Module 2: Demand for Cigarettes and Instrumental Variables

Part 2: Instrumental Variables

Ian McCarthy | Emory University
Econ 470 & HLTH 470

What is instrumental variables

Instrumental Variables (IV) is a way to identify causal effects using variation in treatment particpation that is due to an *exogenous* variable that is only related to the outcome through treatment.

Why bother with IV?

Two reasons to consider IV:

1. Selection on unobservables
2. Reverse causation

Either problem is sometimes loosely referred to as *endogeneity*

Simple example

- $y = \beta x + \varepsilon(x)$,
  where $\varepsilon(x)$ reflects the dependence between our observed variable and the error term.

- Simple OLS will yield
  $\frac{dy}{dx} = \beta + \frac{d\varepsilon}{dx} \neq \beta$

What does IV do?

- The regression we want to do:
  $y_i = \alpha + \delta D_i + \gamma A_i + \epsilon_i$,
  where D_i is treatment (think of schooling for now) and A_i is something like ability.

- A_i is unobserved, so instead we run:
  $y_i = \alpha + \beta D_i + \epsilon_i$

- From this "short" regression, we don't actually estimate δ. Instead, we get an estimate of
  $\beta = \delta + \lambda_{ds}\gamma \neq \delta$,
  where λ_{ds} is the coefficient of a regression of A_i on D_i.

Intuition

IV will recover the "long" regression without observing underlying ability

IF our IV satisfies all of the necessary assumptions.

More formally

- We want to estimate
  $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$

- With instrument Z_i that satisfies relevant assumptions, we can estimate this as
  $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$

- In words, this is effect of the instrument on the outcome ("reduced form") divided by the effect of the instrument on treatment ("first stage")

Derivation

Recall "long" regression: $Y = \alpha + \delta S + \gamma A + \epsilon$.

$$\begin{aligned}
COV(Y,Z) &= E[YZ] - E[Y]E[Z] \\
&= E[(\alpha + \delta S + \gamma A + \epsilon) \times Z] - E[\alpha + \delta S + \gamma A + \epsilon)]E[Z] \\
&= \alpha E[Z] + \delta E[SZ] + \gamma E[AZ] + E[\epsilon Z] \\
&\quad - \alpha E[Z] - \delta E[S]E[Z] - \gamma E[A]E[Z] - E[\epsilon]E[Z] \\
&= \delta(E[SZ] - E[S]E[Z]) + \gamma(E[AZ] - E[A]E[Z]) \\
&\quad + E[\epsilon Z] - E[\epsilon]E[Z] \\
&= \delta C(S,Z) + \gamma C(A,Z) + C(\epsilon, Z)
\end{aligned}$$

Derivation

Working from $COV(Y,Z) = \delta COV(S,Z) + \gamma COV(A,Z) + COV(\epsilon, Z)$, we find

$$\delta = \frac{COV(Y,Z)}{COV(S,Z)}$$

if $COV(A,Z) = COV(\epsilon, Z) = 0$

IVs in practice

Easy to think of in terms of randomized controlled trial...

Measure	Offered Seat	Not Offered Seat	Difference
Score	-0.003	-0.358	0.355
% Enrolled	0.787	0.046	0.741
Effect			0.48

Angrist *et al.*, 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management.*

What is IV *really* doing

Think of IV as two-steps:

1. Isolate variation due to the instrument only (not due to endogenous stuff)
2. Estimate effect on outcome using only this source of variation

In regression terms

Interested in estimating δ from $y_i = \alpha + \beta x_i + \delta D_i + \varepsilon_i$, but D_i is endogenous (no pure "selection on observables").

Step 1: With instrument Z_i, we can regress D_i on Z_i and x_i,
$D_i = \lambda + \theta Z_i + \kappa x_i + \nu$,
and form prediction $\hat{D}_i$.

Step 2: Regress y_i on x_i and $\hat{D}_i$,
$y_i = \alpha + \beta x_i + \delta \hat{D}_i + \xi_i$

Derivation

Recall $\hat{\theta} = \frac{C(Z,S)}{V(Z)}$, or $\hat{\theta} V(Z) = C(Y, Z)$. Then:

$$
\begin{aligned}
\hat{\delta} &= \frac{COV(Y, Z)}{COV(S, Z)} \\
&= \frac{\hat{\theta} C(Y, Z)}{\hat{\theta} C(S, Z)} = \frac{\hat{\theta} C(Y, Z)}{\hat{\theta}^2 V(Z)} \\
&= \frac{C(\hat{\theta} Z, Y)}{V(\hat{\theta} Z)} = \frac{C(\hat{S}, Y)}{V(\hat{S})}
\end{aligned}
$$

In regression terms

But in practice, *DON'T* do this in two steps. Why?

Because standard errors are wrong…not accounting for noise in prediction, $\hat{D}_i$.
The appropriate fix is built into most modern stats programs.

Formal IV Assumptions

Key IV assumptions

1. *Exclusion:* Instrument is uncorrelated with the error term
2. *Validity:* Instrument is correlated with the endogenous variable
3. *Monotonicity:* Treatment more (less) likely for those with higher (lower) values of the instrument

Assumptions 1 and 2 sometimes grouped into an *only through* condition.

Exclusion

Conley et al (2010) and "plausible exogeneity", union of confidence intervals approach

- Suppose extent of violation is known in $y_i = \beta x_i + \gamma z_i + \varepsilon_i$, so that $\gamma = \gamma_0$
- IV/TSLS applied to $y_i - \gamma_0 z_i = \beta x_i + \varepsilon_i$ works
- With γ_0 unknown…do this a bunch of times!
    - Pick $\gamma = \gamma^b$ for $b = 1, \ldots, B$
    - Obtain $(1 - \alpha)$ % confidence interval for β, denoted $CI^b(1 - \alpha)$
    - Compute final CI as the union of all CI^b

Exclusion

Kippersluis and Rietveld (2018), "Beyond Plausibly Exogenous"

- "zero-first-stage" test
- Focus on subsample for which your instrument is not correlated with the endogenous variable of interest
  1. Regress the outcome on all covariates and the instruments among this subsample
  2. Coefficient on the instruments captures any potential direct effect of the instruments on the outcome (since the correlation with the endogenous variable is 0 by assumption).

Validity

Just says that your instrument is correlated with the endogenous variable, but what about the **strength** of the correlation?

Why we care about instrument strength

Recall our schooling and wages equation,

$$y = \beta S + \epsilon.$$

Bias in IV can be represented as:

$$Bias_{IV} \approx \frac{Cov(S,\epsilon)}{V(S)}\frac{1}{F+1} = Bias_{OLS}\frac{1}{F+1}$$

- Bias in IV may be close to OLS, depending on instrument strength
- **Bigger problem:** Bias could be bigger than OLS if exclusion restriction not *fully* satisfied

Testing strength of instruments

Single endogenous variable

- Stock & Yogo (2005) test based on first-stage F-stat (homoskedasticity only)
  - Critical values in tables, based on number of instruments
  - Rule-of-thumb of 10 with single instrument (higher with more instruments)
  - Lee et al (2022): With first-stage F-stat of 10, standard "95% confidence interval" for second stage is really an 85% confidence interval
  - Over-reliance on "rules of thumb", as seen in Anders and Kasy (2019)

Testing strength of instruments

Single endogenous variable

- Stock & Yogo (2005) test based on first-stage F-stat (homoskedasticity only)
- Kleibergen & Paap (2007) Wald statistic
- Effective F-statistic from Olea & Pflueger (2013)

Testing strength of instruments: First-stage

Single endogenous variable

1. Homoskedasticity
   - Stock & Yogo, effective F-stat
2. Heteroskedasticity
   - Effective F-stat

Many endogenous variables

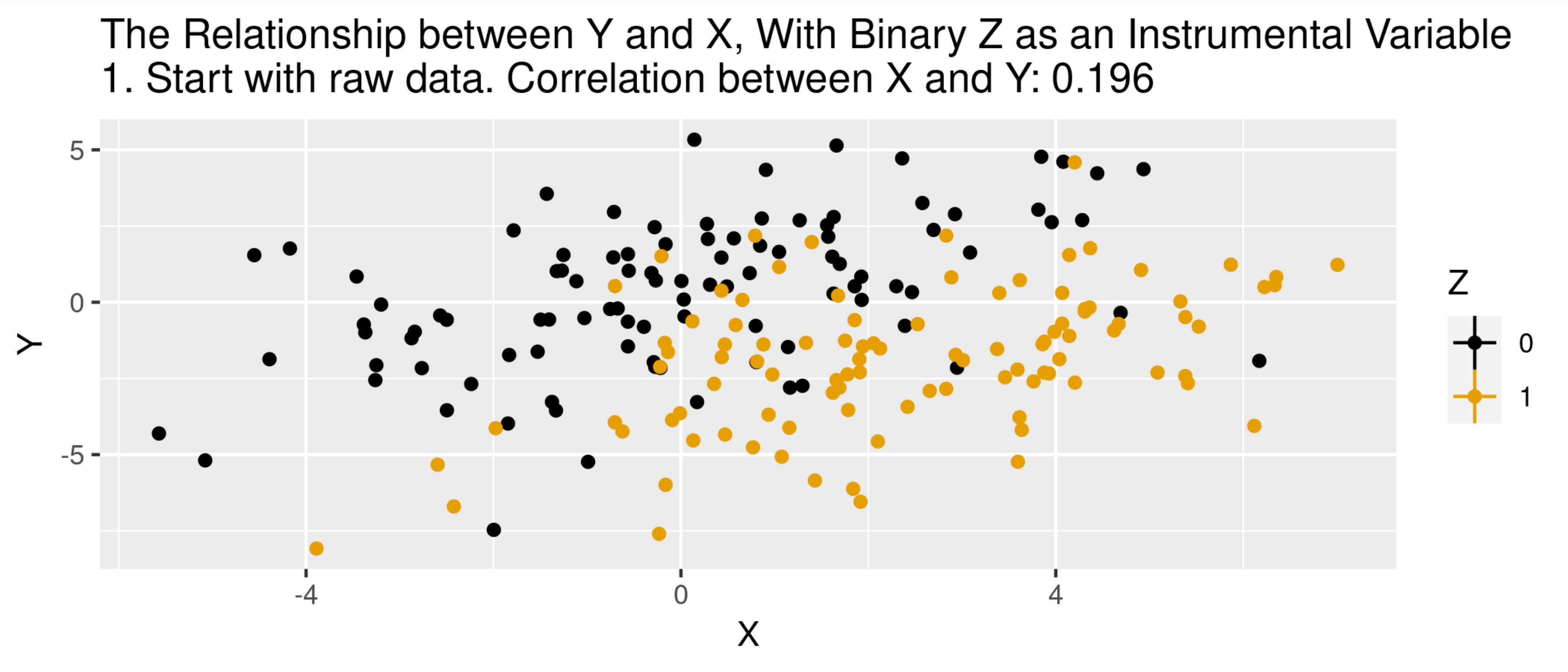
1. Homoskedasticity
   - Stock & Yogo with Cragg & Donald statistic, Sanderson & Windmeijer (2016), effective F-stat
2. Heteroskedasticity
   - Kleibergen & Papp Wald is robust analog of Cragg & Donald statistic, effective F-stat

Making sense of all of this...

- Test first-stage using effective F-stat (inference is harder and beyond this class)
- Many endogenous variables problematic because strength of instruments for one variable need not imply strength of instruments for others

IV with Simulated Data

Animation for IV



Simulated data

```
n ← 5000
b.true ← 5.25
iv.dat ← tibble(
  z = rnorm(n,0,2),
  eps = rnorm(n,0,1),
  d = (z + 1.5*eps + rnorm(n,0,1) >0.25),
  y = 2.5 + b.true*d + eps + rnorm(n,0,0.5)
)
```

- endogenous `eps`: affects treatment and outcome
- `z` is an instrument: affects treatment but no direct effect on outcome

Results with simulated data

Recall that the *true* treatment effect is 5.25

```
##
## Call:
## lm(formula = y ~ d, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8090 -0.6703 -0.0104  0.6898  3.7293
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.08422    0.01977   105.4   <2e-16 ***
## dTRUE        6.16211    0.02914   211.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.027 on 4998 degrees of freedom
## Multiple R-squared:  0.8994,    Adjusted R-squared:  0.8994
## F-statistic: 4.471e+04 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## ivreg(formula = y ~ d | z, data = iv.dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -4.182290 -0.736445 -0.009663  0.726962  4.167480
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.45751    0.02881    85.3   <2e-16 ***
## dTRUE        5.35060    0.05264   101.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 1.104 on 4998 degrees of freedom
## Multiple R-Squared: 0.8838,    Adjusted R-squared: 0.8838
## Wald test: 1.033e+04 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Checking instrument

- Check the 'first stage'

```
##
## Call:
## lm(formula = d ~ z, data = iv.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11348 -0.32880 -0.01652  0.32969  1.12071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.463461   0.005666   81.79   <2e-16 ***
## z           0.150129   0.002868   52.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4007 on 4998 degrees of freedom
## Multiple R-squared:  0.354,    Adjusted R-squared:  0.3539
## F-statistic:  2739 on 1 and 4998 DF,  p-value: < 2.2e-16
```

- Check the 'reduced form'

```
##
## Call:
## lm(formula = y ~ z, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1588 -2.1484 -0.0716  2.1998  9.1674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.93730    0.03993  123.64   <2e-16 ***
## z            0.80328    0.02021   39.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 2.823 on 4998 degrees of freedom
## Multiple R-squared:  0.2401,    Adjusted R-squared:  0.2399
## F-statistic:  1579 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Two-stage equivalence

```r
step1 <- lm(d ~ z, data=iv.dat)
d.hat <- predict(step1)
step2 <- lm(y ~ d.hat, data=iv.dat)
summary(step2)
```

```
##
## Call:
## lm(formula = y ~ d.hat, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1588 -2.1484 -0.0716  2.1998  9.1674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.45751    0.07369   33.35   <2e-16 ***
## d.hat        5.35060    0.13465   39.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.823 on 4998 degrees of freedom
## Multiple R-squared:  0.2401,    Adjusted R-squared:  0.2399
## F-statistic:  1579 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Interpretation

Heterogenous TEs

- In constant treatment effects, $Y_i(1) - Y_i(0) = \delta_i = \delta, \ \forall i$
- Heterogeneous effects, $\delta_i \neq \delta$
- With IV, what parameter did we just estimate? Need **monotonicity** assumption to answer this

Monotonicity

Assumption: Denote the effect of our instrument on treatment by π_{1i}.
Monotonicity states that $\pi_{1i} \geq 0$ or $\pi_{1i} \leq 0, \ \forall i$.

- Allows for $\pi_{1i} = 0$ (no effect on treatment for some people)
- All those affected by the instrument are affected in the same "direction"
- With heterogeneous ATE and monotonicity assumption, IV provides a "Local Average Treatment Effect" (LATE)

LATE and IV Interpretation

- LATE is the effect of treatment among those affected by the instrument (compliers only).
- Recall original Wald estimator:

$$\delta_{IV} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|\text{complier}]$$

- Practically, monotonicity assumes there are no defiers and restricts us to learning only about compliers

Is LATE meaningful?

- Learn about average treatment effect for compliers
- Different estimates for different compliers
  - IV based on merit scholarships
  - IV based on financial aid
  - Same compliers? Probably not

LATE with defiers

- In presence of defiers, IV estimates a weighted difference between effect on compliers and defiers (in general)
- LATE can be restored if subgroup of compliers accounts for the same percentage as defiers and has same LATE
- Offsetting behavior of compliers and defiers, so that remaining compliers dictate LATE