(1 - S_j\right) \widetilde{W}_{ij} \Gamma_j + \xi_i \right] \\
\widetilde{Z}_i &= \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \widetilde{W}_{ij} \beta_j + \left[\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \left(1 - S_j\right) \widetilde{W}_{ij} \beta_j + \psi_i \right] \\
\widetilde{T}_i &= \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j \widetilde{W}_{ij} \delta_j + \left[\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \left(1 - S_j\right) \widetilde{W}_{ij} \delta_j + \omega_i \right]
\end{aligned}$$

where all of these error terms are iid across i

Need

$$cov(\psi_i, \xi_i) = 0$$
$$cov(\psi_i, \omega_i) \neq 0$$

The Newest Part

Now we no longer assume that S_j is i.i.d. but allow the distribution of $(\Gamma_j, \beta_j, \sigma_j^2, \lambda_j)$ to depend on S_j

Specifically we allow

$$\begin{aligned} E(\Lambda_j \Gamma_j \mid S_j = 0) =& \rho_{\lambda\gamma} E(\Lambda_j \Gamma_j \mid S_j = 1) \\ E(\Lambda_j \beta_j \mid S_j = 0) =& \rho_{\lambda\beta} E(\Lambda_j \beta_j \mid S_j = 1) \\ E(\sigma_j^2 \Gamma_j^2 \mid S_j = 0) =& \rho_{\sigma\gamma} E(\sigma_j^2 \Gamma_j^2 \mid S_j = 1) \\ E(\sigma_j^2 \Gamma_j \beta_j \mid S_j = 0) =& \rho_{\sigma\gamma\beta} E(\sigma_j^2 \Gamma_j \beta_j \mid S_j = 1) \end{aligned}$$

where the empirical researcher can explore the robustness of the results to various choices of the ρs.

Note if these are all one, we are back in the observables like the unobservables case.

We define

$$P_{s0} \equiv \Pr(S_j = 1) \approx \frac{K}{K^*}$$

where K is the observable number of covariates.

Now that we have a model it is just a matter of estimating it

It turns that for our estimator we need K^* grows at a slower rate than N, so that in practice

$$\frac{K^*}{N} \to 0$$

We do asymptotics taking joint limits

I suspect that we could allow K^* and N to grow at the same rate. We would have a few bias terms we would have to adjust the estimator to account for.

In general the model is not point identified

Thus we do not obtain a point estimate but rather estimate a set of which α_0 will be an element

As a reminder in the case when the ρs are all 1:

$$\frac{\phi_u}{\phi} \approx \frac{(1 - P_s)\, A}{(1 - P_s)\, A + \sigma_\xi^2}$$

so

- If $\sigma_\xi = 0$, we would have $\phi_u \approx \phi$
- If $P_s = 1$ we would have $\phi_u \approx 0$
- For the cases in between, because of attenuation bias it is straight forward to show that one gets something in between

In practice we estimate 3 parameters:

$$\theta \equiv \left(\alpha, P_S, \sigma_\xi^2\right)$$

with two equations (explained below):

$$q^1(\theta_0) = 0$$
$$q^2(\theta_0) = 0$$

with the additional restrictions that

$$0 < P_{s0} \leq 1$$
$$\sigma_{\xi 0}^2 \geq 0$$

Thus the identified set will be the set of α's that are consistent with these conditions.

Typically one end will occur when $P_{S0} = 1$ (IV) and the other when $\sigma_{\xi 0}^2 = 0$ (obs. like uno.)

Let $q(\theta) = \left[\begin{array}{cc} q^1(\theta) & q^2(\theta) \end{array}\right]'$, then

$$Q(\theta) = q(\theta)'\Omega q(\theta),$$

is the objective function

We find the set of θ that minimize this objective function.

The process works in three steps:

Stage 1	Estimate factor structure-this part does not depend on θ
Stage 2	Given θ, estimate slope coefficients Γ
Stage 3	Calculate $Q(\theta)$

Stage 1: Factor model estimation

First we estimate the renormalized parameter $\lambda \equiv \sqrt{P_{S_0}}\Lambda$ as well as σ^2_{vj}.

It turns out we can get a closed form estimate of λ as

$$\widehat{\lambda}_j = \frac{\frac{K}{K-1}\sum_{\ell \neq j} \frac{1}{N}\sum_{i=1}^{N} \widetilde{W}_{ij}\widetilde{W}_{i\ell}}{\sqrt{\frac{1}{(K-1)}\sum_{\ell_1}\sum_{\ell_2} \frac{1}{N}\sum_{i=1}^{N} \widetilde{W}_{i\ell_1}\widetilde{W}_{i\ell_2}}}$$

For each $\sigma_{j_1}^2$ we only have one moment equation and use the obvious estimator, for for each $j = 1, .., K$,

$$\widehat{\sigma_j^2} = \frac{1}{N} \sum_{i=1}^{N} \left(\widetilde{W_{ij}} \right)^2 - \frac{\widehat{\lambda}_j^2}{K}.$$

Stage 2

The estimator we will use is the following. We are estimating the 3 parameters $\theta = \left(\alpha, P_s, \sigma_\xi^2\right)$ with true values $\theta_0 = \left(\alpha_0, P_{s0}, \sigma_{\xi 0}^2\right)$.

Without getting into details it turns out that

$$\widehat{\gamma}(\theta) \equiv \left[\frac{P_s + (1 - P_s)\,\rho_{\lambda\gamma}}{KP_s}\widehat{\lambda}\widehat{\lambda}' + \widehat{\Sigma}\right]^{-1} \frac{1}{N}\widetilde{W}'\left(\widetilde{Y} - \alpha\widetilde{T}\right)$$

is a good estimator of Γ

(note that if there is no factor loading, $\lambda = 0$ or $P_s = 1$ this is analogous to OLS)

Stage 3

We are estimating the 3 parameters $\theta = \left(\alpha, P_s, \sigma_\xi^2\right)$ with true values $\theta_o = \left(\alpha_0, P_{s0}, \sigma_{\xi 0}^2\right)$.

We show that there are only 2 moments that provide identifying information about the three parameters Why?

We define our estimator based on the following system of equations.

$$
\begin{aligned}
q^1_{N,K^*}(\theta) =& \frac{1}{N}\sum_{i=1}^{N} \widetilde{Z}_i \left(\widetilde{Y}_i - \alpha \widetilde{T}_i\right) - \left(\frac{P_s + (1-P_s)\,\rho_{\lambda\gamma}}{P_s}\frac{\widehat{\gamma}(\theta)'\,\widehat{\lambda}}{\sqrt{K}}\right)\left(\frac{P_s + (1-P_s)\,\rho_{\lambda\beta}}{P_s}\frac{\widehat{\beta}'\widehat{\lambda}}{\sqrt{K}}\right) \\
& - \left(\frac{P_s + (1-P_s)\,\rho_{\sigma\gamma\beta}}{P_s}\right)\widehat{\beta}'\widehat{\Sigma}\widehat{\gamma}(\theta) \\
q^2_{N,K^*}(\theta) =& \frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{Y}_i - \alpha \widetilde{T}_i\right)^2 - \left(\frac{P_s + (1-P_s)\,\rho_{\lambda\gamma}}{P_s}\frac{\widehat{\gamma}(\theta)'\,\widehat{\lambda}}{\sqrt{K}}\right)^2 \\
& - \left(\frac{P_s + (1-P_s)\,\rho_{\sigma\gamma}}{P_s}\right)\widehat{\gamma}(\theta)'\widehat{\Sigma}\widehat{\gamma}(\theta) - \sigma^2\xi
\end{aligned}
$$

Note that the first expression is like the standard moment condition you would have in IV, and the second equation is basically the R^2 of the regression

Our set estimate is

$$\widehat{\Theta} \equiv \{\theta : Q(\theta) \approx 0\}$$

Consistency (In progress)

Theorem

Under our Assumptions, $\widehat{\Theta}$ converges to the identified set

Asymptotic Distribution

We next show that the distribution of q is normal and derive the variance covariance matrix

Theorem

Assuming our factor model for W, and the Assumptions above and that $K^{*3}/N^2 \to 0$, $\sqrt{K^*} q_{N,K^*}(\theta_0)$ is asymptotically normal and we derive its complicated Var/Cov matrix

Concluding Thoughts

We think this approach will be useful in many applications

We also think of this as just the beginning. The basic idea of using observables to say something about unobservables can be extended to other models and one can try alternative assumptions.

Note that it is not a panecea.

- When there is little selection on the observables (as in the Public 8th grade sample) it will give tight bounds
- When there is a lot of selection on the observables (as is the case for Catholic as an instrument) it will give wide bounds