Session 10: Instrumental variables

MGT 581 | Introduction to econometrics

Michaël Aklin

PASU Lab | EPFL

Last time...

- Randomized controlled trial
- Design
- Statistical power

Today:

- Endogeneity
- Instrumental variables

Readings:

- Stock and Watson (2011), ch 12
- Verbeek (2018), ch 5.3, 5.5
- Angrist and Pischke (2008) ch 4
- Morgan and Winship (2014) ch 9

Endogeneity

Source of endogeneity

- Recall that a critical assumption is the conditional mean independence of u: $E[\varepsilon|\mathbf{X}] = 0$ (**exogeneity**)
- Necessary for consistency (and unbiasness) of estimates of β
- What could introduce endogeneity? Three sources...

1. Omitted variable bias (open backdoor) (see previous slides)

2. Simultaneous causality: D causes Y, and Y causes D

- Example: supply and demand...

$$P = a + b * Q$$
$$Q = c + d * P$$

3. Errors-in-variables: treatment X is measured with errors

- Suppose you're interested in X but only measure $\tilde{X} = X + \mu$.
- Then:

$$\begin{aligned} Y &= \alpha + \beta\tilde{X} + \tilde{\varepsilon} \\ &= \alpha + \beta\tilde{X} + [\beta(X - \tilde{X}) + \varepsilon] \end{aligned}$$

- Recall that $E[\varepsilon|X] = 0$ implies $Cov(u, X) = 0$. Yet:

$$\begin{aligned} Cov(\tilde{X}, \tilde{u}) &= Cov(\tilde{X}, \beta(X - \tilde{X}) + \varepsilon) \\ &= \beta Cov(\tilde{X}, X - \tilde{X}) + Cov(\tilde{X}, \varepsilon) \end{aligned}$$

- **Classical measurement error**: $Cov(\tilde{X}, X - \tilde{X}) \neq 0$
- **Attenuation bias**: $\widehat{\beta} = \beta\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\mu^2}$

- Regardless of source: all lead to $E[\varepsilon|\mathbf{X}] \neq 0$ and thus inconsistency
- How can we solve this?
- Ideally: introduce exogenous and random source of variation via RCT
- But as we saw: not always feasible/desirable
- Alternative idea: can we identify variation in X that is plausibly exogenous?
- Quasi-experiments: attempt to find variation that is *as if* generated in an experiment. Many quasi-experimental methods, including instrumental variables.
- Idea behind **instrumental variable** approach: breaks down X in two sets: one **endogenous** (correlated with error term, X) and one **exogenous** (uncorrelated, W)

Instrumental variables

- Starting point:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

- Consider a variable Z such that...
  - $Cor(Z_i, D_i) \neq 0$: it is **relevant**
  - $Cor(Z_i, \varepsilon_i) = 0$: it is **exogenous** (also referred to as **exclusion restriction**) (and thus in the set W)

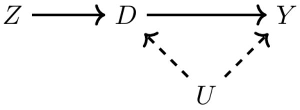

Figure 1: DAG representation of a good instrument. Source: mixtape.

Z causes D (relevant) and Z and ε are independent conditional on D (collider)

- Why do we need these assumptions? How can we recover $\hat{\beta}$?
- To see this, we need to relate β to Z:

$$Cov(Y, Z) = Cov(\alpha + \beta D + \varepsilon, Z)$$

- Then solving for β:

$$\beta = \frac{Cov(Y, Z)}{Cov(D, Z)}$$

- **Relevance** is needed for estimation ($Cov(X, Z) \neq 0$)
- **Exclusion restriction** is needed for consistency ($Cov(\varepsilon, Z) = 0$)

- Why do we need these assumptions? How can we recover $\hat{\beta}$?
- To see this, we need to relate β to Z:

$$Cov(Y, Z) = Cov(\alpha + \beta D + \varepsilon, Z)$$

- Then solving for β:

$$\beta = \frac{Cov(Y, Z)}{Cov(D, Z)}$$

- **Relevance** is needed for estimation ($Cov(X, Z) \neq 0$)
- **Exclusion restriction** is needed for consistency ($Cov(\varepsilon, Z) = 0$)

- Why do we need these assumptions? How can we recover $\hat{\beta}$?
- To see this, we need to relate β to Z:

$$\begin{aligned} Cov(Y, Z) &= Cov(\alpha + \beta D + \varepsilon, Z) \\ &= Cov(\alpha, Z) + Cov(\beta D, Z) + Cov(\varepsilon, Z) \end{aligned}$$

- Then solving for β:

$$\beta = \frac{Cov(Y, Z)}{Cov(D, Z)}$$

- **Relevance** is needed for estimation $(Cov(X, Z) \neq 0)$
- **Exclusion restriction** is needed for consistency $(Cov(\varepsilon, Z) = 0)$

- Why do we need these assumptions? How can we recover $\hat{\beta}$?
- To see this, we need to relate β to Z:

$$\begin{aligned} Cov(Y,Z) &= Cov(\alpha + \beta D + \varepsilon, Z) \\ &= Cov(\alpha, Z) + Cov(\beta D, Z) + Cov(\varepsilon, Z) \\ &= \beta Cov(D, Z) \end{aligned}$$

- Then solving for β:

$$\beta = \frac{Cov(Y,Z)}{Cov(D,Z)}$$

- **Relevance** is needed for estimation ($Cov(X,Z) \neq 0$)
- **Exclusion restriction** is needed for consistency ($Cov(\varepsilon, Z) = 0$)

- In practice: estimation is often done via **two-stage least squares** (TSLS or 2SLS)
- Stage 1: use OLS to estimate...

$$D_i = \pi_0 + \pi_1 Z_i + \mu_i$$

- Compute $\hat{D}$ using $\hat{\pi_0}$, $\hat{\pi}_1$, ...
- Stage 2: use OLS to estimate...

$$Y_i = \alpha + \beta_{tsls} \hat{D}_i + \varepsilon_i$$

- $\widehat{\beta_{tsls}}$ is a consistent estimator of β (assuming no compliance issues)
- Advantage: stats software get standard errors correctly

Example

- Examine the research question: does air pollution reduce GDP per capita?
- Potentially endogenous. Need an instrument.
- (Not so great) candidate: urbanization rate
- D: PM2.5, Z: urbanization, and Y: GDP per capita

```
Call:
ivreg(formula = GDP_Per_Capita ~ log(Population) | pm25 | Urbanization_Rate,
    data = data_combined)

Residuals:
   Min     1Q Median     3Q    Max 
-55782 -26307  -8197  12709 172627 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       7074.1    14732.3   0.480  0.63154    
pm25             -2256.1      417.2  -5.407 1.54e-07 ***
log(Population)   4373.5     1369.6   3.193  0.00159 ** 

Diagnostic tests:
                 df1 df2 statistic p-value    
Weak instruments   1 241     31.05 6.7e-08 ***
Wu-Hausman         1 240     79.50 < 2e-16 ***
Sargan             0  NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35980 on 241 degrees of freedom
Multiple R-Squared: -1.447,	Adjusted R-squared: -1.468 
Wald test: 16.96 on 2 and 241 DF,  p-value: 1.29e-07 
```

Figure 2: Example of IV in R

- Issue: I said that valid IVs yield **consistent** estimates...
- ... not that they yield **unbiased** estimates
- As Stock and Watson (2011) (Appendix 12.5), Wooldridge (2012), and others show:

$$plim\hat{\beta} = \beta + \frac{Cor(Z,\varepsilon)}{Cor(Z,D)}\frac{sd(\varepsilon)}{sd(D)}$$

- 2nd term only disappears if $Cor(Z,D)$ is very large, aka instrument is **strong**
- Typically captured by F statistic on excluded instrument(s). Rule of thumb: $F > 10$ for one instrument.
- Otherwise: problem of **weak instrument** and large bias

```
Call:
ivreg(formula = GDP_Per_Capita ~ log(Population) | pm25 | Urbanization_Rate,
    data = data_combined)

Residuals:
   Min     1Q Median     3Q    Max
-55782 -26307  -8197  12709 172627

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       7074.1    14732.3   0.480  0.63154
pm25             -2256.1      417.2  -5.407 1.54e-07 ***
log(Population)   4373.5     1369.6   3.193  0.00159 **

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1 241     31.05 6.7e-08 ***
Wu-Hausman         1 240     79.50 < 2e-16 ***
Sargan             0  NA        NA      NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35980 on 241 degrees of freedom
Multiple R-Squared: -1.447,	Adjusted R-squared: -1.468
Wald test: 16.96 on 2 and 241 DF,  p-value: 1.29e-07
```

Figure 3: Example of IV in R. Strong instruments!

Inference

- Asymptotically (large sample): distribution of $\hat{\beta}$ from TSLS is normal
- Thus: all we learned (hypothesis testing, confidence intervals) also applies
- However: note that $se(\hat{\beta})$ from the 2nd stage alone are wrong. Need to account for uncertainty from 1st stage.
- Note also that we still (generally) want to use heteroskedastic-robust standard errors (and possibly clustered se)

Generalizing IV

- You can expand this to $k > 0$ endogenous treatments X_1, X_2, ..., X_k
- ... but then you will need more instruments m (Z_1, Z_2, ..., Z_m)
- If $m < k$: **underidentified**. You will need more instruments
- If $m = k$: **exactly identified**
- If $m > k$: **overidentified**. Benefit: you can test whether instruments are valid!

- Advantage of overidentification: with $m > k$ (more instruments than endogenous vars), you can test your instrument's exogeneity
- **J-test** (Anderson-Rubin). Intuition: with multiple instruments for an endogenous variable, you could run separate TSLS. If estimates diverge "a lot," something is wrong
- Step 1: regress Y on W, Z and compute residual u_i
- Step 2: compute the F statistic that all parameters for Z are equal to zero (they should be)
- Step 3: compute the J statistic: $J = mF$.
- Step 4: J is distributed χ^2_{m-k} under H_0 that the instruments are exogenous. If J is large: reject the null and conclude that at least one instrument is in fact endogenous.

Example

- Consider the effect of (log) *cigarette price* (D) on (log) *cigarette consumed* (Y), adjusting for (log) income (X)
- Instrument $\#1$ (Z_1): sales tax as exogenous source of variation of *cigarette price*
- Instrument $\#2$ (Z_2): cigarette tax

```
                                     Y            W             X            W
cig_ivreg_diff1 <- ivreg(packsdiff ~  incomediff + pricediff | incomediff +
      Z
salestaxdiff, data = cig)
coeftest(cig_ivreg_diff1, vcov = vcovHC, type = "HC1")
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.117962   0.068217 -1.7292   0.09062 .
## incomediff   0.525970   0.339494  1.5493   0.12832
## pricediff   -0.938014   0.207502 -4.5205 4.454e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: TSLS with one instrument

```
                                        Y               W                 X                W
cig_ivreg_diff2 <- ivreg(packsdiff ~  incomediff + pricediff | incomediff +
      Z1                 Z2
salestaxdiff + cigtaxdiff, data = cig)
coeftest(cig_ivreg_diff2, vcov = vcovHC, type = "HC1")

##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.052003   0.062488 -0.8322    0.4097
## incomediff   0.462030   0.309341  1.4936    0.1423
## pricediff   -1.202403   0.196943 -6.1053 2.178e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: TSLS with two instrument

- TSLS with $Z =$ sales tax $(m = 1)$

$$\widehat{ln(Q_i^{cigarettes})} = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14}\, ln(P_i^{cigarettes}) + \underset{(0.31)}{0.21}\, ln(Income_i)$$

- TSLS with $Z =$ sales tax & cigarette tax $(m = 2)$

$$\widehat{ln(Q_i^{cigarettes})} = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28}\, ln(P_i^{cigarettes}) + \underset{(0.25)}{0.28}\, ln(Income_i)$$

- Smaller $se(\hat{\beta})$: more (good) instruments means better ability to tease out causal effects

LATE

- Another useful way to apply IV: **local average treatment effects** (LATE)
- Recall earlier experiment: *Encouragement to buy* $\rightarrow$ *Solar panels* $\rightarrow$ *Income*
- *Encouragement* is exogenous and, presumably, meets exclusion restriction
- It's an instrument!
- LATE: effect of solar panels on outcome among compliers (who were encouraged to buy and did so)

Conclusion

- Instrumental variables are a powerful way to address causality issues
- Good instrument allows the estimation of consistent (if not unbiased) estimates of treatment effects
- Great! However...
- Very hard to find convincing instruments, especially that meet **exclusion restriction**.
- Thus: other quasi-experimental methods have been developed to complement IVs

Questions?

References

Angrist, Joshua, and Steffan J. Pischke. 2008. *Mostly Harmless Econometics*. Princeton: Princeton University Press.

Morgan, Steven L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd Edition. Cambridge: Cambridge University Press.

Stock, James H., and Mark W. Watson. 2011. *Introduction to Econometrics, 3rd Edition*. Pearson.

Verbeek, Marno. 2018. *A Guide to Modern Econometrics 5th Edition*. Wiley.

Wooldridge, Jeffrey M. 2012. *Introductory Econometrics*. 5th ed. South-Western Cengage Learning.