Applied Microeconometrics (L4): Instrumental Variables Regression 2

Nicholas Giannakopoulos

University of Patras
Department of Economics

ngias@upatras.gr

November 8, 2015

Overview

Potential outcomes: wages (w) and schooling (s)

$$w_{si} \equiv f_i(s),\ i = 1, ..., N$$
$$f_i(s) = \pi_0 + \pi_1 s + \eta_i$$

- Control variables: A_i="Ability" observed variables
  - $\eta_i = A_i'\gamma + \upsilon i$
  - γ population regression coefficients
  - If A_i is the only reason why η_i and υi are correlated, then $E[S_i\upsilon_i] = 0$
- If A_i were observed then,
  - $w_i = \alpha + \rho S_i + A_i'\gamma + \upsilon_i$
  - Assumption: error term is uncorrelated with schooling
  - If the assumption is correct then we get $\hat{\alpha}$, $\hat{\rho}$ and $\hat{\gamma}$
- BUT, when A_i is unobserved then we need and instrument Z_i
  - Z_i is correlated with S_i but is uncorrelated with any other determinants of the dependent variable, i.e., $Cov(\eta_i, Z_i) = 0$ (Exclusion Restriction)

IV and Causality

- Regression line

$$w_i = \alpha + \rho S_i + A_i'\gamma + \upsilon_i$$

- Given the exclusion restriction

$$\rho = \frac{Cov(w_i, Z_i)}{Cov(S_i, Z_i)} = \frac{Cov(w_i, Z_i)/Var(Z_i)}{Cov(S_i, Z_i)/Var(Z_i)}$$

- Coefficient of interest ρ: ratio of the population regression of w_i on Z_i (*reduced form*) to the population regression of S_i on Z_i (*first stage*)
    1. the instrument Z_i must have a clear effect on S_i
    2. Z_i affects w_i only through S_i
    3. instrument is as good as randomly assigned
    4. the instrument has no effect on outcomes other than through the first-stage channel

IV and Causality

Good instruments come from a combination of three ingredients:

- Good institutional knowledge
- Economic theory
- Original ideas

Usual sources of instruments

- Nature
- Policies
- Choice variables of the agent do not serve as good instruments (e.g., lagged variables as instruments, parental socioeconomic characteristics)

Examples

Returns to schooling

- Quarter of birth (Angrist and Krueger, QJE 1991)
- Laws of compulsory education (Bjrklund et al, QJE 2006)

The effect of family size on children's education

- Twins, gender of the first born, gender of the two first born (Black et al, QJE 2005)

The effect of family size on mother's labour supply in Greece

- gender of the two first born (Daouli et al, EL 2009)

Two-stage Least Squares (2SLS)

- In a model with a single endogenous variable and a single instrument, IV estimates are equivalent to a two stage procedure
- Causal model with covariates

$$w_i = X_i'\alpha + \rho S_i + \eta_i$$

- First-stage equation

$$S_i = X_i'\pi_{10} + \pi_{11} Z_i + \epsilon_{1i} = \hat{S}_i + \epsilon_{1i}$$

- Second-stage equation

$$w_i = X_i'\alpha + \rho \hat{S}_i + [\eta_i + \rho \epsilon_{1i}]$$

- Estimate by OLS

Correct standard errors

With the *manual* two stage procedure, you do not get the *right* standard errors

- The residual that is used to calculate standard errors in second-stage includes an extra error: $w_i - [X_i'\alpha + \rho\hat{S}_i] = [\eta_i + \rho\epsilon_{1i}]$
- remember that $\hat{S}_i$ is a generated regressor and inflates the variance
- Stata ivreg or ivreg2 fixes it by using the original endogenous regressor to construct residuals: $w_i - [X_i'\alpha + \rho S_i] = \eta_i$

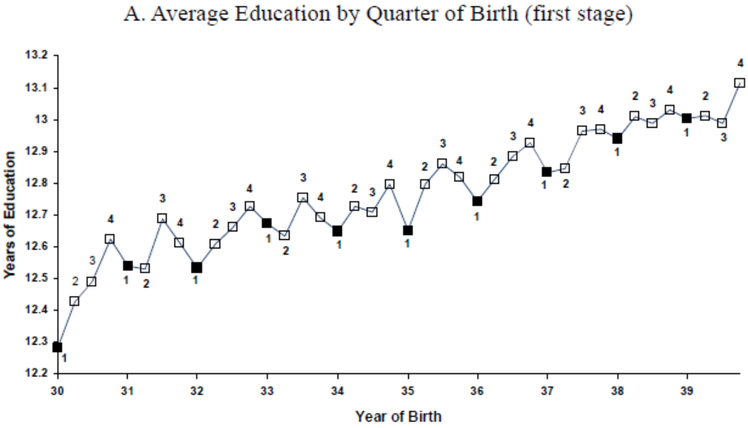
Compulsory School Law, Schooling and Earnings

"Does Compulsory School Attendance Affect Schooling and Earnings" (Angrist and Krueger, QJE 1991)

- quarter of birth as an instrument for schooling
- students enter school in the calendar year in which they turn 6
- compulsory school law requires them to remain in school until they become 16
- people born late in the year are more likely to stay at school longer

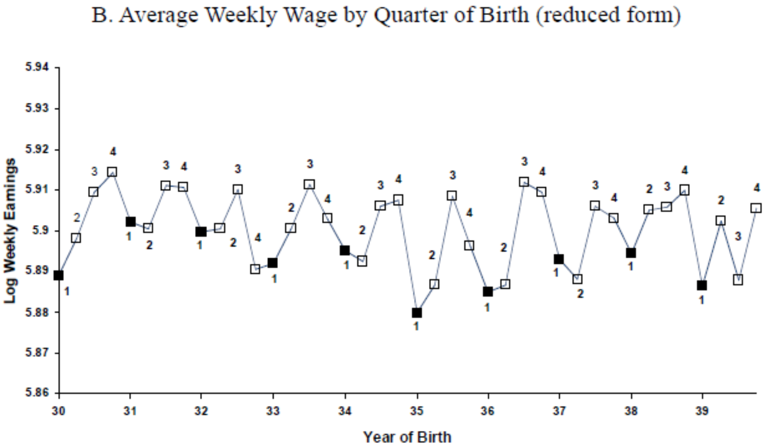
$$Y_i = \alpha X_i' + \rho S_i + \eta_i$$

Compulsory School Law, Schooling and Earnings



Source: Angrist, Joshua D., and Alan B. Krueger (1991): "Does Compulsory Schooling Attendance Affect Schooling and Earnings?" Quarterly Journal of Economics, 106, 976-1014

Compulsory School Law, Schooling and Earnings



Source: Angrist, Joshua D., and Alan B. Krueger (1991): "Does Compulsory Schooling Attendance Affect Schooling and Earnings?" Quarterly Journal of Economics, 106, 976-1014

Compulsory School Law, Schooling and Earnings

What about the exclusion restriction? Is the only reason for the up-and down quarter of birth pattern in earnings the up-and down quarter of birth pattern in schooling?

- Omitted variable background
- Other channels

Compulsory School Law, Schooling and Earnings

Multiple instruments

- we have three instrumental variables: Z_{1i}, Z_{2i} and Z_{3i}
- Angrist and Krueger (1991): dummies for first, second, and third-quarter births
- first-stage

$$S_i = X_i'\pi_{10} + \pi_{11}Z_{1i} + \pi_{12}Z_{2i} + \pi_{13}Z_{3i} + \xi_{1i}$$

- all of the quarter of birth dummies are uncorrelated with η_i in the basic model

Compulsory School Law, Schooling and Earnings

Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.075 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

Source: Angrist Joshua D. and Steffen Pischke. (2009) Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

The Wald estimator

The simplest IV estimator uses a single binary (0-1) instrument Z_i to estimate a model with one endogenous regressor and no covariates

$$y_i = \alpha + \rho S_i + \eta_i$$

- if Z_i equals 1 with probability p, the IV estimator is:
  $$\rho = \frac{Cov(y_i, Z_i)}{Cov(S_i, Z_i)} = \frac{E[y_i|Z_i=1] - E[y_i|Z_i=0]}{E[s_i|Z_i=1] - E[s_i|Z_i=0]}$$
- given that $E[\eta_i|Z_i] = 0$, we get
- $E[y_i|Z_i] = \alpha + \rho E[S_i|Z_i]$ and
- solving for ρ we have the Wald Estimator
- Thus, in the context of a binary instrument, it seems natural to divide the reduced-form difference in means by the corresponding first-stage difference in means

Compulsory School Law, Schooling and Earnings

Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
ln (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

Source: Angrist Joshua D. and Steffen Pischke. (2009) Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Sibling sex composition, employment and fertility in Greece

Table 1
Descriptive statistics for Greek married mothers, aged 21–35 with two or more children

Variables	Label	1991 Census	2001 Census
Children ever born	–	2.29 (.59)	2.28 (.61)
First two children were boys (0/1)	–	.27 (.44)	.27 (.44)
First two children were girls (0/1)	–	.23 (.42)	.23 (.42)
First child was a boy (0/1)	Boy1st	.52 (.49)	.52 (.49)
Second child was a boy (0/1)	Boy2nd	.52 (.49)	.51 (.49)
First two children are of the same sex (0/1)	Samesex	.50 (.49)	.50 (.49)
Mother had more than two children (0/1)	Fertility	.23 (.42)	.21 (.41)
Worked for pay (0/1)	Employment	.25 (.43)	.38 (.48)
Mothers' age	Age	30.51 (3.40)	31.43 (3.03)
Age of mother at first birth	Age at 1st birth	21.37 (3.22)	21.46 (3.53)
Foreign born (0/1)	Foreign	.01 (.11)	.18 (.39)
Number of Observations		28271	18604

Source: IPUMS-International. Standard deviations in parentheses.

Source: Daouli, J., M. Demoussis and N. Giannakopoulos. (2009) Sibling-Sex Composition and its Effects on Fertility and Labour Supply of Greek Mothers. Economics Letters, 102(3):189-191

Sibling sex composition, employment and fertility in Greece

For estimating purposes, we adopt the AE conventional approach which estimates the following two-equation system describing employment (y) and fertility (m):

$$y_i = \mathbf{X}_i \boldsymbol{\gamma} + \beta m_i + u_i \tag{1}$$

$$m_i = \mathbf{Z}_i \boldsymbol{\alpha} + e_i \tag{2}$$

where, $\mathbf{X}$ and $\mathbf{Z}$ are vectors of observed characteristics with $E(\mathbf{X}_i, u_i) = E(\mathbf{Z}_i, e_i) = 0$. The coefficient of the fertility variable in the employment equation (β) estimates the average change in the employment probability with regard to increased fertility (more than two children). The adopted instrument $z, z \quad \mathbf{Z} \forall i$ is a combined indicator with regard to the sex of the higher order first two born children (*samesex*in AE)[3] which takes the following form:

$$z_i = b_{1i} \cdot b_{2i} + (1 - b_{1i}) \cdot (1 - b_{2i}) = (2b_{2i} - 1) b_{1i} + (1 - b_{2i}) \tag{3}$$

where, b_1 and b_2 are indicators for boy-first and boy-second born children, respectively, for the i^{th} mother. For identification purposes, a Wald-type estimate (β_{Wald}) is derived, based on the calculation of the average effect of fertility on labor supply, for those women whose fertility has been affected by the adopted instrument, (i.e., *samesex*).

Source: Daouli, J., M. Demoussis and N. Giannakopoulos. (2009) Sibling-Sex Composition and its Effects on Fertility and Labour Supply of Greek Mothers. Economics Letters, 102(3):189-191

Sibling sex composition, employment and fertility in Greece

- Wald-type estimates, for 1991 is equal to 0.161 (0.010/0.062) with a standard error of 0.082, while for 2001 is equal to 0.093 with a standard error of 0.153.
- These estimates imply that married mothers in 1991 with more than two children exhibit reduced (by 16 percentage points) employment rates as a result of exogenous variations in family size.
- Accordingly, this reduction comes to almost 10 percentage points in 2001, even though the effect does not differ statistically from zero.

Source: Daouli, J., M. Demoussis and N. Giannakopoulos. (2009) Sibling-Sex Composition and its Effects on Fertility and Labour Supply of Greek Mothers. Economics Letters, 102(3):189-191

Sibling sex composition, employment and fertility in Greece

Table 2
The effects of fertility on employment outcomes of Greek married mothers aged 21–35 with two or more children (OLS and 2SLS-IV)

	1991 Census			2001 Census		
	(OLS)	(2SLS-IV)		(OLS)	(2SLS-IV)	
	Employment	Fertility	Employment	Employment	Fertility	Employment
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	−.554	.342	−.535	−.636	.347	−.638
	(.024)	(.023)	(.038)	(.037)	(.032)	(.067)
Boy1st	−.017	−.019	−.018	.004	−.021	.004
	(.005)	(.004)	(.005)	(.007)	(.006)	(.007)
Boy2nd	−.008	−.019	−.009	.002	.001	.002
	(.005)	(.004)	(.005)	(.007)	(.006)	(.007)
Age	.013	.020	.014	.020	.014	.020
	(.001)	(.001)	(.002)	(.001)	(.001)	(.002)
Age at 1st birth	.020	−.033	.018	.019	−.027	.019
	(.001)	(.001)	(.003)	(.001)	(.001)	(.004)
Foreign	−.079	.032	−.077	.026	−.042	.026
	(.022)	(.021)	(.022)	(.009)	(.007)	(.010)
Fertility	−.083	–	−.136	−.105	–	−.100
	(.006)		(.079)	(.008)		(.150)
Samesex	–	.063	–	–	.045	–
		(.004)			(.005)	
F-test	287.32	168.05	255.44	190.30	61.22	165.52
Partial-R^2	–	0.0059	–	–	0.0033	–
R^2	0.0575	–	0.0550	0.0578	–	0.0578
DWH-X^2-value		0.446			0.001	
Observations		28271			18604	

Source: IPUMS-International. Standard errors in parentheses.
Note: The model was also estimated pooling data from the two censuses. The obtained OLS coefficient estimate of the effect of "fertility" on "employment" is −.092 with a standard error of .005. The first stage estimate of the 2SLS-IV regarding the effect of the "samesex" on "fertility" is .056 with a standard error of .004, while the second stage estimate of the effect of the instrumented "fertility" on "employment" is equal to −.125 with a standard error of .072.

Source: Daouli, J., M. Demoussis and N. Giannakopoulos. (2009) Sibling-Sex Composition and its Effects on Fertility and Labour Supply of Greek Mothers. Economics Letters, 102(3):189-191

Local Average Treatment Effects: LATE

- With heterogenous treatment effects, endogeneity creates severe problems for identification of population averages. Population average causal effects are only estimable under very strong assumptions on the effect of the instrument on the endogenous regressor ("identification at infinity, or under the constant treatment effect assumptions). Without such assumptions we can only identify average effects for subpopulations that are induced by the instrument to change the value of the endogenous regressors. We refer to such subpopulations as compliers, and to the average treatment effect that is point identifed as the local average treatment effect.

Local Average Treatment Effects: LATE

$$Y_i = \alpha_0 + \rho_1 D_i + \eta_i$$

- Where D_i is a binary endogenous treatment variable
- Outcome in the absence of treatment is $Y_{0i} = \alpha_0 + \eta_i$
- The causal effect of treatment for individual i is $Y_{1i} - Y_{0i}$

Local Average Treatment Effects: LATE

- Constant effects model is an excellent starting point
- What if $Y_{1i} - Y_{0i}$ is not the same for every i?
- Examples: cancer treatment, foster care...

Local Average Treatment Effects: LATE

- In a design based heterogenous world, we recognize the difference between internal validity and external validity
- A good instrument captures an internally valid causal effect. This is the causal effect of group subject to (quasi) experimental manipulation (i.e. affected by the instrument)
- The external validity of this estimate is its predicted value in populations other than the one for which the experiment is observed

Local Average Treatment Effects: LATE

- What does IV estimate if $Y_{1i} - Y_{0i}$ is not the same for everyone?
- LATE = Local Average Treatment Effect
- Let $Y_i(d, z)$ denote the potential outcome for individual i whose treatment status is $D_i = d$ and instrument value $Z_i = z$
- We assume causal chain: instrument (Z_i) affects treatment (D_i) which in turn affects outcome (Y_i).

Local Average Treatment Effects: LATE

- D_{1i} is treatment status when $Z_i = 1$
- D_{0i} is treatment status when $Z_i = 0$
- Observed treatment status is
$$D_i = D_{0i} + (D_{1i} - D_{0i})Z_i$$
- For all i we have
- Potential outcomes: $Y_i(0,0)$, $Y_i(1,0)$, $Y_i(0,1)$, $Y_i(1,1)$
- Potential treatments: $D_{0i} = 0$,$D_{0i} = 1$,$D_{1i} = 0$,$D_{1i} = 1$
- Potential assignments: $Z_i = 0$, $Z_i = 1$

Local Average Treatment Effects: LATE

Classification of individuals according to treatment and assignment

		$Z_i = 0$	
		D_{0i}=0	D_{0i}=1
$Z_i = 1$	D_{1i}=0	Never-taker	Defier
	D_{1i}=1	Complier	Always taker

Local Average Treatment Effects: LATE

LATE assumptions

1. Independence: instrument is as good as randomly designed
2. Exclusion Restriction: affects outcome through single know channel
3. First Stage: $E[D_{1i} - D_{0i}] \neq 0$
4. Monotonicity: $D_{1i} \geq D_{0i}$ for everyone (or vice versa). All those who are affected are affected in the same way.

The last one is a necessary technical assumptions that is needed for IV to have LATE interpretation

Local Average Treatment Effects: LATE

LATE

- If the LATE assumptions hold
$$\rho = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]$$
- The IV estimates the impact of treatment for those whose behavior changed because of the instrument

Local Average Treatment Effects: LATE

Why do we need a monotonicity condition in model with heterogenous treatment effects?

- A failure of monotonicity means that the instrument pushes some people into treatment, while pushing others out

$$\begin{aligned} E(Y_i|Z_i=1)-E(Y_i|Z_i=0) &= E[(Y_{i1}-Y_{i0})(D_{i1}-D_{i0})] \\ &= E[Y_{i1}-Y_{i0}|D_{i1}>D_{i0}]P[D_{i1}>D_{i0}] \\ &\quad -E[Y_{i1}-Y_{i0}|D_{i1}<D_{i0}]P[D_{i1}<D_{i0}] \end{aligned}$$

- It may be that treatment effects are positive but the reduced form is zero since the effects on compliers are cancelled out by effects on defiers
- This does not come up in constant effects models (reduced form is always constant effect times the first stage)

IV, 2SLS and GMM

- Just-identified case: The number of instruments exactly equals to the number of regressors, $(\hat{\beta}_{IV})$ is the IV estimator
- Not-identified case (under-identified): fewer instruments than regressors, $(\hat{\beta}_{IV})$ is not consistent
- Over-identified case: more instruments than regressors, $(\hat{\beta}_{2SLS})$ is the an efficient estimator. In the just-defined case $(\hat{\beta}_{IV}=\hat{\beta}_{2SLS})$
- General estimator: Generalized Methods of Moments (GMM) Estimator

IV, 2SLS and GMM

- Starting point: Instrument is correlated with the regressor and is uncorrelated with the disturbance term (conditional moment restriction)
- The conditional moment restriction can be tested (in the case of over-identification)
- The stronger the association between the instrument and the regressor the stronger the identification
- When the instrument is weak the estimation becomes less precise and s.e. become larger, thus t-statistic is smaller (than in OLS).
- But even is the IV estimators are consistent, they may provide very poor approximation to the actual sampling distribution in typical finite-sample sizes.
- More instruments implies larger small-sample bias.

Specification Tests

Specification Tests

Testing for Endogeneity - Wu-Hausman Test

- Since OLS is preferred to IV (or TSLS) if we do not have an endogeneity problem, we'd like to be able to test for endogeneity
- If we do not have endogeneity, both OLS and IV are consistent, but IV is inefficient
- Idea of Hausman test is to see if the estimates from OLS and IV are different
- Auxilliary regression is easiest way to do this test

- Consider the following regression:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \varepsilon_i$$
- With Z_{1i} and Z_{2i} as additional exogenous variables (i.e. additional instruments)
- If $X_1$is uncorrelated with Y we should estimate this equation by OLS
- Hausman (1978) suggested comparing the OLS and TSLS estimates and determining whether the differences are significant. If they differ significantly, we conclude that X_1 is an endogenous variable.
- This can be achieved by estimating the first stage regression:
$$X_{1i} = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \alpha_3 W_{1i} + \alpha_4 W_{2i} + v_i$$

- Since each instrument is uncorrelated with ε_i, X_{1i} is uncorrelated with ε_i only if v_i is uncorrelated with ε_i.
- To test this, we run the following regression using OLS:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \delta_1 \hat{v}_i + error$$
- And test whether $\delta_1 = 0$ using a standard t-test (If we reject the null hypothesis we conclude that X_1 is endogenous, since v_i and ε_i will be correlated).

Hausman test-estat endogenous

$$m = \frac{\left(\hat{\beta}^{IV} - \hat{\beta}^{OLS}\right)^2}{var(\hat{\beta}^{OLS}) - var(\hat{\beta}^{IV})}$$

Over-identification Test-estat overid

Testing Overidentifying Restrictions

- IV must satisfy two conditions:

  (1) *relevance*: $Cov(z, x) \neq 0$

  (2) *exogeneity*: $Cov(z, \varepsilon) = 0$
- We cannot test (2) because it involves a correlation between the IV and an unobserved error.
- If we have more than one instrument however, we can effectively test whether some of them are uncorrelated with the structural error.
- Consider the above example:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \varepsilon_i$$

- With Z_1 and Z_2 as additional exogenous variables (i.e. additional instruments)
- Estimate this equation by IV using only Z_1 as an instrument, and compute the residuals, $\hat{\varepsilon}_i$.
- We can now test whether Z_2 and $\hat{\varepsilon}_i$ are correlated. If they are, $Z_2$is not a valid instrument.
- This tells us nothing about whether Z_1 and $\hat{\varepsilon}_i$ are correlated (in fact, for this test to be relevant we have to assume that they are not)
- If however, the two instruments are chosen using the same logic (e.g. mother's and father's education levels) finding that Z_2 and $\hat{\varepsilon}_i$ are correlated casts doubt on the use of Z_1 as an instrument.
- Note: if we have a single instrument then there are no overidentifying restrictions and we cannot use this test; if we have two IVs for X_1 we have one overidentifying restriction; if we have three we have two overidentifying restrictions, and so on.