Université Paris-Saclay
Master of Economics – 1st year
Advanced Econometrics – Tutorial #5
Instrumental variable approach[*]

March 22nd 2021

This tutorial is dedicated to the **instrumental variables (IV) approach**, by reproducing the results of the article of Acemoglu et al. (2001). We first present the context of the article and the econometric challenge.

Tutorial context

Origins of differences in economic development between countries is debated in history and development economics. Three main sources are usually confronted: **geography** to the extent that climate, natural resources or physical constraints influence technological diffusion and agricultural production (see, Bruhn and Gallego (2011)); **international trade** which favors resource and technological transfers between countries; and **institutions** as a set of rules and norms which frame individuals actions and decisions — Acemoglu et al. (2005) and Rodrik et al. (2004).

The article by Acemoglu et al. (2001) investigates **the role of institutions on economic development among countries colonized by Europeans**. In their article, they measure the quality of institutions through the security of property rights. The main mechanisms they want to underline is that countries with more secure property rights will invest more in physical and human capital, and will use these factors more efficiently to achieve a greater level of income. However, investigating the role of institutions on development is challenging, as the relationship between institutions and income is **endogeneous**. Indeed, there is potential **reverse causality** of income on institutions, that is "rich economies choose or can afford better institutions", and there are **omitted variables** as economies may differ in their institutions and their income for a variety of unobserved reasons. Thus, to estimate the impact of institutions on economic performance, we need a source of exogenous variation in institutions, which can be done using an instrumental variable approach.

The approach developed by Acemoglu et al. (2001) is the following. They present historical arguments for the present institutions to be inherited from past institutions established by colonizing countries. They show that institutions imposed by the colonial state persisted and lasted even after the independance of colonized countries. They also state that the colonization strategy and the established institutions depend on the feasability of settlements. In places where the environment is favorable for European settlement (disease-free and with similar climate), colonized countries were more likely to establish institutions replicating

[*] Contact: thibault.richard(at)ens-paris-saclay.fr

those of Europe, while in countries with low feasability of settlements colonized countries established poor protection of property rights to extract the resources from the countries.

Thus, they use settler mortality, at the time of the colonization, as an instrument for current institutions in these countries. Note that there is no selection bias here as they are interested in the effect of colonization policy conditional on being colonized, thus focusing only on the sample of countries which were colonized. In their econometric approach, they thus regress in a first stage current institutions by settler mortality rates and in a second stage they regress current performance on the predicted outcome of their first regression. This is the standard **two stage least square** (2SLS) approach, that we are going to present, implement and question during this tutorial.

1 Data description and preliminary analysis (can be skiped)

Data file: *data_AJR_2001.dta*

We measure current economic performance and current quality of institutions using World Bank data observed in 1995. Our dataset contains 163 observations, among which 64 concern past-colonized countries. The variables that we are going to use are the following:

`shortnam` = three-letter country code

`euro1900` = Percentage of European settlers in the population in 1900.

`avexpr` = Risk of expropriation of private foreign investment by government, from 0 to 10, where a higher score means less risk. Mean value for all years from 1985 to 1995.

`logpgp95` = Log GDP per capita 1995, Purchasing Power Parity Basis, from World Bank

`cons1` and `cons00a` = Constraint on executive in first year of independence and in 1900. Seven-category scale, from 1 to 7, with a higher score indicating more constraints. Score of 1 indicates unlimited authority; score of 3 indicates slight to moderate limitations; score of 5 indicates substantialimitations; score of 7 indicates executive parity or subordination. Equal to 1 if country was not independent at that date.

`democ1` and `democ00a` = Democracy in first year of independence and in 1900. An 11-category scale, from 0 to 10, with a higher score indicating more democracy. Points from three dimensions: Competitiveness of Political Participation (from 1 to 3); Competitiveness of Executive Recruitment (from 1 to 2, with a bonus of 1 point if there is an election); and Constraints on Chief Executive (from 1 to 4). Equal to 1 if country not independent at that date.

`logem4` = estimated log of settlers mortality (standardized measure corresponds to the number of soldiers, missionaries, sailors who died over 1000 men strength)

`loghjypl` = Log output per worker, 1988

`baseco` = Colonial dummies. Dummy indicating whether country was a British, French, German, Spanish, Italian, Belgian, Dutch, or Portuguese colony.

`lat_abst` = Absolute value of the latitude of the country (i.e., a measure of distance from the equator), scaled to take values between 0 and 1, where 0 is the equato

`rich4` = variable binaire indiquant les "néo-europes", territoires au climat tempéré colonisés par les européens

`catho80`, `muslim80` and `no_cpm80` = Percent of population that belonged to the three most widely spread religions of the world in 1980 (or for 1990-1995 for countries formed more recently). The four classifications are: Roman Catholic, Protestant, Muslim, and "other.

`sjlofr` = French legal origin dummy: Legal origin of the company law or commercial code of each country. Our base sample is all French Commercial Code or English Common Law Origin

`avelf` = Ethnolinguistic fragmentation: Average of five different indices of ethnolinguistic fragmentation

`temp` = Temperature variables: Average temperature, minimum monthly high, maximum monthly high, minimum monthly low, and maximum monthly low, all in centigrade

`humid` = Humidity variables: Morning minimum, morning maximum, afternoon minimum, and afternoon maximum, all in percent

`deslow`, `stepmid`, `desmid`, `drystep` and `drywint` = Soil quality: Dummies for steppe (low latitude), desert (low latitude), steppe (middle latitude), desert (middle latitude), dry steppe wasteland, desert dry winter, and highland

`goldm`, `iron`, `silv`, `zinc` and `oilres` = Natural resources: Percent of world gold reserves today, percent of world iron reserves today, percent of world zinc reserves today, number of minerals present in country, and oil resources (thousands of barrels per capita.

`yellow` = Yellow fever: Dummy equal to 1 if yellow fever epidemics before 1900 and 0 otherwise.

`landlock` = Dummy for landlocked: Equal to 1 if country does not adjoin the sea.

`malfal94` = Malaria in 1994: Population living where falciporum malaria is endemic (percent)

`leb95` = Life expectancy: Life expectancy at birth in 1995.

`lt100km` = Distance from the coast: Proportion of land area within 100 km of the seacoast.

1. First produce a table of summary statistics of the most important variables to take into account in our tutorial. We are here interested in whether we can observe significant differences in mean between colonized countries and the rest of the world, and between colonized countries grouped by their level of settlers mortality.

   (a) Construct a variable which identifies the quartile of settler mortality of colonized countries to which belong each country. *Note: Be careful of deriving this quartile for the subsample of colonized countries. You can re-use the approach implemented in tutorial 3, to generate the exit of firms. Other approach: use the* ***`rank`*** *command to create a variable which ranks countries according to their level of settler mortality. Then compute the percentile using your rank variable and the total number of observations using the* count *command.*

   Solution: See STATA do-file for replication code.

   (b) Compute the mean value and standard deviation for the whole sample, the subsample of colonized countries and the subsamples of colonized countries by quartiles of settler mortalities, of the variables `logpgp95`, `loghjypl`, `avexpr`, `cons1` and `cons00a`, `democ00a` and `extmort4`.

   Solution: See STATA do-file for replication code.

   (c) Using this table, what can you conclude on the differences of economic performances, institutions and settler mortality between countries? Does your table support the mechanisms highlighted by the authors?

Solution: See the article of **?**, p.1377-1378, section II.A. for a detailled discussion. Note the relatively few number of observations between groups of countries by quartiles of settler mortality. This also means that when implementing OLS regressions, we will not have a lot of observations to make robust inference regarding our estimations.

2 The naive econometric analysis

2. While we know that it is likely that the estimation of a linear model of economic performance as a function of the measure of the quality of institutions will suffer from endogeneity issues, we still implement this approach as a first step. It is an interesting first step to better investigate the correlation between our variables and how it is affected by some control variables.

   (a) Consider the simple model where the log of income of country i, denoted y_i (measured using the log of GDP per capita) is a linear function of a constant, the risk of expropriation R_i and an error term ε_i, as follows:

   $$y_i = \mu + \alpha R_i + \varepsilon \tag{1}$$

   How do you interprete the coefficient α in Equation (1)?

   Solution: α cannot be interpreted as the causal effect of institutions on economic performance. The error term includes several variables, correlated with the institutions which also influence the economic performance, so we can not at all pretend capturing a causal effect of institution on income. We thus need to limit our interpretation to a correlation between economic performance and institution.

   (b) Now, consider the following model:

   $$y_i = \mu + \alpha R_i + X_i^{'} \gamma + \varepsilon \tag{2}$$

   where X_i is a vector of other covariates, such as geographical variables (dummy for continents and lattitude variables). How do you interprete the coefficient α using this second model?

   Solution: We are now measuring the effect of institutions on income, for countries with similar lattitudes and continent (all else equal). We are thus accounting for the effect of climate/geography on economic performance.

   Recall that this α coefficient is determined as follows:

   $$\alpha = \frac{Cov\left(y_i, \tilde{R}_i\right)}{V(\tilde{R}_i)} \tag{3}$$

where $\tilde{R}_i$ is the residuals of the regression of R_i over all other covariates X_i. We are thus partialling out the correlation of expropriation risks with the other covariates when studying its effect on economic performance.

? underlines that this coefficient have been interpreted as a causal effect of institutions on economic performance in past contributions on the topic. The authors insist on the fact that this coefficient underlines a correlation. The rest of the analysis will underline that this coefficient is not causal, because of endogeneity issues: reverse causality, and omitted variables, and errorment measure (biases in seeing better institutions in rich places, since the institution measure is constructed *ex post*).

(c) Now, reproduce the OLS estimations presented the Table 2 of Acemoglu et al. (2001), and interprete your results regarding your estimated $\hat{\alpha}$ across estimations.

Solution: You obtain the following table:

	(1)	(2)	(3)	(4)	(5)	(6)
`avexpr`	0.532	0.522	0.463	0.390	0.468	0.401
	(0.0406)	(0.0612)	(0.0555)	(0.0507)	(0.0642)	(0.0591)
`lat_abs`			0.872	0.333	1.577	0.875
			(0.488)	(0.445)	(0.710)	(0.628)
`africa`				-0.916		-0.881
				(0.166)		(0.170)
`asia`				-0.153		-0.577
				(0.155)		(0.231)
`other`				0.304		0.107
				(0.375)		(0.382)
Constant	4.626	4.660	4.873	5.851	4.728	5.737
	(0.301)	(0.409)	(0.328)	(0.340)	(0.397)	(0.398)
Observations	111	64	111	111	64	64
R^2	0.611	0.540	0.623	0.715	0.574	0.714

Standard errors in parentheses

Column (1) and (2) measures the simple correlation between the measure of the security of property rights and current income for the whole sample and the sample of colonized countries (respectively). These correlations are quite similar between samples, indeed there is no statistical difference between these two estimations.

Regarding continental dummies, America is the ommitted group. These coefficients must then be interpreted as the difference in mean of the economic performance between the given continent and America, for countries with equivalent lattitude and risks of expropriation. These continental dummies are satitistically significant (for Africa, and Asia on the subsample of colonized countries). That is, even when taking into account the effect of institution effect on economic performance (which remains statistically significant), countries in Asia and in Africa are poorer in average.

Overall $\hat{\alpha}$ is relatively stable accross specifications. Potential heteroskedastiticity which can be accounted for using the `robust` option. Overall, this coefficient is a biased measure of the causal effect of institutions on economic performance.

3 The instrumental variable approach

To instrument current institution quality, the authors use the rate of settlers of mortality at the time of colonization. We are going to show that this can be considered as a valid instrument, that is an exogeneous source of variation whose effect on current economic performance can only be assigned through past and current institutions.

3. We first investigate the relationship between settlers mortality and current economic performance. Briefly recall what is the serie of mechanisms between these two variables. Then, reproduce the Figure 1 in Acemoglu et al. (2001).

Solution: Just rephrase what is said in introduction. You need to underline that current isntitutions are inherited from institutions established by colonizing countries. And the establishment of institutions has been determined on the feasability of settlements in the colonized countries. The feasability of settlements depends on several dimensions (climate, geography...) among which the disease-free environment and thus the risk of mortality.

See STATA do-file for replication of Figure 1. We observe a strong negative relationship as colonies where Europeans faced higher mortality rates are today substantially poorer than countries that were healthy for Europeans. We are thus going to show that this relationship reflects the effect of settler mortality working through the institutions brought by Europeans.

Equation (2) describes the relationship between current institutions and log GDP. In addition we have

$$R_i = \lambda_R + \beta_R C_i + X_i^{'} \gamma_R + \nu_{Ri} \tag{4}$$

$$C_i = \lambda_C + \beta_C S_i + X_i^{'} \gamma_C + \nu_{Ci} \tag{5}$$

$$S_i = \lambda_S + \beta_S m_i + X_i^{'} \gamma_S + \nu_{Si} \tag{6}$$

where C is the measure of early institutions, S is the measure of European settlements in the colony (fraction of the population with European descent in 1900), and m is the logarithm of mortality rates faced by settlers. X is a vector of covariates that affect all variables.

4. To document our IV strategy we consider the models in Equations (4), (5) and (6).

   (a) Estimate the models in Equations (4), (5) and (6) on the sample of colonized countries and using latitude as a control variable.

Solution: See STATA do-file for replication code.

We obtain the following table:

	(1)	(2)	(3)
	`avexpr`	`cons00a`	`euro1900`
`lat_abs`	2.346	0.335	89.91
	(1.435)	(1.999)	(20.33)
`cons00a`	0.255		
	(0.0903)		
`euro1900`		0.0532	
		(0.0102)	
`logem4`			-7.108
			(2.137)
Observations	60	60	63
R^2	0.239	0.447	0.472

Standard errors in parentheses

Note that results correspond to the ones in Panel A. column (2), Panel B. columns (2) and (10) in Table 3 of Acemoglu et al. (2001) (close to a 100 scaling factor when there is variable `euro1900`)

   (b) How do the results can be interpreted with regards to the hypothesized serie of mechanisms linking settlers mortality and current institutions?

Solution: See **?**, Section IV.A. p.1383-1384.

The R^2 gives the percentage of variance of the pheneomenon explained by the model. These results provide evidence that early institutions were shaped, at least in part, by settlements, and that settlements were affected by mortality of settlers.

5. We first reproduce the two stage least square (2-SLS) approach "by hand". Protection against expropriation variable, R_i, is treated as endogenous, and modeled as

$$R_i = \zeta + \beta m_i + X_i'\delta + \nu_i \tag{7}$$

We then obtain predicted values of the protection against expropriation risks variable $\hat{R}_i$ from the estimation of the model in Equation (7). These are the variations of institutions assigned through the variations of the settlers mortality and partialling out any other endogeneous variations in institutions. We then incorporate this predicted value in replacement of R_i in the model of Equation (2), which is now an exogeneous regressor.

(a) Estimate the first stage of the 2-SLS approach, that is the estimation of the model in Equation (7) where you control for the lattitude of country i. Store your predicted values in a new variable.

Solution: See STATA do-file for replication. We obtain the following estimation:

	(1) `avexpr`
`logem4`	-0.510
	(0.141)
`lat_abst`	2.002
	(1.337)
Observations	64
R^2	0.296

Standard errors in parentheses

We observe the satistically significant negative effect of past settlers mortality on current institutions quality. 29 percent of the variance of the risk of expropriation is explained by the first stage model.

(b) Estimate the second stage of the 2-SLS approach. Interprete your estimation of the $\hat{\alpha}$ coefficient. How does it compare to the estimation in the naive approach in question 2.(c)?

Solution:

See STATA do-file for replication. We obtain the following estimation:

	(1) `logpgp95`
`avêxpr`	0.996
	(0.165)
`lat_abst`	-0.647
	(0.996)
Observations	64
R^2	0.500

Standard errors in parentheses

We notice that the coefficient is higher than in the naive approach. The endogeneity in the naive approach thus introduces a downward bias, which is likely to be attributed to measurement errors.

Also, the latitude variable now has the "wrong" sign and is insignificant. This result suggests that many previous studies may have found latitude to be a significant determinant of economic performance because it is correlated with institutions (or with the exogenous component of institutions caused by early colonial experience).

(c) Why does this 2-SLS approach “by-hand” is not valid?

Solution: As we are doing a two-stage estimation process, the variance of the regressors inserted in the second stage estimation are not the true-variance, but an estimated variance from the first stage. Second stage standard errors are wrong because of the incorporation of predicted regressors in the second stage, which generates a sampling error. Indeed, the correct residual variance estimator uses the original endogenous regressor R_i to construct residuals and not the first-stage fitted values, $\hat{R}_i$.

When comparing the results to the ones of **?** in Table 4, we notice that our estimated coefficients are the same, but our standard errors are different, and lower.

6. STATA has a valid built-in command to implement IV regression using the 2-SLS approach. The command is `ivreg` where you need to precise the instrumental variable of your endogeneous variable. That is, if you want to regress economic performance on current institutions by instrumenting the latter with the settlers mortality the syntax is: `ivreg logpgp95 (avexpr=logem4)`. Replicate the estimation done in question 5 using this STATA command. What do you observe?

Solution: See STATA do-file for replication.

We obtain the following outcome:

	(1) logpgp95
avexpr	0.996
	(0.222)
lat_abst	-0.647
	(1.335)
Observations	64
R^2	0.102

Standard errors in parentheses

We now retrieve the results as shown in Table 4 of **?**. Standard errors are now correct.

We thus prefer this second approach. Our results are still statistically significant.

4 Validity of the approach (doesn't have to be prepared)

4.1 The exclusion restriction

The validity of the 2SLS results using the IV approach depends on the assumption that settler mortality in the past has no direct effect on current economic performance. This is also called the exclusion restriction. In other words, conditional on the controls included in the regression, the mortality rates of European settlers more than 100 years ago must have no effect on GDP per capita today, other than their effect through institutional development. Thus, the mortality of settlers must be excluded from the causal model of interest.

7. We check the exclusion restriction assumption of our instrument in the following questions.

   (a) What variables (available in our dataset) can plausibly be correlated with both settler mortality and economic outcome?

   Solution: Many potential variables:

   - the identity of the main colonizing country, since it determines current institutions, but weakly correlated with settlers mortality
   - climate and geographical characteristics: affect both economic outcome and settlers mortality.
   - the presence of disease: some diseases affected settler mortality and may still affect economic development. In this case, the instrumental-variable estimation may be assigning the effect of diseases on income to institutions.

   (b) Add these variables as control variables in your 2SLS estimation done in question 6, and check whether the addition of these variables affects your estimates.

   Solution: See STATA do-file for replication code.

   Adding temperature and humidity variables. See Table 6 for results.

   See Table 7 to seee whether the instrument could be capturing the general effect of disease on development.

4.2 The effect of outliers

8. One may argue that results are driven by some countries with very specific institutions and economic performance in the sample, not corresponding to the mechanisms at hand for the the majority of the sample. Indeed, richer colonized countries, such as the United States, Australia or New Zealand, have very different characteristics to the other countries in the sample. Replicate the estimation done in question 6 when excluding these countries (identified with the dummy variable `rich4`)). How does it affect your results?

Solution: See STATA do-file for replication.

Interpretation of the result is given p.1387