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?

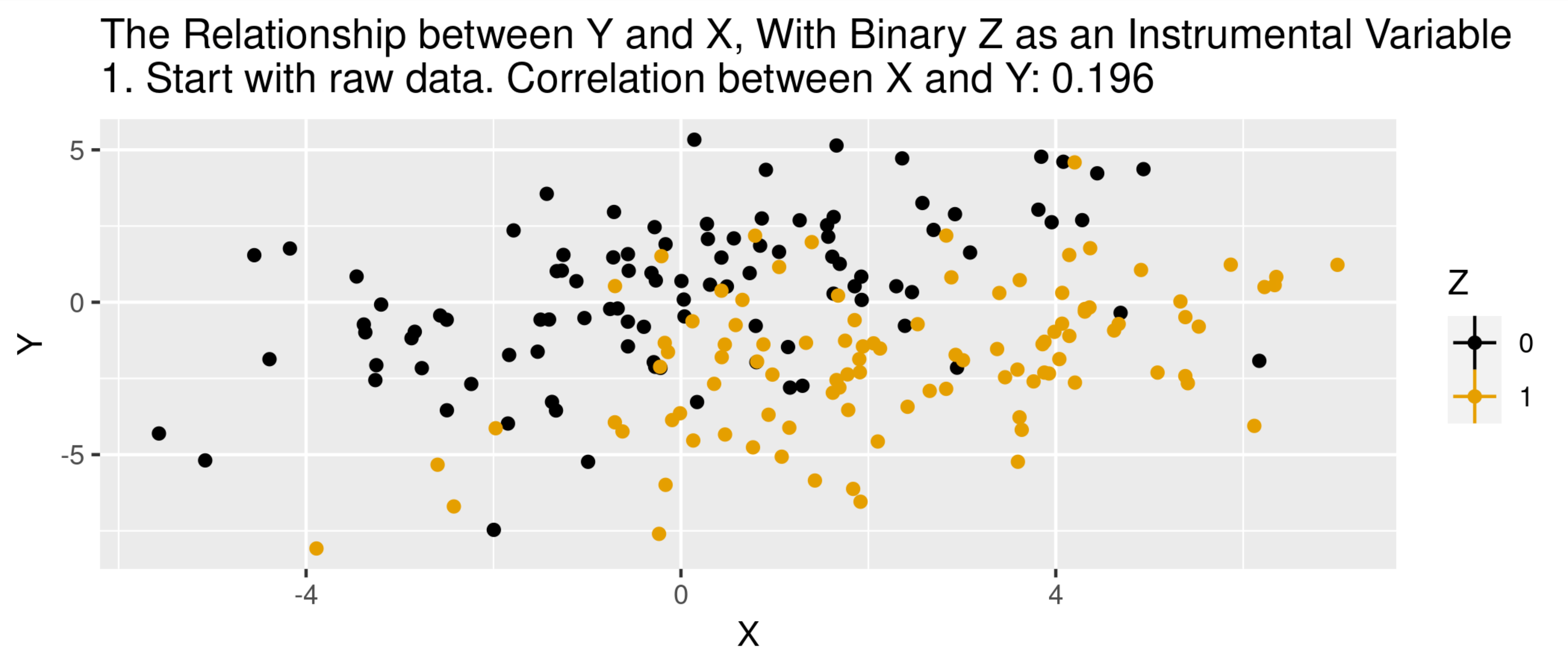
Because standard errors are wrong...not accounting for noise in prediction, $\hat{D}_i$. The appropriate fix is built into most modern stats programs.

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.

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
```

Do we need IV?

- Let's run an "augmented regression" to see if our OLS results are sufficiently different than IV

```
d.iv <- lm(d ~ z, data=iv.dat)
d.resid <- residuals(d.iv)
haus.test <- lm(y ~ d + d.resid, data=iv.dat)
summary(haus.test)
```

```
##
## Call:
## lm(formula = y ~ d + d.resid, data = iv.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2972 -0.6308 -0.0150  0.6771  3.6037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.45751    0.02564   95.83   <2e-16 ***
## dTRUE        5.35060    0.04686  114.19   <2e-16 ***
## d.resid      1.25628    0.05830   21.55   <2e-16 ***
```

Testing exclusion

- Exclusion restriction says that your instrument does not directly affect your outcome
- Potential testing ideas:
  - "zero-first-stage" (subsample on which you know the instrument does not affect the endogenous variable)
  - augmented regression of reduced-form effect with subset of instruments (overidentified models only)

Testing exogeneity

- Only available in over-identified models
- Sargan or Hansen's J test (null hypothesis is that instruments are correlated with residuals)

Testing strength of instruments

Single endogenous variable

- F-test of instruments (rule of thumb critical value of 10)
- Partial R^2

Many endogenous variables

- More complicated

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
- Over-reliance on "rules of thumb", as seen in Anders and Kasy (2019)

LATE and IV Interpretation

- With monotonicity assumption (all those affected by the instrument are affected in the same "direction")
- In the face of **heterogeneous treatment effects**, IV provides a "Local Average Treatment Effect"
- LATE: Effect of treatment among those affected by the instrument (compliers only)
- Why does this matter? Let's discuss Medicaid expansion in Oregon