Lecture 3. Instrumental Variables

Olivier Godechot

Sciences Po

M2. Sociology Master

Causes and effect, I know,
Our little hates and blames,
We are born and grow,
As the seeds we sow,
And right and wrong are – names
Cause and effect, I know

Philip Green Wright, "Revulsion", *The dial of heart*, 1905.
(Inventor of instrumental variables)

Invention of instrumental variables

- Wright. 1928. *The Tariff on Animal and Vegetable Oils.*
- Economist, poet, mathematician, also sociology professor…
- Appendix B
- Estimation of supply and demand curves
- His son, a biologist, might have helped (or written the appendix)

1. The problem: endogeneity

The limits of OLS

- Linear model

  $y_i=\beta_0+\beta_1.x_{1i}+\beta_2.x_{2i}+\ldots+\beta_k.x_{ki}+u_i$

  where i represents individual i

  or

  $y=\beta_0+\beta_1.x_1+\beta_2.x_2+\ldots+\beta_k.x_k+u$

  or

  $y = X.\ \beta + u$

- Method
  - Least square of errors
  - We estimate parameters $\beta_0, \beta_1, \beta_2 \ldots \beta_k$ in such way that $\Sigma_i u_i^2$ is minimal

OLS 6 hypotheses

- Linearity
- Full rank Matrix and absence of auto-correlation between independent variables
- Homeoscedasticity
- Absence of auto-correlation of residuals
- Normality of residuals
- Absence of correlation between independent variables and the residual *in the theoretical model.*

The endogeneity problem

- We speak of endogeneity in a broad sense when there's a violation of the last hypothesis.
- These problems can lead to mistakes in parameters interpretation.
- Instrumental variables can offer a correction technique.

Three problems and their effects on parameters

- Case 1: we don't measure correctly the independent variable
  - We underestimate the absolute value of the parameter
- Case 2: an independent variable is missing and we know a) this variable is positively correlated to the dependent variable b) is positively (resp. negatively) correlated to one of the independent variables
  - We over-estimate (resp. underestimate) the absolute value of the parameter
- Case 3: the independent variable also depends on the dependent variable
  - More complex effect. No evident intuition

Independence of independent variables and residuals: $Cov(x_i,u)=0$

- It may be possible that in our *true model* $Cov(x_i,v)\neq 0$.
  - $y_i = a_{true} + b_{true}.\, x_i + v_i \quad (1)$
- In an empirical OLS estimation, by construction, $Cov(x_i,u)=0$.
  - $y_i = a_{est} + b_{est}.\, x_i + u_i \quad (2)$
- If that's the case, then OLS empirical parameter estimates will not be the one we look for.
  - $E(a_{est}) \neq a_{true}$
  - $E(b_{est}) \neq b_{true}$
- Empirical estimation is not "false" *per se*. But it's wrong to interpret the estimated parameters as that of the true model!

Measurement error on an independent variable

- Our <u>true model</u> is the following

  $y = a_{true} + b_{true}.\ x_{true} + u$ with $cov(x_{true}, u) = 0$ (1)

- We don't measure correctly x_{true}

  $x_{prox} = x_{true} + e$ with $cov(x_{prox}, e) \neq 0$ (2)

- Our <u>modified true model</u> will be the following

  $y = a_{true} + b_{true}.\ x_{prox} + u'$ with $u' = u - b_{true}.\ e$ (3)

- In this modified true model, the proxy variable is correlated with the new residual u'

  $cov(x_{prox}, u') = cov(x_{prox}, u) + cov(x_{prox}, -b_{true}.\ e)$

  $= -b_{true}.\ cov(x_{prox}, e) = -b_{true}.\ cov(x_{true} + e, e) \neq 0$

- Hence, our empirical model will not allow us to estimate the parameters a_{true} and b_{true}

  $y = a_{est} + b_{est}.\ x_{prox} + v$ (4)

- Model 4 is not wrong, but it's wrong to interpret model 4 as an estimation of the parameters of model 3.

Measurement error: Nature of the bias

- We can calculate the relation between the OLS estimate and the true parameter

$$b_{est} \xrightarrow{P} \frac{b_{true}}{1+\frac{V(e)}{V(x_{true})}}$$

- We under-estimate the true parameter
- The estimated parameter is all the worse that the error measurement is large

Omitted variable or unobserved heterogeneity

- Our <u>true model</u> is the following

$$y = a_{true} + b_{true}.\, x + c_{true}.\, z + u \qquad \text{with } cov(x,\, u) = 0 \text{ and } cov(z,\, u) = 0 \qquad (1)$$

- We don't observe the variable z and $cov(x,z) \neq 0$
- Our <u>modified true model</u> will be the following

$$y = a_{true} + b_{true}.\, x + u' \quad \text{where } u' = u + c_{true}.\, z \qquad (3)$$

- In this modified true model, variable x is correlated with the new residual u'

$$cov(x,u') = cov(x,\, u + c_{true}.\, z\,) = cov(x,\, u) + cov(x,\, c_{true}.\, z) = -c.\, cov(x,\, z) \neq 0$$

- Hence, our empirical model does not allow us to estimate the parameters a_{true} and b_{true}

$$y = a_{est} + b_{est}.\, x + v \qquad (4)$$

- Model 4 is not wrong, but it's wrong to interpret model 4 as an estimation of the parameters of model 3

Nature of bias with unobserved heterogeneity

- We can calculate the relation between the OLS estimated parameter and the true parameter

$$b_{est} \xrightarrow{P} b_{true} + c\,\frac{cov(x,z)}{V(x)}$$

- The estimated parameter is all the more biased that
  - variables x and z are strongly correlated ($cov(x,z)$ effect)
  - z has a strong impact on y (parameter c effect)

Simultaneity

- Classical case: prices and quantities on a market ; educational aspirations and educational level
- Our true model is the following

$y = a.x + b + u \quad (1)$

$x = c.y + d + v \quad (2)$

- In this model $cov(x,u) \neq 0$ and $cov(y,v) \neq 0$. We can show this by replacing y in equation 2 by equation 1:

$x = cax + cb + cu + d + v$

$x = (cb + cu + d + v)/(1 - ca)$

$cov(x,u) = cov((cb + cu + d + v)/(1 - ca), u) = (c/(1 - ca)).cov(u, u)$

- Hence, our empirical model does not allow us to estimate the true parameters a and b

$y = a_{est} + b_{est}.x + w \quad (3)$

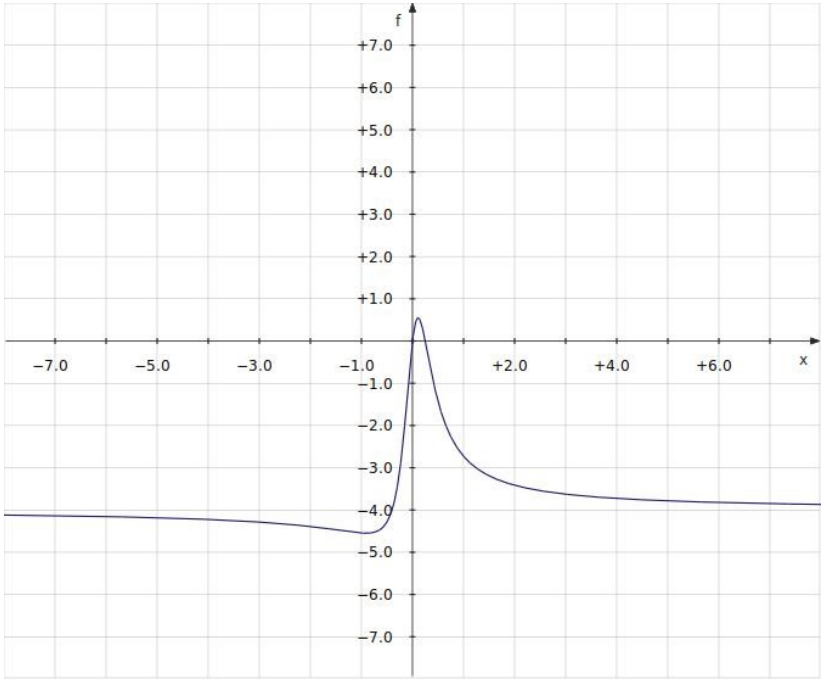
- Model 3 is not wrong, but it's wrong to interpret model 3 as an estimation of the parameters of model 1

Nature of the bias with simultaneity

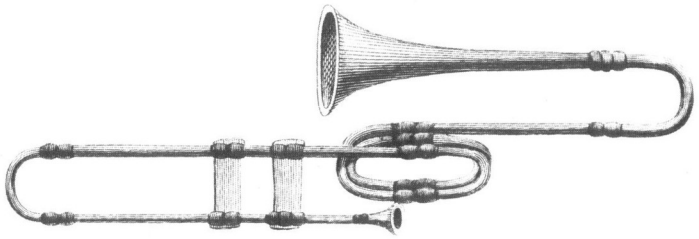
- We can calculate the relation between the OLS parameter estimate and the true parameter

$$a_{est} \xrightarrow{P} \frac{a.V(v)+c.V(u)}{V(v)+c^2.V(u)}$$

- A more complex bias. Cf. Example of the bias variation depending v :



2. The solution: instruments



Ex: local neighborhood composition and school performance

- Two endogeneity problems: simultaneity and unobserved heterogeneity

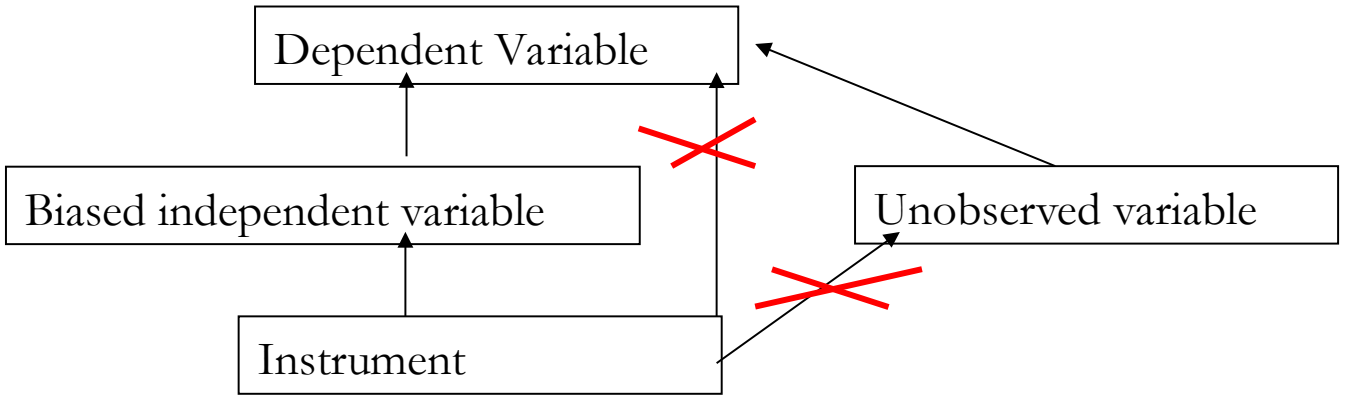
  $Educ_i = Educ\_neighbor_k + exogenous + Omitted + u$

  $Educ_i = Educ\_neighbor_k(Educ_i) + exogenous + Omitted + u$ with $cov(Educ\_neighbor_k, Omitted) \neq 0$

  - Endogeneity of the model. OLS estimates may be biased.

- Solution : instrumental variables
  - Find an exogenous instrumental variable: a variable which impacts the dependent variable ONLY through its impact on the endogenous independent variable.
  - Idea of Goux and Maurin (2005, 2007)
    - The month of birth impacts neighbors' Educ level.
    - I don't select neighbors and neighborhood based on the month of birth of its children.
    - The neighbors' month of birth impact my Educ level only through its impact on their own Educ level.

Instrumental variables in one graph



Instrumental variables: the method

- Imagine we have an endogeneity problem with the variable x_{endo} in our true model (for one of the three aforementioned reasons).

$$y = a_{true} + b_{true} x_{endo} + c_{true} . x_2 + u$$

$$cov(x_{endo}, u) \neq 0$$

- Therefore, the OLS parameters does not allow us to estimate the neither the true parameter, b_{true}, nor a_{true} and c_{true}

$$E(b_{est.OLS}) \neq b_{true} \quad ; \quad E(a_{est.OLS}) \neq a_{true} \quad ; \quad E(c_{est.OLS}) \neq c_{true}$$

- We can correct this problem if we have an instrument z_{inst} such as:

$$cov(z_{inst}, x_{endo}) \neq 0$$

$$cov(z_{inst}, u) = 0$$

A two-stage procedure

- True model : $y = a_{true} + b_{true} x_{endo} + c_{true} . x_2 + u$ with $cov(x_{endo}, u) \neq 0$
- **First stage :** we regress the endogenous variable both on the instrument and all other independent variables. NB: we use all the second stage independent variables in the first stage, even if they are meaningless.

  $$\boldsymbol{x_{endo}} = a_0 + a_1 . z_{inst} + a_2 . x_2 + u_{first}$$

- We keep from this first regression x'_{endo}, the prediction of the endogenous variable x_{endo} :

  $$x'_{endo} = a_0 + a_1 . z_{inst} + a_2 . x_2 = x_{endo} - u_{first}$$

- **Second stage:** we replace in the regression x_{endo} by its prediction x'_{endo}

  $$y = a_{est} + \boldsymbol{b_{est} x'_{endo}} + c_{est} . x_2 + u_{secon}$$

- As z_{inst} and x_2 are not correlated with the residual u, then x'_{endo} is not correlated with u either. Therefore, the instrumental variable enables to estimate without bias b_{true} (and also a_{true} and c_{true})
  - $E(b_{est}) = b_{true}$; $E(a_{est}) = a_{true}$; $E(c_{est}) = c_{true}$

Example: the IV correction of measurement errors

- Our <u>modified true model</u> was the following

$$y = a_{true} + b_{true}.\, x_{prox} + u' \text{ with } u' = u - b_{true}.\, e \qquad (1)$$

- We estimate the first stage with our instrument z_{inst}

$$x_{prox} = dz_{inst} + v \qquad \text{with } cov(z_{inst}, v) = 0 \quad (2)$$

- Second stage : we replace x_{prox} with x'_{prox} , its first stage prediction.

$$y = a_{true} + b_{true}.\, x'_{prox} + w \qquad (3)$$

- In this modified true model (3), x'_{prox}, is not anymore correlated to the new residual $(u - b_{true}.\, e + b_{true}.v)$

$$cov(x'_{prox}, u - b_{true}.\, e + b_{true}.\, v) = cov(dz_{inst}, u) + cov(dz_{inst}, -b_{true}.\, e) + cov(dz_{inst}, b_{true}.\, v) = 0$$

- Hence, OLS empirical modified model enables now to estimate without bias parameters a_{true} and b_{true}

Goux and Maurin (*Revue économique*, 2005)

Tableau 6. *Une analyse de l'effet de contexte endogène par la technique des variables instrumentales*

Modèle à probabilité linéaire	Variables dépendantes		
	% voisins en retard	Être en retard à 15 ans	
	Première étape	MCO	VI
% voisins en retard		0.22 (0.01)	0.20 (0.11)
[Garçon = 1]	.001 (.004)	0.11 (0.01)	0.11 (0.01)
Distribution des mois de naissance .			
% Janvier-juin	– .14 (.01)		
% Juillet-novembre	– .07 (.01)		
% Décembre	Ref.		
R^2	0.001	0.04	0.02
Nombre d'observations	24 367	24 367	24 367

Champ : pour l'enquête *t*, enfants nés en *t* – 15 résidant dans le voisinage depuis plus d'un an.
Source : INSEE, Enquêtes « Emploi », 1991-2002.

Replication: OLS "naïve" model

```
Call:
lm(formula = RET15 ~ VRET15 + S1, data = gm2)

residuals:
    Min      1Q  Median      3Q     Max
-0.5928 -0.4043 -0.2771  0.5355  0.7453

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.254656   0.005871   43.38   <2e-16 ***
VRET15      0.224473   0.009309   24.11   <2e-16 ***
S1          0.113680   0.006228   18.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual standard error: 0.4818 we 23948 degrees of freedom
  (405 observations deleted due to missingness)
Multiple R-squared: 0.03663,    Adjusted R-squared: 0.03655
F-statistic: 455.2 we 2 and 23948 DF,  p-value: < 2.2e-16
```

Reduced form. Direct impact of instruments on the dependent variable

(not presented in 2005 article)

```
Call:
lm(formula = RET15 ~ S1 + VJANJUN + VJULNOV, data = gm2)

residuals:
    Min      1Q  Median      3Q     Max
-0.4947 -0.4553 -0.3435  0.5433  0.6576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.381750   0.017307  22.057   <2e-16 ***
S1           0.112928   0.006304  17.914   <2e-16 ***
VJANJUN     -0.039376   0.018725  -2.103   0.0355 *
VJULNOV     -0.036094   0.019145  -1.885   0.0594 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

residual standard error: 0.4876 we 23940 degrees of freedom
Multiple R-squared:  0.01339,        Adjusted R-squared:  0.01326
F-statistic: 108.3 we 3 and 23940 DF,  p-value: < 2.2e-16
```

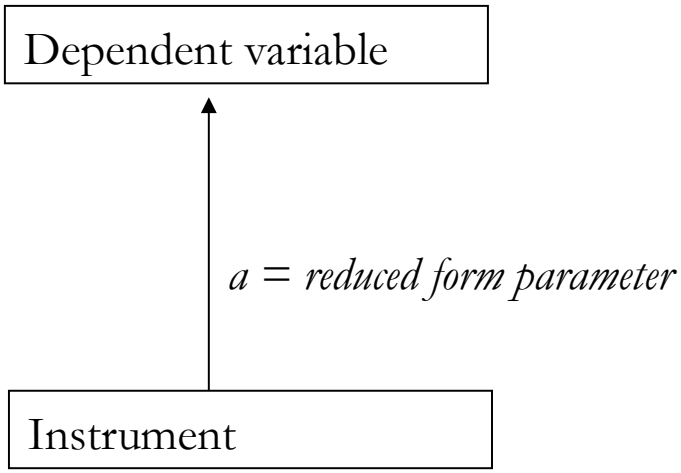
Goux and Maurin. First stage regression

```
2SLS estimates for 'eq1' (equation 1)
Model Formula: VRET15 ~ VJANJUN + VJULNOV + S1
Instruments: ~VJANJUN + VJULNOV + S1

              Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)  0.50992739  0.01182131  43.13629 < 2.22e-16 ***
VJANJUN     -0.14299532  0.01278992 -11.18032 < 2.22e-16 ***
VJULNOV     -0.06400824  0.01307643  -4.89493  9.898e-07 ***
S1          -0.00293271  0.00430569  -0.68112     0.4958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual standard error: 0.333023 we 23940 degrees of freedom
Number of observations: 23944 Degrees of Freedom: 23940
SSR: 2655.052533 MSE: 0.110904 Root MSE: 0.333023
Multiple R-Squared: 0.008394 Adjusted R-Squared: 0.008269
```

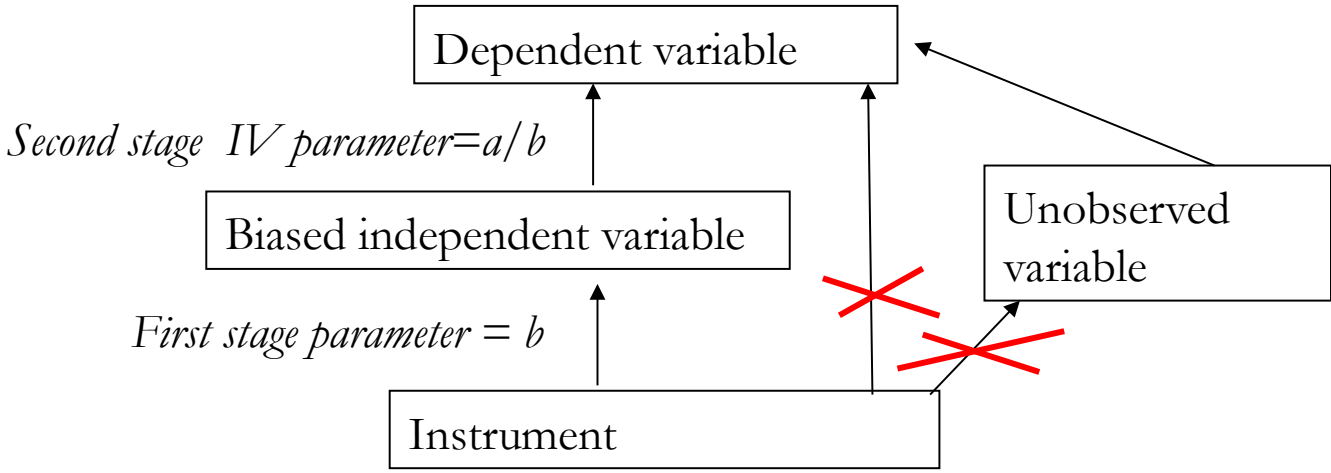
Goux and Maurin. Second stage regression

```
2SLS estimates for 'eq2' (equation 2)
Model Formula: RET15 ~ VRET15 + S1
Instruments: ~VJANJUN + VJULNOV + S1

              Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 0.28178700 0.04207210  6.69772 2.1639e-11 ***
VRET15      0.15861092 0.10186035  1.55714    0.11945
S1          0.11330863 0.00624303 18.14963 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual standard error: 0.48229 we 23941 degrees of freedom
Number of observations: 23944 Degrees of Freedom: 23941
SSR: 5568.769685 MSE: 0.232604 Root MSE: 0.48229
Multiple R-Squared: 0.034594 Adjusted R-Squared: 0.034514
```

Reduced form and two stage estimation



Reduced form



Instrumental variables estimator
(Two stage least squares)

Instrumental variables reduced to its simplest expression

(inspired from Goux and Maurin, 2007, Table 3-3 and Angrist and Pischke 2008)

- >50% of neighbors born during the 1er semester
  - 4.4 percentage points of neighbors one-year behind at the age of 15
  - 2.3 percentage point chance for ego to be one-year behind at the age of 16

Differences in mean	(Reduced form)	(First stage)	(IV Estimator)
Dependent variable	Probability for ego to be one-year behind at 16	Proportion of neighbors one-year behind at 15	Proba. ego one-year behind Diff col. 1 / Diff. col. 2
Independent Variable			
Neighbors born 1st sem. >50%	0.552	0.383	
Neighbors born 1st sem. <=50%	0.575	0.426	
Difference	-0.023**	-0.044 ***	+0.512**

- If the effect of neighbors born during the first semester on ego's repeating a grade is only due to the fact that these neighbors will be less often one year behind
  - => If 100% of neighbors are one-year behind at age 15: -0.023/-0.044 = 0.512 : 51.2 percentage points chances of being one-year behind at age 16.

	« Naive » OLS	*Reduced form*	*First stage*	*Instrumental Variable*
	P(one year behint at 16)	P(one year behint at 16)	Prop neighbors one year behing at 15	P(one year behind at 16)
Prop. neighbors one year behing at 15	0.224***			0.382*
Prop. neighbors born between jan. and may		-0.0283*	-0.074***	
Controls	Oui	Oui	Oui	Oui

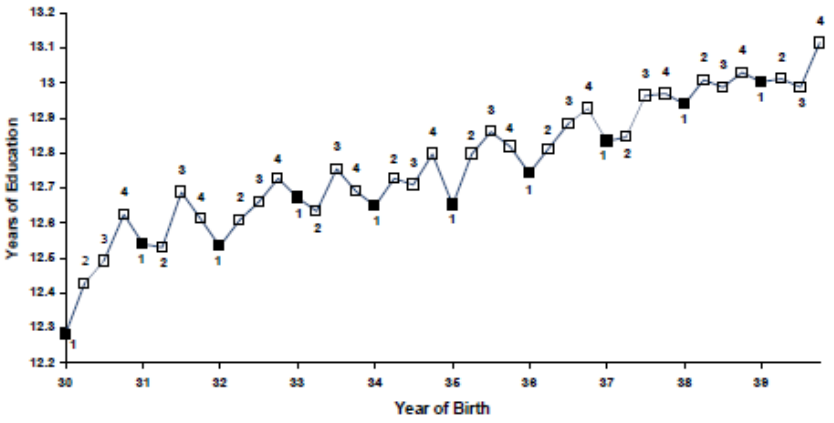
- When there's only one endogenous variable and one instrument, the IV estimator: 0.382 = reduced form effect/first stage effect= (-0.0283/-0.074)

A canonical exemple (Angrist, Krueger, 1991)

- Measure of the school effect on income.
  - The number of years of education don't measure correctly the school institutional effect as it also depends on students' initial capacities.
  - Instrumental variable: Using compulsory schooling rules
  - In the United-States
    - School compulsory: you need to be 6 years old before January 1st.
    - School compulsory up to 16 exactly (or 17, or 18 depending the states). It is possible to drop out during the year.
    - Consequences: Kids born during the beginning of the year start school later than those born at the end year. But both can stop at the same age.
    - Kids born during the beginning of the year can go less to school.
    - Exogenous variation of school length not depending on "intelligence".

An effect small but obvious

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



B. Average Weekly Wage by Quarter of Birth (reduced form)

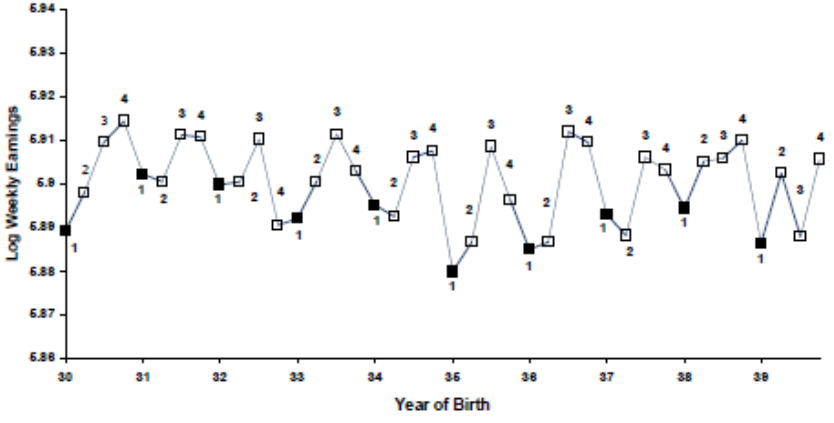


Figure 4.1.1: Graphical depiction of first stage and reduced form for IV estimates of the economic return to schooling using quarter of birth (from Angrist and Krueger 1991).

Instrumental variables reduced to their most simple expression (Angrist & Pischke 2008)

- Born during the 1st semester
  - =>-0.15 year of school (1.8 month of school)
  - => -1.3% less wage

Mean differences	(Reduced Form)	(First stage)	(IV estimator)
Dependent bariable			Weakly log wage.
Independent Variable	Weakly Log Wage	Number of education years	Diff col. 1 / Diff. col. 2
Born 1st semester	5.892	12.69	
Born 2nd semester	5.905	12.84	
Difference	-0.013***	-0.15 ***	+0.087***

- If all the birth semester effect on wages goes only through education length (and not through any other channel)
  - => 1 year school more: -0.013/-0.15 = +8.7% wage

Estimation of the effect through the instrumental variable technique

- Dependent Variable
  - log of weakly wage
- Endogenous Variable
  - number of years of education
- Instruments
  - trimesters of birth* year of birth.

TABLE IV

OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1920–1929: 1970 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0802	0.0769	0.0802	0.1310	0.0701	0.0669	0.0701	0.1007
	(0.0004)	(0.0150)	(0.0004)	(0.0334)	(0.0004)	(0.0151)	(0.0004)	(0.0334)
Race (1 = black)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
SMSA (1 = center city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
Married (1 = married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region of residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
Age-squared	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
χ^2 [dof]	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

a. Standard errors are in parentheses. Sample size is 247,199. Instruments are a full set of quarter-of-birth times year-of-birth interactions. The sample consists of males born in the United States. The sample is drawn from the State, County, and Neighborhoods 1 percent samples of the 1970 Census (15 percent form). The dependent variable is the log of weekly earnings. Age and age-squared are measured in quarters of years. Each equation also includes an intercept.

3. Validating the solution

A technique that is not easy to validate

- The validity of this technique holds first and foremost on the quality of the argument
  - Convince the reader that the instrument influences the biased independent variable and that it influences only the latter
- There are some statistical tests. But these tests presuppose that instruments are valid.
- The validity of tests is at best (generally) a necessary condition for showing the quality of a regression with instrumental variables.
- But it is not sufficient

The argumentative strategy

- 1. The argument
- 2. Showing that the instrument is as good as a random assignment
  - Non correlation with other independent variables
- 3. Is the variable suspected of being endogenous really endogenous?
  - Wu-Hausman Endogeneity Test
- 4. Are the instruments really exogenous?
  - Sargan exogeneity test of instruments (or of joint validity of two instruments)
- 5. Are the instruments powerful enough to correct the bias?
  - Weakness of instruments

2. As good as random

- We study the correlations (with correlation coefficients, simple Student T-Tests or regression) between the instrument(s) and the other non-biased observable independent variables in order to show there's no correlation
  - => Showing that the instrument is as good as a random assignment
  - Ex: Is the distribution of neighbors' month of birth linked to ego's characteristics?
- If the instrument is correlated to observable variables (which is not a problem per se as introducing them as control variable would correct the problem), the instrument is also likely to be correlated to unobservable variables.
- But the fact that the instrument is not correlated to observable variables does not prove it is unrelated to unobservable variables…
- And sometimes, there can be correlations with some observable variables for good reasons (random assignment within clusters)

The month of birth as good as random?

Table A4

Adolescents' Characteristics and Neighbours' Dates of Birth

Individual characteristics	Distribution of dates of birth of 15-*year-old* neighbours	
	Prop. born January–May	Prop. born June–November
Father college grad.	42.4 (0.7)	49.7 (0.7)
Father not college grad.	42.3 (0.3)	49.6 (0.3)
Born January–May	42.7 (0.4)	49.2 (0.4)
Born June–November	42.0 (0.4)	49.9 (0.4)
Boy	42.3 (0.6)	49.0 (0.6)
Girl	42.2 (0.6)	50.6 (0.6)
French	42.3 (0.2)	49.6 (0.2)
non-French	41.7 (1.1)	50.9 (1.1)

Source: LFS $t = 1991, \ldots 2002$, Insee. *Sample*: 15-year-old respondents, observed at t and $t + 1$, who have been living in their neighbourhood for more than one year.
Note: The average proportion of peers born in January–May is 42.3% for boys and 42.2% for girls.

The month of birth as good as random? (2)

Table A3

Relationships Between an Adolescent's Characteristics and the Distribution of Dates of Birth of Other Adolescents in the Neighbourhood

Independent variables	Dependent variables: Proportion neighbours born January–May	Proportion neighbours born June–December	Neighbours' average month of Birth
Date of birth (continuous specification	–	–	0.005 (0.005)
Date of birth (dummies)			
Born January–May	0.008 (0.010)	−0.006 (0.010)	–
Born June–November	0.001 (0.010)	0.001 (0.010)	–
December	Ref.	Ref.	–
Boy	−0.008 (0.005)	0.013 (0.05)	0.001 (0.036)
Non-French	0.002 (0.012)	0.001 (0.012)	−0.007 (0.008)
Father's education			
College grad.	−0.004 (0.009)	0.005 (0.009)	0.009 (0.063)
High-school grad.	0.010 (0.010)	−0.011 (0.011)	−0.051 (0.072)
Vocational	Ref.	Ref.	Ref.
No Dip.	−0.010 (0.009)	0.008 (0.009)	0.022 (0.063)
Missing	−0.003 (0.010)	0.007 (0.010)	0.028 (0.070)
Mother's education			
College grad.	−0.000 (0.009)	0.004 (0.009)	0.009 (0.063)
High-school grad.	0.006 (0.009)	−0.003 (0.010)	−0.025 (0.066)
Vocational	Ref.	Ref.	Ref.
No Dip.	−0.002 (0.009)	−0.006 (0.009)	0.050 (0.059)
Missing	−0.002 (0.009)	0.016 (0.009)	0.077 (0.063)
No. Obs.	13,116	13,116	13,116
R^2	0.002	0.003	0.001
Fisher (4 dummies father Educ. = 0)	0.93 (0.42)	0.83 (0.47)	0.46 (0.76)
Fisher (4 dummies mother Educ. = 0)	0.17 (0.91)	0.28 (0.83)	0.42 (0.74)

Source: LFS $t = 1991, \ldots 2002$, Insee. *Sample*: 15-*year-old* respondents, observed at t and $t + 1$, who have been living in their neighbourhood for more than one year. All regressions include a set of eleven years dummies as additional control variables. Standard deviation in parenthesis.

3. Wu-Hausman endogeneity test

- The test is the following: we compare OLS estimation with Instrumental Variable regression. If the parameter estimates are not different than there's no endogeneity problem.
- Simple implementation with augmented regression technique
  - Instead of replacing the endogenous variable with its first stage prediction, we introduce in the first stage residual as supplementary control variable along with the biased variable
  - First stage : $x_{endo} = a_0 + a_1.z_{inst} + a_2.x_2 + u_{prem}$
  - Second stage : $y = a_{est} + b_{est}x_{endo} + c_{est}.x_2 + \boldsymbol{d_{est}.u_{prem}} + u_{deux}$

Wu-Hausman Test (2)

- If the first stage residual u_{prem} is significant in the second stage equation, the suspected endogenous variable was really endogenous, and we had good reason to instrument
- If the residual is not significant, this suspected endogenous was not endogenous.
- It's better to use OLS estimates rather than instrumental variable estimates because OLS are more precise

- Limits: In order to conduct this test, we need a good instrument

Goux and Maurin: Wu Hausman Test

```
Call:
lm(formula = RET15 ~ VRET15 + S1 + res, data = gm3)

residuals:
    Min      1Q  Median      3Q     Max
-0.5960 -0.4040 -0.2779  0.5356  0.7521

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.281787   0.042029   6.705 2.06e-11 ***
VRET15      0.158611   0.101755   1.559    0.119
S1          0.113309   0.006237  18.168  < 2e-16 ***
res         0.066532   0.102184   0.651    0.515
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual standard error: 0.4818 we 23940 degrees of freedom
Multiple R-squared: 0.03663,    Adjusted R-squared: 0.03651
F-statistic: 303.4 we 3 and 23940 DF,  p-value: < 2.2e-16
```

4. Sargan's test of instruments joint-validity (or "exogeneity", or "over-identification")

- Are instruments correct? Are they truly exogenous, ie non-correlated with u?
- Condition: we need to have at least one instrument more than the number of endogenous variable.
- When one variable is endogenous, we need two instruments

Sargan's Test of instruments joint validity

- We estimate the following second stage augmented regression

$$y = a_{est} + b_{est}.x_{endo} + c_{est}.x_2 + \boldsymbol{d_{est}.u_{first}} + u_{two}$$

- We get the residual u_{two} and we regress this residual on the instruments z_{inst1} and z_{inst2}.

$$u_{two} = f.z_{inst1} + g.z_{inst2} + w \quad (2)$$

- If one of the variables is significant, it means that it is not exogenous: it is correlated to the residual, which goes against the hypotheses.
- A summary of Sargan Test for all exogenous variable is given by the following statistics for the regression (2) :
  - N*R2 that we compare to a Chi2 law with the following degree of freedom (Nb instruments- Nb endogenous variables)
- If this test is significant, instruments are not (all) exogenous

Limits of Sargan Test

- It's already difficult to find one instrument… finding two is even more difficult.
  - There's some "tips" for obtaining many instruments from one variable like squaring the variable, or cutting the instrument in groups, etc.
- Sargan Test only establishes whether two instruments are correcting a variable the same way
  - A least one instrument should be correct
    - Two bad instruments can pass positively Sargan test ! (Cf. simulation)
  - The compliers (populations reacting to the instrument) should be the same
    - Two good instruments but impacting different compliers may no pass positively Sargan's joint validity test
    - Indeed the IV estimator only estimates Local Average Treatment Effect (LATE) and not the Average Treatment Effect (ATE)
    - Ex: size of a family explained by birth of twins or by the sex ratio of the two first child.

Goux and Maurin: instruments' joint-validity

```
Call:
lm(formula = et2b$residuals ~ gm2$VJANJUN + gm2$VJULNOV)

residuals:
    Min      1Q  Median      3Q     Max
-0.6056 -0.4048 -0.2784  0.5354  0.7547

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.01912    0.01683   1.136    0.256
gm2$VJANJUN -0.01669    0.01850  -0.902    0.367
gm2$VJULNOV -0.02594    0.01892  -1.371    0.170

residual standard error: 0.4818 we 23941 degrees of freedom
Multiple R-squared: 8.91e-05,    Adjusted R-squared: 5.572e-06
F-statistic: 1.067 we 2 and 23941 DF,  p-value: 0.3442
```

Goux and Maurin: Joint Validity of instruments

- We know the R2=0.0000891 and the size N= 23944

```
stat<- 0.0000891*23944
stat
```

=> 2.133

```
1-pchisq(2.133,df=1)
```

=> 0.144

The test is not significant. Instruments are not endogenous. They are correct (or at least jointly valid).

5. Weakness of instruments

- IV estimator enables to estimate the true parameters when we converge to the infinite.
- On a small sample, there can still be a substantial bias.
- The instrumental variables need to have enough explanatory power on the endogenous variable, otherwise on a a finite sample the IV estimator behaves as the OLS estimator but somehow worse (less precise).
- In that last case, we have a **weak instrument** problem.

Weakness of instruments indicator

- Up to now, there's no "statistical test" in a strict sense, but rules of detection.
- Rule of thumb: (F<10)
- If the Fisher statistics of instruments joint nullity in the first stage regression is below 10, then we have weak instruments.

Weakness of instruments

F Test

```
Analysis of Variance Table
Model 1: VRET15 ~ S1
Model 2: VRET15 ~ VJANJUN + VJULNOV + S1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  23942 2677.5
2  23940 2655.1  2    22.419 101.07 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A famous criticism (Bound, Jaeger, Baker, 1995)

- Columns 3 to 6. Angrist and Krueger replication. Weak Instruments
- Column 2 Simpler model with only trimester of birth as instruments

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

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) OLS	(6) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)	.063 (.000)	.060 (.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, 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).

6. Average effect or local effect

- OLS measures Average Treatment Effect (ATE), which are eventually biased
- Instrumental variable regression estimates the effect of the treatment for the compliers who react to the instrument.
- Those unbiased effects are not necessary the average treatment effect but only the Local Average Treatment Effect: LATE.
- Ex. Bound & Alii. Effect of one supplementary year of school: +14% in wage.
- But effect of one supplementary year of school maybe stronger around 16 than afterwards.

A LATE effect

- Instrument
  - Estimated on compliers
  - But not on always takers
  - Or never takers
- LATE= (3+4+6+4+4)/5

  = 4.2
- ATE= (3+2+4+3+6+6+4+4+3)/9

  =3.9

Observation	Y i (0)	Y i (1)	Y i (1) − Yi(0)	di (z=0)	di (z=1)	Type
1	4	7	**3**	0	1	Complier
2	3	5	2	0	0	Never-taker
3	1	5	**4**	0	1	Complier
4	5	8	3	1	1	Always-taker
5	4	10	**6**	0	1	Complier
6	2	8	6	0	0	Never-taker
7	6	10	**4**	0	1	Complier
8	5	9	**4**	0	1	Complier
9	2	5	3	1	1	Always-taker

4. Instrumental Variables. An assessment

Finding instrumental variables: that is the question?

- From wild search of instruments
  - (Unspeakable) data mining research of instruments that pass tests
  - Once found -> telling a story more or less convincing why we should use this instrument
  - Ex: Return to education => instrumenting ego's education by parental education
  - Hypothesis: all the effect of parental education on wages go only through child's education.
  - Wild search instruments is disappearing
- Research of "natural experiments"

Examples of instrumental variables based on natural experiments

- Sex ratio.
  - Random (if there's no gendered abortion)
  - Has an incidence on many possibly endogenous behaviors : Number of children
  - Enables to estimate the effect of the number of children on female activity, divorce, etc.
- Weather
  - Random
  - Has an influence on agricultural output
  - Supply and demand model

Instrumental variables (examples)

- Date of birth
  - Random assignment during the year (not totally: there's some birth cyclicity).
  - Effect on education and other aspects
- Public policy measures
  - Threshold of activation of a public policy
  - People are more or less randomly around the threshold of activation
  - Distance to the threshold can be used as an instrument
  - Ex : Aurelie Ouss (Maurin, Ouss, 2009), effect of one year sentence reduction on repeating the offense
    - 14 July in France, traditional date of automatic sentence reduction.
    - If the end of the sentence is before July 14th, no automatic sentence reduction.
    - After sentence reduction
  - (official date of exit – July 14th) => instrument for sentence reduction

Instrumental variables (examples)

- Spatial localization
  - Spatial position is to some extent exogenous.
  - Influence of slavery in Africa in the 17th/18th on interpersonal trust in 2007(Nunn and Wantchekon 2011)
    - Mechanism: slave trade destroyed interpersonal trust in societies submitted to slavery
    - Inverse causality problem: African groups engaging in slave trade maybe had already very low interpersonal trust.
    - Instrument: Proximity to the seaside as an instrument for slave trade.

Use in sociology – In progress

- Effect of magazines on the development of anti-slavery groups in the US during the 19th (King, Haveman, 2008)
  - problem: reverse causality. The presence of anti-slavery groups may have sponsored the development of the press
  - Instrument: number of post offices
- Effect of the network position on the probability to get a job in academia. (Godechot, Mariot, 2004)
  - Problem: the network position captures the PhD quality that we don't measure well
  - Instrument: position on the network of other doctors of the same supervisor (social capital exogenous to PhD quality).
  - Limit: match PhD student-director could be quality based

Use in sociology

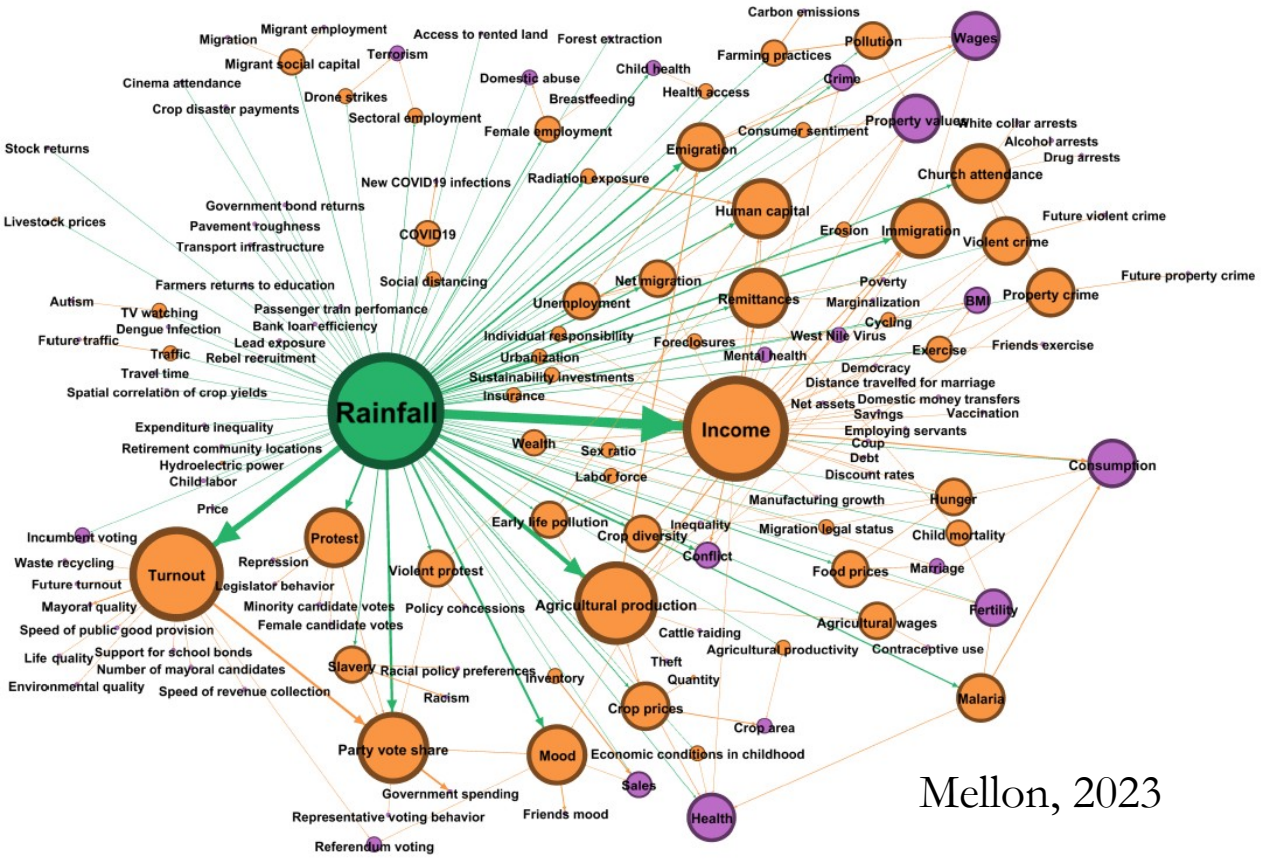
- Nationality acquisition on employment. (Fougère & Safi, 2006)
  - Reverse causality problem: employment -> cause of nationality acquisition
  - Instruments: number of foreigners living in the same département in the census, and the number of strangers of the same origin living in the same département in the census.
  - These two variables impact the length of waiting queues for people applying to French citizenship, and therefore the individual probability of acquiring French citizenship between two census.
- Effect of discrimination feeling on satisfaction (Safi, 2010)
  - Plausible reverse causality problem: Happy people don't feel discriminated
  - Instrument: religious membership to minority religions (Judaism or Islam)
  - Hypothesis: it only affects discrimination feeling. No direct effect on satisfaction
  - Limit: we can discuss whether this last hypothesis is a reasonable proxy.

Instrumental variables: from apology to doubts

- Limits of the technique
  - More complex econometrics
  - Difficult to find
  - "Tip" style of research ("Age of the captain")
  - Not always truly exogenous
  - Exogeneity difficult (impossible) to fully prove
  - Potentially weak
  - Only estimating local effects
  - Fairly unstable and not very powerful on small samples
- Is there an improvement vis-à-vis a biased but consistent OLS regression?
  - Debatable
- Evolution in economics / social sciences
  - From systematic research of instruments
  - To randomized controlled trials.
    - A treated group
    - A control group
  - If the random assignment is unbiased, we measure directly the treatment effect by a simple mean difference and significance with a simple student T test

Doubts

- Rainfall as an instrument (Mellon, 2023) → Exclusion hypothesis
- Lal et al., 2023
  - IV often overestimates OLS event if prediction go otherwise
  - → Exclusion hypothesis ?



Mellon, 2023

5. Programs

With R : ivreg function in AER package

```
#installation of AER package
install.packages("AER")
library("AER")

#Syntax
myreg<-ivreg(y ~ x_endo+x2+x3|instr+x2+x3,data=db)
summary(myreg)

#How to have all tests
summary(myreg,diagnostics=TRUE)

#Limit: does not print first stage regression… -> to estimate separately with
   lm
```

ivreg with diagnostics

```
Call:
ivreg(formula = RET15 ~ VRET15 + S1 | VJANJUN + VJULNOV + S1,
    data = gm2)
residuals:
    Min      1Q  Median      3Q     Max
-0.5537 -0.3951 -0.2976  0.5520  0.7182
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.281787   0.042072   6.698 2.16e-11 ***
VRET15      0.158611   0.101860   1.557    0.119
S1          0.113309   0.006243  18.150  < 2e-16 ***
Diagnostic tests:
                  df1   df2 statistic p-value
Weak instruments    2 23940   101.071  <2e-16 ***
Wu-Hausman          1 23940     0.424   0.515
Sargan              1    NA     2.129   0.145
```

With R : systemfit

```
#installation of package systemfit
install.packages("systemfit")
library("systemfit")

#Syntaxe
first_st <- x_endo ~ instr+x2+x3
second_st <- y ~ x_endo+x2+x3
system <- list( first_st, second_st)
inst <- ~ instr+x2+x3
fit2sls <- systemfit( system, "2SLS",inst, data=db)
summary(fit2sls)
```

With Stata

- Ivregress & ivprobit

  ivregress 2sls y x2 x3 **(**x_endo**=**instr1 instr2**)**

  ivprobit y x2 x3 **(**x_endo**=**instr1 instr2**)**

  ivregress 2sls y x2 x3 **(**x_endo1 x_endo2**=**instr1 instr2**)**
- First stage regression

  ivregress 2sls y x2 x3 **(**x_endo**=**instr1 instr2**), first**
- Endogeneity test

  estat endogenous
- Overidentification test

  estat overid
- Detection of weak instruments

  estat firststage

With SAS

```
proc syslin 2SlS data=mabase FIRST;
model y = x_endo x2 x3 /overid ;
endogenous y x_endo;
instruments z1 z2 x2 x3 ;
run;
```

SPSS

- For SPSS users : 2SLS
  - In script mode :

    2sls y with x w

    / instruments z w

    / constant.

  - With Menus
    - Analyze → Regression → Two-Stage Least Squares
    - DEPENDENT, EXPLANATORY, and INSTRUMENTAL

References

Angrist, Joshua D, and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106 (4): 979–1014.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50.

Fougère, Denis, and Mirna Safi. 2005. "L'acquisition de La Nationalité Française: Quels Effets Sur l'accès à l'emploi Des Immigrés?" In *France, Portrait Social, 2005-2006*, 260–82. Insee.

Godechot, Olivier, and Nicolas Mariot. 2004. "Les Deux Formes Du Capital Social: Structure Relationnelle Des Jurys de Thèses et Recrutement En Science Politique." *Revue Française de Sociologie* 45 (2): 243–82.

Goux, Dominique, and Éric Maurin. 2005. "Composition Sociale Du Voisinage et Échec Scolaire: Une Évaluation Sur Données Françaises." *Revue Économique*, no. 2: 349–61.

Goux, Dominique, and Eric Maurin. 2007. "Close Neighbours Matter: Neighbourhood Effects on Early Performance at School." *The Economic Journal* 117 (523): 1193–1215.

King, Marissa D, and Heather A Haveman. 2008. "Antislavery in America: The Press, the Pulpit, and the Rise of Antislavery Societies." *Administrative Science Quarterly* 53 (3): 492–528.

Lal, Apoorva, Mac Lockhart, Yiqing Xu, and Ziwen Zu. 2023. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on over 60 Replicated Studies." *arXiv Preprint arXiv:2303.11399*.

Maurin, Éric, and Aurélie Ouss. 2009. "Sentence Reductions and Recidivism: Lessons from the Bastille Day Quasi Experiment." IZA Discussion Papers 3990. IZA Institute of Labor Economics.

Mellon, Jonathan. 2023. "Rain, Rain, Go Away: 195 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable." *Available at SSRN 3715610*.

Nunn, Nathan, and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101 (7): 3221–52.

Safi, Mirna. 2010. "Immigrants' Life Satisfaction in Europe: Between Assimilation and Discrimination." *European Sociological Review* 26 (2): 159–76.

Wright, Philip Green. 1928. *The Tariff on Animal and Vegetable Oils*. New York: The MacMillan Company.