Applied Econometrics
Lecture 3

Giovanni Marin[1]

[1]Università di Urbino

Università di Urbino
PhD Programme in Global Studies
Spring 2018

Instrumental variable approach

- A more 'classical' approach to solve the issue of endogeneity (and of selection bias) is to take an instrumental variable approach
- In brief, the idea of instrumental variables is to identify one or more variables (instrumental variables) that do not belong to the theoretical model that you want to test but are correlated with the 'exogenous part' of your endogenous variable
- The use of an instrumental variable allows to exploit only part of the variation of the endogenous variable $\Rightarrow$ the part that is not correlated with the omitted variable
- Given that only part of the variation is used, IV estimates are always less efficient than OLS estimates $\Rightarrow$ less precise, higher standard errors, lower t-stat, higher p-value
- If you believe that your variable of interest is not characterized by endogeneity issues $\Rightarrow$ non-IV estimates should be your baseline (and, eventually, IV should be your robustness check)

The mechanics of the IV

- IV regression allows to keep just the variance of the endogenous variable that is not correlated with the error term
- In linear models, the instrumental variable estimator is estimated with the two stages least squares (2SLS):
- Let the true model be $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ where:
  - x_1 is an endogenous regression $\Rightarrow corr(x_1, \varepsilon) \neq 0$
  - x_2 is an exogenous regression $\Rightarrow corr(x_2, \varepsilon) = 0$
- The objective of the IV approach is to find one or more variables z, that do not belong to the set of covariates in the true model, that are correlated with the endogenous variable x_1 but uncorrelated with the error term ε
- This is equivalent to assuming that the only influence that we should expect of z on y passes through the impact of z on $x_1 \Rightarrow$ exclusion restriction

The mechanics of IVs

- The 2SLS approach works as follows:
  1. In the first stage you estimate an OLS model in which you have as dependent variable the endogenous variable x_1 and as independent variables the set of exogenous variables from the full model (x_2 in our case) and the exogenous instrumental variables (only z in our case) $\Rightarrow x_1 = \gamma + \delta x_2 + \eta z + \xi$
  2. You estimate the predicted value of the endogenous variable as estimated in the first stage $\Rightarrow \hat{x_1} = \hat{\gamma} + \hat{\delta} x_2 + \hat{\eta} z$
  3. You estimate your true model in which you substitute the independent variable x_1 with its predicted value in the first stage $\hat{x_1}$ $\Rightarrow y = \alpha + \beta \hat{x_1} + \gamma x_2 + \varepsilon$
     - By using $\hat{x_1}$ instead of x_1 in the second stage, you will just keep the variance of x_1 that is not correlated with the error term ε and is, in turn, correlated with the IV z
- If your instrument is valid, $\hat{\beta}$ is the unbiased estimate of the true parameter β

The mechanics of IVs

- All explanatory variables included in the first stage are called instruments
  - Some variables (e.g. x_2) are also part of the true model $\Rightarrow$ they should appear both in the first and in the second stage
  - The variable z (one or more) only appear in the first stage $\Rightarrow$ also defined 'excluded IVs'
  - In order for the model to be exactly identified, the number of excluded IVs (the zs) should be equal to the number of endogenous explanatory variables
  - In case you have two endogenous explanatory variables, in order for your model to be exactly identified you need to find two 'excluded IVs'
  - If you have two endogenous explanatory variables but just one excluded IV, the model is not identified (and cannot be estimated)
  - If you have more IVs than endogenous explanatory variables, your model is overidentified
  - Overidentification is particularly useful as it allows to test the validity of the instruments that you are using

The mechanics of IVs

- An alternative way of estimating the instrumental variable model is by means of the Control Function approach
- The procedure is as follows:
  1. As before, in the first stage you estimate an OLS model in which you have as dependent variable the endogenous variable x_1 and as independent variables the set of exogenous variables from the full model (x_2 in our case) and the exogenous instrumental variables (only z in our case) $\Rightarrow x_1 = \gamma + \delta x_2 + \eta z + \xi$
  2. You estimate the predicted value of the residual the first stage $\Rightarrow \hat{\xi} = x_1 - \hat{\gamma} + \hat{\delta} x_2 + \hat{\eta} z$
  3. You estimate your true model in which you include all the structural variables (x_1 and x_2) but also $\hat{\xi} \Rightarrow y = \alpha + \beta x_1 + \gamma x_2 + \psi \hat{\xi} + \varepsilon$
- By adding $\hat{\xi}$ as a control variable, you are explicitly accounting for the unobserved (and endogenous) part of x_1, removing it from the error term

The validity of IVs: strength

- First stage: $x_1 = \gamma + \delta x_2 + \eta z + \xi$
- Second stage: $y = \alpha + \beta \hat{x}_1 + \gamma x_2 + \varepsilon$
- An instrument is valid if it is:
  - Strong
  - Exogenous
- Strong (non-weak) instruments
  - The excluded instrument(s) (z), conditional on other non-excluded instruments (x_2), is strongly correlated with the endogenous explanatory variable(s) in the first stage $\Rightarrow \hat{\eta}$ is significantly different from zero
  - The idea is that the instrument should have a significant and clear impact on the endogenous variable
  - The rule of the thumb is that the F-stat that tests the joint significance of all excluded IVs in the first stage(s) should be at least > 10

The validity of IVs: exogeneity

- First stage: $x_1 = \gamma + \delta x_2 + \eta z + \xi$
- Second stage: $y = \alpha + \beta \hat{x_1} + \gamma x_2 + \varepsilon$
- The instrument is valid if it is not correlated with the error term of the true model
- If the instrument was correlated with the error term of the true model (and if it was strong, i.e. correlated with the endogenous variable x_1), then it would imply that $\hat{x_1}$ is also correlated with the residual of the true model
- Exogeneity can be thought in two (equivalent) ways:
  - Conditional on the 'internal' instruments (i.e. the exogenous variables that belong to the true model, x_2 in our case), the external instrument is not correlated with the residuals
  - The only impact of the excluded instrument on the dependent variable y passes through the impact of the instrument z on the endogenous variable $x_1 \Rightarrow corr(y, z|x_k) = 0$
- Differently from the strength, the exogeneity of the IVs cannot be explicitly tested
- If you have more IVs than endogenous variables (overidentification), you can compare IVs in terms of exogeneity

Tips for doing IV

- Clearly identify the source of endogeneity
- What is the variable that you are omitting (including measurement error)?
- Reason about the direction of the bias that you expect based on:
  - The expected correlation between the omitted variable and the endogenous variable
  - The expected (conditional) correlation of the omitted variable and the dependent variable
- Identify the possible sources of variation that are not correlated with the omitted variable (i.e. that have no direct impact on the dependent variable but are correlated with the endogenous variable)
- Search for actual measures (even imperfect ones) of these sources of variation
- Estimate your 2SLS and cross your fingers...
- VERY IMPORTANT $\Rightarrow$ using IVs should not prevent you for adding control variables to account for (other) omitted variables!

IV in practice

- In order to do IV in an effective way, it is very important to read papers that used IVs in a clever way
- Often you will find nice ideas about IVs by reading papers unrelated to your main research question!
- Economists in the fields of labour economics and political economy are obsessed (too much sometimes...) by causality and, consequently, extremely clever in finding original IVs
- Typical case in labour economics $\Rightarrow$ estimate the returns from education, where innate ability cannot be measured

Example 1: Angrist and Krueger (1991)

Angrist JD, Krueger AB (1991) Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics 106(4):979-1014

- Objective of the paper: estimating the wage returns to schooling
- Source of endogeneity: pupils self select into continuing their education
  - Studying is easier for people with higher innate ability ⇒ those who will continue to study will self-select into education ⇒ unobserved innate ability is correlated with educational attainment
  - For a given level of education, workers with higher ability will earn more ⇒ the omitted variable is also correlated with the outcome variable
- Angrist and Krueger had to identify a source of variation of educational attainment that was not correlated with ability

Example 1: Angrist and Krueger (1991)

- In most US states (like in Italy), compulsory school starts in the calendar year in which the pupil turns six ⇒ some pupils start school (in September) when they are 6 years and 9 months old, other when they are 5 years and 9 months old
- Angrist and Krueger assume that the birth date is randomly assigned
- Students can legally drop out school the day when they turn 16
  - When turning 16, students born in the first quarter will have a lower grade than students born in the fourth quarter
  - The quarter of birth is correlated with the educational attainment (just for students that drop out at 16...) but uncorrelated with ability (and consequently on age)

Figure: First stage

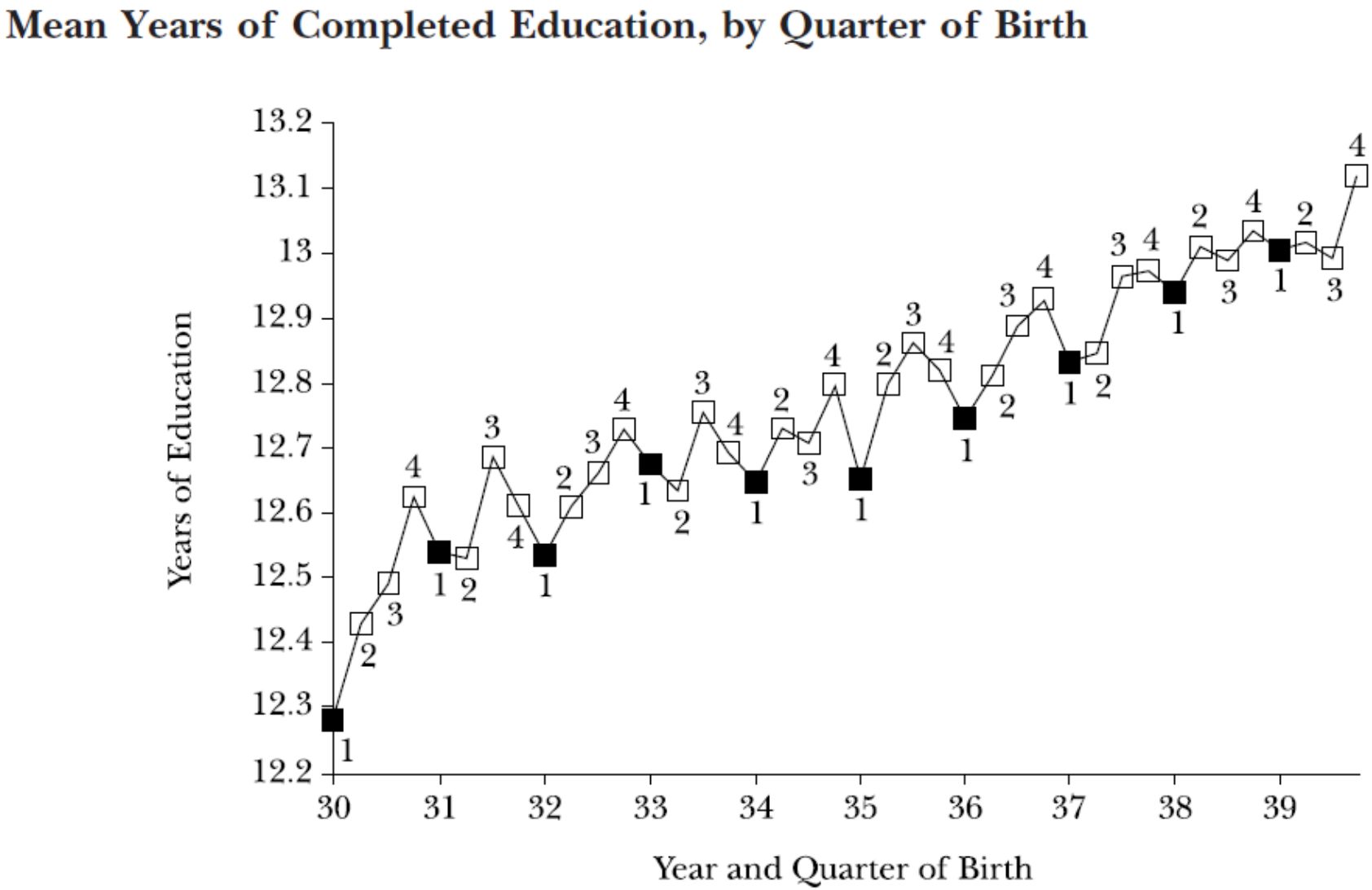

Figure: Reduced form (outcome vs IV)

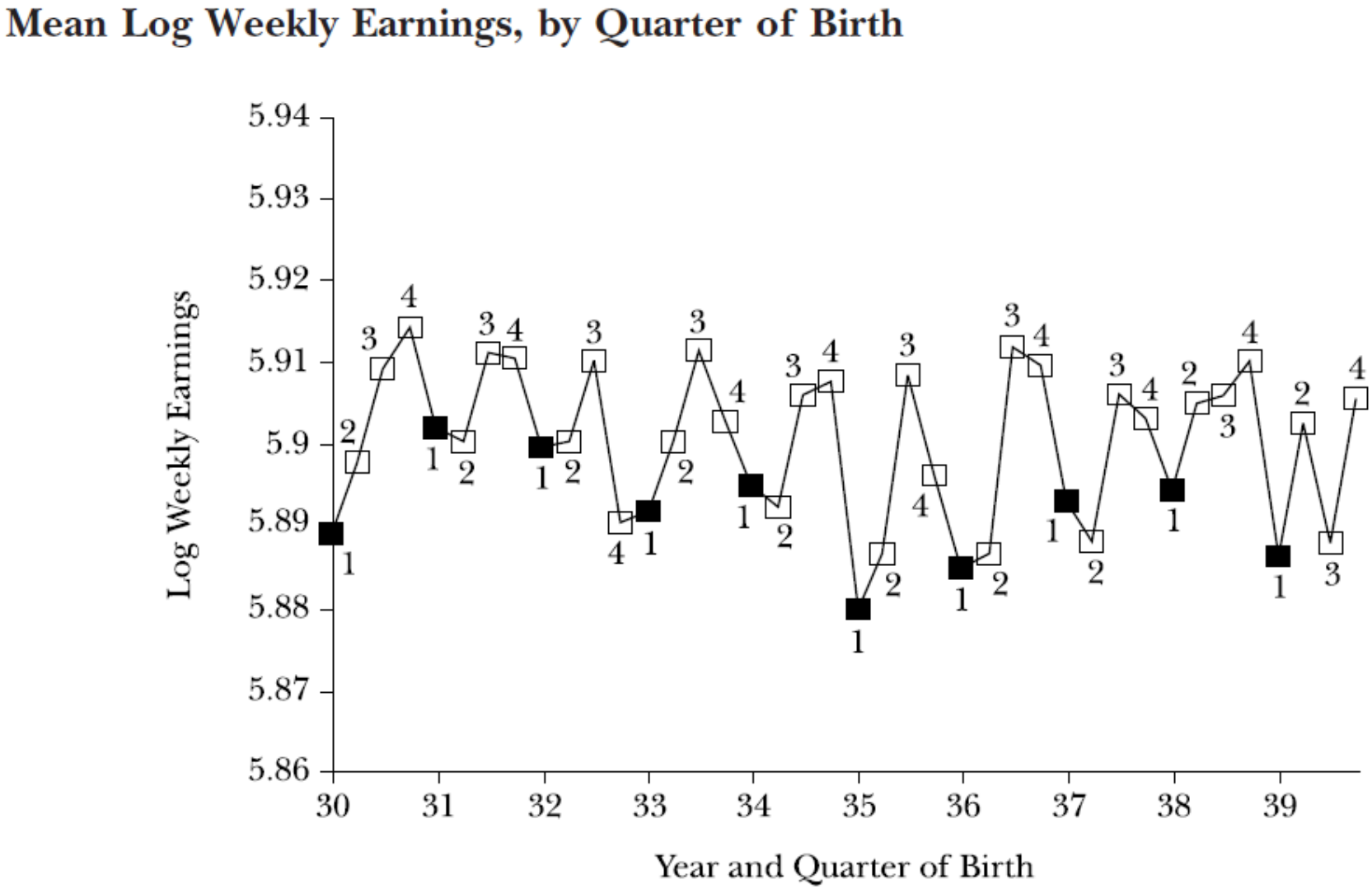

Figure: Returns to schooling: results

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.

Example 2: Acemoglu, Johnson and Robinson, 2001

Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. American Economic Review 91(5):1369-1401

- Objective of the paper: evaluate the extent to which the presence of strong institutions influence current economic performance
- Source of endogeneity: economic development in itself favours the emergence of institutions (reverse causality) and presence of omitted variables (e.g. culture)
- Need to search for a variable that is correlated with current institutions (measured with Average protection against expropriation risk 1985-1995) and not to current economic performance (GDP per capita in PPP in 1995) beyond its impact on institutions
- IV: settler mortality of European colonizers

What is the expected link between settler mortality and current institutions?

(potential) settler mortality $\Rightarrow$ settlements

$\Rightarrow$ early institutions $\Rightarrow$ current institutions

$\Rightarrow$ current performance.

What is the expected link between settler mortality and current institutions?

- Colonizers had to decide whether to settle in the colony or to just extract local resources
- The decision depended on the feasibility of settling ⇒ strongly correlated with the presence of local diseases
- If the colonizer settled in a colony, they would have tried to establish institutions to replicate the conditions of the home country
- Institutions are persistent

⇒ Data on mortality rates of soldiers, bishops and sailors stationed in colonies between 17th and 19th century

Figure: First stage

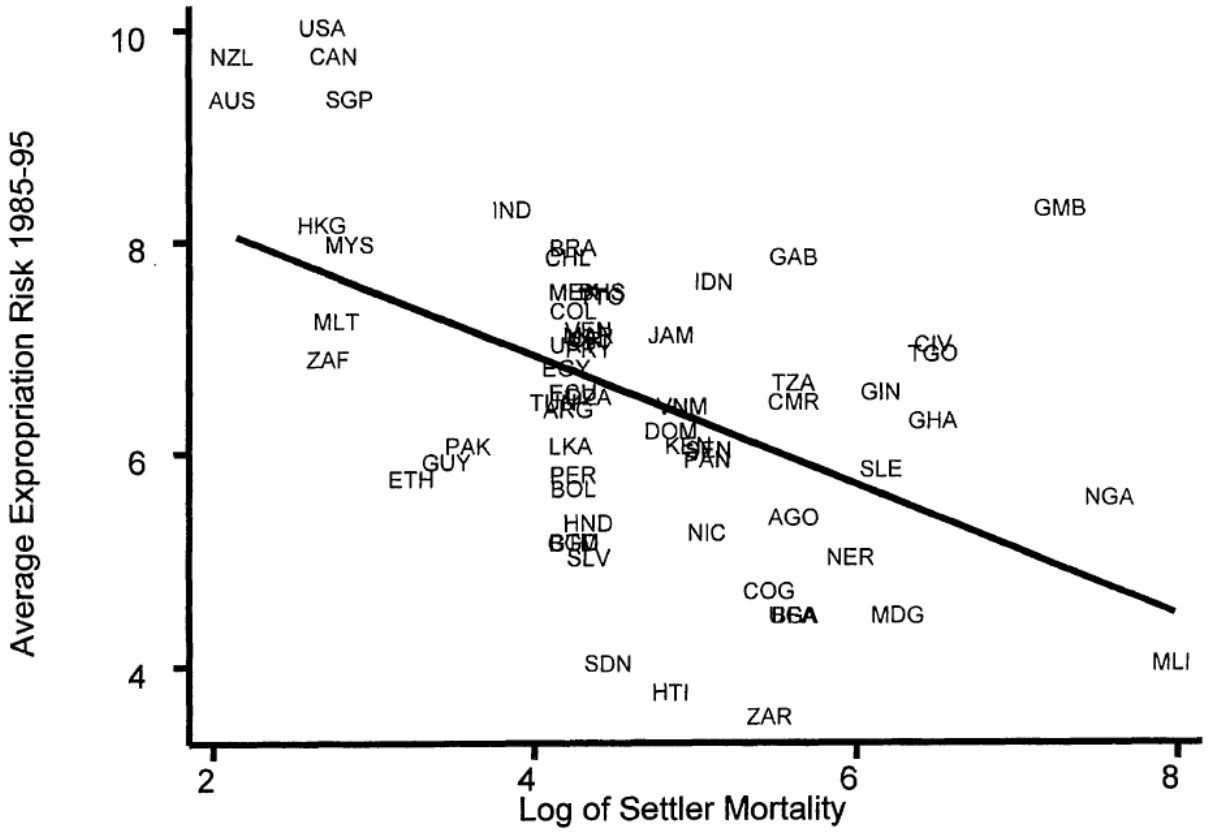

FIGURE 3. FIRST-STAGE RELATIONSHIP BETWEEN SETTLER MORTALITY AND EXPROPRIATION RISK

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk 1985–1995	0.94	1.00	1.28	1.21	0.58	0.58	0.98	1.10	0.98
	(0.16)	(0.22)	(0.36)	(0.35)	(0.10)	(0.12)	(0.30)	(0.46)	(0.17)
Latitude		−0.65		0.94		0.04		−1.20	
		(1.34)		(1.46)		(0.84)		(1.8)	
Asia dummy							−0.92	−1.10	
							(0.40)	(0.52)	
Africa dummy							−0.46	−0.44	
							(0.36)	(0.42)	
"Other" continent dummy							−0.94	−0.99	
							(0.85)	(1.0)	
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	−0.61	−0.51	−0.39	−0.39	−1.20	−1.10	−0.43	−0.34	−0.63
	(0.13)	(0.14)	(0.13)	(0.14)	(0.22)	(0.24)	(0.17)	(0.18)	(0.13)
Latitude		2.00		−0.11		0.99		2.00	
		(1.34)		(1.50)		(1.43)		(1.40)	
Asia dummy							0.33	0.47	
							(0.49)	(0.50)	
Africa dummy							−0.27	−0.26	
							(0.41)	(0.41)	
"Other" continent dummy							1.24	1.1	
							(0.84)	(0.84)	
R^2	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33	0.28
Panel C: Ordinary Least Squares									
Average protection against expropriation risk 1985–1995	0.52	0.47	0.49	0.47	0.48	0.47	0.42	0.40	0.46
	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Number of observations	64	64	60	60	37	37	64	64	61

Notes: The dependent variable in columns (1)–(8) is log GDP per capita in 1995, PPP basis. The dependent variable in column (9) is log output per worker, from Hall and Jones (1999). "Average protection against expropriation risk 1985–1995" is measured on a scale from 0 to 10, where a higher score means more protection against risk of expropriation of investment by the government, from Political Risk Services. Panel A reports the two-stage least-squares estimates, instrumenting for protection against expropriation risk using log settler mortality; Panel B reports the corresponding first stage. Panel C reports the coefficient from an OLS regression of the dependent variable against average protection against expropriation risk. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable descriptions and sources.