# ECON 2007: Quant Econ and Econometrics
# Instrumental Variables

Dr. Áureo de Paula

Department of Economics
University College London

# Practicalities

- **Office Hours**
  Thursdays, 16:00-18:00 (January), 14:30-15:30 (after January), or by appointment
  228 Drayton House
- **Required Textbook**
  Wooldridge: *Introductory Econometrics: A Modern Approach*
- **Supplementary (non-required) reading:**
  Stock and Watson: *Introduction to Econometrics*

# Practicalities

***The fight is won or lost far away from witnesses – behind the lines, in the gym, and out there on the road, long before I dance under those lights.*** *(Muhammad Ali)*

- **Homeworks (= the "gym"):**
  Three problem sets (. . . but there are many exercises in the book as well!).

- **Exam (= the "lights"):**
  3 hour written exam for entire module in Term 3.

# Practicals

- Practical sessions will take place on Mondays, 10:00-11:00.
- The practical sessions will be led by Gavin Kader.
- Practicals are scheduled for 30th January, 6th February, 20th February, 6th March and 20th March.

# Road Map

**Week 20 (13/01):** IV (W Ch.15.1-15.6, SW Ch.12) (Conway Hall)
**Week 21 (20/01):** IV (W Ch.15.1-15.6, SW Ch.12) (Royal Nat. Hotel Galleon Ste A)

**Week 22 (27/01):** Simultaneous Equations (W Ch.16.1-16.3) (IOE - Logan Hall)

**Week 23 (05/02):** LDV (W Ch.17; SW Ch.11) (IOE - Logan Hall)
**Week 24 (10/02):** LDV (W Ch.17; SW Ch.11) (IOE - Logan Hall)
**Week 26 (24/02):** LDV (W Ch.17; SW Ch.11) (IOE - Logan Hall)
**Week 27 (03/03):** LDV (W Ch.17; SW Ch.11) (IOE - Logan Hall)

**Week 28 (10/03):** Reg with TS (W Ch.10, 11.1-3; SW Ch.14, 15) (IOE - Logan Hall)
**Week 29 (17/03):** Serial Corr. and Heterosk. (W Ch.12; SW Ch.15) (IOE - Logan Hall)
**Week 30 (24/03):** Serial Corr. and Heterosk. (W Ch.12; SW Ch.15) and Selected Further Topics (W Ch. 16.4, Ch. 18; SW Ch.16) (Royal Nat. Hotel Galleon Suite A)

# Outline: Instrumental Variables (IV)

Part I: The basics of IV

- Motivation and basic idea
- IV assumptions and estimator
    - Example

Part II: Issues in IV estimation

- Inference and weak instruments
- 2SLS
    - Multiple explanatory variables
    - Multiple instruments
- Testing for endogeneity
- Overidentification

# Introduction

Remember the classical linear regression model:

$$y = \beta_0 + \beta_1 x + u \quad (\equiv \beta^\top \mathbf{x} + u \text{ using matrix algebra}).$$

For the OLS estimator to be consistent (i.e. to get "close" to $\beta$ as the sample increases) we assume

$$\text{cov}(x, u) = 0.$$

If $\text{cov}(x, u) \neq 0$ we say that $x$ is (econometrically) endogenous.

# What does $\text{cov}(x, u) \neq 0$ mean?

Let's think about this in the context of an example:

$$\ln(wage) = \beta_0 + \beta_1 educ + u.$$

We need to consider what factors are captured by $u$.

$cov(u, educ) = 0$ implies no (linear) association between *educ* and $u$, which may incorporate variables such as ability. But more educated individuals will tend to be more "able".

Remember that if $\text{cov}(educ, u) \neq 0$, OLS estimators for $\beta_0$ and $\beta_1$ are not consistent.

# When could $\text{cov}(x, u) \neq 0$ arise?

Common sources of endogeneity are:

- Ommited variables
- Measurement error (i.e. errors-in-variables)
- Simultaneity

# Small detour: unbiasedness $\neq$ consistency!

- Unbiasedness: small sample; Consistency: large sample.
- Conditions are different:

$$E(u|x) = 0 \quad \Rightarrow \quad cov(x, u) = 0$$

- It is **not** true that $cov(x, u) = 0$ implies $E(u|x) = 0$.
- Intuitively, these two conditions highlight that $x$ and $u$ should be unrelated. Technically, they are nevertheless different!
- If you confuse the two in the exam, you will be penalized.

# IV: The Basic Idea

OLS: Exogenous regressors

$$\begin{tikzcd}
x \arrow[r] & y \\
u \arrow[ur] &
\end{tikzcd}$$

OLS: Omitted variables bias:

$$\begin{tikzcd}
x \arrow[r] \arrow[d, leftrightarrow] & y \\
u \arrow[ur] &
\end{tikzcd}$$

IV: Suppose there exists a variable *z* such that

$$\begin{tikzcd}
z \arrow[r] & x \arrow[r] \arrow[d, leftrightarrow] & y \\
 & u \arrow[ur] &
\end{tikzcd}$$

# IV Assumptions

Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad (\text{2SLS.1})$$

where $cov(u, x) \neq 0$

Assume now that we have an *instrumental variable* $z_i$ that satisfies the following two conditions

- $cov(z_i, u_i) = 0$ (exogeneity or validity)
  $(+E(u_i) = 0 \Rightarrow \text{2SLS.4})$
- $cov(z_i, x_i) \neq 0$ (relevance) $\qquad (\Leftarrow \text{2SLS.3})$

The exogeneity assumption requires that:

- $z_i$ affects $y_i$ only through $x_i$
- $z_i$ is unrelated to $u_i$

and is not testable (since it involves $u_i$).

# IV Assumptions

Instrument relevance requires that $z_i$ affect $x_i$

- It can be verified by estimating the following regression

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

Since $\pi_1 = cov(z_i, x_i)/var(z_i)$, we can (and must!) test relevance:

- $H_0 : \pi_1 = 0$: instrument irrelevant
- $H_1 : \pi_1 \neq 0$: instrument relevant
  - As usual, perform t-test

# IV Estimator

Now we can identify $\beta_1$ by (i) noting that

$$cov(z_i, y_i) = \beta_1 cov(z_i, x_i) + cov(z_i, u_i)$$

and ii) using the IV assumptions to write

$$\beta_1 = \frac{cov(z_i, y_i)}{cov(z_i, x_i)} = \frac{cov(z_i, y_i)/var(z_i)}{cov(z_i, x_i)/var(z_i)}$$

which is the slope coefficient estimator from the *reduced form* divided by the slope coefficient estimator from the *first stage*.

Reduced form: $cov(z_i, y_i)/var(z_i)$

- slope coefficient from a regression of $y_i$ on $z_i$ and an intercept

First stage: $cov(z_i, x_i)/var(z_i)$

- slope coefficient from a regression of $x_i$ on $z_i$ and an intercept

# IV Estimator

The sample analog is the *instrumental variable estimator* of $\beta_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})}$$

The IV estimator of $\beta_0$ is:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Note that $\hat{\beta}_1$ is the OLS estimator of $\beta_1$ when $z_i = x_i$.

2SLS.1, 3-4 + Random Sampling (2SLS.2) $\Rightarrow$ 2SLS is consistent.

# Special Case of IV: Wald Estimator

A common (and simple) example of IV is one where the instrument is binary

$$z_i \in \{0, 1\}$$

Note that

$$\begin{aligned} E[y_i|z_i = 1] &= \beta_0 + \beta_1 E[x_i|z_i = 1] \\ E[y_i|z_i = 0] &= \beta_0 + \beta_1 E[x_i|z_i = 0] \end{aligned}$$

so that

$$E[y_i|z_i = 1] - E[y_i|z_i = 0] = \beta_1(E[x_i|z_i = 1] - E[x_i|z_i = 0])$$

which after rearranging and taking sample analogues gives the Wald estimator:

$$\hat{\beta}_1 = \frac{\overline{y}_{(z=1)} - \overline{y}_{(z=0)}}{\overline{x}_{(z=1)} - \overline{x}_{(z=0)}}$$

where $\overline{x}_{(z=1)} \equiv (|\{i : z_i = 1\}|)^{-1} \sum_{\{i:z_i=1\}} x_i$, etc.

# Example

Consider our example of returns to schooling

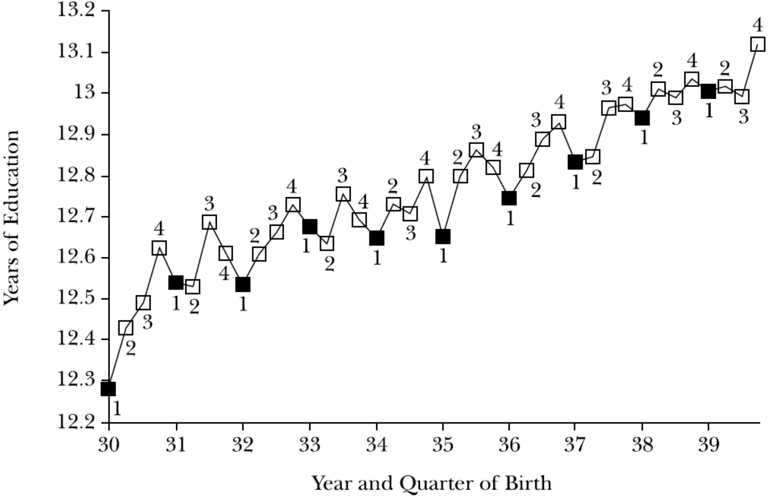
$$ln(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

Angrist and Krueger (1991) came up with an instrument for education in the US: quarter of birth.

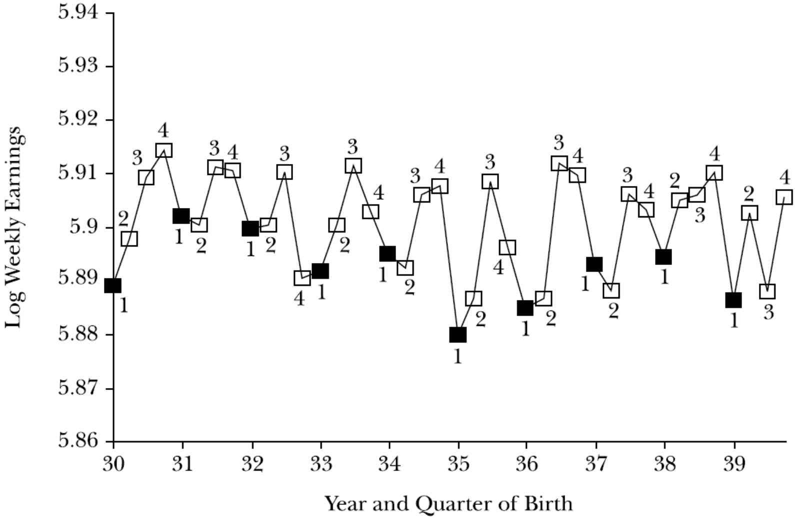
Arguments for instrument:

- Education is compulsory by law until your 16th birthday
- School start in the year you turn 6:
  - children born early in the year begin school at an older age
  - and may therefore leave school with somewhat less education
- Quarter of birth is arguably uncorrelated with unobservables affecting wages (though see Bound, Jaeger and Baker (1995) for arguments otherwise).

# Mean Years of Completed Education, by Quarter of Birth



# Mean Log Weekly Earnings, by Quarter of Birth



# OLS and IV estimates

| | Quarter of birth | | Difference |
|---|---|---|---|
| | 1st | 4th | (2)-(1) |
| | (1) | (2) | (3) |
| ln(weekly wage) | 5.892 | 5.905 | 0.0135 |
| | | | (0.0034) |
| Years of education | 12.688 | 12.839 | 0.151 |
| | | | (0.016) |
| Wald estimate of<br>return to education | | | 0.089<br>(0.021) |
| OLS estimate of<br>return to education | | | 0.070<br>(0.0005) |

# Inference

Assuming homoscedasticity $E[u_i^2|z_i] = \sigma^2$, it can be shown that the estimated (asymptotic) variance of the IV estimator is:

$$\widehat{var}(\hat{\beta}_{IV}) = \frac{\hat{\sigma}^2}{SST_x R^2_{x,z}}$$

- $SST_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$
- $R^2_{x,z}$: the R-squared from a regression of $x_i$ on $z_i$ and an intercept
- $\hat{\sigma}^2 = (n-2)^{-1}\sum_{i=1}^{n}\hat{u}_i^2$ where $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

As before, we compute the standard error and use it to perform tests and derive confidence intervals, but we might need large sample sizes.

# IV vs. OLS

Advantage of IV estimator: Consistent even if $u$ and $x$ are correlated, in which case the OLS estimator is biased and inconsistent.

Disadvantage of IV estimator: less efficient if $u$ and $x$ are uncorrelated.

Assume that $u$ and $x$ are uncorrelated. Then:

$$\widehat{var}(\hat{\beta}_{IV}) = \frac{\hat{\sigma}^2}{SST_x R^2_{x,z}} > \frac{\hat{\sigma}^2}{SST_x} = \widehat{var}(\hat{\beta}_{OLS})$$

and we can see the variance of the IV estimator

- is always larger than the variance of the OLS estimator and
- depends crucially on the correlation between $z$ and $x$.

# Weak Instruments and Bias

Weak instrument means that $z$ and $x$ are only weakly correlated:

- not only lead to imprecise IV estimates,
- but can also give large bias.

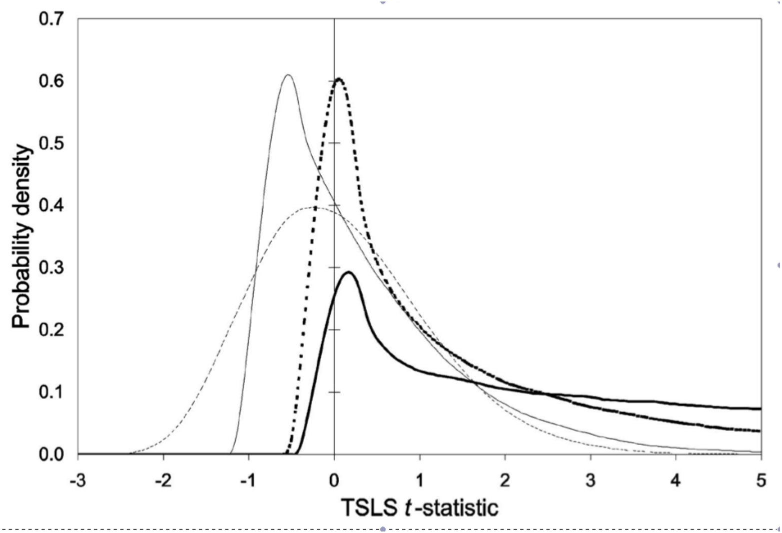
Recall that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} = \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \overline{z})(u_i - \overline{u})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})}$$

Even if $cov(z, u)$ is smaller than $cov(x, u)$, it is not necessarily better when the denominator is small.

Even if $cov(z, u) = 0$, $\hat{\beta}_1$ becomes very unstable and unreliable if $z$ and $x$ are only weakly correlated, so that the denominator is close to zero

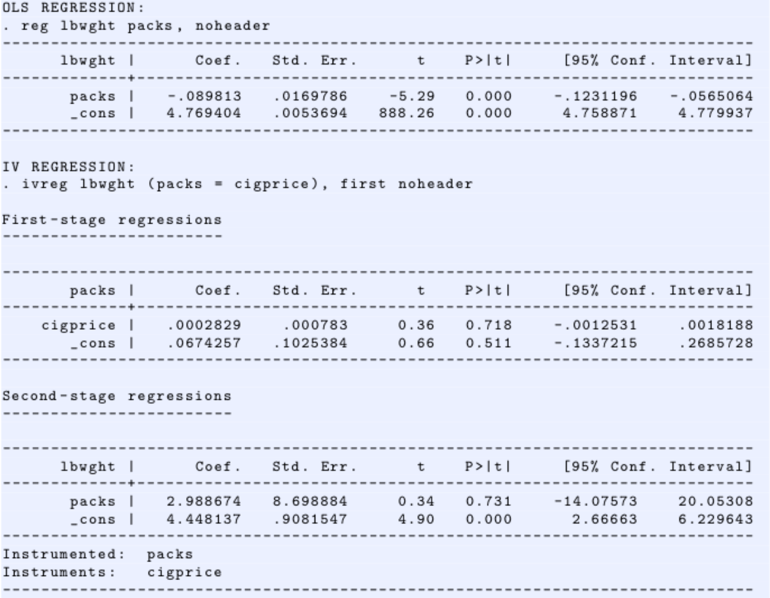
# Weak Instruments and Bias



# Weak Instruments and Bias

- If the IVs are weak, the sampling distribution of the TSLS estimator (and its *t*-statistic) is not well approximated by its large *n* normal approximation.
- IV estimation thus requires fairly strong instruments
- Rule of thumb: F-statistic above 10 (same as t-statistic above $\sqrt{10}$) for the instrument in the first stage

# Example - Weak instrument: Birth weight and smoking



```
OLS REGRESSION:
. reg lbwght packs, noheader
------------------------------------------------------------------------------
      lbwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       packs |   -.089813   .0169786    -5.29   0.000    -.1231196   -.0565064
       _cons |   4.769404   .0053694   888.26   0.000     4.758871    4.779937
------------------------------------------------------------------------------

IV REGRESSION:
. ivreg lbwght (packs = cigprice), first noheader

First-stage regressions
-----------------------


------------------------------------------------------------------------------
       packs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cigprice |   .0002829    .000783     0.36   0.718    -.0012531    .0018188
       _cons |   .0674257   .1025384     0.66   0.511    -.1337215    .2685728
------------------------------------------------------------------------------

Second-stage regressions
------------------------


------------------------------------------------------------------------------
      lbwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       packs |   2.988674   8.698884     0.34   0.731    -14.07573    20.05308
       _cons |   4.448137   .9081547     4.90   0.000      2.66663    6.229643
------------------------------------------------------------------------------
Instrumented:  packs
Instruments:   cigprice
------------------------------------------------------------------------------
```

# IV in the MLR model

We can add an additional explanatory variables $x_2$ to the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

Assume that $x_2$ is uncorrelated with $u$, while $x_1$ is correlated with $u$

- $x_{i2}$: exogenous explanatory variable
- $x_{i1}$: endogenous explanatory variable

# IV in the MLR model

To consistently estimate all the $\beta$'s we use the sample analogues of the moment conditions

$$\mathbb{E}(u_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

$$cov(u_i, z_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) z_{i1} = 0$$

$$cov(u_i, x_{i2}) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) x_{i2} = 0$$

3 equations with 3 unknowns: can be solved as we did for OLS

# IV in the MLR model

As before we need $z_1$ to be correlated with $x_1$, but now over and above $x_2$.

We can test this by estimating the following regression

$$x_{i1} = \pi_0 + \pi_1 z_{i1} + \pi_2 x_{i2} + v_i$$

and instrument relevance is tested as:

$$H_0 : \pi_1 = 0 \text{ vs } H_1 : \pi_1 \neq 0$$

In other words, IV in the MLR model is just as IV in the SLR model except the exogeneity assumption is now:

$$cov(z_i, u_i | \, x_{i2}) = 0$$

(Note that $z_i$ and $x_{i2}$ can be correlated!)

# Two-stage least squares (2SLS)

We still consider the following model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

but now with $M > 1$

$$cov(u_i, z_{im} |\, x_{i2}) = 0 \quad m = 1, \ldots, M$$

We only need to modify the first stage such that:

$$x_{i1} = \pi_0 + \pi_1 z_{i1} + \ldots + \pi_M z_{iM} + \pi_{M+1} x_{i2} + v_i$$

Instrument relevance is tested using an $F$ statistic for

$$H_0 : \pi_1 = \ldots = \pi_M = 0$$

# 2SLS: Step-by-step

1. Estimate the *first-stage* regression:

$$x_{i1} = \pi_0 + \pi_1 z_{i1} + \ldots + \pi_M z_{iM} + \pi_{M+1} x_{i2} + v_i$$

   - regressing the endogenous explanatory variable on the instruments and *all* the other exogenous explanatory variable

2. Compute the predicted value of $x_1$:

$$\hat{x}_{i1} = \hat{\pi}_0 + \hat{\pi}_1 z_{i1} + \ldots + \hat{\pi}_M z_{iM} + \hat{\pi}_{M+1} x_{i2}$$

3. Estimate the *second-stage* regression:

$$y_i = \beta_0 + \beta_1 \hat{x}_{i1} + \beta_2 x_{i2} + e_i$$

   - regressing the outcome variable on $\hat{x}_{i1}$ and *all* the other exogenous explanatory variable

Note that using more than one instrument is not necessary, but it can give you more efficient IV estimates.

# Multiple Endogenous Variables

If there is more than one endogenous variable, e.g.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + u_1$$

we need at least two exogenous variables $z_2$ and $z_3$ that do not appear in the equation above. (This is known as the order condition.)

Both of these variables still need to be relevant though. If $z_2$ does not correlate with either endogenous variable or if both $z_2$ and $z_3$ correlate with only one of the endogenous variables we would not be able to identify the desired parameters. The sufficient condition for identification is called the rank condition. (We will revisit that when we discuss simultaneous equations.)

# Weak IV Revisited

- Weak IV are also a problem with many instruments.
- Adding instruments with low predictive power in the first stage lowers the $F$-statistic and exacerbates the bias in the 2SLS estimator.
- Bound, Jaeger, and Baker (1995) illustrate this using the Angrist and Krueger (1991). AK present results using different sets of IVs (plus other covariates):
  - quarter of birth dummies: $M = 3$ instruments.
  - quarter of birth + (quarter of birth) *x* (year of birth) dummies: $M = 30$ instruments.
  - quarter of birth + (quarter of birth) *x* (year of birth) + (quarter of birth) *x* (state of birth): $M = 180$ instruments.

# Weak IV Revisited

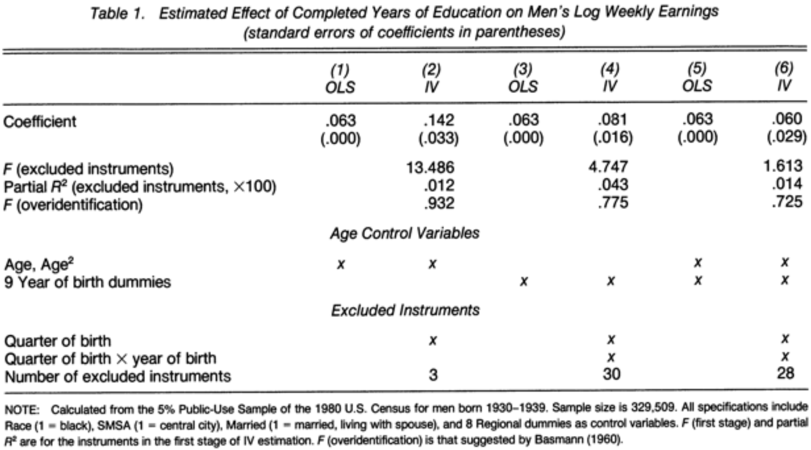


*Table 1. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings (standard errors of coefficients in parentheses)*

| | (1)<br>OLS | (2)<br>IV | (3)<br>OLS | (4)<br>IV | (5)<br>OLS | (6)<br>IV |
|---|---|---|---|---|---|---|
| Coefficient | .063<br>(.000) | .142<br>(.033) | .063<br>(.000) | .081<br>(.016) | .063<br>(.000) | .060<br>(.029) |
| *F* (excluded instruments) | | 13.486 | | 4.747 | | 1.613 |
| Partial $R^2$ (excluded instruments, ×100) | | .012 | | .043 | | .014 |
| *F* (overidentification) | | .932 | | .775 | | .725 |
| | | *Age Control Variables* | | | | |
| Age, $\text{Age}^2$ | x | x | | | x | x |
| 9 Year of birth dummies | | | x | x | x | x |
| | | *Excluded Instruments* | | | | |
| Quarter of birth | | x | | x | | x |
| Quarter of birth × year of birth | | | | x | | x |
| Number of excluded instruments | | 3 | | 30 | | 28 |

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. *F* (first stage) and partial $R^2$ are for the instruments in the first stage of IV estimation. *F* (overidentification) is that suggested by Basmann (1960).

# Weak IV Revisited

With more than one endogenous variable the $F$-statistic in the first stage may not suffice for detection of weak IVs either.

One alternative is to perform a test based on the Cragg-Donald Eigenvalue statistic (see Stock, Wright and Yogo (2002)).

(Alternatively, see Sanderson and Windmeijer, CeMMAP Working Paper CWP58/13.)

# Testing for endogeneity: Hausman test

Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i$$

Test for endogeneity:

- $H_0 : cov(x_{i1}, u) = 0$, both OLS and IV are consistent
- $H_1 : cov(x_{i1}, u) \neq 0$, only IV is consistent and $x_{i1}$

We perform a Hausman test by

1. Calculating the first-stage residual $\hat{v}_i$ (this contains the endogenous part of $x_{i1}$)

$$x_{i1} - \hat{x}_{i1} = \hat{v}_i$$

2. Adding $\hat{v}_i$ to the regression model, and estimate by OLS:

$$y_i \;=\; \beta_0 + \beta_1 x_{i1} + \theta \hat{v}_i + e_i$$

3. Using a t-test to check if $\theta$ is significantly different from zero $\Rightarrow$ reject $H_0 : cov(x_{i1}, u_i) = 0$

# Testing Overidentification Restrictions

- When there are as many IVs as endogenous variables, exogeneity of the instruments is not testable.
- However, when there are more IVs than endogenous variables, we can test whether some of them are uncorrelated with the *u*.
- With two IVs and one endogenous variable, for example, we could compute alternative 2SLS estimates using each of the IVs. If the IVs are both exogenous, the 2SLS will converge to the same parameter and they will differ only by sampling error.

# Testing Overidentification Restrictions

- If the two estimates are statistically different, we would not be able to reject the hypothesis that at least one of the IVs is invalid.
- But we would not be able to ascertain which one!
- Moreover, if they are similar and pass the test it could because both IVs fail the exogeneity requirement.

# Testing Overidentification Restrictions

In practice, *under homoskedasticity* (i.e., $E(u^2|\mathbf{z}) = \sigma^2$),

1. Estimate coefficients by 2SLS and obtain residuals $\hat{u}_i$.
2. Regress $\hat{u}_i$ on *all exogenous* variables. Record the $R^2$.
3. Under the null hypothesis that all IVs are exogenous, $nR^2 \sim \chi^2_{M-1}$.

This test can be made robust to heteroskedasticity.

These slides covered:

Wooldridge 15.1-15.6, Stock and Watson 10.