15.2 Instrumental Variables

Our basic linear regression model is

$$Y = Z\beta + \varepsilon \quad \varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

But what if ε is correlated with Z? This situaton is best illustrated using the following example: consider a study on the relationship between earnings and number of years spent in school.The basic model for this would be

$$\begin{aligned} Y &= \log(\text{earnings}) \\ S &= \text{number of years in school} \\ Y &= \beta_0 + \beta_1 S + \varepsilon^{(1)} \qquad \text{model (1)} \end{aligned}$$

However this model can lead to probelmatic conclusions. We can't just state that "β_1 is the extra log(money) that a person will earn per extra year in school" (since we are implicitely recommending that people stay in school forever). In real life there are other factors involved, e.g. ability, parental wealth, motivation etc. So the actual model looks more like

$$\begin{aligned} Y &= \log(\text{earnings}) \\ S &= \text{number of years in school} \\ A &= \text{ability} \\ M &= \text{parental wealth...} \\ Y &= \beta_0 + \beta_1 S + \beta_2 A + \beta_3 M + ... + \varepsilon^{(2)} \qquad \text{model (2)} \end{aligned}$$

And we want the β_1 for model (2), not model (1). The obvious problem is that many of these variables (e.g. ability, motiviation etc.) are hard to quantify, let alone find data for.

15.2.1 Model Using the Instrument

Suppose we have a variable W where $corr(W, S) \neq 0$ but $corr(W, \varepsilon^{(1)}) = 0$ or in other words $corr(W, A) = corr(W, M) = ..$(correlations with $\sim$ top 10 variables)$... = 0$. This variable W is called the "Instrumental variable". We can use this variable to find β_1 for model (2):

$$\begin{aligned} cov(Y, W) &= cov(\beta_0 + \beta_1 S + \beta_2 A + \beta_3 M + ... + \varepsilon^{(2)}, W) \\ &\quad [\text{or } cov(\beta_0 + \beta_1 S + \varepsilon^{(1)}, W)] \\ &= \beta_1 cov(S, W) \\ \beta_1 &= \frac{cov(Y, W)}{cov(S, W)} \\ \hat{\beta}_1 &= \frac{\hat{cov}(Y, W)}{\hat{cov}(S, W)} \\ &= \frac{S_{YW}}{S_{SW}} = \frac{\sum_{i=1} n(Y_i - \bar{Y})(W_i - \bar{W})}{\sum_{i=1} n(S_i - \bar{S})(W_i - \bar{W})} = \frac{S_{YW}/S_{WW}}{S_{SW}/S_{WW}} \end{aligned}$$

We can regress Y on W to obtain $\hat{\beta}_{Y\sim W}$ and regress S on W to obtain $\hat{\beta}_{S\sim W}$ which we can use to obtain

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{Y\sim W}}{\hat{\beta}_{S\sim W}}$$

where IV stands for "Instrumental variable". These estimates converge to their counterparts:

$$\begin{aligned} \text{numerator: } & \hat{\beta}_{Y\sim W} \to \beta_{Y\sim W} \\ \text{denominator: } & \hat{\beta}_{S\sim W} \to \beta_{S\sim W} \\ \text{ratio: } & \hat{\beta}_{IV} \to \beta_1 \text{ from model (2)} \end{aligned}$$

15.2.2 Finding Instrumental Variables

How can a variable be correlated to no nuisance variables but only to the variable of interest?

Income vs. Education Example

The following example is taken from Angrist and Kreuger http://www.jstor.org/stable/2696517. The data looked at the income of people in the 1990s who started kindergarten in the 1930s.

$$\begin{aligned} Y &= \text{income} \\ S &= \text{years of education} \\ W &= \text{(calendar) month of birth} \end{aligned}$$

Kindergarten cutoff date was December 31.

Birth month	Kindergarten entry age
OCT NOV DEC	5 ¾ - 6
JAN FEB MAR	6 - 6 ¼
APR MAY JUN	6 ¼ - 6 ½
JUL AUG SEP	6 ½ - 6 ¾

The main idea: people had to go to school until they were at least 16 so W (the month of birth) was correlated with S (the number of years in school), but not with ability, parental wealth etc. However there can be potential problems with this analysis:

- biggest sensitivity will be to drop out vs. not dropping out (as opposed to e.g. 1st year in college)
- data applied to people who grew up in the 1930s, so not necessarily relevant to today
- parents can play an influence on when to start the child in kindergarten

Other Examples

Response	Variable	Instrument
Earnings	Veteran status	Lottery draft number
Health	Surgery post heartache	Distance to cardiac center (this variable is common in medical studies)
Birth weight	Maternal smoking	State cigarette tax

Note that in a desgined experiment, the randomization itself is the instrument (i.e. the outcome of the cointoss for whether a person was selected to be in the control or test group).

15.2.3 Multiple Instrument Variables

More detailed notes are in MacFadden chapter 4 http://eml.berkeley.edu/~mcfadden/e240b_f01/ch4.pdf.

$$\begin{aligned} Y = Z\beta + \varepsilon, \quad & Z \in \mathbb{R}^{n\times k} \\ & W \in \mathbb{R}^{n\times j}\ j \geq k \end{aligned}$$

W is uncorrelated with ε (i.e. it is a "clean instrument") and $cov(W, Z)$ is a rank k matrix, i.e. W and Z are fully correlated. Note that W's can be correlated to each other as well.
We choose $R \in \mathbb{R}^{j\times k}$ which has rank k and we want to reduce the number of instruments to k. We can find linear combinations of our current instruments that still give instruments (only in a few cases they end up being correlated to other nuisance variables).

$$\begin{aligned} R^T W^T Y &= R^T W^T Z\beta + R^T W^T Z \\ &\doteq R^T W^T Z\beta \text{ (since } W^T \text{ and } Z \text{ are uncorrelated)} \\ \hat{\beta}_W &= (R^T W^T Z)^{-1} R^T W^T Y \end{aligned}$$

If $j = k$ (i.e. R is squared) then

$$\begin{aligned} \hat{\beta}_W &= (R^T W^T Z)^{-1} R^T W^T Y \\ &= (W^T Z)^{-1} W^T Y \end{aligned}$$

15.2.4 Extra notes

- Using instrument variables is not quite a causal relationship, but it does give you the correct β_1
- Watch out for "weak instruments" i.e. when $corr(S, W) \to 0$. This can lead to false results!
- The math is robust, but the most difficult (and interesting) part is finding a true instrument (i.e. a variable that is in fact uncorrelated to all the nuisance variables)