Microeconometrics Module

Lecture 4: Selection Bias

Swapnil Singh

Lietuvos Bankas and KTU

Course Link

Health Insurance and Health

- **Question:** Does provision of health insurance improves health outcomes?
- "Other things equal" question: contrasting the health of someone with insurance coverage to the same person
- We arrive at the fundamental problem of causal inference
  - either people are insured or they are not
  - you cannot observe both simultaneously

National Health Interview Survey (NHIS)

- Conducted by National Center for Health Statistics (NCHS)
- Primary source of information on individual's health living in US
- Questions are asked related to medical conditions, health insurance, doctor's office visits, physical activity etc.

National Health Interview Survey (NHIS)

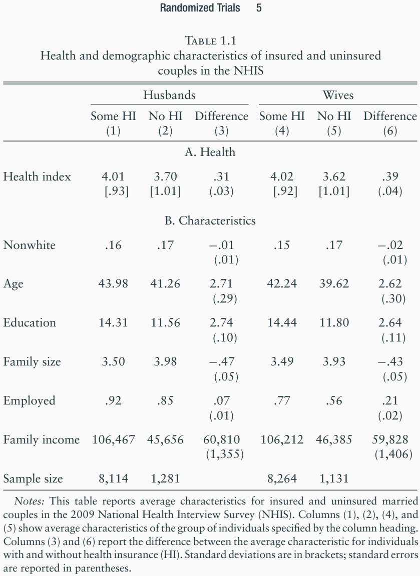

Table 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

From *Mastering 'Metrics: The Path from Cause to Effect*. © 2015 Princeton University Press. Used by permission.

Simpler Example

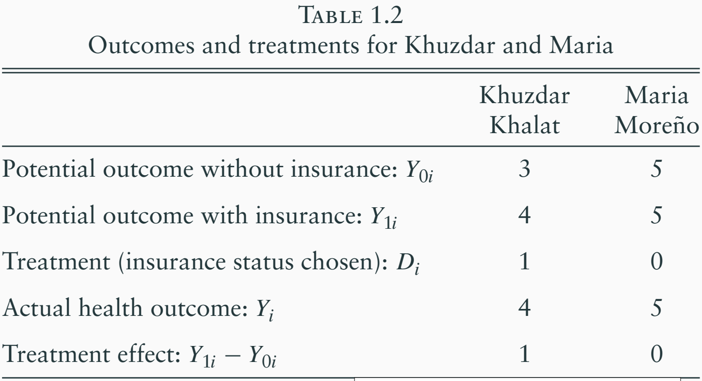

Table 1.2
Outcomes and treatments for Khuzdar and Maria

	Khuzdar Khalat	Maria Moreño
Potential outcome without insurance: Y_{0i}	3	5
Potential outcome with insurance: Y_{1i}	4	5
Treatment (insurance status chosen): D_i	1	0
Actual health outcome: Y_i	4	5
Treatment effect: $Y_{1i} - Y_{0i}$	1	0

From *Mastering 'Metrics: The Path from Cause to Effect*. © 2015 Princeton University Press. Used by permission. All rights reserved.

Simpler Example

- Khuzdar takes the insurance
- Maria does not take the insurance
- Hence:

$$Y_{Khuzdar} = Y_{Khuzdar,1} = 4$$
$$Y_{Maria} = Y_{Maria,0} = 5$$

- $Y_{Khuzdar} - Y_{Maria} = 4 - 5 = -1$
- Health insurance negatively affects health standard
- Wait a minute! What?

Simpler Example

- Comparison between Khuzdar and Maria's health is not a good comparison. Why?

$$
\begin{aligned}
Y_{Khuzdar} - Y_{Maria} &= Y_{Khuzdar,1} - Y_{Maria,0} \\
&= \underbrace{Y_{Khuzdar,1} - Y_{Khuzdar,0}}_{\text{causal effect}} + \underbrace{Y_{Khuzdar,0} - Y_{Maria,0}}_{\text{selection bias}} \\
&= 4 - 3 + 3 - 5 \\
&= 1 - 2 \\
&= -1
\end{aligned}
$$

- So causal effect of health insurance is positive
- But selection bias, which captures the difference in health if both of them had decided to go without insurance, dominates
- In other words: the health of Khuzdar was worse than Maria to begin with. Hence comparing the (treated) health of Khuzdar with (untreated) health of Maria is not a fair comparison

Further Notations and Generalizations

- Suppose there are *n* individuals
- Some get health insurance and some don't
- Construct a dummy variable *D* such that

$$D_i = \begin{cases} 1 & \text{if } i \text{ is insured} \\ 0 & \text{otherwise} \end{cases}$$

- Further assume that access to health insurance affects health standard by κ

$$Y_{1i} = Y_{0i} + \kappa$$

- Then difference in health status of insured versus uninsured is

$$\Rightarrow Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0]$$
$$\Rightarrow Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]$$
$$\Rightarrow Avg_n[\kappa + Y_{0i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]$$
$$\Rightarrow \underbrace{\kappa}_{\text{causal effect}} + \underbrace{Avg_n[Y_{0i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]}_{\text{selection bias}}$$

Summary

- Differences in group means = average causal effect + selection bias
- How can we be sure that there is selection bias?
- Let's have a look at Table 1.1 again

National Health Interview Survey (NHIS)

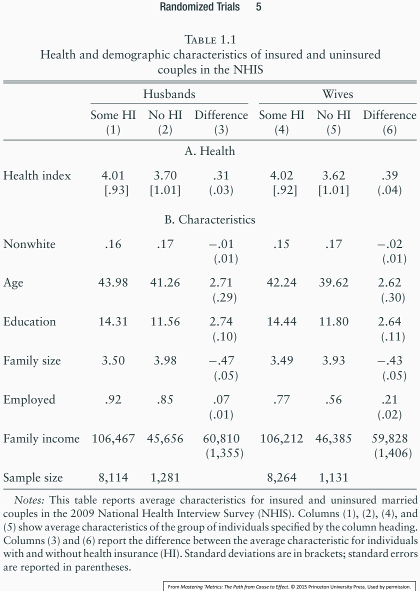

Table 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

From *Mastering 'Metrics: The Path from Cause to Effect*. © 2015 Princeton University Press. Used by permission.

Summary

- Differences in group means = average causal effect + selection bias
- How can we be sure that there is selection bias?
- Let's have a look at Table 1.1 again
  - insured people are also more educated
  - hence, maybe, education is also playing a role
- So, in short, selection bias does not allow us to compare apples to apples
- "Other things" are really not equal

Observed versus Unobserved Differences

- Table 1.1 shows that insured and uninsured people differ in observable ways
- But people can (and do) differ in unobservable ways
  - *A* and *B* are twins
  - Both have same education, age and occupation
  - However, *A* loves sweets but *B* doesn't
  - Sugar eating habits are not observed
- Point is that there can still be selection bias even if observables are not different