Applied Microeconometrics I

Lecture 5: Identification based on observables (continued)

Cristina Bratu

Aalto University

September 19, 2023

`Lecture Slides`

What did we do last time?

- Conditional Independence Assumption (CIA)

$$E[Y_{0i}|D_i = 1, X_i] = E[Y_{0i}|D_i = 0, X_i]$$

- How to condition on X?
  - Matching
  - Regression

What did we do last time?

- Things to worry about
- Omitted variable bias

$$
\begin{aligned}
Y_i &= \alpha^s + \rho^s S_i + u_i \\
Y_i &= \alpha + \rho S_i + \gamma A_i + e_i \\
\hat{\rho}^s_{OLS} &= \tfrac{Cov(Y,S)}{Var(S)} = \rho + \underbrace{\gamma \frac{Cov(A,S)}{Var(s)}}_{OVB}
\end{aligned}
$$

What did we do last time?

- Adding more controls is not always better - **bad controls**
  - Bad controls: X that are themselves caused by D
  - Example: Effect of college $(C_i = 1)$ among white collar workers $(W_i = 1)$
  - Assume that the treatment C_i is randomly assigned:
    $$\{Y_{0i}, Y_{1i}, W_{0i}, W_{1i}\} \perp\!\!\!\perp C_i$$
  - Can we estimate: $E[Y_{1i} - Y_{0i}|W_{1i} = 1]$?
  - Comparing the average outcomes we get:

$$\begin{aligned} & E[Y_i|W_i = 1, C_i = 1] - E[Y_i|W_i = 1, C_i = 0] \\ = \quad & \underbrace{E[Y_{1i} - Y_{0i}|W_{1i} = 1]}_{\text{Causal effect}} + \underbrace{E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]}_{\text{Selection bias}} \end{aligned}$$

Reminder: Interpreting results of a regression

- Probability distribution of the 'true' effect β
- Summarized with two moments of this distribution:
  - Point estimate: expected value
  - Standard error: provides information about the accuracy, or precision, of the estimate
- Stars are sometimes used to report significance levels
- Another useful way to summarize this distribution:
  - 95% confidence interval : point estimate $\pm$ 2*standard error
- $\hat{\beta}$ not statistically different from zero $\nRightarrow$ β is equal to zero
  - Precisely estimated zeros vs. uninformative estimates
- Statistical significance vs. economic significance

Reminder: Interpreting results of a regression

- Some examples - impact of taking this course on your lifetime income (in euros)
  - 100,000 (100,000)
  - 100 (100,000)
  - 100 (30)
  - 100,000 (20,000)
- Corollary: Estimates are useful when they are precise

Identification based on observables (continued)

- Main threats to validity
  1. Omitted variables
  2. Bad controls
  3. Measurement error in the independent variable
  4. Measurement error in the dependent variable?
  5. Measurement error in the controls

Measurement error in the independent variable

- Suppose that the amount of cabbages Y_i produced by a lot i depends on the daily rainfall x_i during spring, which changes from lot to lot because of local weather conditions
- Rainfall is arguably random and we are interested in the causal relationship

$$Y_i = \mu + \tau x_i + v_i$$

- I have a device that gives a daily rainfall measure $\tilde{x}_i$ of the rain falling on the lot, **with a random error** e_i so that $Cov(x, e) = 0$:

$$\tilde{x}_i = x_i + e_i$$

  which implies that:

$$Cov(\tilde{x}, e) = Cov(x + e, e) = Var(e)$$

- Assume also $Cov(v, e) = 0$
- This kind of measurement error is called **classical-errors-in-variables (CSV)**

Measurement error in the independent variable

- Hence, if we plug in:
$$Y_i = \mu + \tau\tilde{x}_i - \tau e_i + v_i = \mu + \tau\tilde{x}_i + (v_i - \tau e_i)$$
- Now the OLS estimate of τ can be written as:
$$\hat{\tau} = \tau\frac{Var(x)}{Var(x)+Var(e)}$$
- If $Var(e) \neq 0$, the term multiplying τ is always less than one $\implies$ **attenuation bias**
- As $Var(e) \to \infty \implies \hat{\tau} \to 0$
- Even if CIA applies, measurement error in the independent variable would still bias our estimates

Measurement error in the dependent variable

- Assume now that the only imperfect measure we are dealing with is that of the dependent variable:

$$Y_i^* = \mu + \tau x_i + v_i$$

- and we can only observe Y, which is an imperfect measure of Y^*, such that $Y = Y^* + e$
- What is relevant is how e is correlated with other regressors. Let us plug Y into the regression

$$Y_i - e_i = \mu + \tau x_i + v_i$$
$$Y_i = \mu + \tau x_i + (e_i + v_i)$$

- If e is uncorrelated with x, we can consistently estimate our regression and all the usual statistics are valid for inference.

Measurement error in the control variable

- The idea behind the CIA is that we can control for selection
- This suggests that the sensitivity of our estimate of our parameter of interest to additional controls is an indication of selection bias
- Example

$$\begin{aligned} Y_i &= \beta^s S_i + e_i^s \\ Y_i &= \beta S_i + \gamma X_i + e_i \\ X_i &= \rho S_i + v_i \end{aligned}$$

- OVB formula tells us that

$$\beta^s - \beta = \gamma\rho$$

Measurement error in the control variable

- But often we are forced to use proxies as controls (e.g. ability proxies, coarse geographical identificatiors, proxies for parental background etc.)
- The use of these kinds of proxies introduces measurement error in control variables
- The case of classical measurement error:

$$\tilde{X}_i = X_i + u_i$$

  where $E(u_i) = 0$ and $Cov(X, u) = 0$
- In addition we assume that $Cov(X, S) = Cov(\tilde{X}, S)$

Measurement error in the control variable

- Now we run
$$Y = \beta^m S_i + \gamma^m \tilde{X}_i + e_i^m$$
- The OLS estimates of β and γ are:

$$\begin{aligned}
\gamma^m &= \Lambda\gamma \\
\beta^m &= \beta + \gamma\rho(1-\Lambda) \\
\text{where } \Lambda &= \frac{Var(S)Var(X) - Cov(X,S)^2}{[Var(X)+Var(u)]Var(S) - Cov(X,S)^2} < 1
\end{aligned}$$

- Note that when $Var(u) = 0$ then $\Lambda = 1$ which implies that $\gamma^m = \gamma$ and $\beta^m = \beta$

Measurement error in the control variable

- But if $Var(u) > 0$, then $\gamma^m < \gamma$ and

$$\beta^s - \beta^m = \gamma\rho\Lambda < \gamma\rho = \beta^s - \beta$$

- We are underestimating the sensitivity of β^s to controls because our estimate of γ is attenuated due to measurement error
- Notice, however, that we can always estimate

$$\tilde{X} = \rho^m S_i + u_i + e_i$$

- Since measurement error is now in the dependent variable, we get an unbiased estimate of ρ:

$$\rho^m = \frac{Cov(X,S)}{Var(S)}$$

- Hence, a test for $\rho^m = 0$ is still a valid test for selection bias

Example: Returns to education with controls for family background and ability

Zhuan Pei, Steve Pischke, and Hannes Schwandt (2019):
'Poorly Measured Confounders are More Useful on the Left Than on the Right',
Journal of Business & Economic Statistics, Vol. 37., No. 2, 205-16.

Pei et al

- Illustrative example of the consequences of introducing noisy controls to account for selection bias
- Classic question: What are the returns to education
- Selection bias: Unobserved ability and family background
- Pei et al use a well-known American data set to estimate returns to education, controlling for an ability proxy (KWW score) and introducing variables that might proxy for family background:
    - Mother's years of education
    - Library card at age 14
    - Body height

Pei et al: Results

Table 2. Regressions for returns to schooling and specification checks controlling for the KWW score

	Log hourly earnings					Mother's years of education	Library card at age 14	Body height in inches
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.0609	0.0596	0.0608	0.0603	0.0591	0.2500	0.0133	0.0731
	(0.0059)	(0.0060)	(0.0059)	(0.0059)	(0.0060)	(0.0422)	(0.0059)	(0.0416)
KWW score	0.0070	0.0068	0.0069	0.0069	0.0067	0.0410	0.0076	0.0145
	(0.0015)	(0.0016)	(0.0016)	(0.0015)	(0.0016)	(0.0107)	(0.0016)	(0.0117)
Mother's years of education		0.0053			0.0048			
		(0.0037)			(0.0037)			
Library card at age 14			0.0097		0.0045			
			(0.0215)		(0.0216)			
Body height in inches				0.0078	0.0075			
				(0.0034)	(0.0034)			
p-values								
Coefficient comparison test		0.161	0.651	0.156	0.084			

Pei et al: Conclusions

- Adding noisy controls for family background has only a small effect on the coefficient of years of education
- Based on this we might erroneously conclude that there is no selection bias
- However, the correlation of years of education and family background proxies is highly significant which implies that there is selection

Example: Returns to university quality

Stacy Berg Dale and Alan B. Krueger (2002):
'Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables',
The Quarterly Journal of Economics, Vol. 117., 4 (2), 1491-1527.

Dale and Krueger: Motivation

- Does attending a more selective university lead to higher earnings?
- Students who attend more selective universities may have greater earnings capacity regardless of which university they attend
- Most studies try to control for differences in student attributes that are correlated with earnings and the selectivity of universities
- But in many countries, and especially in the U.S., college entry is determined by characteristics that are not observed by researchers

Dale and Krueger: Identification strategy

- Compare university selectivity and earnings among students who are accepted and rejected by a comparable set of universities
- Example in the Finnish context
  - Take two students, A and B, that are accepted to Turku and Tampere but rejected by Aalto
  - A decides to go to Turku and B decides to go to Tampere
  - Compare the earnings of these students
- University admission is based on:
  - Factors that are observable to the researcher (grades etc.)
  - Factors that are observable to the university but not to the researcher (enrtance exam, interviews etc.)
- Looking within matched set of students can help overcome the bias due to unobservables

Dale and Krueger: Identification strategy (from Angrist-Pischke textbook)

Table 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Dale and Krueger: Identification strategy

- Assume that the admission to a university is determined by observable characteristics X_{i1}, unobservable characteristics X_{i2}, and idiosyncratic luck
- Each university j accepts the applicant i based on his or her latent quality Z_{ij} if:

$$Z_{ij} = \gamma_1 X_{i1} + \gamma_2 X_{i2} + e_{ij} > C_j$$

  and rejects otherwise
- Earnings are determined by:

$$W_i = \beta_0 + \beta_1 Q_j + \beta_2 X_{1i} + \beta_3 X_{i2} + \epsilon_i$$

  where Q is the university quality

Dale and Krueger: Identification strategy

- Since X_{2i} is unobservable, we are forced to estimate:

$$W_i = \beta_0^{'} + \beta_1^{'} Q_j + \beta_2^{'} X_{1i} + u_i$$

- Since applicants that are admitted to higher quality universities have on average higher values of X_{2i} it is likely that $Cov(Q, u) > 0$ and $\hat{\beta}_1^{'} > \hat{\beta}_1$
- Introducing a full set of dummies to control for groups of students who received the same admission decision will absorb differences in X_{i2}
- Run:

$$W_i = \beta_0^{'} + \beta_1^{'} Q_j + \beta_2^{'} X_{1i} + \sum_{j=1}^{J} \beta_3 D_{ij} + u_i$$

Dale and Krueger: Results

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135 (.055)	.095 (.052)	.086 (.034)	.007 (.038)	.003 (.039)	.013 (.025)
Own SAT score ÷ 100		.048 (.009)	.016 (.007)		.033 (.007)	.001 (.007)
Log parental income			.219 (.022)			.190 (.023)
Female			−.403 (.018)			−.395 (.021)
Black			.005 (.041)			−.040 (.042)
Hispanic			.062 (.072)			.032 (.070)
Asian			.170 (.074)			.145 (.068)
Other/missing race			−.074 (.157)			−.079 (.156)
High school top 10%			.095 (.027)			.082 (.028)
High school rank missing			.019 (.033)			.015 (.037)
Athlete			.123 (.025)			.115 (.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Dale and Krueger: Results

Table 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212 (.060)	.152 (.057)	.139 (.043)	.034 (.062)	.031 (.062)	.037 (.039)
Own SAT score ÷ 100		.051 (.008)	.024 (.006)		.036 (.006)	.009 (.006)
Log parental income			.181 (.026)			.159 (.025)
Female			−.398 (.012)			−.396 (.014)
Black			−.003 (.031)			−.037 (.035)
Hispanic			.027 (.052)			.001 (.054)
Asian			.189 (.035)			.155 (.037)
Other/missing race			−.166 (.118)			−.189 (.117)
High school top 10%			.067 (.020)			.064 (.020)
High school rank missing			.003 (.025)			−.008 (.023)
Athlete			.107 (.027)			.092 (.024)
Average SAT score of schools applied to ÷ 100				.110 (.024)	.082 (.022)	.077 (.012)
Sent two applications				.071 (.013)	.062 (.011)	.058 (.010)
Sent three applications				.093 (.021)	.079 (.019)	.066 (.017)
Sent four or more applications				.139 (.024)	.127 (.023)	.098 (.020)

Dale and Krueger: Checking the identifying assumption

- Key assumption behind the Dale-Krueger approach is that the university students choose among the set of universities to which they were admitted is unrelated to unobservables X_2
- Can we test this assumption?
- We can check the relationship between the proxies for X_2 and private school attendance, once we control for the universities that the student was admitted to
- Identical to testing for $\rho^m = 0$ in the Pei at al paper

Dale and Krueger: Balancing test

TABLE 2.5
Private school effects: Omitted variables bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165 (.196)	1.130 (.188)	.066 (.112)	.128 (.035)	.138 (.037)	.028 (.037)
Female		−.367 (.076)			.016 (.013)	
Black		−1.947 (.079)			−.359 (.019)	
Hispanic		−1.185 (.168)			−.259 (.050)	
Asian		−.014 (.116)			−.060 (.031)	
Other/missing race		−.521 (.293)			−.082 (.061)	
High school top 10%		.948 (.107)			−.066 (.011)	
High school rank missing		.556 (.102)			−.030 (.023)	
Athlete		−.318 (.147)			.037 (.016)	
Average SAT score of schools applied to ÷ 100			.777 (.058)			.063 (.014)
Sent two applications			.252 (.077)			.020 (.010)
Sent three applications			.375 (.106)			.042 (.013)
Sent four or more applications			.330 (.093)			.079 (.014)

Dale and Krueger: why application dummies work

Assume short regression includes no controls (col. 1 in Table 2.3)
Assume long regression adds SAT scores as controls (col. 2)
Recall $OVB = \beta^s - \beta = \gamma\rho = .212 - .152 = .06 = .051 \times 1.165$

Now assume short regression includes only application dummies (col. 4 in Table 2.3)
Assume long regression adds SAT scores as controls + application dummies (col. 5)
Recall $OVB = \beta^s - \beta = \gamma\rho = .034 - .031 = .003 = 0.36 \times .066$

The effect of the omitted SAT on earnings falls from .051 to .036 in the regression with application dummies
The relationship between SAT and private school attendance goes from 1.165 to 0.066 in the regression with application dummies

$\Longrightarrow$ conditional on the application dummies, students who go to private vs. public are not very different in terms of their SAT scores

Dale and Krueger: Conclusions

- Simply controlling for observables suggests that returns to university quality are substantial
- However, if we control for the set of universities that the applicant is admitted to, the returns are zero
- This suggests that positive returns are driven by selection bias

What did we do last time?

$$
\begin{aligned}
& E[Y_i|W_i = 1, C_i = 1] - E[Y_i|W_i = 1, C_i = 0] \\
= \; & E[Y_{1i}|W_{1i} = 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0] \\
= \; & E[Y_{1i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1] \\
= \; & E[Y_{1i}|W_{1i} = 1] - E[Y_{0i}|W_{1i} = 1] \\
& + E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1] \\
= \; & \underbrace{E[Y_{1i} - Y_{0i}|W_{1i} = 1]}_{\text{Causal effect}} + \underbrace{E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]}_{\text{Selection bias}}
\end{aligned}
$$

Bias due to measurement error in the independent variable

$$
\begin{aligned}
\frac{Cov(Y,\tilde{x})}{Var(\tilde{x})} &= \frac{Cov(\mu+\tau x+v, x+e)}{Var(x+e)} \\
&= \tau\frac{Var(x)}{Var(x+e)} + \frac{\tau Cov(x,e)}{Var(x+e)} \\
&+ \frac{Cov(x,v)}{Var(x+e)} + \frac{Cov(v,e)}{Var(x+e)}
\end{aligned}
$$

Recall the assumptions:
$Cov(x,e)=0; Cov(x,v)=0; Cov(v,e)=0$

$$
\begin{aligned}
&= \tau\frac{Var(x)}{Var(x)+Var(e)+2Cov(x,e)} \\
&= \tau\frac{Var(x)}{Var(x)+Var(e)}
\end{aligned}
$$

Measurement error in the control variable

- We have that

$$\begin{aligned}\hat{\beta}^s &= \frac{Cov(Y,S)}{Var(S)} \\ &= \frac{Cov(\beta S + \gamma X + e, S)}{Var(S)} \\ &= \beta + \gamma \frac{Cov(S,X)}{Var(S)} \\ &= \beta + \gamma \rho\end{aligned}$$

- So it follows that:

$$\hat{\beta}^s - \beta = \gamma \rho$$

Measurement error in the control variable

- Denote $Cov(X,Y) = \sigma_{XY}$ and $Var(X) = \sigma^2_X$
- Use OLS formula for the case of two independent variables

$$
\begin{aligned}
\hat{\beta}^m &= \frac{\sigma^2_{\tilde{X}}\sigma_{YS} - \sigma_{\tilde{X}S}\sigma_{Y\tilde{X}}}{\sigma^2_{\tilde{X}}\sigma^2_S - (\sigma_{\tilde{X},S})^2} \\
&= \frac{[\sigma^2_X + \sigma^2_u][\beta\sigma^2_S + \gamma\sigma_{XS}] - \sigma_{XS}[\beta\sigma_{XS} + \gamma\sigma^2_X]}{[\sigma^2_X + \sigma^2_u]\sigma^2_S - (\sigma_{XS})^2} \\
&= \beta + \gamma\frac{\sigma^2_u\sigma_{SX}}{[\sigma^2_X + \sigma^2_u]\sigma^2_S - (\sigma_{XS})^2} \\
&= \beta + \gamma\rho\frac{\sigma^2_u\sigma^2_S}{[\sigma^2_X + \sigma^2_u]\sigma^2_S - (\sigma_{XS})^2}
\end{aligned}
$$

Measurement error in the control variable

$$
\begin{aligned}
\hat{\gamma}^m &= \frac{\sigma^2_{\tilde{S}}\sigma_{Y\tilde{X}} - \sigma_{\tilde{X}S}\sigma_{YS}}{\sigma^2_{\tilde{X}}\sigma^2_S - (\sigma_{\tilde{X},S})^2} \\
&= \frac{\sigma^2_S[\beta\sigma_{XS} + \gamma\sigma^2_X] - \sigma_{XS}[\beta\sigma^2_S + \gamma\sigma_{XS}]}{[\sigma^2_X + \sigma^2_u]\sigma^2_S - (\sigma_{XS})^2} \\
&= \frac{\sigma^2_S\sigma^2_X - (\sigma_{SX})^2}{[\sigma^2_X + \sigma^2_u]\sigma^2_S - (\sigma_{XS})^2}\gamma \\
&= \Lambda\gamma
\end{aligned}
$$

Measurement error in the control variable

$$
\begin{aligned}
1-\Lambda &= \frac{[\sigma_X^2+\sigma_u^2]\sigma_S^2-(\sigma_{XS})^2-\sigma_x^2\sigma_S^2+(\sigma_{XS})^2}{[\sigma_X^2+\sigma_u^2]\sigma_S^2-(\sigma_{XS})^2} \\
&= \frac{\sigma_u^2\sigma_S^2}{[\sigma_X^2+\sigma_u^2]\sigma_S^2-(\sigma_{XS})^2}
\end{aligned}
$$

- Hence, we have that:

$$\hat{\beta}^m = \beta + \gamma\rho(1-\Lambda)$$

OLS with two independent variables

- The regression:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$$

- Choosing the estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ to minimize the sum of squared residuals yields the OLS estimators:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z} \\
\hat{\beta}_1 &= \frac{Cov(x,y)Var(z) - Cov(z,y)Cov(x,z)}{Var(x)Var(z) - Cov(x,z)^2} \\
\hat{\beta}_2 &= \frac{Cov(z,y)Var(x) - Cov(x,y)Cov(x,z)}{Var(x)Var(z) - Cov(x,z)^2}
\end{aligned}$$