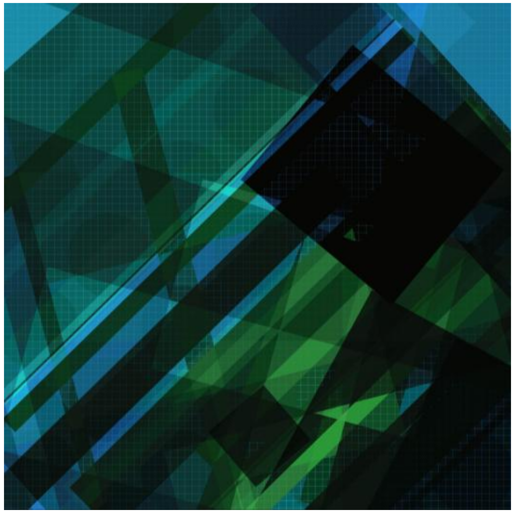
Intermediate Econometrics

IMQF 2025/26



Aleksandra
Nojković

Endogeneity and Instrumental Variables

(Verbeek, Chapter 5)

Gauss-Markov conditions and OLS

Recall the *Gauss-Markov conditions* for the linear model

$$y_i = x_i'\beta + \varepsilon_{i,} \qquad (4.1)$$

which state:

(A1) Error terms have mean zero: $E\{\varepsilon_i\}=0$

(A2) *All* error terms **are independent** of *all* x variables:

$\{\varepsilon_i ,\ldots \varepsilon_N\}$ is independent of $\{x_1,\ldots x_N\}$

(A3) All error terms have the same variance (homoskedasticity): $V\{\varepsilon_i\} = \sigma^2$.

(A4) The error terms are mutually uncorrelated (no autocorrelation): $cov\{\varepsilon_{i,},\varepsilon_j\} = 0, \; i \neq j$

Gauss-Markov conditions

- Denoting the *N*-dimensional vector of all error terms by ε, and the entire matrix of explanatory variables by X, the two essential implications of the Gauss-Markov conditions are:

$$E\{ \varepsilon \mid X\} = 0 \qquad (4.3)$$

and

$$V\{\varepsilon \mid X\} = \sigma^2 I, \qquad (4.4)$$

where *I* is the *N*x*N* identity matrix.

- This says: the distribution of error terms given X has **means of zero** and **constant variances** and **zero covariances (**spherical correlation matrix)

Stochastic regressors (overview)

- Three scenarios in the case of stochastic regressors:

1. Still uncorrelated with the error term of the model:

$$\text{cov}(X_i, \varepsilon_j)=0 \text{ for all } i \text{ and } j$$

- OLS estimator is unbiased (BLUE)

2. Correlated with the error term for different observations:

$$\text{cov}(X_i, \varepsilon_j)\neq 0, \text{ for } i\neq j$$

- OLS estimator is biased, but consistent

3. Correlated with the error term for same observation:

$$\text{cov}(X_i, \varepsilon_j)\neq 0, \text{ for } i=j$$

- **<u>OLS estimator is biased, but inconsistent</u>**

The linear regression model

$$y_t = x_t'\beta + \varepsilon_t$$

- Until now, it was assumed that the error term ε_t and the explanatory variables x_t were **contemporaneously uncorrelated:**

$$E\{ \varepsilon_t \, x_t \} = 0$$

or even **independent of all** explanatory variables

- As a result, the regression model is describing ***a conditional expectation*** $E\{ y_t \mid x_t \} = x_t'\beta$
- In general, OLS is fine (i.e., consistent) to estimate a conditional expectation
- However, behavioral relationships **not necessarily correspond** to conditional expectations

When can we expect E{ ε_t x_t } ≠ 0?

- **Measurement error** in x

- **Omitted variable bias**: some unobservable (or omitted) variable affects both y and x

- If a relevant variable is omitted, OLS becomes biased if the omitted variable is correlated with the included ones

- This is particularly problematic if we **wish to attach a causal interpretation** to our model

- For example, in a wage equation including schooling, omitted factors capturing a person's "ability" may be correlated with schooling. Persons with higher ability have higher wages, but also more schooling

Unobserved heterogeneity

- Stochastic regressor: E{ $\varepsilon_i \, x_i$ } ≠ 0

- In order to estimate "returns to education":

$$\log(\widehat{wages}) = \beta_1 + \beta_2 \text{educ} + \ldots$$

- But more able individuals are both, more successful in labor marker and attend school longer (they find it easier?), then years of schooling will be correlated with the omitted variable, innate ability, and the OLS estimator of education will be biased

- Innate ability is very difficult to measure!

Endogeneity and omitted variable bias

- Consider a wage equation

$$y_i = x'_{1i}\beta_1 + x_{2i}\beta_2 + u_i\gamma + v_i,$$

where x_{2i} denotes years of schooling, and ***u_i* is an unobserved variable** reflecting "ability"

- Estimating β by OLS yields

$$b = \beta + \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i u_i \gamma + \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i v_i$$

showing a bias if u_i and x_i are correlated (the 2nd term does not have mean or plim zero).

When can we expect E{ ε_t x_t } ≠ 0? (II)

- **Simultaneity and reverse causality**.

This happens if x_t not only has an impact on y_t, but at the same time y_t has an impact on x_t

- Consider a Keynesian consumption function:

$$y_t = \beta_1 + \beta_2 x_{2t} + \varepsilon_t \,,$$

where β_2 denotes the marginal propensity to consume

- However, aggregate income (X_{2t}) is not exogenous. For example:

$$x_{2t} = y_t + z_{2t} \,,$$

where z_{2t} denotes investment

Simultaneity and reverse causality

- This implies that income x_{2t} and error term ε_t are correlated

- This can be shown be deriving the **"reduced form",** which describes y_t and x_{2t} as a function of exogenous variable(s) and error terms

- In particular:

$$x_{2t} = \frac{\beta_1}{1-\beta_2} + \frac{1}{1-\beta_2} z_{2t} + \frac{1}{1-\beta_2} \varepsilon_t,$$
$$y_t = \frac{\beta_1}{1-\beta_2} + \frac{\beta_2}{1-\beta_2} z_{2t} + \frac{1}{1-\beta_2} \varepsilon_t.$$

Simultaneity and reverse causality (II)

- From the first of these two equations, it follows that

$$\operatorname{cov}\{x_{2t}, \varepsilon_t\} = \frac{1}{1-\beta_2}\operatorname{cov}\{z_{2t}, \varepsilon_t\} + \frac{1}{1-\beta_2}V\{\varepsilon_t\} = \frac{\sigma^2}{1-\beta_2}$$

which is nonzero. Accordingly, **OLS is inconsistent** for estimating the marginal propensity to consume β_2

- Note, again, that the consumption function **does not correspond** to a conditional expectation.

An alternative estimator

- Let us, for simplicity, consider the simple model

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$$

where E{ $\varepsilon_t x_t$ } ≠ 0. So, OLS is inconsistent. Now, suppose we can find an ***instrumental variable*** $\boldsymbol{z_t}$ satisfying both (**valid instrumental variable**):

1. **Exogeneity**: E{ $\varepsilon_t z_t$ } = 0 (instrument uncorrelated to error term), **and**
2. **Relevance:** cov{ x_t , z_t } ≠ 0 (instrument correlated with endogenous regressor)

An alternative estimator (II)

- Let us now **take the covariance with z_t** on both sides of

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$$

to get

$$\text{cov}\{ y_t , z_t \} = \beta_2 \text{cov}\{ x_t , z_t \} + \text{cov}\{ \varepsilon_t , z_t \}.$$

- So we can write

$$\beta_2 = \text{cov}\{ y_t , z_t \} / \text{cov}\{ x_t , z_t \}$$

- This is (theoretically) defining β_2 . How to estimate it?

An alternative estimator (III)

- Simply replace the population covariances by the sample covariances. Thus, we obtain:

$$\hat{\beta}_{2,IV} = \frac{\frac{1}{T}\sum_t (z_t - \bar{z})(y_t - \bar{y})}{\frac{1}{T}\sum_t (z_t - \bar{z})(x_t - \bar{x})}$$

or

$$\hat{\beta}_{2,IV} = \frac{\sum_t (z_t - \bar{z})(y_t - \bar{y})}{\sum_t (z_t - \bar{z})(x_t - \bar{x})}$$

- Note that **this reduces to OLS** if $z_t = x_t$

IV Estimator

- **Consistent**

  $\mathbf{b}_{IV} = (\mathbf{Z'X})^{-1}\mathbf{Z'y}$

  $= (\mathbf{Z'X}/n)^{-1} (\mathbf{Z'X}/n)\boldsymbol{\beta} + (\mathbf{Z'X}/n)^{-1}\mathbf{Z'}\boldsymbol{\varepsilon}/n$

  $= \boldsymbol{\beta} + (\mathbf{Z'X}/n)^{-1}\mathbf{Z'}\boldsymbol{\varepsilon}/n \rightarrow \boldsymbol{\beta}$

- Asymptotically normal (same approach to proof as for OLS)
- **Inefficient!**

The General Result

- By construction, the **IV estimator is consistent.** So, we have an estimator that is consistent when least squares is not!

Properties of estimator (OLS vs IV)

Method	Exogenous	Endogenous
OLS	consistent, efficient	inconsistent
IV	consistent, inefficient	consistent

IV Estimator properties

- The instrumental variables estimator **is a consistent** estimator for β_2 provided **the instruments are valid**

- This requires that they are both:

- *Exogenous*, i.e., $E\{ \varepsilon_t z_t \} = 0$

and

- *Relevant*, i.e., $\text{cov}\{ x_t , z_t\} \neq 0$.

- Typically, it cannot not be shown that the IV estimator **is unbiased** (small sample properties **are unknown**)

More generally

- Consider the model

$$y_t = x_t'\beta + \varepsilon_t$$

were

$$E\{\varepsilon_t x_t\} \neq 0$$

for some elements of x_t

- Suppose we can find a **vector of instruments z_t**, having **the same dimensions as x_t** such that

$$E\{\varepsilon_t z_t\} = 0$$

More generally (II)

- Then the IV estimator based on these instruments is given by:

$$\hat{\beta}_{IV} = \left(\sum_{t=1}^{T} z_t x_t'\right)^{-1} \left(\sum_{t=1}^{T} z_t y_t\right) \qquad (5.43)$$

- **Its (asymptotic) covariance matrix** is given by

$$V\{\hat{\beta}_{IV}\} = \sigma^2 \left[\left(\sum_{t=1}^{T} x_t z_t'\right)^{-1} \left(\sum_{t=1}^{T} z_t z_t'\right)^{-1} \left(\sum_{t=1}^{T} z_t x_t'\right)\right]^{-1}$$

which can be estimated fairly easily (to get standard errors etc.)

Finding instruments

- Often this is hard
- Why? Statistical theory is of little help here. We need economic arguments to motivate them
- Why? If we drop E{ ε_t x_t } = 0 the **model is not identified** unless we impose other identifying assumptions, i.c.,

$$E\{ \varepsilon_t \, z_t \} = 0$$

- Because ***ε_t* is unobservable**, we cannot statistically identify **which of these two restrictions makes more sense**. And to estimate ε_t (i.e., to get a residual), we need a consistent estimator for β first

Finding instruments (II)

- Instruments need **to be uncorrelated** with the unobservable affecting y

- E.g., we want to estimate a wage equation explaining earnings from schooling and other variables

- Which factors **affect schooling but not earnings directly**? I.e., what affects schooling but not unobserved ability /intelligence that is determining wages?

- *Parents' education? Distance to school? Quarter of birth???*

Estimating the returns to schooling (Example 1)

- Estimating the causal effect of schooling upon earnings has attracted substantive attention in the literature

- **Causal**: what is the effect on earnings of an exogenous increase in schooling?

- OLS estimates tend to be biased, because they reflect differences in unobserved characteristics of individuals that have attained different levels of schooling

- This is referred to as “ability bias”.

  (Another cause of biased OLS estimates is measurement error in schooling.)

Data

- Taken from Card (1995), based on the National Longitudinal Survey of Young Men
- 3010 men, wages in 1976
- We observe individual characteristics, incl. experience, race, region, family background etc.
- We choose a fairly simple specification
- First step: **always do (and report) OLS**. Provides a benchmark for what follows

Table 5.1 Wage equation estimated by OLS

Dependent variable: log(*wage*)			
Variable	Estimate	Standard error	t-ratio
constant	4.7337	0.0676	70.022
schooling	0.0740	0.0035	21.113
exper	0.0836	0.0066	12.575
$exper^2$	−0.0022	0.0003	−7.050
black	−0.1896	0.0176	−10.758
smsa	0.1614	0.0156	10.365
south	−0.1249	0.0151	−8.259

$s = 0.374 \quad R^2 = 0.2905 \quad \bar{R}^2 = 0.2891 \quad F = 204.93$

The Effect of Education on LWAGE

$$LWAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXP + \beta_4 EXP^2 + \ldots + \varepsilon$$

$$What\ is\ \varepsilon? \quad Ability, \ldots + everything\ else$$

$$EDUC \ = \ f(BLACK, SMSA, SOUTH, Ability, \ldots, u)$$

What Influences LWAGE?

$$\mathbf{LWAGE} = \beta_1 + \beta_2 \mathbf{EDUC}(\mathbf{X}, \text{Ability}, \ldots)$$
$$+ \beta_3 \mathbf{EXP} + \beta_4 \mathbf{EXP}^2 + \ldots$$
$$+ \varepsilon(\text{Ability})$$

Increased Ability is associated with increases in **EDUC**(**X**, Ability,...,u) and ε(Ability)

What looks like an effect due to increase in **EDUC** may be an increase in Ability. The estimate of β_2 picks up the effect of **EDUC** and the hidden effect of Ability.

An Exogenous Influence

$$\begin{aligned}\mathbf{LWAGE} = \beta_1 + \beta_2 &\mathbf{EDUC}(\mathbf{X}, \mathbf{Z}, \text{Ability}, \ldots) \\ &+ \beta_3 \mathbf{EXP} + \beta_4 \mathbf{EXP}^2 + \ldots \\ &+ \varepsilon(\text{Ability})\end{aligned}$$

Increased **Z** is associated with increases in $\mathbf{EDUC}(\mathbf{X}, \mathbf{Z}, \text{Ability}, \ldots, u)$ and not $\varepsilon(\text{Ability})$

An effect due to the effect of an increase **Z** on **EDUC** will only be an increase in **EDUC**. The estimate of β_2 picks up the effect of **EDUC** only.

Z is an Instrumental Variable

Reduced form for schooling, estimated by OLS

Dependent variable: *schooling*

Variable	Estimate	Standard error	t-ratio
constant	−1.8695	4.2984	−0.435
age	1.0614	0.3014	3.522
age^2	−0.0188	0.0052	−3.386
black	−1.4684	0.1154	−12.719
smsa	0.8354	0.1093	7.647
south	−0.4597	0.1024	−4.488
lived near college	0.3471	0.1070	3.244

$s = 2.5158 \quad R^2 = 0.1185 \quad \bar{R}^2 = 0.1168 \quad F = 67.29$

Wage equation estimated by IV

Dependent variable: log(*wage*)

Variable	Estimate	Standard error	t-ratio
constant	4.0656	0.6085	6.682
schooling	0.1329	0.0514	2.588
exper	0.0560	0.0260	2.153
$exper^2$	−0.0008	0.0013	−0.594
black	−0.1031	0.0774	−1.333
smsa	0.1080	0.0050	2.171
south	−0.0982	0.0288	−3.413

Instruments: *age*, age^2, *lived near college*
used for: *exper*, $exper^2$ and *schooling*

Issues

- Any IV estimate requires a choice of instruments that should be motivated. Always mention this choice

- Reduced form explaining endogenous regressors from exogenous regressors and instruments, should show significant effect of the instruments. (If weak: **weak instruments problem**)

- IV estimates are (much) less accurate than OLS (how much depends upon their correlation with the endogenous regressors)

- It is possible to use more instruments than required (**overidentification**)

The IV / 2SLS estimator

- The resulting estimator is referred to as the **instrumental variable's estimator**

- It is also known by the name two-stage least squares estimator (2SLS). Why?

- The **same estimator can be obtained in two-steps**:

1. Estimate reduced forms (by OLS) that explain x_i from z_i. Take the fitted values from these regressions (these are interpreted as best linear approximations)
2. Estimate the original model (by OLS) replacing the **endogenous regressors by the fitted values** from 1. (Do not replace them by the instruments!!!)

Important remarks (summary)

- Instruments should be **exogenous**, i.e., uncorrelated with the equation's error term
- They should also be **relevant**, i.e., correlated with the regressors that they are supposed to be instrumenting
- This means that in the reduced form, where we explain x_i from z_i, the instruments should be **"sufficiently important".** (for example, lived-near-college should have a non-negligible impact upon schooling, conditional upon the other exogenous variables/instruments)
- Otherwise, we may have a **"weak instruments" problem**

Testing for endogeneity?

- It is **possible to test** whether one or more regressors are endogenous (correlated with the error term), provided **we are willing to assume that the instruments are valid** (i.e. ,assuming $E\{ \varepsilon_i z_i \} = 0$ we can test whether $E\{ \varepsilon_i x_i \} = 0$)

- Under the null, both the OLS and IV estimator are consistent. They should differ by sampling error only. Under the alternative hypothesis, **only the IV estimator is consistent** (and OLS is inconsistent)

- Hausman based a test on the difference between the two estimators

Hausman test

- We formally test:

H_0: E{ ε_i x_i } = 0 (d=0) → x_i *is exogenous*

H_1: E{ ε_i x_i } ≠ 0 (d ≠ 0) → x_i *is endogenous,*

where

$$d = \hat{\beta}_{IV} - \hat{\beta}_{OLS}$$

- Decision is based on Wald test-statistic:

$$H = d'[Asim.Var.(d)]^{-1}d$$

that has Chi-squared distribution (DF = # variables that we test for endogeneity).

Hausman test (II)

- As those estimates are independent, we use:

$$\text{Asim. Var. } (d) = Asim.Var.(\hat{\beta}_{IV})\text{- } Asim.Var.(\hat{\beta}_{OLS})$$

- Provided that **the instruments are valid!**

Testing for endogeneity?

- A simple version is obtained by running an auxiliary regression, **where we augment the original model with the residual(s) from the reduced form equations** (also known as Durbin-Wu-Hausman, DWH or just Wu test)

- Estimation of this auxiliary regression by OLS reproduces the IV estimator. Under the null hypothesis (x_i is exogenous) the added residual(s) should be irrelevant

- The Hausman test for endogeneity is based on the *t*-statistic (of *F*-statistic) on the reduced form residuals

Hausman test (DWH version)

- Let us consider the simple model:

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

In which we test potential endogeneity of x_i, *i.e.*, $E\{ \varepsilon_i x_i \} \neq 0$

- Now, suppose we can find two valid instrumental variables (z_{1i} *and* z_{2i}) for x_i

Hausman test (DWH version;II)

1) From the estimated auxiliary regression:

$$x_i = \alpha_1 + \alpha_2 z_{1i} + \alpha_3 z_{2i} + v_i$$

we get the residuals: $\hat{v}_i$.

2) We estimate original model estimated with residual(s) from 1):

$$y_i = \beta_1 + \beta_2 x_i + \gamma \hat{v}_i + \varepsilon_i$$

- We formally test:

H_0: $\gamma = 0 \rightarrow x_i$ is exogenous ($E\{ \varepsilon_i x_i \} = 0$)

H_1: $\gamma \neq 0 \rightarrow x_i$ is endogenous ($E\{ \varepsilon_i x_i \} \neq 0$)

- Test of endogeneity is based on t-statistic(s) (F-statistic) for reduced form residual(s).

Testing instrument validity?

1. **Exogeneity**: E{ $\varepsilon_i\, z_i$ } = 0 (instrument uncorrelated to error term), **and**

2. **Relevance:** cov{ x_i , z_i} $\neq$ 0 (instrument correlated with endogenous regressor)

Testing relevance of instrument (checking for weak instruments)?

1. Stock-Watson test – one simple rule of thumb: you do not need to worry about instruments if the first stage F-statistic exceeds 10

2. Stock-(Wright)-Yogo test (based on Cragg-Donald test-statistics **which is safely large enough to conclude the instruments are strong**) – null hypothesis (instruments are weak) can not be rejected if value of test statistic is below critical values

J-test of instruments exogeneity (Sargan-Hansen)

1) We estimate baseline model:

$$y_i = \alpha_1 + \alpha_2 x_i + \varepsilon_i$$

we save the residuals: e_t; z_{1t} and z_{2t} are potential instruments for x_t

2) We use residuals from 1) for the auxiliary regression:

$$e_i = \alpha_1 + \alpha_2 z_{1i} + \alpha_3 z_{2i} + v_i$$

- We formally test:

H_0: $\alpha_2 = \alpha_2 = 0 \rightarrow instruments$ is exogenous (E{ ε_i z_i } = 0)

H_1: $\alpha_2 \neq \alpha_2 \neq 0 \rightarrow instruments$ is endogenous (E{ ε_i z_i } $\neq$ 0)

J-test of instruments exogeneity (II)

4) Test statistic: *T* multiplied by R^2 of the auxiliary regression 2. Has Chi-squared distribution (# DF = **difference in number** of instruments minus number of explanatory variables in baseline regression)

5) We will reject the null hypothesis (*instruments are exogenous*) if value of J test-statistic is larger than critical value.

Note: Test is valid only for the overidentified case!

Estimating the returns to schooling (Example 2)

- Data taken from Wooldridge textbook (Moroz data, 2012)

- Wages of 428 married women

- We choose a fairly simple specification (log(wages) as a function of education and…)

- We check for validity potential instruments for education- education of parents + husband education

OLS estimates of wages for married women

Dependent Variable: LWAGE
Method: Least Squares
Date: 01/15/22 Time: 18:22
Sample (adjusted): 1 428
Included observations: 428 after adjustments
Huber-White-Hinkley (HC1) heteroskedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.522041	0.201650	-2.588840	0.0100
EDUC	**0.107490**	0.013219	8.131471	0.0000
EXPER	0.041567	0.015273	2.721561	0.0068
EXPERSQ	-0.000811	0.000420	-1.931083	0.0541

R-squared	0.156820	Mean dependent var	1.190173
Adjusted R-squared	0.150855	S.D. dependent var	0.723198
S.E. of regression	0.666420	Akaike info criterion	2.035509
Sum squared resid	188.3052	Schwarz criterion	2.073445
Log likelihood	-431.5990	Hannan-Quinn criter.	2.050492
F-statistic	26.28616	Durbin-Watson stat	1.960988
Prob(F-statistic)	0.000000	Wald F-statistic	27.29936
Prob(Wald F-statistic)	0.000000		

The Ultimate Source of Endogeneity

- **LWAGE** = f(**ED,**

  EXP,EXPSQ,…) + ε

- **ED** = f(**MED,FED)** + u

Remove the Endogeneity

– **LWAGE** = f(**ED,**
EXP,EXPSQ,…)

+ u + ε

– Strategy
  - Estimate u
  - **Add u to the equation.** ED is **uncorrelated** with ε when u is in the equation.

RES1 from auxiliary regression

Dependent Variable: EDUC
Method: Least Squares
Date: 11/08/21 Time: 22:54
Sample: 1 753
Included observations: 753

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.975657	0.225668	39.77374	0.0000
MOTHEDUC	0.183279	0.026217	6.990911	0.0000
FATHEDUC	0.183418	0.024714	7.421766	0.0000

R-squared	0.244970	Mean dependent var	12.28685
Adjusted R-squared	0.242956	S.D. dependent var	2.280246
S.E. of regression	1.984002	Akaike info criterion	4.212085
Sum squared resid	2952.198	Schwarz criterion	4.230508
Log likelihood	-1582.850	Hannan-Quinn criter.	4.219182
F-statistic	121.6689	Durbin-Watson stat	1.961485
Prob(F-statistic)	0.000000		

Test for endogeneity of EDUC

Dependent Variable: LWAGE
Method: Least Squares
Date: 01/15/22 Time: 18:39
Sample (adjusted): 1 428
Included observations: 428 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.026967	0.378398	0.071266	0.9432
EDUC	**0.063404**	0.029483	2.150561	0.0321
EXPER	0.041537	0.013146	3.159686	0.0017
EXPERSQ	-0.000841	0.000393	-2.140500	0.0329
RES1	0.056370	0.033097	1.703166	**0.0893**

R-squared	0.162563	Mean dependent var	1.190173
Adjusted R-squared	0.154644	S.D. dependent var	0.723198
S.E. of regression	0.664931	Akaike info criterion	2.033348
Sum squared resid	187.0226	Schwarz criterion	2.080768
Log likelihood	-430.1365	Hannan-Quinn criter.	2.052076
F-statistic	20.52819	Durbin-Watson stat	1.931082
Prob(F-statistic)	0.000000		

Test for endogeneity of EDUC (II)

Endogeneity Test
Equation: EQ01
Endogenous variables to treat as exogenous: EDUC
Specification: LWAGE C EDUC EXPER EXPERSQ
Instrument specification: C FATHEDUC MOTHEDUC EXPER EXPERSQ
Null hypothesis: EDUC are exogenous

	Value	df	Probability
Difference in J-stats	2.780836	1	**0.0954**

J-statistic summary:

	Value
Restricted J-statistic	3.164752
Unrestricted J-statistic	0.383916

IV estimates of wages for married women

Dependent Variable: LWAGE
Method: Two-Stage Least Squares
Date: 01/15/22 Time: 18:26
Sample (adjusted): 1 428
Included observations: 428 after adjustments
White heteroskedasticity-consistent standard errors & covariance
Instrument specification: **FATHEDUC MOTHEDUC** EXPER EXPERSQ
Constant added to instrument list

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.048100	0.429798	0.111914	0.9109
EDUC	**0.061397**	0.033339	1.841608	0.0662
EXPER	0.044170	0.015546	2.841202	0.0047
EXPERSQ	-0.000899	0.000430	-2.090220	0.0372

R-squared	0.135708	Mean dependent var	1.190173
Adjusted R-squared	0.129593	S.D. dependent var	0.723198
S.E. of regression	0.674712	Sum squared resid	193.0200
F-statistic	8.140709	Durbin-Watson stat	1.945659
Prob(F-statistic)	0.000028	Second-Stage SSR	212.2096
J-statistic	**0.374538**	Instrument rank	5
Prob(J-statistic)	**0.540541**		

Test for weak instruments (Stock-Watson test - from auxiliary regression)

Dependent Variable: EDUC
Method: Least Squares
Date: 11/08/21   Time: 22:54
Sample: 1 753
Included observations: 753

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.975657	0.225668	39.77374	0.0000
MOTHEDUC	0.183279	0.026217	6.990911	0.0000
FATHEDUC	0.183418	0.024714	7.421766	0.0000

R-squared	0.244970	Mean dependent var	12.28685
Adjusted R-squared	0.242956	S.D. dependent var	2.280246
S.E. of regression	1.984002	Akaike info criterion	4.212085
Sum squared resid	2952.198	Schwarz criterion	4.230508
Log likelihood	-1582.850	Hannan-Quinn criter.	4.219182
F-statistic	**121.6689**	Durbin-Watson stat	1.961485
Prob(F-statistic)	0.000000		

Test for weak instruments (Stoc-Yogo test)

Weak Instrument Diagnostics
Equation: EQ01

Cragg-Donald F-stat: 55.40030

Stock-Yogo bias critical values not available for models with less than 3 instruments.

Stock-Yogo critical values (size):

10%	19.93
15%	11.59
20%	8.75
25%	7.25

Moment selection criteria:

SIC-based:	-5.684586
HQIC-based:	-3.246608
Relevant MSC:	-15.98390

Estimating the returns to schooling (Example 3)

- Angrist and Krueger, „Does Compulsory School Affect Schooling and Earnings“, *Quarterly Journal of Economics*, 1991 (AK model).

- Model estimated on U.S.Census data (329,000 obs.):

$$(1) \qquad E_i = X_i\pi + \Sigma_c Y_{ic}\delta_c + \Sigma_c \Sigma_j Y_{ic} Q_{ij} \theta_{jc} + \epsilon_i$$

$$(2) \qquad \ln W_i = X_i\beta + \Sigma_c Y_{ic}\xi_c + \rho E_i + \mu_i,$$

- Using the individual's quarter of birth as instrumental variable!

Estimating the returns to schooling (Example 3, II)

- Famous example or "Scary Regression" for labor economists (Stock and Watson, 2003) – illustration for weak instruments!
- Krueger suggested a creative way to find out: replace each individual quarter of birth by fake quarter of birth, randomly generated
- **Re-analysis using fake instruments** is published in Bound, Jaeger and Baker (1995).
- **TSLS estimates based on real data are just as unreliable** as those based on the fake data!
- **Problem:** Instrument are very weak in some AK regressions (the first-stage F-statistics is less than 2; btw/ returns to education are about 8% - somewhat greater than OLS estimates)