Intermediate Econometrics

Lecture 8

Vincenzo Gioia

11 November 2026

Instrumental Variable Regression

IV regression

- We can eliminate the bias when $\mathbb{E}[\varepsilon|X] \neq 0$ using an instrumental variable (that we will denote as Z)

- Cases when $\mathbb{E}[\varepsilon|X] \neq 0$

1. **Errors-in-variables bias**, when the variables are measured with errors
2. **Omitted variable bias**, due to a variable correlated with X
3. **Simultaneity bias**, that happens when X causes Y and Y causes X

Instrumental Variable Regression

Our three examples

Example	Endogenous X	Source	Candidate instrument Z
Wages and ability (wage2)	IQ	measurement error	KWW score
Returns to schooling (Card, 1995)	education	omitted ability	grew up near a college
Fulton Fish Market (Graddy, 1995)	price	simultaneity	wave height at sea

- In each case we need a variable that moves X but has nothing to do with the error of the equation of interest

IV regression or OLS regression

OLS regression

- Estimate the effect of X on Y controlling for other predictors: **It works only if X is uncorrelated with errors**
- It requires exogeneity

$$\text{cov}(X, \varepsilon) = 0$$

- It is suitable when **there is no endogeneity (or it is negligible)** or you have good control variables (e.g. a proxy of the omitted variable) to include in the regression

IV regression

- Estimate the effect of X on Y using an instrument Z
- It is important to consider when the X variable is endogenous, that is correlated with the errors
- It should be used when **there is a strong (and reported) source of endogeneity and exist a valid instrument** (not always easy to detect)

Instrumental Variable Estimator

General Idea of IV

- The instruments allow to get an exogenous source of variation of the covariate, that is a source of variation that has nothing to do with the process of interest

1. do not have a direct effect on the response
2. have an indirect effect on the response because of their correlation with the endogenous covariates

Conditions on the instrument (Z)

1. **Relevance**: The instrument must affect X, that is

$$\text{cov}(Z, X) \neq 0$$

2. **Exclusion**: The instrument must be uncorrelated with the errors (ε), that is

$$\text{cov}(Z, \varepsilon) = 0$$

- If at least one of the conditions is missing, the IV estimate is not valid

Instrumental Variable Estimator

Conditions on the instrument: our examples

1. **IQ**: KWW is correlated with IQ because both reflect ability (relevance); the error of KWW is unrelated to the error of IQ and KWW has no direct effect on the wage once ability is accounted for (exclusion)
2. **Schooling**: living near a college lowers the cost of studying (relevance); but it should not affect wages in other ways, and should be unrelated to ability (exclusion)
3. **Fish**: stormy weather at sea reduces the catch and raises the price (relevance); it should not change the demand of New York buyers (exclusion)

- **Relevance can be tested** (first stage regression); **exclusion cannot** (it is an assumption that must be defended with economic arguments), unless there are more instruments than endogenous variables (Sargan test)

Simple Instrumental Variable Estimator

A key difference on variables

- **Endogenous**: variable correlated with the error ε
- **Exogenous**: variable uncorrelated with the error ε
- IV focuses on cases where the covariate X is endogenous and the instrument Z is exogenous

Framework: Simple instrumental variable estimator

- Let us consider the regression

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

- Let z_i an instrumental variable, that is correlated with x_i but it is uncorrelated with ε_i

- We consider the **exactly/just identified case** (the number of instruments, one in this case, is equal to the number of regressors, one in this case)

Simple Instrumental Variable Estimator

Where does the IV estimator come from?

- Take the covariance of both sides of $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ with z_i:

$$\operatorname{cov}(z, y) = \beta_2 \operatorname{cov}(z, x) + \operatorname{cov}(z, \varepsilon) = \beta_2 \operatorname{cov}(z, x)$$

  because $\operatorname{cov}(z, \varepsilon) = 0$ (exclusion)

- Hence, if $\operatorname{cov}(z, x) \neq 0$ (relevance),

$$\beta_2 = \frac{\operatorname{cov}(z, y)}{\operatorname{cov}(z, x)} \qquad \Longrightarrow \qquad \hat{\beta}_2^{IV} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

- Dividing numerator and denominator by $\operatorname{var}(z)$: $\hat{\beta}_2^{IV}$ is the ratio between the coefficient of z in the regression of y on z (**reduced form**) and the coefficient of z in the regression of x on z (**first stage**)

- With a weak instrument the denominator is close to zero and the ratio is unstable

Simple Instrumental Variable Estimator

Two Stage Least Squares (TSLS) Estimator

- Stage 1: Regress X on Z via OLS

$$x_i = \pi_1 + \pi_2 z_i + \varepsilon_i^*$$

and calculate the predicted values

$$\hat{x}_i = \hat{\pi}_1 + \hat{\pi}_2 z_i$$

- Stage 2: In the regression of interest, substitute the observed x_i with the predicted $\hat{x}_i$,

$$y_i = \beta_1 + \beta_2 \hat{x}_i + \varepsilon_i$$

- If the model contains **exogenous controls** W (experience, day dummies...), they enter both stages and are used as instruments for themselves

Instrumental Variable Estimator

Two Stage Least Squares Estimator

$$Y = X\beta + \varepsilon$$

- Let's denote with Z the model matrix of the instruments (including the exogenous regressors) and with X the model matrix of the regressors

- The estimator $\hat{\beta}^{TSLS}$ is

$$\hat{\beta}^{TSLS} = (X^\top P_Z X)^{-1} X^\top P_Z Y$$

where $P_Z = Z(Z^\top Z)^{-1} Z^\top$

- It is clear the reason of substituting $\hat{X}$ onto X. Indeed, one of the results of the OLS is that $\hat{X} = P_Z X$ and, since P_Z is symmetric and idempotent,

$$\hat{\beta}^{TSLS} = (\hat{X}^\top \hat{X})^{-1} \hat{X}^\top Y$$

Instrumental Variable Estimator

Two Stage Least Squares Estimator

1. Consistent ($\hat{\beta} \xrightarrow{p} \beta$)
2. The estimate of the variance of the estimator is, assuming spherical disturbances, equal to

$$\hat{\mathbb{V}}(\hat{\beta}^{TSLS}) = \hat{\sigma}^2 (X^\top P_Z X)^{-1}$$

where

$$\hat{\sigma}^2 = \frac{1}{n - \dim(\beta)} \sum_{i=1}^{n} e_i^2$$

with $e_i = y_i - X_i \hat{\beta}^{TSLS}$ (residuals computed with the **original** X, not with $\hat{X}$)

- From $\hat{\mathbb{V}}(\hat{\beta}^{TSLS})$ we can obtain the estimate of the standard error of the β's

Instrumental Variable Estimator

A special case: the Wald estimator

- The Wald estimator is the special case of the instrumental variable estimator where the instrument is a binary variable and therefore defines two groups ($z = 0$ and $z = 1$).

- The Wald estimate is then

$$\hat{\beta}_2 = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

- It is the effect of the instrument on the outcome (**reduced form**), rescaled by the effect of the instrument on the covariate (**first stage**)

Instrumental Variable Regression

Weak instruments

- Instruments should be not only uncorrelated with the error of the model, but they should also be correlated with the covariates
- Therefore, if the correlation between the endogenous variable and the instruments is weak, i.e., if the IV is performed using weak instruments, the estimator will not only be highly imprecise, but it will also be seriously biased, in the direction of the OLS estimator
- While performing IV estimation it is therefore important to check that the instruments are sufficiently correlated with the endogenous covariate
- This can be performed using an F test for the first stage regression, comparing the fit for the regression of the endogenous covariates on the set of the exogenous covariates, and on the external instruments.
- A rule of thumb often used is that the F statistic should be at least equal to 10 (a less strict rule is at least equal to 5)

Econometric Example

Recap: the OLS fits

- The three OLS regressions that we want to compare with IV:

```
library(ivreg)
library(wooldridge)
```

```
## Warning: il pacchetto 'wooldridge' è stato creato con R versione 4.5.3
```

```
data("SchoolingReturns", package = "ivreg")
data("fish")
data("wage2")

ols_ed   <- lm(log(wage) ~ education + experience + I(experience^2) +
                  ethnicity + smsa + south, data = SchoolingReturns)
ols_fish <- lm(ltotqty ~ lavgprc + mon + tues + wed + thurs, data = fish)
fit_iq   <- lm(lwage ~ educ + exper + tenure + married + south + urban + black
               data = wage2)
```

Econometric Example

Wages and ability: IV regression for the measurement error

- IQ is the mismeasured regressor and KWW is its instrument; the other regressors (education included) are exogenous and are used as instruments for themselves

```
library(ivreg)
iv_iq <- ivreg(lwage ~ educ + exper + tenure + married + south +
                 urban + black + IQ |
                 educ + exper + tenure + married + south +
                 urban + black + KWW, data = wage2)
cf <- function(m) summary(m)$coefficients[c("educ", "IQ"), 1:2]
round(cbind(OLS = cf(fit_iq), IV = cf(iv_iq)), 4)
```

```
##      Estimate Std. Error Estimate Std. Error
## educ   0.0544     0.0069    0.025     0.0166
## IQ     0.0036     0.0010    0.013     0.0049
```

Econometric Example

Wages and ability: the output of `ivreg`

```
summary(iv_iq, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + tenure + married + south +
##     urban + black + IQ | educ + exper + tenure + married + south +
##     urban + black + KWW, data = wage2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.095561 -0.224017  0.006572  0.234495  1.398914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.592453   0.325781  14.097  < 2e-16 ***
## educ         0.025032   0.016607   1.507  0.13206
## exper        0.014420   0.003321   4.342 1.56e-05 ***
## tenure       0.010456   0.002601   4.020 6.30e-05 ***
## married      0.200690   0.040678   4.934 9.56e-07 ***
## south       -0.051553   0.031128  -1.656  0.09803 .
## urban        0.176706   0.028212   6.264 5.75e-10 ***
## black       -0.022561   0.073960  -0.305  0.76040
## IQ           0.013047   0.004934   2.644  0.00832 **
##
## Diagnostic tests:
##                  df1 df2 statistic  p-value
## Weak instruments   1 926    43.024 8.98e-11 ***
## Wu-Hausman         1 925     4.267   0.0391 *
## Sargan             0  NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3807 on 926 degrees of freedom
## Multiple R-Squared:  0.19,   Adjusted R-squared: 0.183
## Wald test: 36.96 on 8 and 926 DF,  p-value: < 2.2e-16
```

Econometric Example

Wages and ability: how to read the output of `ivreg`

- **Coefficients** (Estimate, Std. Error, t value, Pr): the 2SLS estimates and their standard errors, computed with the residuals $y - X\hat{\beta}_{IV}$ (with the *observed* IQ, not the fitted one)

- **Weak instruments**: F test of the first stage, H_0: KWW has no effect on IQ, given the other regressors. A large F (rule of thumb > 10) means a **relevant** instrument
- **Residual standard error**: $\hat{\sigma}$ computed from the IV residuals. **R-squared** has no usual interpretation in IV (it can even be negative): do not use it to compare with OLS
- **Wald test**: joint significance of all the coefficients except the intercept
- **Wu-Hausman** and **Sargan** (next lecture)

Econometric Example

Wages and ability: how to read the result

- The IQ coefficient **increases** from 0.0036 to 0.0130 (a factor of 3.6): this is what we expect if OLS was attenuated by the measurement error
- The education coefficient **decreases** from 0.054 to 0.025: once ability is controlled for in a better way, the part of the effect that was wrongly attributed to education disappears
- The first-stage F of KWW is 43: the instrument is strong
- **Caution**: validity requires that KWW has no direct effect on wages once ability is accounted for, which is debatable. The price of IV is **precision**: the IQ standard error is about 5 times the OLS one, and educ is no longer significant ($p = 0.13$, 95% CI roughly $(-0.008, 0.058)$, which contains both 0 and the OLS value). IV removes (part of) the bias but at the cost of a **larger variance**: we cannot conclude that education has no effect, only that this estimate is imprecise

Econometric Example

Returns to schooling: the Wald estimator

- Instrument: `nearcollege` (binary). Compare men who grew up near a college and men who did not

```
m <- aggregate(cbind(logwage = log(wage), education) ~ nearcollege,
               data = SchoolingReturns, FUN = mean)
m
```

```
##   nearcollege  logwage education
## 1          no 6.155494  12.69801
## 2         yes 6.311401  13.52703
```

```
(m$logwage[2] - m$logwage[1]) / (m$education[2] - m$education[1])
```

```
## [1] 0.1880626
```

- Those near a college have 0.83 more years of schooling and a log wage higher by 0.156: the implied return is about 0.19 per year (without controls)

Econometric Example

Returns to schooling: first stage

- Relevance: regress education on the instrument and on the exogenous controls

```
fs_ed <- lm(education ~ nearcollege + experience + I(experience^2) +
              ethnicity + smsa + south, data = SchoolingReturns)
round(summary(fs_ed)$coefficients["nearcollegeyes", ], 4)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     0.3373     0.0825     4.0887     0.0000
```

```
fs_ed0 <- update(fs_ed, . ~ . - nearcollege)
anova(fs_ed0, fs_ed)[2, c("F", "Pr(>F)")]
```

```
##         F    Pr(>F)
## 2 16.718 4.452e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Growing up near a college is associated with 0.34 more years of schooling; $F = 16.7$ (> 10): the instrument is not weak

Econometric Example

Returns to schooling: IV regression

```
iv_ed <- ivreg(log(wage) ~ education + experience + I(experience^2) +
                  ethnicity + smsa + south |
                  nearcollege + experience + I(experience^2) +
                  ethnicity + smsa + south, data = SchoolingReturns)
rbind(OLS = summary(ols_ed)$coefficients["education", 1:2],
      IV  = summary(iv_ed)$coefficients["education", 1:2])
```

```
##      Estimate  Std. Error
## OLS 0.0740090 0.003505435
## IV  0.1322888 0.049233236
```

```
confint(iv_ed)["education", ]
```

```
##      2.5 %     97.5 %
## 0.03575455 0.22882311
```

- IV: one more year of schooling raises wages by about 13%, against 7.4% with OLS; the IV estimate is much less precise (standard error 0.049 against 0.004)

Econometric Example

Returns to schooling: how to read the result

- The IV estimate is **larger** than the OLS one, which is the opposite of what the simple ability story ($\hat{\beta}^{OLS} > \beta_2$) predicts

- Possible explanations discussed in the literature:

1. measurement error in schooling (attenuation of OLS)
2. the IV estimates the effect for those whose education is changed by the instrument (people who go to college only because it is near), who may have high returns
3. the instrument is not perfectly exogenous (e.g. families living near colleges are different)

- The confidence interval of the IV estimate is wide, and it includes the OLS estimate: with these data we cannot rule out that the two are equal (we will see with the Hausman test)

Econometric Example

Returns to schooling: TSLS by hand

- Using $\hat{\beta}^{TSLS} = (X^\top P_Z X)^{-1} X^\top P_Z Y$:

Exercise: get by hand

```
X <- model.matrix(~ education + experience + I(experience^2) +
                    ethnicity + smsa + south, data = SchoolingReturns)
Z <- model.matrix(~ nearcollege + experience + I(experience^2) +
                    ethnicity + smsa + south, data = SchoolingReturns)
y <- log(SchoolingReturns$wage)

XPZ  <- t(X) %*% Z %*% solve(t(Z) %*% Z) %*% t(Z)
beta <- solve(XPZ %*% X, XPZ %*% y)

e      <- y - X %*% beta                          # residuals with the original X
sigma2 <- sum(e^2) / (nrow(X) - ncol(X))
se     <- sqrt(diag(sigma2 * solve(XPZ %*% X)))
cbind(estimate = beta, se = se)[2, ]
```

```
##                     se
## 0.13228883 0.04923324
```

Econometric Example

Fulton Fish Market: first stage

- Instruments for the price: `wave2`, `wave3` (two instruments for one endogenous variable: **overidentified** model)

```
fs_fish  <- lm(lavgprc ~ mon + tues + wed + thurs + wave2 + wave3, data = fish)
fs_fish0 <- lm(lavgprc ~ mon + tues + wed + thurs, data = fish)
anova(fs_fish0, fs_fish)[2, c("F", "Pr(>F)")]
```

```
##      F    Pr(>F)
## 2 19.1 1.219e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Joint test on the two instruments: $F = 19.1$ (p-value $\approx 10^{-7}$): rough weather at sea raises the price, the instruments are strong

Econometric Example

Fulton Fish Market: IV regression

```
iv_fish <- ivreg(ltotqty ~ lavgprc + mon + tues + wed + thurs |
                    mon + tues + wed + thurs + wave2 + wave3, data = fish)
rbind(OLS = summary(ols_fish)$coefficients["lavgprc", 1:2],
      IV  = summary(iv_fish)$coefficients["lavgprc", 1:2])
```

```
##       Estimate Std. Error
## OLS -0.5246553  0.1761115
## IV  -0.8158181  0.3274372
```

```
confint(iv_fish)["lavgprc", ]
```

```
##      2.5 %     97.5 %
## -1.4662318 -0.1654044
```

- The IV estimate of the price elasticity of demand is -0.82 (OLS: -0.53): a 1% increase of the price lowers the quantity demanded by about 0.8%
- IV is again less precise: the 95% interval is $(-1.46, -0.17)$, and it contains the OLS estimate

Econometric Example

Fulton Fish Market: what about the supply curve?

- To estimate the supply curve we need variables that shift the **demand** only: the day-of-the-week dummies are candidates

```
iv_sup <- ivreg(ltotqty ~ lavgprc | mon + tues + wed + thurs, data = fish)
summary(iv_sup, diagnostics = TRUE)$diagnostics["Weak instruments", ]
```

```
##         df1         df2  statistic    p-value
##  4.0000000 92.0000000  0.2031537  0.9360477
```

```
round(summary(iv_sup)$coefficients["lavgprc", 1:2], 3)
```

```
##   Estimate Std. Error
##      3.048      4.368
```

- The first-stage $F \approx 0.2$: the day dummies hardly move the price, they are **weak instruments**, and the supply elasticity cannot be estimated (huge standard error)
- Identification of the two curves requires a shifter for each of them

Econometric Example

References

- Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press.

- Graddy, K. (1995). Testing for imperfect competition at the Fulton Fish Market. *RAND Journal of Economics* 26 (1): 75–92.

- Blackburn, M. and Neumark, D. (1992). Unobserved ability, efficiency wages, and interindustry wage differentials. *Quarterly Journal of Economics* 107 (4): 1421–1436.