Instrumental Variables and Systems of Equations

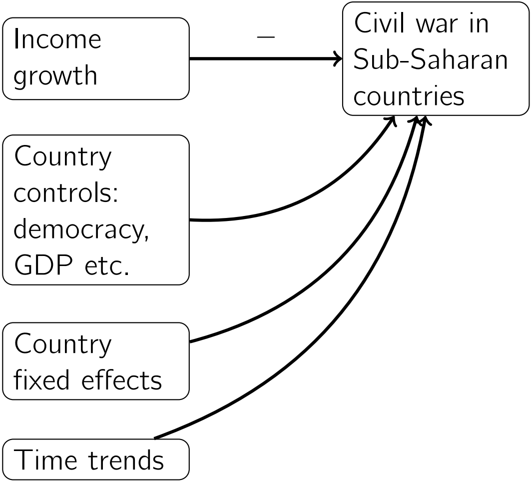
Philip Leifeld

GV903: Advanced Research Methods, Week 10

University of Essex

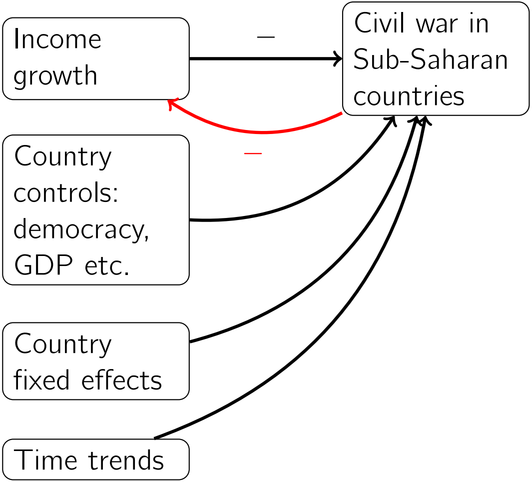
1. Conceptual Introduction

Miguel et al (2004): Economic Shocks and Civil Conflict



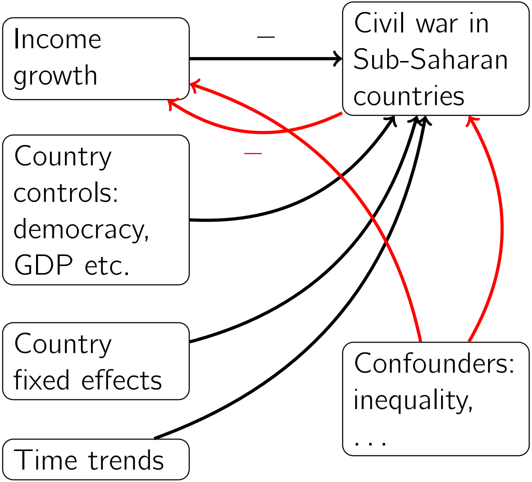
$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

Miguel et al (2004): Economic Shocks and Civil Conflict



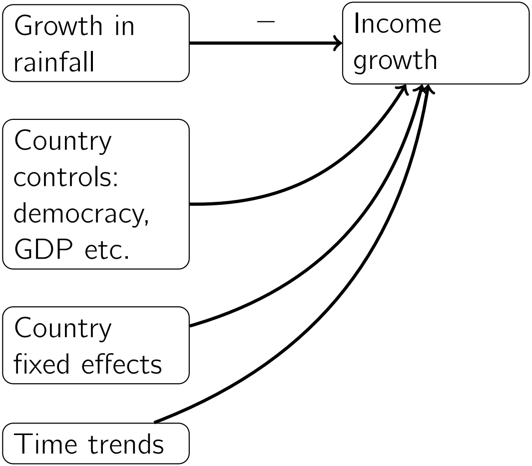
$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\beta + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

Miguel et al (2004): Economic Shocks and Civil Conflict



$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\beta + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

Instrumental Variable: Growth in Rainfall



$$\text{growth}_{it} = a_i + \mathbf{X}_{it}^{\top}\mathbf{b} + c_1 \Delta R_{it} + c_2 \Delta R_{i,t-1} + d_i \text{year}_t + e_{it}$$

Sources of Bias for Independent Variables

More generally:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + u$$

β_1 is biased because...

Sources of Bias for Independent Variables

More generally:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + u$$

β_1 is biased because...

- y_2 is partially caused by y_1 ("reverse causality").

More generally:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + u$$

β_1 is biased because...

- y_2 is partially caused by y_1 ("reverse causality").
- y_2 is affected by systematic measurement error.

Sources of Bias for Independent Variables

More generally:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + u$$

β_1 is biased because...

- y_2 is partially caused by y_1 ("reverse causality").
- y_2 is affected by systematic measurement error.
- y_1 and y_2 are jointly caused by a confounding variable.

Sources of Bias for Independent Variables

More generally:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + u$$

β_1 is biased because...

- y_2 is partially caused by y_1 ("reverse causality").
- y_2 is affected by systematic measurement error.
- y_1 and y_2 are jointly caused by a confounding variable.

As a consequence, y_2 is correlated with the error term.

Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.

Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.
   z is correlated with the endogenous variable.

Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.
   z is correlated with the endogenous variable.
2. Instrument exogeneity: $\text{Cov}(z, u) = 0$.

Two Conditions for Instrumental Variables

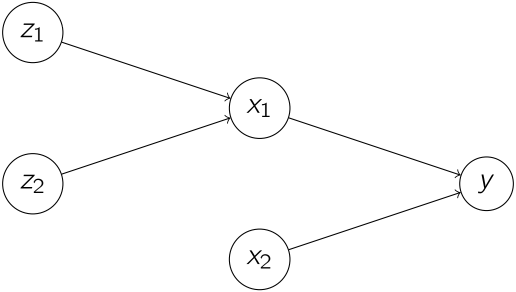
Instrumental variables z can help. They require two assumptions:

1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.
   z is correlated with the endogenous variable.
2. Instrument exogeneity: $\text{Cov}(z, u) = 0$.
   z *only* affects y_1 through y_2 and no other causal route.

Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

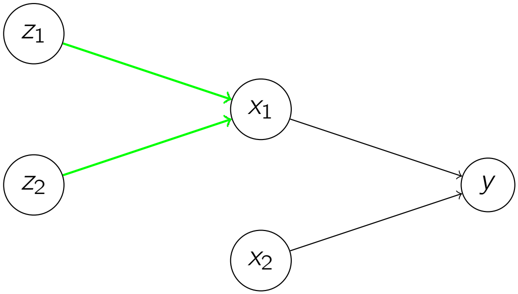
1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.
   z is correlated with the endogenous variable.
2. Instrument exogeneity: $\text{Cov}(z, u) = 0$.
   z *only* affects y_1 through y_2 and no other causal route.



Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

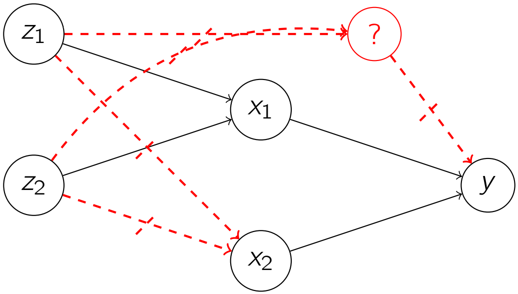
1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.
   z is correlated with the endogenous variable.
2. Instrument exogeneity: $\text{Cov}(z, u) = 0$.
   z *only* affects y_1 through y_2 and no other causal route.



Two Conditions for Instrumental Variables

Instrumental variables z can help. They require two assumptions:

1. Instrument relevance: $\text{Cov}(z, y_2) \neq 0$.
   z is correlated with the endogenous variable.
2. Instrument exogeneity: $\text{Cov}(z, u) = 0$.
   z *only* affects y_1 through y_2 and no other causal route.



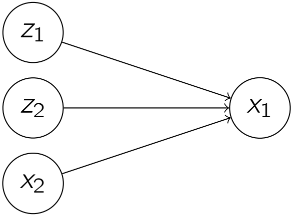
Two-Stage Least Squares IV Estimation (2SLS)

Two-Stage Least Squares IV Estimation (2SLS)

Stage 1: Regress the endogenous variable on the instruments and the other independent variables.

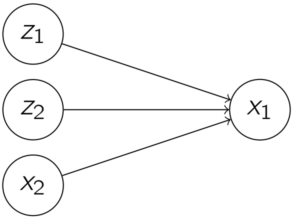
Two-Stage Least Squares IV Estimation (2SLS)

Stage 1: Regress the endogenous variable on the instruments and the other independent variables.



Two-Stage Least Squares IV Estimation (2SLS)

Stage 1: Regress the endogenous variable on the instruments and the other independent variables.

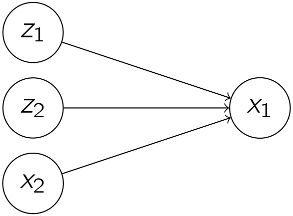


We can express this as:

$$x_1^* = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 x_2$$

Two-Stage Least Squares IV Estimation (2SLS)

Stage 1: Regress the endogenous variable on the instruments and the other independent variables.



We can express this as:

$$x_1^* = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 x_2$$

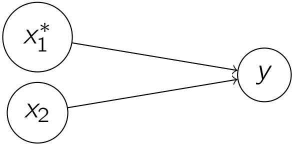
We do this to produce a predicted version of x_1 that is "purged" of any confounding (i. e., correlations with the error term u).

Two-Stage Least Squares IV Estimation (2SLS)

Stage 2: Generate predicted values from Stage 1 and regress DV on predicted values along with the independent variables.

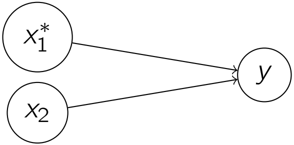
Two-Stage Least Squares IV Estimation (2SLS)

Stage 2: Generate predicted values from Stage 1 and regress DV on predicted values along with the independent variables.



Two-Stage Least Squares IV Estimation (2SLS)

Stage 2: Generate predicted values from Stage 1 and regress DV on predicted values along with the independent variables.

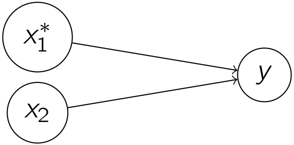


We can express this as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2$$

Two-Stage Least Squares IV Estimation (2SLS)

Stage 2: Generate predicted values from Stage 1 and regress DV on predicted values along with the independent variables.



We can express this as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2$$

This is like the original OLS estimation, but with the predicted version of x_1 that is not subject to endogeneity (i. e., "purged").

Let's Reconsider the Rainfall and Conflict Example...

The endogenous equation (endogenous term highlighted):

$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

Let's Reconsider the Rainfall and Conflict Example...

The endogenous equation (endogenous term highlighted):

$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

- $\text{growth}_{i,t-1}$ is exogenous because conflict at time t cannot influence growth at $t-1$ ("retrocausality"...).

Let's Reconsider the Rainfall and Conflict Example...

The endogenous equation (endogenous term highlighted):

$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^\top \boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

- $\text{growth}_{i,t-1}$ is exogenous because conflict at time t cannot influence growth at $t-1$ ("retrocausality"...).
- year_t is exogenous because conflict (or anything else) does not cause time.

Let's Reconsider the Rainfall and Conflict Example...

The endogenous equation (endogenous term highlighted):

$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^\top \boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

- $\text{growth}_{i,t-1}$ is exogenous because conflict at time t cannot influence growth at $t-1$ ("retrocausality"...).
- year_t is exogenous because conflict (or anything else) does not cause time.
- Variables in $\mathbf{X}$ are *assumed* to be exogenous. (Theoretical arguments needed...)

Let's Reconsider the Rainfall and Conflict Example...

The endogenous equation (endogenous term highlighted):

$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

- $\text{growth}_{i,t-1}$ is exogenous because conflict at time t cannot influence growth at $t-1$ ("retrocausality"...).
- year_t is exogenous because conflict (or anything else) does not cause time.
- Variables in $\mathbf{X}$ are *assumed* to be exogenous. (Theoretical arguments needed...)

Stage 1:

$$\text{growth}_{it}^{*} = \hat{a}_i + \mathbf{X}_{it}^{\top}\hat{\mathbf{b}} + \hat{c}_1 \Delta R_{it} + \hat{c}_2 \Delta R_{i,t-1} + \hat{d}_i \text{year}_t$$

Let's Reconsider the Rainfall and Conflict Example...

The endogenous equation (endogenous term highlighted):

$$\text{conflict}_{it} = \alpha_i + \mathbf{X}_{it}^\top \boldsymbol{\beta} + \gamma_1 \text{growth}_{it} + \gamma_2 \text{growth}_{i,t-1} + \delta_i \text{year}_t + u_{it}$$

- $\text{growth}_{i,t-1}$ is exogenous because conflict at time t cannot influence growth at $t-1$ ("retrocausality"...).
- year_t is exogenous because conflict (or anything else) does not cause time.
- Variables in $\mathbf{X}$ are *assumed* to be exogenous. (Theoretical arguments needed...)

Stage 1:

$$\text{growth}^*_{it} = \hat{a}_i + \mathbf{X}_{it}^\top \hat{\mathbf{b}} + \hat{c}_1 \Delta R_{it} + \hat{c}_2 \Delta R_{i,t-1} + \hat{d}_i \text{year}_t$$

Stage 2:

$$\widehat{\text{conflict}}_{it} = \hat{\alpha}_i + \mathbf{X}_{it}^\top \hat{\boldsymbol{\beta}} + \hat{\gamma}_1 \text{growth}^*_{it} + \hat{\gamma}_2 \text{growth}_{i,t-1} + \hat{\delta}_i \text{year}_t$$

Explanatory Variable	Dependent Variable: Civil Conflict ≥25 Deaths						Dependent Variable: Civil Conflict ≥1,000 Deaths
	Probit (1)	OLS (2)	OLS (3)	OLS (4)	IV-2SLS (5)	IV-2SLS (6)	IV-2SLS (7)
Economic growth rate, t	−.37	−.33	−.21	−.21	−.41	−1.13	−1.48*
	(.26)	(.26)	(.20)	(.16)	(1.48)	(1.40)	(.82)
Economic growth rate, $t-1$	−.14	−.08	.01	.07	−2.25**	−2.55**	−.77
	(.23)	(.24)	(.20)	(.16)	(1.07)	(1.10)	(.70)
Log(GDP per capita), 1979	−.067	−.041	.085		.053		
	(.061)	(.050)	(.084)		(.098)		
Democracy (Polity IV), $t-1$	.001	.001	.003		.004		
	(.005)	(.005)	(.006)		(.006)		
Ethnolinguistic fractionalization	.24	.23	.51		.51		
	(.26)	(.27)	(.40)		(.39)		
Religious fractionalization	−.29	−.24	.10		.22		
	(.26)	(.24)	(.42)		(.44)		
Oil-exporting country	.02	.05	−.16		−.10		
	(.21)	(.21)	(.20)		(.22)		
Log(mountainous)	.077**	.076*	.057		.060		
	(.041)	(.039)	(.060)		(.058)		
Log(national population), $t-1$	.080	.068	.182*		.159*		
	(.051)	(.051)	(.086)		(.093)		
Country fixed effects	no	no	no	yes	no	yes	yes
Country-specific time trends	no	no	yes	yes	yes	yes	yes
R^2	…	.13	.53	.71	…	…	…
Root mean square error	…	.42	.31	.25	.36	.32	.24
Observations	743	743	743	743	743	743	743

How Can We Find Good Instruments?

How Can We Find Good Instruments?

- Do not violate the relevance assumption.

How Can We Find Good Instruments?

- Do not violate the relevance assumption.
- Do not use weak instruments.

How Can We Find Good Instruments?

- Do not violate the relevance assumption.
- Do not use weak instruments.
- Test the relevance of your instruments in Stage 1, possibly with an F test if there are several instruments.

How Can We Find Good Instruments?

- Do not violate the relevance assumption.
- Do not use weak instruments.
- Test the relevance of your instruments in Stage 1, possibly with an F test if there are several instruments.
- In practice, the F test should have a p value of 0.001 or less (i. e., a t ratio of 3.2 or more; see Wooldridge).

How Can We Find Good Instruments?

- Do not violate the relevance assumption.
- Do not use weak instruments.
- Test the relevance of your instruments in Stage 1, possibly with an F test if there are several instruments.
- In practice, the F test should have a p value of 0.001 or less (i. e., a t ratio of 3.2 or more; see Wooldridge).
- Do not violate the exogeneity assumption.

How Can We Find Good Instruments?

- Do not violate the relevance assumption.
- Do not use weak instruments.
- Test the relevance of your instruments in Stage 1, possibly with an F test if there are several instruments.
- In practice, the F test should have a p value of 0.001 or less (i. e., a t ratio of 3.2 or more; see Wooldridge).
- Do not violate the exogeneity assumption.
- Only theoretical thinking can help here. Which variables z cause y *only* through x?

How Can We Find Good Instruments?

- Do not violate the relevance assumption.
- Do not use weak instruments.
- Test the relevance of your instruments in Stage 1, possibly with an F test if there are several instruments.
- In practice, the F test should have a p value of 0.001 or less (i. e., a t ratio of 3.2 or more; see Wooldridge).
- Do not violate the exogeneity assumption.
- Only theoretical thinking can help here. Which variables z cause y *only* through x?
- Make sure there are at least as many instruments as there are endogenous variables. Only then the model is *identified.*

Exercise

1. In the growth and conflict example, is the *relevance* condition met? Why? How can you test this empirically?
2. Is the *exogeneity* assumption met? Are there possible alternative causal pathways between rainfall and conflict? How can you assess whether they could be problematic?

Instrumental Variables and Outcome Distributions

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.
- Other outcome distributions (logit etc) yield biased results.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.
- Other outcome distributions (logit etc) yield biased results.
- However, the bias may be smaller than not using IVs at all.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.
- Other outcome distributions (logit etc) yield biased results.
- However, the bias may be smaller than not using IVs at all.
- There are other estimation strategies (MLE-based – more complicated) if you want to use binary or other variables at one or both stages.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.
- Other outcome distributions (logit etc) yield biased results.
- However, the bias may be smaller than not using IVs at all.
- There are other estimation strategies (MLE-based – more complicated) if you want to use binary or other variables at one or both stages.
- In the conflict example, the authors used 2SLS with OLS at both stages although conflict is a binary variable.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.
- Other outcome distributions (logit etc) yield biased results.
- However, the bias may be smaller than not using IVs at all.
- There are other estimation strategies (MLE-based – more complicated) if you want to use binary or other variables at one or both stages.
- In the conflict example, the authors used 2SLS with OLS at both stages although conflict is a binary variable.
- Application of OLS to binary data is called the linear probability model.

Instrumental Variables and Outcome Distributions

- 2SLS IV estimation works only with the linear model.
- Both stages must be linear.
- Other outcome distributions (logit etc) yield biased results.
- However, the bias may be smaller than not using IVs at all.
- There are other estimation strategies (MLE-based – more complicated) if you want to use binary or other variables at one or both stages.
- In the conflict example, the authors used 2SLS with OLS at both stages although conflict is a binary variable.
- Application of OLS to binary data is called the linear probability model.
- It is not ideal (predictions outside [0, 1] etc), but can be regarded as an acceptable fix because having biased estimates due to endogeneity would be worse.

Limitations of the Instrumental Variables Approach

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.
- No way to test the exogeneity assumption properly.

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.
- No way to test the exogeneity assumption properly.
- Weak instruments.

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.
- No way to test the exogeneity assumption properly.
- Weak instruments.
- Does not fix all kinds of endogeneity, such as complex non-independence of observations (think networks, multilevel data...).

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.
- No way to test the exogeneity assumption properly.
- Weak instruments.
- Does not fix all kinds of endogeneity, such as complex non-independence of observations (think networks, multilevel data...).
- A sample correction for standard errors is necessary. So we better rely on a ready-made implementation in `R`...

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.
- No way to test the exogeneity assumption properly.
- Weak instruments.
- Does not fix all kinds of endogeneity, such as complex non-independence of observations (think networks, multilevel data...).
- A sample correction for standard errors is necessary. So we better rely on a ready-made implementation in `R`...
- No natural interpretation of R^2, thus no F test.

Limitations of the Instrumental Variables Approach

- Works best in conjunction with OLS.
- No way to test the exogeneity assumption properly.
- Weak instruments.
- Does not fix all kinds of endogeneity, such as complex non-independence of observations (think networks, multilevel data. . .).
- A sample correction for standard errors is necessary. So we better rely on a ready-made implementation in `R`. . .
- No natural interpretation of R^2, thus no F test.
- In practice, it is *hard* to come up with relevant and exogenous instruments.

2. Estimation in `R`

Instrumental Variables Estimation in R

Let's load some panel data on cigarette consumption for the 48 continental US States from 1985 to 1995.

```r
library("AER")
data("CigarettesSW")
```

price	Average price during year, including tax.
cpi	Consumer price index.
income	State personal income.
population	State population.
tax	Average tax per year, federal, state, and local.
taxs	Average tax per year, state level.

Some Data Preparation. . .

We need to create some additional variables for the analysis. . .

```
# real prices
CigarettesSW$rprice <- with(CigarettesSW, price / cpi)

# real income per capita
CigarettesSW$rincome <- with(CigarettesSW, income / population / cpi)

# real state tax relative to everywhere
CigarettesSW$tdiff <- with(CigarettesSW, (taxs - tax) / cpi)
```

Instrumental Variables Estimation in R

Cigarette consumption is a function of price and income:

$$\log packs = \beta_0 + \beta_1 \log rprice + \log rincome + u,$$

Instrumental Variables Estimation in R

Cigarette consumption is a function of price and income:

$$\log \mathit{packs} = \beta_0 + \beta_1 \log \mathit{rprice} + \log \mathit{rincome} + u,$$

where $\log \mathit{rprice}$ is endogenous and instrumented as follows:

$$\log \mathit{rprice}^* = \hat{\pi}_0 + \hat{\pi}_1 \log \mathit{rincome} + \hat{\pi}_2 \mathit{tdiff} + \hat{\pi}_3 \frac{\mathit{tax}}{\mathit{cpi}} + e$$

Instrumental Variables Estimation in R

Cigarette consumption is a function of price and income:

$$\log packs = \beta_0 + \beta_1 \log rprice + \log rincome + u,$$

where $\log rprice$ is endogenous and instrumented as follows:

$$\log rprice^* = \hat{\pi}_0 + \hat{\pi}_1 \log rincome + \hat{\pi}_2 tdiff + \hat{\pi}_3 \frac{tax}{cpi} + e$$

$\log rincome$ appears in both stages because assumed exogenous.

Instrumental Variables Estimation in R

Cigarette consumption is a function of price and income:

$$\log packs = \beta_0 + \beta_1 \log rprice + \log rincome + u,$$

where $\log rprice$ is endogenous and instrumented as follows:

$$\log rprice^* = \hat{\pi}_0 + \hat{\pi}_1 \log rincome + \hat{\pi}_2 tdiff + \hat{\pi}_3 \frac{tax}{cpi} + e$$

$\log rincome$ appears in both stages because assumed exogenous.

The idea is that prices are correlated with high vs low excise taxes and that excise taxes have no other causal path towards cigarette consumption.

Instrumental Variables Estimation in R

Cigarette consumption is a function of price and income:

$$\log packs = \beta_0 + \beta_1 \log rprice + \log rincome + u,$$

where $\log rprice$ is endogenous and instrumented as follows:

$$\log rprice^* = \hat{\pi}_0 + \hat{\pi}_1 \log rincome + \hat{\pi}_2 tdiff + \hat{\pi}_3 \frac{tax}{cpi} + e$$

$\log rincome$ appears in both stages because assumed exogenous.

The idea is that prices are correlated with high vs low excise taxes and that excise taxes have no other causal path towards cigarette consumption.

In R with the `ivreg` function:

```
m <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
  tdiff + I(tax / cpi), data = CigarettesSW, subset = year == "1995")
```

Cigarette Demand Results

```
summary(m)
##
## Call:
## ivreg(formula = log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
##     tdiff + I(tax/cpi), data = CigarettesSW, subset = year ==
##     "1995")
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.6006931 -0.0862222 -0.0009999  0.1164699  0.3734227
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.8950     1.0586   9.348 4.12e-12 ***
## log(rprice)   -1.2774     0.2632  -4.853 1.50e-05 ***
## log(rincome)   0.2804     0.2386   1.175    0.246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1879 on 45 degrees of freedom
## Multiple R-Squared: 0.4294,Adjusted R-squared: 0.4041
## Wald test: 13.28 on 2 and 45 DF,  p-value: 2.931e-05
```

```r
# heteroskedasticity-consistent SEs + diagnostics; Inf: z- or chi^2 test
summary(m, vcov = sandwich, df = Inf, diagnostics = TRUE)
##
## Call:
## ivreg(formula = log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
##     tdiff + I(tax/cpi), data = CigarettesSW, subset = year ==
##     "1995")
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.6006931 -0.0862222 -0.0009999  0.1164699  0.3734227
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    9.8950     0.9288  10.654  < 2e-16 ***
## log(rprice)   -1.2774     0.2417  -5.286 1.25e-07 ***
## log(rincome)   0.2804     0.2458   1.141    0.254
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2  44   228.738  <2e-16 ***
## Wu-Hausman         1  44     3.823  0.0569 .
## Sargan             1  NA     0.333  0.5641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1879 on Inf degrees of freedom
```

F Test for Restricted IV Specification

Let's compute a restricted model with only one instrument:

```
m2 <- ivreg(log(packs) ~ log(rprice) | tdiff, data = CigarettesSW,
            subset = year == "1995")
```

F Test for Restricted IV Specification

Let's compute a restricted model with only one instrument:

```
m2 <- ivreg(log(packs) ~ log(rprice) | tdiff, data = CigarettesSW,
            subset = year == "1995")
```

We can now use an F test to check if the additional instrumental variable and income add anything. F tests for linear models (with or without IV) can generally be run using the `anova` function.

F Test for Restricted IV Specification

Let's compute a restricted model with only one instrument:

```r
m2 <- ivreg(log(packs) ~ log(rprice) | tdiff, data = CigarettesSW,
            subset = year == "1995")
```

We can now use an F test to check if the additional instrumental variable and income add anything. F tests for linear models (with or without IV) can generally be run using the `anova` function.

```r
anova(m, m2)
## Analysis of Variance Table
##
## Model 1: log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff +
##     I(tax/cpi)
## Model 2: log(packs) ~ log(rprice) | tdiff
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 1.5880
## 2     46 1.6668 -1 -0.078748 1.3815  0.246
```

F Test for Restricted IV Specification

Let's compute a restricted model with only one instrument:

```
m2 <- ivreg(log(packs) ~ log(rprice) | tdiff, data = CigarettesSW,
            subset = year == "1995")
```

We can now use an F test to check if the additional instrumental variable and income add anything. F tests for linear models (with or without IV) can generally be run using the `anova` function.

```
anova(m, m2)
## Analysis of Variance Table
##
## Model 1: log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff +
##     I(tax/cpi)
## Model 2: log(packs) ~ log(rprice) | tdiff
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 1.5880
## 2     46 1.6668 -1 -0.078748 1.3815  0.246
```

Here, we conclude that the simpler model is sufficient.

Manual Steps for 2SLS in R

We can replicate the `ivreg` results using `lm`:

```
cig <- subset(CigarettesSW, subset = year == "1995")
s1 <- lm(log(rprice) ~ log(rincome) + tdiff + I(tax / cpi), data = cig)
cig$pred1 <- predict(s1)
s2 <- lm(log(packs) ~ pred1 + log(rincome), data = cig)
```

Manual Steps for 2SLS in R

We can replicate the ivreg results using lm:

```
cig <- subset(CigarettesSW, subset = year == "1995")
s1 <- lm(log(rprice) ~ log(rincome) + tdiff + I(tax / cpi), data = cig)
cig$pred1 <- predict(s1)
s2 <- lm(log(packs) ~ pred1 + log(rincome), data = cig)
```

For multiple endogenous variables: repeat Stage 1 multiple times and replace each endogenous variable with its predicted values in Stage 2.

Manual Steps for 2SLS in `R`

We can replicate the `ivreg` results using `lm`:

```r
cig <- subset(CigarettesSW, subset = year == "1995")
s1 <- lm(log(rprice) ~ log(rincome) + tdiff + I(tax / cpi), data = cig)
cig$pred1 <- predict(s1)
s2 <- lm(log(packs) ~ pred1 + log(rincome), data = cig)
```

For multiple endogenous variables: repeat Stage 1 multiple times and replace each endogenous variable with its predicted values in Stage 2.

Here, $log(rincome)$ is assumed to be exogenous, so not necessary.

Let's Compare the Manual Results to `ivreg`...

```
library("texreg")
screenreg(list(m, s2), single.row = TRUE)
##
## =================================================
##                Model 1           Model 2
## -------------------------------------------------
## (Intercept)    9.89 (1.06) ***   9.89 (1.14) ***
## log(rprice)   -1.28 (0.26) ***
## log(rincome)   0.28 (0.24)       0.28 (0.26)
## pred1                           -1.28 (0.28) ***
## -------------------------------------------------
## R^2            0.43              0.34
## Adj. R^2       0.40              0.31
## Num. obs.     48                48
## =================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Same coefficients. SEs differ and would need adjustment. R^2 lower (can be negative in principle; no natural interpretation).

Instrumental Variables with Panel Data in R

In the `ivreg` example, we used only one time point. What if we want to use IVs with panel data?

Instrumental Variables with Panel Data in `R`

In the `ivreg` example, we used only one time point. What if we want to use IVs with panel data?

The `plm` function in the `plm` package (for panel data; fixed effects etc) can also deal with instrumental variables.

Instrumental Variables with Panel Data in R

In the `ivreg` example, we used only one time point. What if we want to use IVs with panel data?

The `plm` function in the `plm` package (for panel data; fixed effects etc) can also deal with instrumental variables.

The syntax is like in `ivreg`:

```
formula = y ~ x1 + x2 + x3 | x3 + z1 + z2
```

Instrumental Variables with Panel Data in R

In the `ivreg` example, we used only one time point. What if we want to use IVs with panel data?

The `plm` function in the `plm` package (for panel data; fixed effects etc) can also deal with instrumental variables.

The syntax is like in `ivreg`:

```
formula = y ~ x1 + x2 + x3 | x3 + z1 + z2
```

Several instrument variable transformation methods are available. See `?plm`.

Exercise

- Can you come up with a new research example in political science where instrumental variables would make sense?
- Write down the equations for both stages.
- What is the dependent variable?
- What is the endogenous variable that is problematic?
- What is the nature of the endogeneity? Why is the variable not exogenous?
- What control variables would you include?
- What are possible instruments? Discuss their relevance and exogeneity.

3. Manual Computation Using Matrix Algebra

Manual Computation of IV using Matrix Algebra

See also: https://stats.stackexchange.com/questions/265780/
See also slides on Heteroskedasticity earlier this semester.

First, we create a projection matrix for the space spanned by $\mathbf{Z}$, where $\mathbf{Z}$ is the instrument design matrix including exogenous variables and actual instruments):

$$\mathbf{P}_Z = \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}\mathbf{Z}^\top$$

Manual Computation of IV using Matrix Algebra

See also: `https://stats.stackexchange.com/questions/265780/`
See also slides on Heteroskedasticity earlier this semester.

First, we create a projection matrix for the space spanned by $\mathbf{Z}$, where $\mathbf{Z}$ is the instrument design matrix including exogenous variables and actual instruments):

$$\mathbf{P}_Z = \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}\mathbf{Z}^\top$$

This would now in principle permit us to generate predicted values if we wanted to (Stage 1):

$$\hat{\mathbf{X}}_i = \mathbf{P}_Z \mathbf{X}_i$$

Manual Computation of IV using Matrix Algebra

See also: `https://stats.stackexchange.com/questions/265780/`
See also slides on Heteroskedasticity earlier this semester.

First, we create a projection matrix for the space spanned by **Z**, where **Z** is the instrument design matrix including exogenous variables and actual instruments):

$$\mathbf{P}_Z = \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}\mathbf{Z}^\top$$

This would now in principle permit us to generate predicted values if we wanted to (Stage 1):

$$\hat{\mathbf{X}}_i = \mathbf{P}_Z \mathbf{X}_i$$

Second, we run 2SLS (both stages in one go), where **X** is the regressor design matrix including exogenous and endogenous variables):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{P}_Z \mathbf{y}$$

Manual Computation of IV using Matrix Algebra

See also: `https://stats.stackexchange.com/questions/265780/`
See also slides on Heteroskedasticity earlier this semester.

First, we create a projection matrix for the space spanned by **Z**, where **Z** is the instrument design matrix including exogenous variables and actual instruments):

$$\mathbf{P}_Z = \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top$$

This would now in principle permit us to generate predicted values if we wanted to (Stage 1):

$$\hat{\mathbf{X}}_i = \mathbf{P}_Z \mathbf{X}_i$$

Second, we run 2SLS (both stages in one go), where **X** is the regressor design matrix including exogenous and endogenous variables):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{P}_Z \mathbf{y}$$

You will recognise this as WLS where the weight matrix $\mathbf{W} := \mathbf{P}_Z$.

Calculation of Variances and SEs

See also: https://stats.stackexchange.com/questions/265780/
See also slides on Heteroskedasticity earlier this semester.

Calculate the VCOV matrix:

$$\mathrm{Cov}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{P}_Z \mathbf{\Omega} \mathbf{P}_Z \mathbf{X} \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1}$$

where $\mathbf{\Omega} = \mathrm{Cov}(\mathbf{y}) = \sigma_2 \mathbf{I}$ (or plug in residuals to do FGLS. . .).

Calculation of Variances and SEs

See also: `https://stats.stackexchange.com/questions/265780/`
See also slides on Heteroskedasticity earlier this semester.

Calculate the VCOV matrix:

$$\mathrm{Cov}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{P}_Z \boldsymbol{\Omega} \mathbf{P}_Z \mathbf{X} \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1}$$

where $\boldsymbol{\Omega} = \mathrm{Cov}(\mathbf{y}) = \sigma_2 \mathbf{I}$ (or plug in residuals to do FGLS. . .).

If $\boldsymbol{\Omega} = \sigma_2 \mathbf{I}$, we can more simply compute the variances as:

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{P}_Z \mathbf{X})^{-1}$$

where $\hat{\sigma}^2 = \mathrm{Var}(u)$ (i. e., the error variance in Stage 2).

Calculation of Variances and SEs

See also: `https://stats.stackexchange.com/questions/265780/`

See also slides on Heteroskedasticity earlier this semester.

Calculate the VCOV matrix:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{P}_Z \boldsymbol{\Omega} \mathbf{P}_Z \mathbf{X} \left(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}\right)^{-1}$$

where $\boldsymbol{\Omega} = \mathrm{Cov}(\mathbf{y}) = \sigma_2 \mathbf{I}$ (or plug in residuals to do FGLS. . .).

If $\boldsymbol{\Omega} = \sigma_2 \mathbf{I}$, we can more simply compute the variances as:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{P}_Z \mathbf{X})^{-1}$$

where $\hat{\sigma}^2 = \mathrm{Var}(u)$ (i. e., the error variance in Stage 2).

We can then extract the standard errors as the square root of $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ or of the diagonal elements of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$.

4. Simultaneous Equation Models

Simultaneous Equation Models (SEM)

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + u \tag{1}$$

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + u \tag{1}$$

$$x_1 = \gamma_0 + \gamma_1 y + \gamma_2 z_3 + v \tag{2}$$

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + u \tag{1}$$

$$x_1 = \gamma_0 + \gamma_1 y + \gamma_2 z_3 + v \tag{2}$$

- Each equation is identified if there is at least one exogenous variable in the respective other equation.

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + u \quad (1)$$

$$x_1 = \gamma_0 + \gamma_1 y + \gamma_2 z_3 + v \quad (2)$$

- Each equation is identified if there is at least one exogenous variable in the respective other equation.
- This is true because each equation is essentially used as a Stage 1 model in 2SLS for the respective other equation.

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + u \tag{1}$$

$$x_1 = \gamma_0 + \gamma_1 y + \gamma_2 z_3 + v \tag{2}$$

- Each equation is identified if there is at least one exogenous variable in the respective other equation.
- This is true because each equation is essentially used as a Stage 1 model in 2SLS for the respective other equation.
- Systems with more than two equations are possible.

Simultaneous Equation Models (SEM)

- If there is bidirectional causality (endogeneity) between the DV and an independent variable, we can also model this using a SEM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + u \tag{1}$$

$$x_1 = \gamma_0 + \gamma_1 y + \gamma_2 z_3 + v \tag{2}$$

- Each equation is identified if there is at least one exogenous variable in the respective other equation.
- This is true because each equation is essentially used as a Stage 1 model in 2SLS for the respective other equation.
- Systems with more than two equations are possible.
- Look at Reuveny and Li on the reading list for an interesting example. They have three equations to model joint democracy and conflicts between countries simultaneously.

Simultaneous Equation Models (SEM)

- In addition to 2SLS, there are more complicated estimation procedures for SEMs.
- SEMs have been implemented in several `R` packages:
  - `lavaan`
  - `systemfit`
  - `sem`
- If you do not carefully consider the issue of identification by exogenous variables, they will throw error messages.

Exercise

- Can you come up with a political science example of simultaneous equations that would be identified?
- What are the endogenous variables and exogenous instruments, respectively? Write down the equations.
- Are your exogenous variables really exogenous? Are they relevant? Discuss.
- The example can be based on your own research, the previous IV exercise, or something completely new.