# Endogeneity and Instrumental Variables Estimation

Applied Econometrics for Researchers, PhD

Vera Rocha, CBS-SI, vr.si@cbs.dk

23rd November 2022

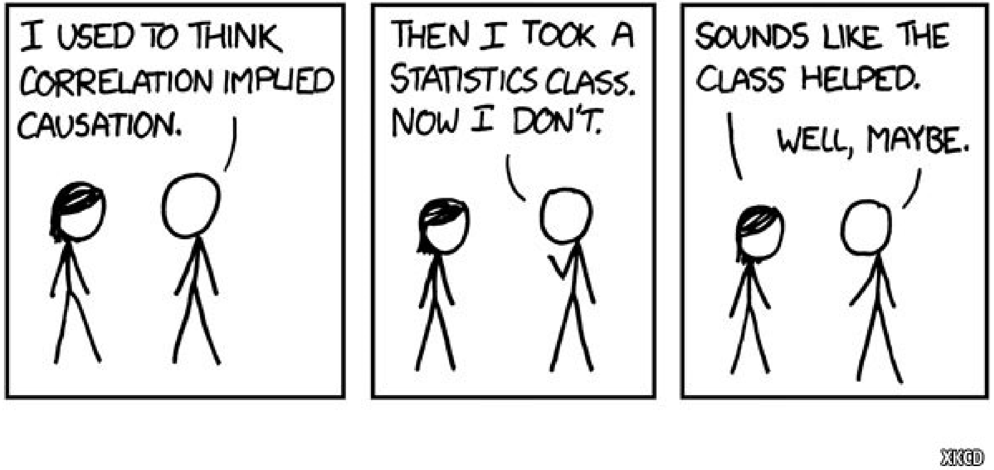
# Agenda

1. The source(s) of the endogeneity problem
2. One possible solution: Instrumental Variables
3. Requirements: What makes a good instrument?
4. Key tests when using IV estimation
5. Examples, examples, examples (including Stata examples ☺)

**Suggested readings:**

- **Wooldridge**, ”Introductory econometrics, a modern approach”, **Chapter 15:** IV estimation and 2SLS
    - or S. Cunningham, Chapter 7 on Instrumental Variables (The Mixtape book)
- Hill et al. (2021), ”Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research”, *Journal of Management*, 47(1): 105-143.

# Researchers care about causality!




# It is important to distinguish between…

| | Heckman (sample selection) models | Matching Models (e.g., PSM) | Instrumental Variables **(today)** |
|---|---|---|---|
| **When** | Y is missing in some cases | X is a binary intervention/choice | X is endogeneous (correlated with unobservables) |
| **Problem** | The missings in Y are driven by a "selection process" | T & C groups are very different | Four possible causes that make X correlated with the error term |
| **Stata commands** | *heckman, (twostep)* | *teffects psmatch, tebalance, teffects overlap* | *ivreg2, (first)* *ivendog, overid* |
| **Key tests** | Significance of the Inverse Mills Ratio or *rho* | Balancing and overlapping conditions | Relevance and validity of the instruments |
| **Attention!** | Need for valid exclusion restrictions; selection bias important when IMR/rho significant and X predicts selection | T & C only matched on observable characteristics. If unobservables matter **PSM does not provide causal effects → IV** | Validity only possible to assess ("imperfectly") when the model is overidentified; bad IVs are worse than OLS |
| **First stage** | Probit predicting selection into the sample (Y ≠ missing) | Probit predicting probability of being treated (X) | OLS predicting the endogenous variable (X) |



**General equation:** $y_i = \beta_0 + \beta_1 X1_i + \cdots + u_i$

# What is the (identification) problem?

Assume we want to estimate the **causal effect** of some variable $x$ on another ($y$), *holding everything else constant:*

- ***Example 1)*** *What is the effect on wages from having a longer education? ("What is the return to schooling?")*
- ***Example 2)*** *What is the effect of wealth on the propensity to become an entrepreneur? ("Is entrepreneurship limited by access to capital markets?")*

We have measurements of both $x$ and $y$ so we can compute a correlation or the regression coefficients (by regressing $y$ on $x$).

Does any of these coefficients identify the **causal effect of $x$ on $y$**?

# What is the (identification) problem?

- - - -> ***Does x cause y?***

Yes! If $x$ is an **exogenous** variable

No! If $x$ is an **endogenous** variable

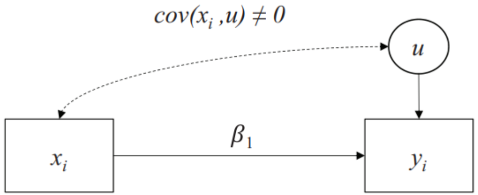
What is **endogeneity**?

What is the **consequence** of endogeneity?

How to **detect** endogeneity?

How to **address** endogeneity?

# What is endogeneity exactly?




We want **explanatory variables** to be **exogenous**, but **what if they are not**?

Recall the classical linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

which requires $\boxed{E(u_i|x_i) = E(u_i) = 0}$

The expected value of $u_i$ does not depend on $x_i$ ⇔ $u_i$ & $x_i$ are independent ⇔

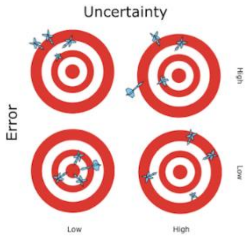
necessary condition for OLS estimator to be consistent (i.e., unbiased)

If, for **any reason**, $x_i$ is correlated with $u_i$, this condition is violated and $x_i$ is said to be an ***endogenous*** variable → $cov(x_i, u_i) \neq 0$

# The Different Sources of Endogeneity

**1. Mis-specification of the model (omitted variable bias)**

– The most common case




**2. Measurement error (in an explanatory variable)**

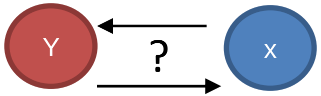
– The classical errors-in-variables case

**3. Self-selection (into *treatment*)**

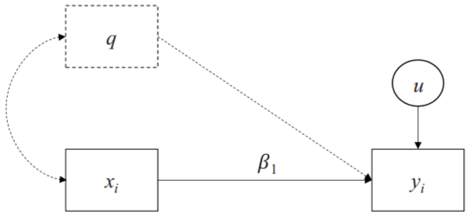
– e.g., Individuals self-select into certain behaviors or programs

***4. Simultaneity (or reverse causality)***

– One or more explanatory variables are jointly determined with the dependent variable



# 1. Omitted Variable Bias




Consider a wage function. We want to examine how the length of a person's education affects her wage (*returns to schooling*)

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + \varepsilon \qquad [1]$$

We suspect that *educ* is correlated with *ability*. Why? Theory!

Assume **ability is unobservable**: an ***omitted variable*** captured by the error term

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u \qquad [2]$$

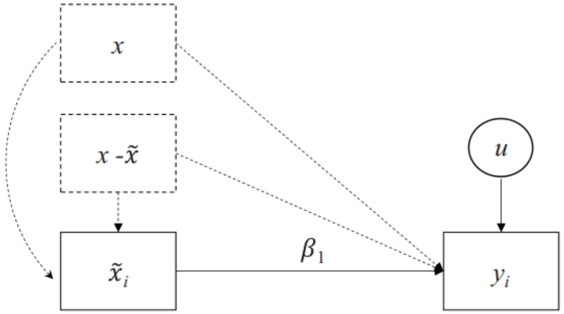
where $u = \beta_3 abil + \varepsilon$

***educ*** is **endogenous** because of its correlation with the error term $u$ (via *abil*).

The OLS estimator of $\beta_1$ (and $\beta_2$) is biased. Correlation between a single explanatory variable and the error generally results in **all** OLS estimates being biased.

*If panel data are available, consider a model with fixed effects to capture this permanent unobserved heterogeneity (next week!).*

# 2. Measurement Error Bias




Consider estimating the effect of family income on college grades

$$colGPA = \beta_0 + \beta_1 \boldsymbol{faminc}^* + \beta_2 hsGPA + $$

where $\boldsymbol{faminc}^*$ is the actual (true) value of family income. Let $faminc$ be the observed value of family income, and

$$\boldsymbol{faminc} = \boldsymbol{faminc}^* + e_1$$ where $e_1$ is a measurement error.

Equation [1] can be rewritten as

$$colGPA = \beta_0 + \beta_1 \boldsymbol{faminc} + \beta_2 hsGPA + (\varepsilon - \boldsymbol{\beta_1 e_1})$$

It can be easily shown that (observed) $faminc$ is correlated with $(\varepsilon - \beta_1 e_1)$, causing bias in the OLS estimator of $\beta_1$ (and $\beta_2$).

*Other examples:*
*Firm reputation → stock price*
*(a survey on firm reputation may systematically overrate firms with a high stock price)*

# 3. Self-selection Bias

Recall matching lecture & don't confuse with Heckman models!

Consider estimating the **effect of size** of current firm on the **probability that a wage worker becomes self-employed**

$$Prob(Self = 1) = \beta_0 + \beta_1 sfirm + u$$

- $Self$ is an outcome variable (1/0: worker becomes self-employed or not)
- $sfirm$ is a binary variable (1 if an individual worked in a firm smaller than N employees)

Self-selection arises because workers with **higher entrepreneurial preference** are more willing to work for small firms. *Entrepreneurial preference* is captured by the **error term, u**.

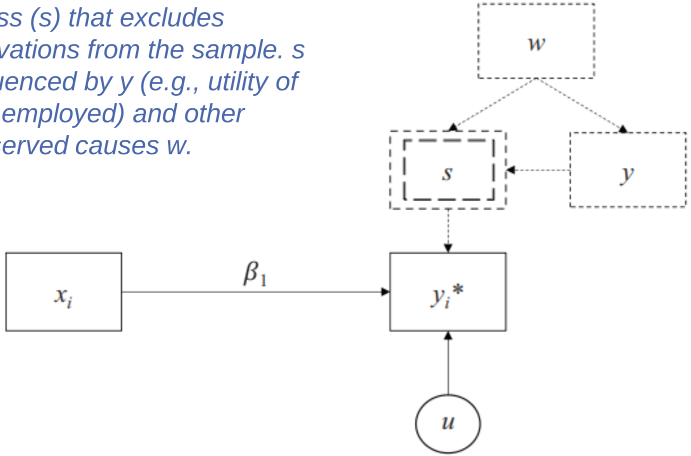
Therefore → $E(u|sfirm = 0) \neq E(u|sfirm = 1)$ or $cov(sfirm, u_i) \neq 0$

The correlation between $sfirm$ and $\varepsilon$ causes bias in the OLS estimator of $\beta_1$. **Self-selection boils down to the omitted variable problem.**

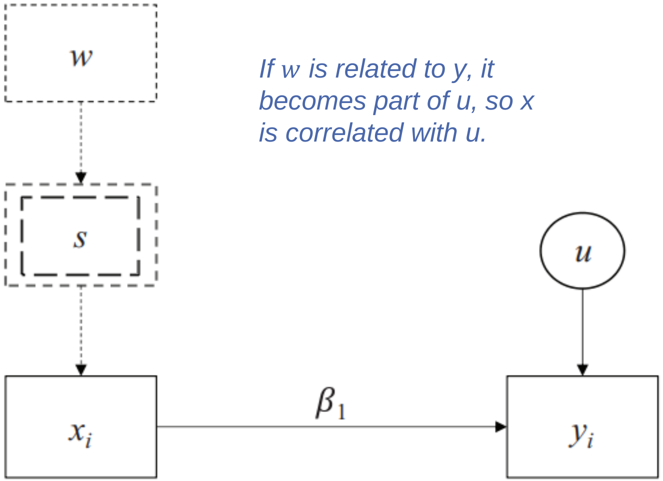
# Selection into sample vs. into treatment!

## Sample selection (Heckman models)

*$y^*$ is affected by a selection process (s) that excludes observations from the sample. s is influenced by y (e.g., utility of being employed) and other unobserved causes w.*

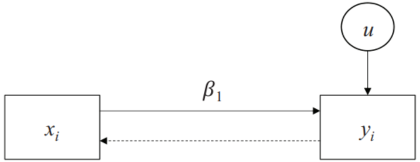



## Selection of treatment (based on *unobservables*): **today's focus**




*If w is related to y, it becomes part of u, so x is correlated with u.*

# 4. Simultaneity Bias




A city wants to determine if adding officers to its police force will reduce the crime rate

$$crimepc = \beta_{10} + \alpha_1 policepc + \beta_{11} incpc + \varepsilon_1 \qquad [1]$$

A city's spending on law enforcement is also partly determined by its (expected) crime rate

$$policepc = \beta_{20} + \alpha_2 crimepc + \beta_{21} csize + \varepsilon_2 \qquad [2]$$

Insert equation [1] into equation [2] and solving:

$$policepc = \frac{\beta_{20} + \alpha_2 \beta_{10}}{1 - \alpha_2 \alpha_1} + \frac{\alpha_2 \beta_{11}}{1 - \alpha_2 \alpha_1} incpc + \frac{\beta_{21}}{1 - \alpha_2 \alpha_1} csize + \frac{\alpha_2 \boldsymbol{\varepsilon_1} + \varepsilon_2}{1 - \alpha_2 \alpha_1}$$

We see that $cov(policepc, \varepsilon_1) \neq 0$. Applying OLS to [1] will give a biased estimate of $\alpha_1$. What happens if $\alpha_2 = 0$?

*Other examples:*
*Alcohol consumption ↔ unemployment*
*R&D expenditures ↔ firm performance*
*...*

# Solution: Instrumental Variables Estimation

Recall the wage function example:

$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e$ ["true" equation]

problem: $educ$ and $abil$ are correlated & $abil$ is not observed

$\log(wage) = \beta_0 + \beta_1 educ + u$ [estimated equation]

$u$ now includes unmeasured $abil$

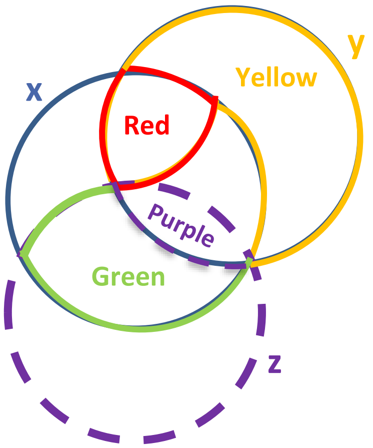
The OLS estimator of $\beta_1$ is biased because

$cov(educ, u) \neq 0$

Solution: **Instrumental Variable, $z$, for *educ***

Requirements: $cov(z, u) = 0$ & $cov(z, educ) \neq 0$

- Note: *z* is “extra information” and observable
- What does the *IV*, $z$, do to remove the bias?




[1] **Red** area: correlation between $x$ and the error term

[2] Using OLS for the intersection between $x$ and $y$ makes estimates inconsistent, because of [1].

[3] Find a variable ***z***, which is **not correlated with the error term (e.g., ability), but with *x*** (z is an instrument for *x*).

[4] Find the estimate of the overlapping area of *x* and *z*, $\hat{x}$, by regressing *x* on *z* (First stage).

[5] Regress *y* on $\hat{x}$. The **purple** area is used to form the IV estimator of $\beta_1$ (Second stage).

# IV is one of the most common solutions

*Self*-selection or *selection into treatment*!
(not *sample* selection)

**Summary of Identified Endogeneity Sources and Methods**

| | Omitted Variable | Simultaneity | Measurement Error | Selection | Unclear | Total |
|---|---|---|---|---|---|---|
| Experiment | 2 | 2 | 0 | 4 | 2 | 10 |
| Quasi-experiment | 4 | 1 | 0 | 5 | 1 | 11 |
| Design choices | 2 | 2 | 0 | 1 | 3 | 8 |
| Matching sample | 4 | 2 | 0 | 10 | 4 | 20 |
| Measurement | 0 | 1 | 5 | 0 | 0 | 6 |
| Control variables | 10 | 1 | 0 | 3 | 5 | 19 |
| Panel | 11 | 11 | 1 | 2 | 13 | 38 |
| Instrumental variable | 41 | 42 | 0 | 94 | 48 | 225 |
| Dynamic panel | 6 | 16 | 0 | 3 | 12 | 37 |
| Other | 6 | 7 | 0 | 10 | 14 | 37 |
| **Total** | **86 (20.9%)** | **85 (20.7%)** | **6 (1.5%)** | **132 (32.1%)** | **102 (24.8%)** | **411** |

Next week

*Source: Hill et al. (2021) Journal of Management*

# **Implementing IV Estimation:**
# One Endogenous Variable & One Instrument

Just-identified model

$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$ [structural equation]

- $y$: $\log(wage)$ $x$: $educ$ **(Endogenous Variable)** $z_1$: $exper$ **(Exogenous Variable)**
- Omitted variable $ability$ (correlated with $x$, therefore $x$ is correlated with $u$)

**First Stage**

- Find an instrument, $\boldsymbol{z_2}$ (*e.g.* number of siblings) assumed to satisfy conditions:
  - $cov(sibl, u) = 0$
  - $cov(sibl, educ) \neq 0$

Convincing?

- Estimate the following regression: $x = \pi_0 + \pi_1 z_1 + \pi_2 \boldsymbol{z_2} + v_2$ [reduced form equation]
- Obtain the fitted value of $x$: $\hat{x}$ is now "clean" from endogenous variation

**Second Stage**

- $y = \beta_0 + \beta_1 \hat{x} + \beta_2 z_1 + u$
- IV estimators $(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$ are consistent (all terms in the right-hand side are exogenous)

# Implementing IV Estimation:
## <u>One</u> Endogenous Variable & <u>Multiple</u> Instruments

Over-identified model

$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$ [structural equation]

- $y$: $\log(wage)$ $x$: $educ$ **(Endogenous Variable)** $z_1$: $exper$ **(Exogenous Variable)**
- Omitted variable $ability$ (correlated with $x$, therefore $x$ is correlated with $u$)

**Instrumental variables available:** $sibl$ ($\boldsymbol{z_2}$), $motheduc$ ($\boldsymbol{z_3}$), $fatheduc$ ($\boldsymbol{z_4}$)

- $cov(\boldsymbol{z_2}, u) = 0, cov(\boldsymbol{z_3}, u) = 0, cov(\boldsymbol{z_4}, u) = 0$
- $cov(\boldsymbol{z_2}, educ) \neq 0, cov(\boldsymbol{z_3}, educ) \neq 0, cov(\boldsymbol{z_4}, educ) \neq 0$

(more) convincing?

**Best IV:** Linear combination of all instruments is likely to be highly correlated with $x$ ($educ$)

**First Stage:** $x = \pi_0 + \pi_1 z_1 + \pi_2 \boldsymbol{z_2} + \pi_3 \boldsymbol{z_3} + \pi_4 \boldsymbol{z_4} + v_2$ [reduced form equation]

- Obtain the fitted value of $x$: $\hat{x} = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 \boldsymbol{z_2} + \hat{\pi}_3 \boldsymbol{z_3} + \hat{\pi}_4 \boldsymbol{z_4}$

**Second Stage**

- $y = \beta_0 + \beta_1 \hat{x} + \beta_2 z_1 + u$
- IV estimators ($\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$) are consistent and often called the **two stage least square (2SLS) estimators**.

# Implementing IV Estimation:
## Multiple Endogenous Explanatory Variables

$y = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u$ [structural equation]

- **Endogenous variable:** $x_2, x_3$; **Exogenous variable:** $z_1, z_2, z_3$

**Order Condition for Identification:**

- We need at least **as many** excluded exogenous variables (i.e, **instruments**) as there are endogenous explanatory variables
- **2SLS estimator not feasible if there is only 1 instrument** – e.g. $\boldsymbol{z_4}$ for both $x_2$ and $x_3$

**Two instruments:** ($\boldsymbol{z_4}$, $\boldsymbol{z_5}$) for ($x_2$,$x_3$) – although in theory: $\boldsymbol{z_4}$ for $x_2$, $\boldsymbol{z_5}$ for $x_3$

- Two reduced forms:
- $x_2 = \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 + \pi_{23} z_3 + \pi_{24} \boldsymbol{z_4} + \pi_{25} \boldsymbol{z_5} + v_2$
- $x_3 = \pi_{30} + \pi_{31} z_1 + \pi_{32} z_2 + \pi_{33} z_3 + \pi_{34} \boldsymbol{z_4} + \pi_{35} \boldsymbol{z_5} + v_3$

# Two Requirements for Instruments

### Instrument Relevance

$$cov(z, x) \neq 0$$

We need the **Purple** and **Green** area as large as possible (**a strong instrument**) to have sufficient information to explain variations in *Y.*

**How to test?** First-stage reduced form equation (including all exogenous variables)

$$x = \pi_0 + \pi_1 z + \cdots + v$$

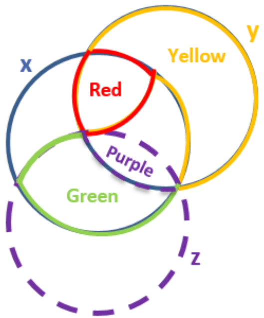
Relevance holds if $\hat{\pi}_1 \neq 0$.

### Instrument Validity (i.e., Exogeneity)

$$cov(z, u) = 0$$

We want the **Purple** area to solely explain variations in Y, independent of the error term.

**How to test?** Impossible, because $u$ is unobserved. Rely on intuition or theory!




*All in all, we want $cov(z, u)$ to be as **small** as possible (in principle, zero) and $cov(z, x)$ as **large** as possible.*

# Key Tests: The Presence of Endogeneity

**Is IV/2SLS estimation really necessary?** *[Note: 2SLS estimator is less efficient than OLS, due to larger standard errors]*

We could directly **compare $\hat{\boldsymbol{\beta}}^{OLS}$ and $\hat{\boldsymbol{\beta}}^{IV}$**, and determine whether the difference is statistically significant (Hausman Test). If it is, endogeneity is likely to be present!

Other **alternative tests**:

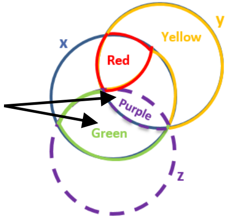
- Run the reduced-form equation: $\boldsymbol{x} = \pi_0 + \pi_1 \boldsymbol{z} + \cdots + v$ *(if there is a single endogenous variable)*
- Obtain the residuals $\hat{\boldsymbol{v}}$ by OLS, and include $\hat{\boldsymbol{v}}$ as an additional regressor in the structural equation for $y$:
- $y = \beta_0 + \beta_1 \boldsymbol{x} + \cdots + \delta_1 \hat{\boldsymbol{v}} + e_1$ $\rightarrow$ **test $\boldsymbol{H_0}$: $\boldsymbol{\delta_1 = 0}$** using the ***t***-statistic
- **If rejected, endogeneity is present and must be solved!** We cannot trust $\hat{\beta}^{OLS}$.

For **multiple endogenous variables**

- Obtain the residuals by OLS from the first-stage reduced-form for **each** suspected endogenous variable
- Include **each** residual in the structural equation for $y$, and perform an ***F*-test** to test their joint significance
- **If jointly significant, endogeneity is present and we must find valid and relevant instruments.**

# Key Tests: The Relevance of the Instrument(s)

We want to check the **relevance** of the instrument(s) ⇔ do we have **weak or strong** instruments?




Test the significance of the **green** and **purple** overlaps between the variation of $x$ and $z$.

*Just-identified model*

1. One endogenous regressor $\boldsymbol{x}$, and one instrument, $\boldsymbol{z}$ → test significance of $\hat{\pi}_1$ *(recall slide 17)*

2. One endogenous regressor $\boldsymbol{x}$, and two instruments, $\boldsymbol{z_3}$ and $\boldsymbol{z_4}$.

$$x = \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 + \pi_{23} \boldsymbol{z_3} + \pi_{24} \boldsymbol{z_4} + v_2$$

*Remember, $z_1$ and $z_2$ are exogeneous regressors that also enter the structural equation!*

(followed by **F-test** on the significance of $\boldsymbol{z_3}$ and $\boldsymbol{z_4}$)

3. Two endogenous regressors $\boldsymbol{x_2}, \boldsymbol{x_3}$, and two instruments, $\boldsymbol{z_3}$ and $\boldsymbol{z_4}$.

   – Two reduced forms. Same logic (**F-test**), but also applied across equations (Stock-Yogo). Ideally, we want to obtain smaller p-values than usual to guarantee good performance

$$x_2 = \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 + \pi_{23} z_3 + \pi_{24} \boldsymbol{z_4} + \pi_{25} \boldsymbol{z_5} + v_2$$

$$x_3 = \pi_{30} + \pi_{31} z_1 + \pi_{32} z_2 + \pi_{33} z_3 + \pi_{34} \boldsymbol{z_4} + \pi_{35} \boldsymbol{z_5} + v_3$$

# Key Tests: Overidentifying Restrictions

**Can we somehow test for the correlation between the error term and the instruments?**
*(recall the validity/exogeneity requirement)*

- Only possible, if we have ”extra” instruments for the endogeneous explanatory variable(s) *(i.e., more z's than endogeneous x's)*. In that case, we can test whether *some of them* are uncorrelated with the structural error.

**A formal test:**

- Estimate the structural equation by 2SLS, and obtain the 2SLS residuals $\hat{u}$:
- $\hat{u} = y - \hat{\beta}_0 - \hat{\beta}_1\hat{x} - \hat{\beta}_2 z_1 - \hat{\beta}_3 z_2$ (Note: $\hat{x}$! $z_1$ and $z_2$ are exogenous determinants of $y$)
- Then regress $\hat{u}$ on ***all* exogenous variables** (all $z$'s, including the instruments).
- Obtain the *R*-squared and perform the following test:
- $H_0$: *all* IVs are uncorrelated with $u$
- Key statistic: $\mathrm{n}R^2 \sim \chi_q^2$ (Note: $q = \#\, instruments - \#\, endogenous\ variables$ (= # extra instruments))
- If $\mathrm{n}R^2$ exceeds (say) the 5% critical value in the distribution, we reject $H_0$, and conclude that at least one of the IVs is *not* exogenous.

# Searching for (Good) Instruments: Examples

## Example 1. **Financial constraints and entry into self-employment**

- RQ: Do prospective entrepreneurs face limited access to the capital market? If so, personal wealth should play an important role in business creation:

$$self_i = \beta_0 + \beta_1 wealth_i + \beta_2 X_i + u_i, \qquad \text{expect } \beta_1 > 0$$

- What is the potential endogeneity problem?
    - An unmeasured omitted variable: (entrepreneurial) ability
    - **People with higher (entrepreneurial) ability are more likely to a) enter self-employment; b) accumulate more wealth**
    - $cov(wealth, u) > 0$ through (unobserved) ability. There might be a **(positive) bias in OLS**
- Solution? Find an **instrumental variable, $z$ for $wealth$**, such that:
    - $cov(z, wealth) \neq 0$, and $cov(z, u) = 0$
    - Can $z$ be any variable? No. Theoretically, $z$ is an exogenous variable that does not enter the structural model. But it must affect wealth.

# Searching for (Good) Instruments: Examples

## Example 1. Financial constraints and entry into self-employment (cont.)

**Candidates for instruments?**

- Inheritance and gifts (Blanchflower and Oswald, 1991, JOLE)
- Lottery winnings (Lindh and Ohlsson, 1996, EJ)

$\widehat{wealth}_i = \hat{\pi}_0 + \hat{\pi}_1 \boldsymbol{z_i} + \hat{\pi}_2 X_i$ [reduced form eq]

$self_i = \beta_0 + \beta_1 \widehat{wealth}_i + \beta_2 X_i + u_i$ [structural eq]

- **Criticisms:**
  - Inheritance and gifts are *indirectly* correlated with ability
  - The behavior of purchasing lottery tickets is also related to personalities
  - Both are unobservable to econometricians
  - Can we really claim that $cov(z, u) = 0$?

**Alternative instruments?**

Outcomes of random economic shocks:

- Capital gains from housing price change (Hurst and Lusardi 2004, JPE)
- Unforeseen change in income tax policy
- …

# Searching for (Good) Instruments: Examples

## Example 2. Returns to education (in wage equations)

**Education is correlated with unobserved ability → Candidates for instruments?**

- Distance to educational institutions? *(but what if areas with no education institutions are also low-income areas?)*
- Institutional constraints – e.g., compulsory schooling laws?
- Any **natural experiment?** – e.g., Angrist-Krueger (1991) – **individual's quarter of birth!**

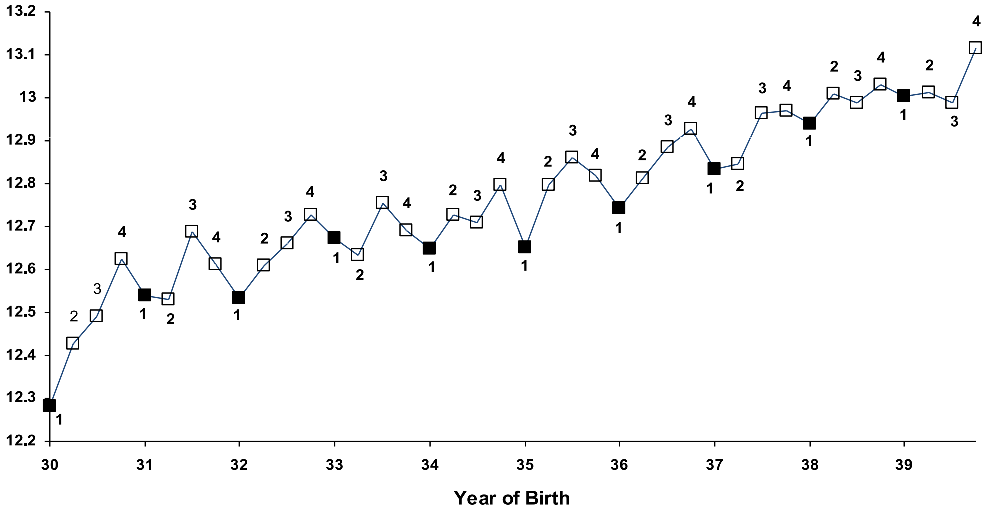
- In the US, <u>most states require students to enter school in the calendar year in which they turn 6</u>
- School start age is therefore a function of the date of birth:
- If born in January (December) → starts school in August at age 6 (5)
- Legal dropout rate: after 16
- This creates different education levels for those born in different dates
- Birth date arguably unrelated to ability!

***Food for thought for the workshop!***

# Searching for (Good) Instruments: Examples

## Example 2. Returns to education (in wage equations) (cont.)

A. Average Education by Quarter of Birth (first stage)




Source: Angrist & Krueger (1991)

**Criticisms:**

- Exogeneous variation is only created by a small subset of kids who leave school early – generalizable?
- **Is the instrument relevant enough?** (Recall the need to maximize the green & purple areas)
- **Family background** can be **associated with kids' quarter of birth and future earnings potential**: e.g., Incidence of schizophrenia depend on the season in which kids are born (!) [Bound, Jaeger, Baker, 1995]

# Searching for (Good) Instruments: Examples

## Example 3. Effect of children in income gender gap among inventors

- Sample of inventors: R&D workers with patents
- Gender gap still significant in this narrow and highly-skilled subsample of the population!
- Could **parenthood** account for (part of) the **income differential between male and female inventors**?
- But **having children is a choice** (often made jointly with labor market-related decisions ⇔ **simultaneity bias**?); **Unobserved factors** can also matter (*what if ambitious women choose to have fewer children in order to focus on their career, thereby possibly earning higher wages?)*

**Instruments for fertility (#children):**

- Abortion legislation (Bloom et al. 2009)
- Occurrence of twins at first birth (Angrist & Evans, 1998)
- <u>Whether individuals dedicate time to religious and spiritual activities in their leisure time</u> (Hoisl and Mariani, 2016)
- Why? Religious beliefs tend to be developed early and prior to determination of labor market outcomes and parenthood. They also affect one's opinion about birth control, contraception, abortion, and family plans.

Source: Hoisl & Mariani (2016), "Income and the gender gap in industrial research", *Management Science.*

# Yes, finding good IVs is challenging!

# Instrumental Variables in STATA

Requires installation! Type: **ssc install "command"**

Basic syntax is simple:

```
ivreg2 depvar [varlist1] (varlist2 = varlist_iv), first
```

***depvar*** is the dependent variable ($y$) – e.g., wages

***varlist1*** are the exogenous regressors of the structural equation – e.g., exper

***varlist2*** are the endogenous regressors that are being instrumented ($x$) – e.g., educ

***varlist_iv*** are the exogenous variables excluded from the structural regression (instruments $z$) – e.g., parents' education

Option `first` gives us the output of the reduced form equation.

Lots of options, and lots of outputs! Check **`help ivreg2`** for very complete guidance.

**`ivendog`** and **`overid`** post-estimation commands will be useful to run key tests.

# STATA example: Medical expenditures

"mus06.dta": these data will be used to study some determinants of expenditures on prescribed medications, using a sample of individuals older than 65 years old.

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| ldrugexp | 10391 | 6.479668 | 1.363395 | 0 | 10.18017 |
| hi_empunion | 10391 | .3796555 | .4853245 | 0 | 1 |
| totchr | 10391 | 1.860745 | 1.290131 | 0 | 9 |
| age | 10391 | 75.04639 | 6.69368 | 65 | 91 |
| female | 10391 | .5797325 | .4936256 | 0 | 1 |
| blhisp | 10391 | .1703397 | .3759491 | 0 | 1 |
| linc | 10089 | 2.743275 | .9131433 | -6.907755 | 5.744476 |

**hi_empunion**: dummy = 1 if the individual holds either employer or union-sponsored health insurance. It's a choice variable. Possibly endogenous! Why?

# Instrumenting Health Insurance

**Two potential instruments** for *hi_empunion*:

**ssiratio:** ratio of an individual's social security income to the overall income from all sources

**lowincome:** dummy = 1 if the individual is a low-income status person

Are they **relevant instruments**? Testable. →

Are they **valid instruments**? It cannot be tested. We must rely on <u>theoretical arguments</u> – i.e., we need to assume they can be omitted from the medical expenditures equation, since the direct role of income is already captured by the regressor *linc*. The only way these variables will affect expenditures is through "hi_empunion".

```
. pwcorr hi_empunion ssiratio lowincome, star (0.01)

             | hi_emp~n ssiratio lowinc~e
-------------+---------------------------
 hi_empunion |   1.0000
    ssiratio |  -0.1963*  1.0000
   lowincome |  -0.1144*  0.2498*  1.0000
```

```
. reg hi_empunion ssiratio lowincome totchr age female blhisp linc

      Source |       SS       df       MS              Number of obs =   10089
-------------+------------------------------           F(  7, 10081) =  122.58
       Model |  186.860924     7  26.6944176           Prob > F      =  0.0000
    Residual |  2195.38192 10081  .21777422            R-squared     =  0.0784
-------------+------------------------------           Adj R-squared =  0.0778
       Total |  2382.24284 10088  .236146197           Root MSE      =  .46666

------------------------------------------------------------------------------
 hi_empunion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ssiratio |  -.1787027    .014337   -12.46   0.000    -.2068061   -.1505993
   lowincome |  -.0633991   .0124133    -5.11   0.000    -.0877317   -.0390665
      totchr |   .0128507   .0036181     3.55   0.000     .0057586    .0199428
         age |  -.0085393   .0007124   -11.99   0.000    -.0099357   -.0071429
      female |  -.0729811   .0094819    -7.70   0.000    -.0915675   -.0543947
      blhisp |   -.062178   .0127532    -4.88   0.000    -.0871767   -.0371792
        linc |   .0448206   .0057127     7.85   0.000     .0336225    .0560186
       _cons |   1.036267   .0573559    18.07   0.000     .9238382    1.148696
------------------------------------------------------------------------------
```

*F-test confirms joint significance.*

# IV estimates (1 instrument): Manually vs. ivreg2

Recall the steps behind IV estimator:

1. Regress ***hi_empunion*** on ***ssiratio*** and <u>all the other exogenous variables in the structural equation</u> (gender, race, income, #chronical conditions)
2. Predict the estimated values of hi_empunion (`predict hi_hat, xb`)
3. Use these predicted values as a regressor in the structural equation (**hi_hat** instead of **hi_empunion**)

(**ivreg2** does it automatically for you) →

```
IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =    10089
                                                      F(  6, 10082) =   333.25
                                                      Prob > F      =   0.0000
Total (centered) SS     =  18715.11622                Centered R2   =   0.0640
Total (uncentered) SS   =  442534.2012                Uncentered R2 =   0.9604
Residual SS             =  17518.21658                Root MSE      =    1.318
```

| ldrugexp | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| hi_empunion | -.8975913 | .2211268 | -4.06 | 0.000 | -1.330992 | -.4641908 |
| totchr | .4502655 | .0101969 | 44.16 | 0.000 | .43028 | .470251 |
| age | -.0132176 | .0029977 | -4.41 | 0.000 | -.0190931 | -.0073421 |
| female | -.020406 | .0326114 | -0.63 | 0.531 | -.0843232 | .0435113 |
| blhisp | -.2174244 | .0394944 | -5.51 | 0.000 | -.294832 | -.1400167 |
| linc | .0870018 | .0226356 | 3.84 | 0.000 | .0426368 | .1313668 |
| _cons | 6.78717 | .2688453 | 25.25 | 0.000 | 6.260243 | 7.314097 |

*Supplementary-insured individuals have medical expenses that are 90% lower than those without insurance.*

# Test: Presence of Endogeneity

Recall slide 20

1. Run the reduced form equation for **hi_empunion** (**including the instrument(s)**) – predict the **residuals**
2. Include those residuals in the **second stage** and **test the significance of the coefficient**
3. **If statistically significant, endogeneity is present and must be solved**.

Other alternatives:

- Run **OLS and IV** and compare the estimates – **significantly different**? [Hausman test] If yes, endogeneity is likely present!
- Or simply **run ivreg2** and run the post-estimation command **`ivendog hi_empunion`**. H0 states that the regressor is exogeneous. If rejected, endogeneity is present.

# Test: Presence of Endogeneity

| Source | SS | df | MS |
|---|---|---|---|
| Model | 3350.66632 | 7 | 478.666618 |
| Residual | 15364.4499 | 10081 | 1.52409978 |
| Total | 18715.1162 | 10088 | 1.85518599 |

| | | |
|---|---|---|
| Number of obs | = | 10089 |
| F( 7, 10081) | = | 314.07 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.1790 |
| Adj R-squared | = | 0.1785 |
| Root MSE | = | 1.2345 |

| ldrugexp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| hi_res | .9891978 | .1965648 | 5.03 | 0.000 | .6038916 | 1.374504 |
| hi_empunion | -.8975913 | .1947955 | -4.61 | 0.000 | -1.279429 | -.5157533 |
| totchr | .4502655 | .0097613 | 46.13 | 0.000 | .4311315 | .4693996 |
| age | -.0132176 | .0026935 | -4.91 | 0.000 | -.0184973 | -.0079379 |
| female | -.020406 | .0295501 | -0.69 | 0.490 | -.0783301 | .0375181 |
| blhisp | -.2174244 | .0362335 | -6.00 | 0.000 | -.2884492 | -.1463995 |
| linc | .0870018 | .020625 | 4.22 | 0.000 | .0465728 | .1274308 |
| _cons | 6.78717 | .2393123 | 28.36 | 0.000 | 6.31807 | 7.25627 |

. hausman IV_1inst ols

| | Coefficients (b) IV_1inst | Coefficients (B) ols | (b-B) Difference | sqrt(diag(V_b-V_B)) S.E. |
|---|---|---|---|---|
| hi_empunion | -.8975913 | .0738788 | -.9714701 | .2062727 |
| totchr | .4502655 | .4403807 | .0098848 | .0041119 |
| age | -.0132176 | -.0035295 | -.0096881 | .0021698 |
| female | -.020406 | .0578055 | -.0782115 | .0190166 |
| blhisp | -.2174244 | -.1513068 | -.0661176 | .0187806 |
| linc | .0870018 | .0104815 | .0765202 | .0170289 |

b = consistent under Ho and Ha; obtained from ivreg2
B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 22.18
Prob>chi2 = 0.0011

. ivendog hi_empunion

Tests of endogeneity of: hi_empunion
H0: Regressor is exogenous

| | | | |
|---|---|---|---|
| Wu-Hausman F test: | 25.32531 | F(1,10081) | P-value = 0.00000 |
| Durbin-Wu-Hausman chi-sq test: | 25.28190 | Chi-sq(1) | P-value = 0.00000 |

*Endogeneity is present indeed!*

# Test: Overidentifying Restrictions

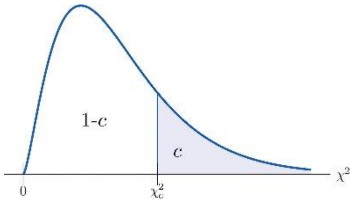
```
. reg uhat ssiratio lowincome totchr age female blhisp linc

      Source |       SS       df       MS              Number of obs =   10089
-------------+------------------------------           F(  7, 10081) =    0.96
       Model |  11.1820729     7  1.59743898           Prob > F      =  0.4587
    Residual |  16775.5778 10081  1.66407874           R-squared     =  0.0007
-------------+------------------------------           Adj R-squared = -0.0000
       Total |  16786.7599 10088   1.6640325           Root MSE      =    1.29

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ssiratio |   .0525874   .0396317     1.33   0.185    -.0250986    .1302734
   lowincome |  -.0832573   .0343141    -2.43   0.015    -.1505197   -.0159948
      totchr |  -.0004007   .0100014    -0.04   0.968    -.0200054     .019204
         age |  -.0001267   .0019692    -0.06   0.949    -.0039867    .0037332
      female |  -.0006952   .0262107    -0.03   0.979    -.0520734     .050683
      blhisp |  -.0003397   .0352535    -0.01   0.992    -.0694436    .0687641
        linc |    .000949   .0157916     0.06   0.952    -.0300056    .0319036
       _cons |  -.0044991   .1585483    -0.03   0.977    -.3152854    .3062872
------------------------------------------------------------------------------
```

```
.
. * obtain the test statistic:
. scalar overidtest = e(r2)*e(N)

. di overidtest
6.7205306
```




For c=0.05, the critical value is 3.84

**H0: All IVs are uncorrelated with *u* (the structural error).**

We reject H0 when using these 2 instruments ⇔ so at least one instrument is not valid.

From the ivreg2 output:

```
Sargan statistic (overidentification test of all instruments):     6.721
                                           Chi-sq(1) P-val =    0.0095
```

Post-estimation test:

```
. overid

Tests of overidentifying restrictions:
Sargan N*R-sq test          6.721  Chi-sq(1)     P-value = 0.0095
Basmann test                6.720  Chi-sq(1)     P-value = 0.0095
```

# One last time…

| | Heckman (sample selection) models | Matching Models (e.g., PSM) | Instrumental Variables |
|---|---|---|---|
| **When** | Y is missing in some cases | X is a binary intervention/choice | X is endogeneous (correlated with unobservables) |
| **Problem** | The missings in Y are driven by a "selection process" | T & C groups are very different | Four possible causes that make X correlated with the error term |
| **Stata commands** | *heckman, (twostep)* | *teffects psmatch, tebalance, teffects overlap* | *ivreg2, (first)* *ivendog, overid* |
| **Key tests** | Significance of the Inverse Mills Ratio or *rho* | Balancing and overlapping conditions | Relevance and validity of the instruments |
| **Attention!** | Need for valid exclusion restrictions; selection bias important when IMR/rho significant and X predicts selection | T & C only matched on observable characteristics. If unobservables matter **PSM does not provide causal effects → IV** | Validity only possible to assess ("imperfectly") when the model is overidentified; bad IVs are worse than OLS |
| **First stage** | Probit predicting selection into the sample (Y ≠ missing) | Probit predicting probability of being treated (X) | OLS predicting the endogenous variable (X) |



**General equation:** $y_i = \beta_0 + \beta_1 X1_i + \cdots + u_i$

# Good luck with your instruments! ☺