REGRESSION ANALYSIS AND CAUSALITY WITH R

Instrumental Variables

João Cerejeira[1] Miguel Portela[1,2,3]

[1]NIPE – UMinho
[2]IZA, Bonn
[3]Banco de Portugal

October 26, 2021

Introduction

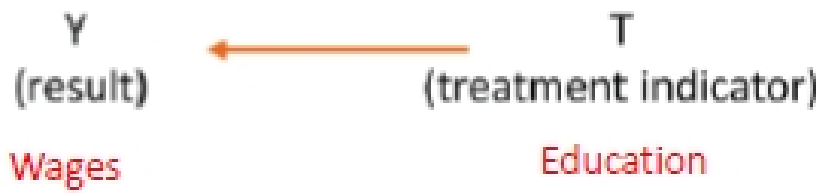
How to get a consistent estimator of β if X is correlated with ε?

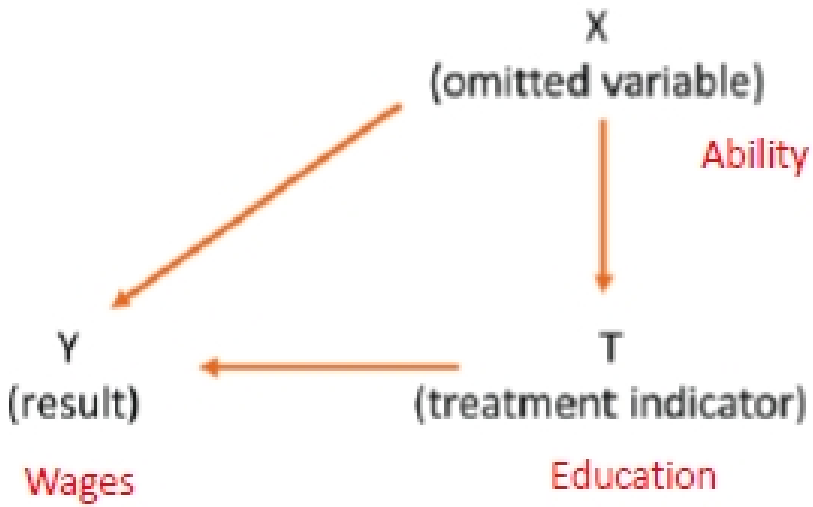
$$Y = X\beta + \varepsilon, \tag{1}$$

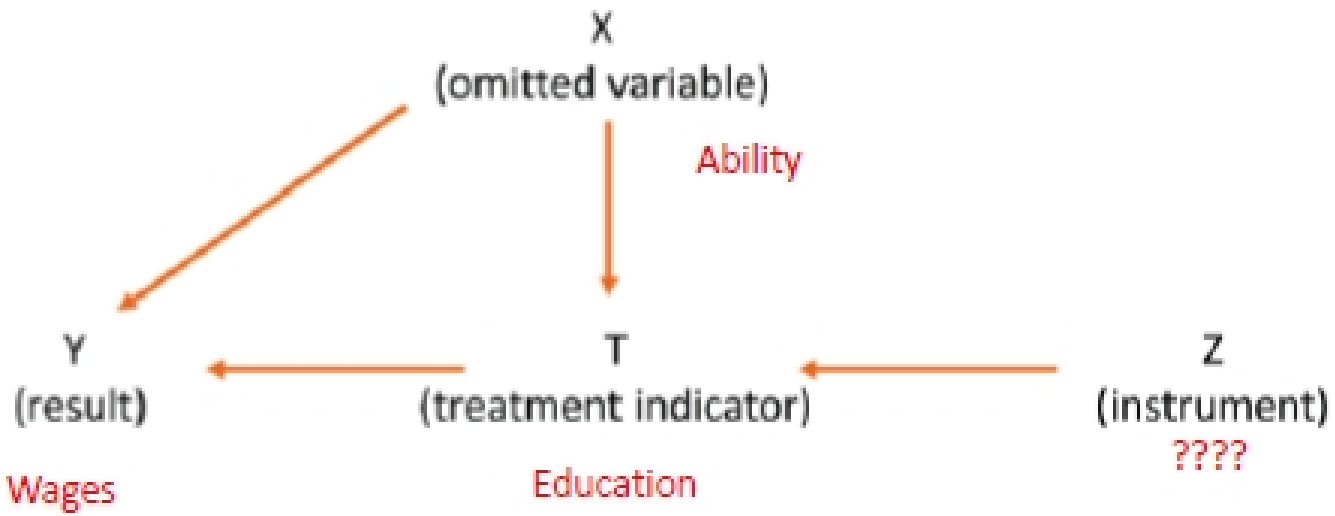
Intuition: total variation of X can be divided in two parts:

- one part correlated with ε;
- other uncorrelated with ε.

IV strategy is to use just the second part of X variation.


Y
(result)
T
(treatment indicator)
Wages
Education


X
(omitted variable)
Ability
Y
(result)
Wages
T
(treatment indicator)
Education


X
(omitted variable)
Ability
Y
(result)
Wages
T
(treatment indicator)
Education
Z
(instrument)
????

Notation

Z $(N \times l)$ is the instruments matrix;
X $(N \times k)$ is the explanatory variables matrix;
If $k = l$, just-identified case;
If $k < l$, over-identified case;
If $k > l$ under-identified case - cannot be estimated.
All exogeneous variables in X should be included in Z.
All other variables in X are endogenous.
Variables in Z not included in X are called exogenous variables.

Conditions for instrument validity

Must be correlated with X - testable:

$$COV(Z_i, X_i) \neq 0,$$

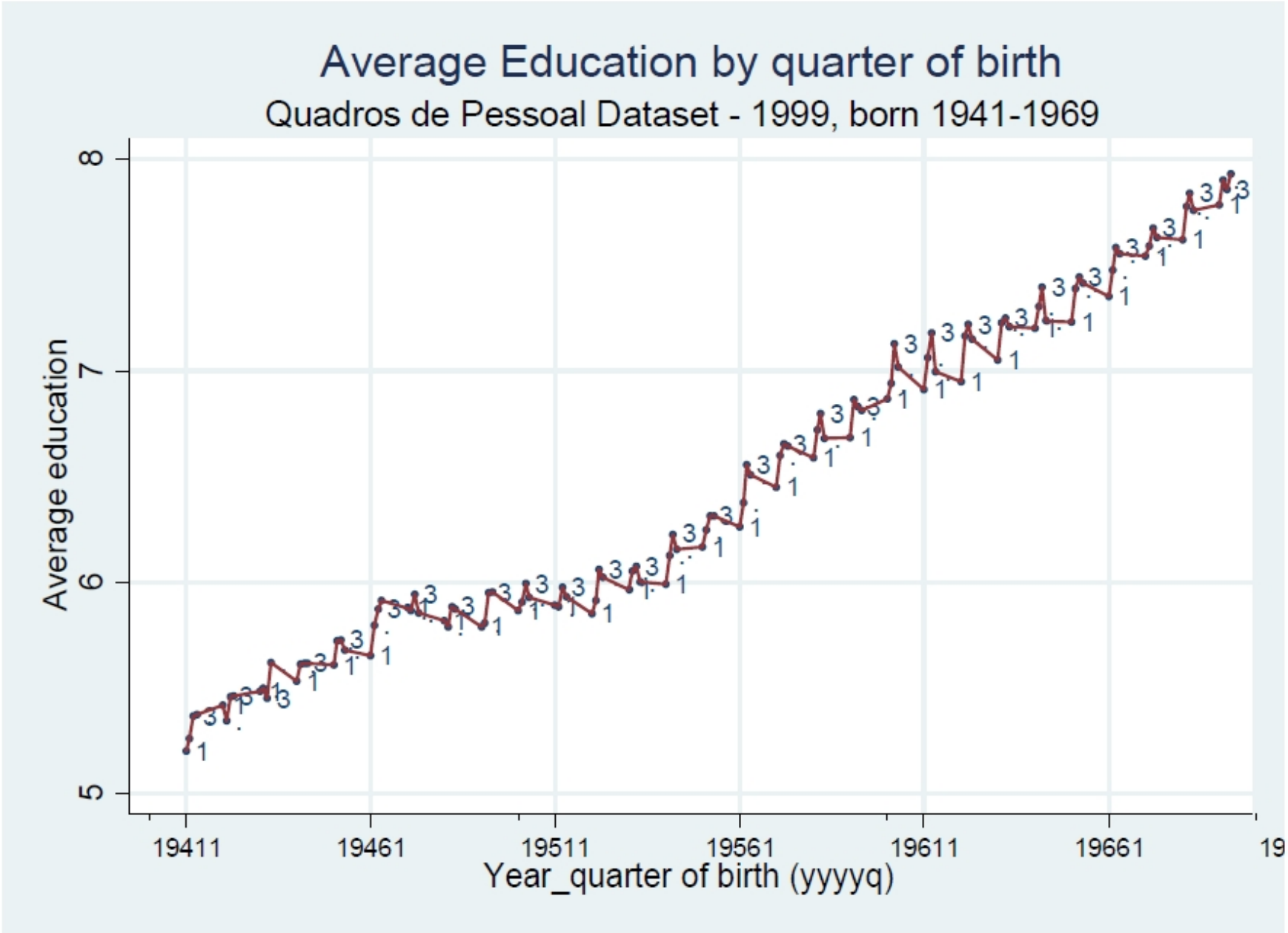
Must be uncorrelated with 'error' – untestable if $k = l$ – have to argue case for this assumption:

$$COV(Z_i, \varepsilon_i) = 0.$$

These conditions guaranteed with instrument for experimental data. But more problematic for data from quasi-experiments.

Example: Causal effect of education on wages
Instrument: quarter of birth
(QP dataset, 1999, born 1940-1970)

Quarter of birth	Educ Mean	Freq.
1	6.555	383,593
2	6.669	360,218
3	6.749	352,951
4	6.684	367,036
Total	6.662	1,463,798


Average Education by quarter of birth
Quadros de Pessoal Dataset - 1999, born 1941-1969
Average education
5
6
7
8
19411
19461
19511
19561
19611
19661
Year_quarter of birth (yyyyq)

2SLS (Two stage least squares method)

First Stage

Regress X on Z:

$$X = Z\Pi + v, \tag{2}$$

therefore

$$\widehat{\Pi} = (Z'Z)^{-1}Z'X, \tag{3}$$

predict $\widehat{X}$:

$$\widehat{X} = Z\widehat{\Pi} = Z(Z'Z)^{-1}Z'X, \tag{4}$$

2SLS (Two stage least squares method)

Second Stage

Regress Y on $\widehat{X}$:

$$Y = \widehat{X}\beta_{IV} + u, \tag{5}$$

$$\begin{aligned}
\widehat{\beta}_{IV} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y = \\
&= ((Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X))^{-1}(Z(Z'Z)^{-1}Z'X)'Y = \\
&= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z')Y = \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y
\end{aligned} \tag{6}$$

2SLS (Two stage least squares method)

If $l = k$, then $X'Z$ is an invertible square matrix:

$$\begin{aligned} \widehat{\beta}_{IV} &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'Y = && (7) \\ &= (Z'X)^{-1}Z'Y && (8) \end{aligned}$$

Assimptotic variance of IV estimator

$$Y = \widehat{X}\beta_{IV} + u, \tag{9}$$

and

$$\widehat{X} = Z(Z'Z)^{-1}Z'X, \tag{10}$$

then:

$$\begin{aligned} Var(\widehat{\beta}_{IV}) &= \widehat{\sigma}^2(\widehat{X}'\widehat{X})^{-1} = \widehat{\sigma}^2((Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X))^{-1} = \\ &= \widehat{\sigma}^2(X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1} = \widehat{\sigma}^2(X'Z(Z'Z)^{-1}Z'X \end{aligned}$$

If $l = k$:

$$Var(\widehat{\beta}_{IV}) = \widehat{\sigma}^2(X'Z)^{-1}Z'Z(Z'X)^{-1}.$$

Assimptotic variance of IV estimator

The variance of $\widehat{\beta}_{IV}$ will be large if the correlation between X with Z is small. This is the"weak instruments" problem. Note: if $X = Z$, then $Var(\widehat{\beta}_{IV}) = Var(\widehat{\beta}_{OLS})$. Therefore $Var(\widehat{\beta}_{IV}) \geq Var(\widehat{\beta}_{OLS})$.
Note that $\widehat{\sigma}^2$ is estimated using:

$$\widehat{\sigma}^2 = (Y - X\widehat{\beta}_{IV})'((Y - X\widehat{\beta}_{IV}))/(N - K),$$

not:

$$\widehat{\sigma}^2 = (Y - \widehat{X}\widehat{\beta}_{IV})'((Y - \widehat{X}\widehat{\beta}_{IV}))/(N - K).$$

The Problem of Weak Instruments

- Say that instruments are 'weak' if correlation between X and Z low (after inclusion of other exogenous variables).
- Rule of thumb - if F-statistic on instruments in first-stage less than 10 ($t < 3.33$) then may be problem.
- A whole range of problems tend to arise if instruments are weak.

The Problem of Weak Instruments

- Asymptotic problems:

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.
  - Small departures from instrument exogeneity lead to big inconsistencies.

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.
  - Small departures from instrument exogeneity lead to big inconsistencies.
- Finite-Sample Problems:

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.
  - Small departures from instrument exogeneity lead to big inconsistencies.
- Finite-Sample Problems:
  - Small-sample distirbution may be very different from asymptotic one.

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.
  - Small departures from instrument exogeneity lead to big inconsistencies.
- Finite-Sample Problems:
  - Small-sample distirbution may be very different from asymptotic one.
  - May be large bias.

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.
  - Small departures from instrument exogeneity lead to big inconsistencies.
- Finite-Sample Problems:
  - Small-sample distirbution may be very different from asymptotic one.
  - May be large bias.
  - Computed variance may be wrong.

The Problem of Weak Instruments

- Asymptotic problems:
  - High asymptotic variance.
  - Small departures from instrument exogeneity lead to big inconsistencies.
- Finite-Sample Problems:
  - Small-sample distirbution may be very different from asymptotic one.
  - May be large bias.
  - Computed variance may be wrong.
  - Distribution may be very different from normal.

Testing overidentification restrictions

If the number of excluded instruments is equal to the number of endogenous variables it is not possible to test if these instruments are correlated with the error term.
In the overidentified case, we can test if some instruments are correlated with the error, but we don't know which.
To construct this test you:
- estimate the model by IV and get the residuals;
- run a regression of these residuals on all exogenous variables and get $n \times R^2$.
- under the null that all instruments are not correlated with the disturbance term, the LM statistic of $n \times R^2$ follows a $\chi^2(l-k)$ distribution where $l-k$ is the number of restrictions of overidentification (the number of instruments in excess).