PLSC 503 – Spring 2022
Simultaneity and Endogeneity

February 20, 2023

Endogeneity

Consider:

$$Y_1 = \mathsf{X}_1\boldsymbol{\beta}_1 + \gamma_1 Y_2 + \mathsf{u}_1$$

$$Y_2 = \mathsf{X}_2\boldsymbol{\beta}_2 + \gamma_2 Y_1 + \mathsf{u}_2$$

Rewrite:

$$\begin{aligned}
Y_1 &= \mathsf{X}_1\boldsymbol{\beta}_1 + \gamma_1[\mathsf{X}_2\boldsymbol{\beta}_2 + \gamma_2 Y_1 + \mathsf{u}_2] + \mathsf{u}_1 \\
&= \mathsf{X}_1\boldsymbol{\beta}_1 + \gamma_1(\mathsf{X}_2\boldsymbol{\beta}_2) + \gamma_1\gamma_2 Y_1 + \gamma_1\mathsf{u}_2 + \mathsf{u}_1 \\
Y_1 - \gamma_1\gamma_2 Y_1 &= \mathsf{X}_1\boldsymbol{\beta}_1 + \gamma_1(\mathsf{X}_2\boldsymbol{\beta}_2) + \gamma_1\mathsf{u}_2 + \mathsf{u}_1 \\
(1 - \gamma_1\gamma_2) Y_1 &= \mathsf{X}_1\boldsymbol{\beta}_1 + \gamma_1(\mathsf{X}_2\boldsymbol{\beta}_2) + \gamma_1\mathsf{u}_2 + \mathsf{u}_1 \\
Y_1 &= \mathsf{X}_1\left(\frac{1}{1-\gamma_1\gamma_2}\boldsymbol{\beta}_1\right) + \mathsf{X}_2\left(\frac{\gamma_1}{1-\gamma_1\gamma_2}\boldsymbol{\beta}_2\right) + \left(\frac{\gamma_1\mathsf{u}_2 + \mathsf{u}_1}{1-\gamma_1\gamma_2}\right) \\
&= \Delta_1\mathsf{X}_1 + \Delta_2\mathsf{X}_2 + \mathsf{e}
\end{aligned}$$

"Reduced Form"

$$Y_1 = X_1 \left(\frac{1}{1-\gamma_1\gamma_2} \boldsymbol{\beta}_1 \right) + X_2 \left(\frac{\gamma_1}{1-\gamma_1\gamma_2} \boldsymbol{\beta}_2 \right) + \left(\frac{\gamma_1 u_2 + u_1}{1-\gamma_1\gamma_2} \right)$$

means

$$\frac{\partial Y_1}{\partial X_\ell} = \frac{\beta_\ell}{1-\gamma_1\gamma_2}.$$

But

$$\hat{\Delta}_1 \neq \hat{\boldsymbol{\beta}}_1.$$

For (e.g.)

$$Y_1 = \mathsf{X}_1\boldsymbol{\beta}_1 + \gamma_1 Y_2 + \mathsf{u}_1$$

we have:

$$\mathsf{E}(Y_2, \mathsf{u}_1) = \frac{\gamma_2}{1 - \gamma_1\gamma_2}\sigma_{\mathsf{u}}^2$$

Result:

- Bias (unless $\gamma_2 = 0$)
- Inconsistency

What To Do

- OLS
- Lagged Variables
- Instrumental Variable Design / Two-Stage Least Squares (2SLS)
- Systems of Equations / 3SLS / etc.

IV Design

Recall that a simple linear model:

$$\mathsf{Y} = \mathsf{X}\boldsymbol{\beta} + \mathsf{u}$$

gives us:

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'\mathsf{u}.$$

Suppose $\mathrm{Cov}(\mathsf{X}, \mathsf{u}) \neq 0$, but we have Z with

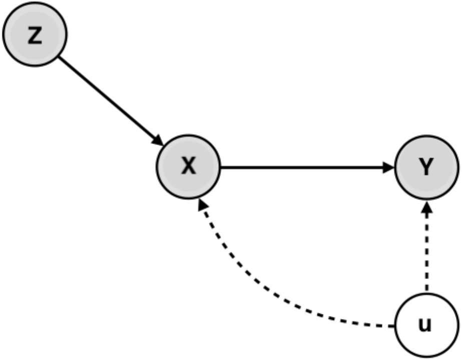
- $\mathrm{Cov}(\mathsf{Z}, \mathsf{X}) \neq 0$ and
- $\mathrm{Cov}(\mathsf{Z}, \mathsf{u}) = 0$.

Then:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{IV} &= (\mathsf{Z}'\mathsf{X})^{-1}\mathsf{Z}'\mathsf{Y} \\ &= (\mathsf{Z}'\mathsf{X})^{-1}\mathsf{Z}'(\mathsf{X}\boldsymbol{\beta} + \mathsf{u}) \\ &= \boldsymbol{\beta} + (\mathsf{Z}'\mathsf{X})^{-1}\mathsf{Z}'\mathsf{u} \end{aligned}$$

is consistent.

The Diagram



- Z is correlated with X (instrumental relevance)
- Z and u are conditionally independent
  - Z is quasi-randomly assigned
  - Z does not have a direct effect on Y beyond the channel through X (referred to as exclusion restriction)

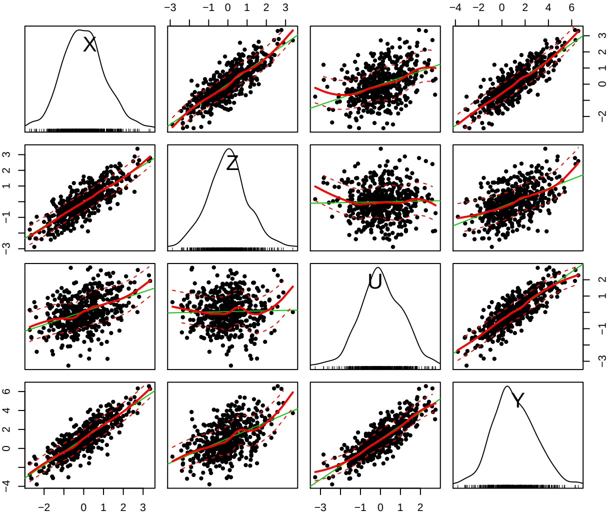
Some Examples of IV Designs in Political Science

- Miguel, Satyanah, Sergenti (2004) examine the effect of economic growth on civil conflict and use rainfall as an instrument for economic growth
- Zhu (2017) examines corruption and uses weighted geographic closeness to economic centers as an instrument for the activities of multinational corporations
- Vernby (2013) uses historical immigration levels to instrument the current number of noncitizen residents
- Overall, researchers use geography, climate, weather, and even historical data in their IV designs.
    - Weather seems to be a very popular instrument to examine a wide variety of outcomes which might imply possible exclusion restriction violations

IV Design: How-To

- Regress endogenous Xs variables on $\{Z, X\}$
- Generate $\hat{X}$s
- Regress Y on $\hat{X}$ to get β_{2SLS}.
- Adjust standard error estimates

```
library(MASS)
library(sem)
library(car)

seed<-1337
set.seed(seed)

mu<-c(0,0,0) # <== X, Z, U
Sigma<-matrix(c(1,0.8,0.4,0.8,1,0,0.4,0,1),
             nrow=3,byrow=TRUE)              # Cor(X,Y)=0.8, etc.
Vars<- mvrnorm(500,mu,Sigma)
colnames(Vars)<-c("X","Z","U")
Vars<-data.frame(Vars)

Vars$Y<- 1 + Vars$X + Vars$U
```

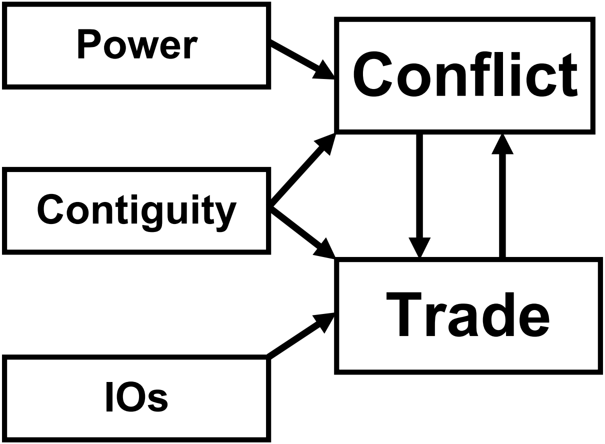

X
Z
U
Y
−3 −1 0 1 2 3
−4 −2 0 2 4 6
−2 0 1 2 3
−3 −1 0 1 2
−2 0 1 2 3
−3 −1 1 3
−3 −1 1 2
−4 0 2 4 6

Plain Old OLS...

```
> OLS<- lm(Y~X,data=Vars)
> summary(OLS)

Call:
lm(formula = Y ~ X, data = Vars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3809 -0.6058 -0.0102  0.6320  2.9470

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04770    0.04209   24.89   <2e-16 ***
X            1.40254    0.04005   35.02   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9413 on 498 degrees of freedom
Multiple R-squared:  0.7112,Adjusted R-squared:  0.7106
F-statistic:  1226 on 1 and 498 DF,  p-value: < 2.2e-16
```

Two-Stage Least Squares

```
> TSLS<-tsls(Y~I(X),data=Vars,instruments=~Z)
> summary(TSLS)

2SLS Estimates

Model Formula: Y ~ I(X)

Instruments: ~Z

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.29300 -0.68210 -0.06139  0.00000  0.76270  2.70300

             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 1.0491828  0.0456017 23.00754 < 2.22e-16 ***
I(X)        1.0302012  0.0536909 19.18763 < 2.22e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.0196738 on 498 degrees of freedom
```


Power
Conflict
Contiguity
Trade
IOs

```
> summary(IRData)
     dyadid        logdisputes          logtrade             IOs
 Min.   :  2020   Min.   :-0.6931   Min.   :-0.6931   Min.   : 4.579
 1st Qu.:135155   1st Qu.:-0.6931   1st Qu.: 2.4079   1st Qu.:19.500
 Median :220484   Median :-0.6931   Median : 5.5786   Median :27.704
 Mean   :275526   Mean   :-0.2627   Mean   : 4.6518   Mean   :30.891
 3rd Qu.:385710   3rd Qu.: 0.0000   3rd Qu.: 7.1248   3rd Qu.:39.289
 Max.   :900920   Max.   : 3.4965   Max.   :11.5037   Max.   :93.700
   contiguity          capratio           GDPgrowth
 Min.   :0.0000   Min.   :   1.081   Min.   :-9.0800
 1st Qu.:0.0000   1st Qu.:   4.849   1st Qu.:-0.2923
 Median :0.0000   Median :  26.577   Median : 0.8363
 Mean   :0.3207   Mean   : 196.310   Mean   : 0.5097
 3rd Qu.:1.0000   3rd Qu.: 144.035   3rd Qu.: 1.7106
 Max.   :1.0000   Max.   :7451.982   Max.   : 7.0460
```

Ordinary Regression

```
> OLSWar<-lm(logdisputes~logtrade+contiguity+capratio,data=IRData)
> summary(OLSWar)

Call:
lm(formula = logdisputes ~ logtrade + contiguity + capratio,
    data = IRData)

Residuals:
   Min     1Q Median     3Q    Max
-0.828 -0.326 -0.269 -0.090  3.455

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4253192  0.0602014   -7.06  3.5e-12 ***
logtrade     0.0085581  0.0105739    0.81    0.419
contiguity   0.4622674  0.0712406    6.49  1.5e-10 ***
capratio    -0.0001296  0.0000647   -2.00    0.045 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.853 on 813 degrees of freedom
Multiple R-squared:  0.083,Adjusted R-squared:  0.0796
F-statistic: 24.5 on 3 and 813 DF,  p-value: 3.35e-15
```

2SLS "By-Hand" (stage one)

```
> ITrade<-lm(logtrade~contiguity+IOs+capratio)
> summary(ITrade)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0385 -1.7666  0.4139  1.6154  7.6029

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7319793  0.1912570   3.827 0.000140 ***
contiguity   1.3386037  0.1816041   7.371 4.17e-13 ***
IOs          0.1218373  0.0055313  22.027  < 2e-16 ***
capratio    -0.0013913  0.0001626  -8.555  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.239 on 813 degrees of freedom
Multiple R-squared: 0.5535,Adjusted R-squared: 0.5519
F-statistic: 335.9 on 3 and 813 DF,  p-value: < 2.2e-16
```

2SLS "By-Hand" (stage two)

```
> IVWarByHand<-with(IRData, lm(logdisputes~capratio+contiguity+
+                        (ITrade$fitted.values)))
> summary(IVWarByHand)

Call:
lm(formula = logdisputes ~ capratio + contiguity + (ITrade$fitted.values))

Residuals:
   Min     1Q Median     3Q    Max
-1.006 -0.362 -0.278 -0.049  3.530

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.1515180  0.0832287   -1.82  0.06905 .
capratio            -0.0002664  0.0000705   -3.78  0.00017 ***
contiguity           0.6263774  0.0788444    7.94  6.5e-15 ***
ITrade$fitted.values -0.0558374  0.0171921   -3.25  0.00121 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.848 on 813 degrees of freedom
Multiple R-squared:  0.094,Adjusted R-squared:  0.0907
F-statistic: 28.1 on 3 and 813 DF,  p-value: <2e-16
```

2SLS, Automagically

```
> library(AER)
> TwoSLSWar<-ivreg(logdisputes~contiguity+capratio+I(logtrade),
   instruments=~contiguity+capratio+IOs)
> summary(TwoSLSWar)

Call:
ivreg(formula = logdisputes ~ contiguity + capratio + I(logtrade) |
    contiguity + capratio + IOs, data = IRData)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1515180  0.0856218   -1.77  0.07717 .
contiguity   0.6263774  0.0811114    7.72  3.4e-14 ***
capratio    -0.0002664  0.0000725   -3.67  0.00025 ***
I(logtrade) -0.0558374  0.0176864   -3.16  0.00165 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.872 on 813 degrees of freedom
Multiple R-Squared: 0.0412,Adjusted R-squared: 0.0376
Wald test: 26.6 on 3 and 813 DF,  p-value: <2e-16
```

Weak Instruments

```
> OLSTrade<-lm(logtrade~logdisputes+contiguity+IOs)
> summary(OLSTrade)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2467 -2.2067  0.4275  1.6659  6.1264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.191111   0.182875   1.045    0.296
logdisputes 0.408116   0.095067   4.293 1.98e-05 ***
contiguity  1.357557   0.193109   7.030 4.38e-12 ***
IOs         0.133778   0.005614  23.831  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.312 on 813 degrees of freedom
Multiple R-squared: 0.5241,Adjusted R-squared: 0.5223
F-statistic: 298.4 on 3 and 813 DF,  p-value: < 2.2e-16
```

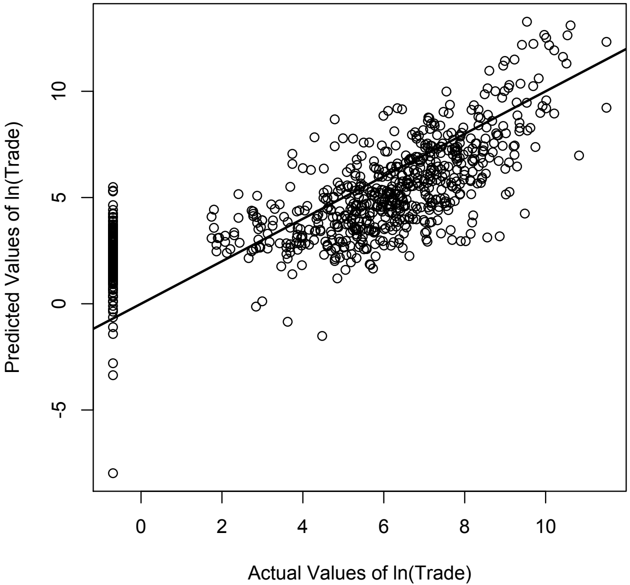
Weak Instruments (continued)

```
> TwoSLSTrade<-ivreg(logtrade~contiguity+IOs+I(logdisputes),
    instruments=~contiguity+capratio+IOs)
> summary(TwoSLSTrade)

Call:
ivreg(formula = logtrade ~ contiguity + IOs + I(logdisputes) |
    contiguity + capratio + IOs, data = IRData)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.1501     0.8512    2.53   0.0117 *
contiguity      -2.7276     1.5262   -1.79   0.0743 .
IOs              0.1720     0.0205    8.41   <2e-16 ***
I(logdisputes)   7.3712     2.4520    3.01   0.0027 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.37 on 813 degrees of freedom
Multiple R-Squared: -2.62,Adjusted R-squared: -2.63
Wald test: 41.5 on 3 and 813 DF,  p-value: <2e-16
```

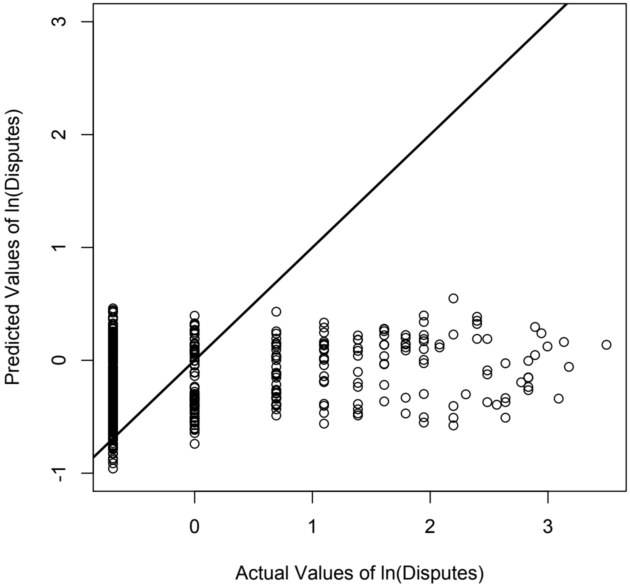
Side-By-Side...

	Dependent variable:			
	logdisputes		logtrade	
	OLS	*IV*	*OLS*	*IV*
logtrade	0.009			
	(0.011)			
logdisputes			0.408^{***}	
			(0.095)	
contiguity	0.462^{***}	0.626^{***}	1.358^{***}	-2.728^{*}
	(0.071)	(0.081)	(0.193)	(1.526)
capratio	-0.0001^{**}	-0.0003^{***}		
	(0.0001)	(0.0001)		
I(logtrade)		-0.056^{***}		
		(0.018)		
IOs			0.134^{***}	0.172^{***}
			(0.006)	(0.020)
I(logdisputes)				7.371^{***}
				(2.452)
Constant	-0.425^{***}	-0.152^{*}	0.191	2.150^{**}
	(0.060)	(0.086)	(0.183)	(0.851)
Observations	817	817	817	817
R^2	0.083	0.041	0.524	-2.616
Adjusted R^2	0.080	0.038	0.522	-2.630
Residual Std. Error (df = 813)	0.853	0.872	2.312	6.372
F Statistic (df = 3; 813)	24.530^{***}		298.400^{***}	
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Pretty Good Instrument (Trade)



Crappy Instrument (War)



Wrapping Up

- Standard linear regression performs poorly when there is endogeneity bias, omitted variable bias, or measurement bias
- IV estimation / 2SLS used as a tool to address these issues and uncover the causal effect of a variable on the outcome
- Finding a strong and valid instrument, that affects the treatment variable but do not have a direct effect on the outcome variable, is difficult.