REPLICATION EXERCISE 3: ACEMOGLU, JOHNSON, AND ROBINSON (2001)

VOJTĚCH BARTOŠ

Joint classroom discussion: AJR (2012)

Read the introduction of the paper. The paper asks a fundamental question in economics: do institutions *cause* economic performance?

(1) First, why do we believe that institutions affect economic growth?
(2) Describe briefly the theoretical framework used in the paper, namely through which channels the authors explain the differences in the current performance of countries.
(3) Now let's look at some history of how institutions might have evolved. What is the link between mortality rates in the colonies and the types of European settlements? What are the AJR arguments for why in some countries the early institutions were "extractive", while in other countries they were "inclusive" (replicating European institutions)? And what do they mean by extractive and inclusive institutions? Section I.A and I.B can be of help.
    (a) This will become important later on, but comment briefly on how exactly are the mortality rates collected in this paper. Why do the authors use a different methodology for Latin American countries than for the rest of the world? Explain the two main sources of mortality data AJR use. See Section III.B. We'll discuss how they make the sources comparable later on.
(4) Explain the reasoning for why institutions developed in the 19th century should persist until today, despite the fact that all former colonies gained independence during the 20th century and hence the original rulers who set up the institutions left. Section I.C can be of help here.
(5) What we could do is to examine the link between institutions today and todays' economic performance. What are the reasons why we can't interpret the link between todays institutions and economic performance as a causal one? There are three reasons the authors propose in the last paragraph of section II. What are these, explain all briefly.
(6) The authors propose the use of instrumental variables to overcome the endogeneity problem. Reconstruct the model that AJR propose.
    (a) In order for a variable to be a proper instrumental variable, two assumptions need to be made. What are these?
    (b) What are the authors' arguments that support that the exclusion restriction is valid in their case? The introduction and section III.A could be of help here.

(c) What is the relationship between the mortality rates of the settlers and of the local population? Why is this important for the use of instrumental variable?

Group work: AJR (2012)

Now we are in good shape to open the data. I assembled the complete dataset that can be used to replicate all the results presented in the original paper. Always store the results of all regressions using `outreg2` command (I am usually using the `dec(2) se` options to make the regression output more tractable).

(1) Load *AJR2001_replication.dta* (you can download it from my website).
(2) Open the "Data Browser" and the "Variables Manager" to familiarize yourself with the data. We will mainly be working with the GDP data, with the data on settler mortality, and with the institutional data (average expropriation risk and sometimes with the data on constraint on executive powers). Section II.A and Table 1 can be helpful.
(3) AJR use the data on settler mortality rates per thousand for a sample of 75 countries. The first step is to show that indeed, settler mortality long time ago is linked to todays economic performance of the country.
    (a) Create a graph plotting the log of settler mortality on the horizontal axis and the log per capita GDP in 1995 on the vertical axis (see `help graph twoway scatter`). Use only the restricted sample of ex-colonies. This will give you the Figure 1 in AJR. Comment briefly.
        (i) Note: If you want to make it fancy, use `mstyle(none) mlabel(shortnam) mlabgap(-3)` options to change the dots to short country names as in the paper.
(4) The variable in the dataset only presents a logarithmic transformation of the original settler mortality data.
    (a) To get a better understanding of the magnitudes, transform the variable back into levels. Create a new variable.
    (b) Summarize the new variable and briefly comment on some interesting cases. How is it possible that some variables are larger than 1000? (note: we'll get back to the extreme cases later...)
(5) Now that we have established that there is some relation between the mortality of settlers in the 19th century and current economic performance (we call this a *reduced form* relationship) in the country, we should start examining the actual question we care about. Namely, the relationship between the institutional quality and economic performance.
    (a) AJR use the variable called index of protection against expropriation (they take an average value between 1985-1995) collected by Political Risk Services. Briefly comment on why we can believe that this variable is a good measure

of institutional setting. You can also refer to explanations in Hall and Jones (1999) paper (p. 97) to support your argument.[1]

(b) To get an idea of how current institutional setting is linked to economic performance, plot a graph with the average expropriation risk on the horizontal axis and the log per capita GDP in 1995 on the vertical axis (use ex-colonies subsample only). You should get Figure 2. Comment briefly.

(c) In the previous step, we see that there is some link. Let's examine it using a regression analysis. Take the entire sample run a regression $GDP_i = \alpha + \beta\ avgexpropriation_i + \varepsilon_i$ (variables are GDP per capita and the average expropriation risk, respectively). This gives almost the same results as column 1 of Table 2.[2]

(d) Our main interest, however, is the ex-colonies.

(i) To restrict our sample accordingly, we need to create a dummy variable that lists countries that are 1) ex-colonies, 2) have data on settler mortality, 3) on protection against expropriation risk, and 4) on GDP. As a check, this variable should give you 64 observations marked with a "1". This is the "base sample".

(ii) Let's now run the restricted model from the previous point on the sample of ex-colonies. Comment on the results.

(iii) More specifically, try to understand the predictive power of the model using a numerical example by examining the case of Uganda (UGA) and Mexico (MEX). These countries are around the lower and upper quartiles of the institutional quality in our sample, respectively. How large a difference in GDP would the countries' institutional quality predict? And what is the true difference in GDP?

(6) Going back to your responses in point (5) of our classroom discussion, we know that we cannot interpret these results causally. In order to do this, we go back to the model we reconstructed in point (6).

(a) We need to show that there is a link between settler mortality and current institutions. Plot a graph with (log) settler mortality on the horizontal axis and the average expropriation risk (1985-1995) on the vertical axis. Comment.[3]

(b) To see that there is indeed a relationship between current day institutions and the institutions in 1900, percentage of population of European descent in 1900, and of settler mortality, let's construct a model with current institutions on

[1] There are also alternative measures that could be used? We might try to test them if we have time: you can download the: Polity data, or Corruption perceptions index.

[2] AJR admit themselves in a comment to the posted data that this gives 111 observations instead of 110. If you want, you can try to find the culprit yourself. I don't know which country it is.

[3] You can also plot graphs for 1) settler mortality and number of Europeans living in the colony in 1900, 2) number of Europeans living in the colony in 1900, and the constrains to executive power in 1900 or democracy in 1900 (as a proxy for early institutions), and 3) constrains to executive power in 1900 or democracy in 1900, and average expropriation risk (1985-1995). This would complete the model we constructed in point (6) of our classroom discussion.

the left hand side and the respective other variables on the right hand side. Run a simple OLS model on a subsample of ex-colonies. This should produce results as in columns 1, 3, 5, 7, and 9 in Panel A of Table 3 (reconstruct the right hand side variables from the table to match the results). What have we learned?[4]

(c) It seems that the model we proposed at the beginning is plausible. We can thus try to use the settler mortality as an instrument (exogenous source of variation) for the current day institutions and estimate its effect on current day economic performance.
   (i) Run the 2SLS regression on a subsample of excolonies. Comment briefly.[5]
   (ii) Take again the example of Uganda and Mexico and predict the difference in their economic performance based on the results of the 2SLS regression. Compare the results to those you made in 5.d.iii., comment. Are we getting more realistic results now?
   (iii) Thinking back of the responses you gave in point (5) of our classroom discussion, can you tell which of the reasons seems most plausible based on the comparison of OLS and 2SLS results? Section IV.B can help, but explain a bit more.

(d) Now we can enrich our model by including a control variable representing absolute distance from equator (on scale from 0 to 1). This is motivated by the paper of Hall and Jones (1999) who used the distance from equator as an instrument for institutional quality.
   (i) First, run the original model from Hall and Jones (1999): $GDP_i = \alpha + \beta\ absdistance_i + \varepsilon_i$. Comment.
   (ii) What is the correlation coefficient between the absolute distance from equator and the average expropriation risk (1985-1995)? Is it significant? This command can help: `pwcorr` (option `sig`). Comment.
   (iii) Run the model $GDP_i = \alpha + \beta\ avgexpropriation_i + \gamma\ absdistance_i + \varepsilon_i$ and comment on the significance of the variables.

(e) Among other potential candidates, one variable is of a particular concern: prevalence of malaria today could affect todays economic outcomes. If we assume that malaria prevalence today is correlated with malaria prevalence earlier, this would have caused increased settler mortality.
   (i) Let's run an OLS model $GDP_i = \alpha + \beta\ malaria_i + \varepsilon_i$ on a subsample of ex-colonies. The malaria variable is a share of population exposed to plasmodium falciparum malaria. Comment.
   (ii) Now let's enrich our IV model by including the malaria variable to the model we estimated in point 6.c.i. Comment on both coefficients. This gives you the same results as column 1 in Panel A of Table 7 in AJR. See also the comments by the authors in section V.A on Table 7 results.

[4]If you have time, you can replicate also Panel B of Table 3.

[5]Use the command `ivregress 2sls` with the option `first` to reveal also the results of the first stage.

(7) Briefly comment on the results obtained so far. What have we learned? Can you think of policy advice based on the results obtained?

Joint classroom discussion: Albouy (2012)

Let's now move to year 2012 when David Albouy published a comment on the original AJR paper. Read the introduction of Albouy (2012).

(1) One of his main critique is that the data on settler mortality used in AJR come from very different sources in very different countries. Remember, the data are for soldiers in Africa (assembled by Curtin and co-authors) and for bishops in Latin America (assembled by Gutierrez; benchmarked to soldier rates using the fact that they had data on both bishop and soldier rates for Mexico). Why could use of such data be problematic for the AJR story?
   (a) Hint 1: note the differences in institutional quality and wealth between the continents.
   (b) Hint 2: note that the bishop mortality rates are extremely low, see third paragraph in Albouy (2012, p. 3063).
(2) Another issue discussed in Albuoy's critique is that in some cases the mortality rates are extremely large. This is probably caused by epidemics in the period of measurement, making the data rather uncomparable to the data from countries where the measurement was done at "healthier" periods. Recall your responses to the point 4.b in the group work part. Why is this a problem for the econometric analysis? How best would you solve the problem without a need for more historical data?
(3) What are the other critical points Albuoy discusses (just briefly)?

Group work: Albouy (2012)

(1) It seems that using the historical data is tricky. Let's now focus on a conservative specification by dropping all 36 countries Albuoy defines as problematic to test the robustness of results presented. This leaves us with a sample of 28 countries. Use the variable *source0* to select the subsample.
   (a) On a full sample (we'll get to the restricted one soon), estimate the IV regression $GDP_i = \alpha + \beta\ \widehat{avgexpropriation_i} + \varepsilon$ where average expropriation rate (1985-1995) is instrumented using the (log) settler mortality on a subsample of ex-colonies (this gives the same results as the originial AJR paper). Also, run the same regression with absolute distance from equator control variable included.
   (b) Estimate the same models as in part (a) and cluster on the level of log settler mortality. We should cluster as the mortality data are implanted for some

countries, hence errors would be correlated for countries with source and implanted data. Use the `vce(cluster ...)` option. Comment on the differences in the first and second stage.[6]

(c) Estimate the same regressions as in (a) and (b), now restricted to the sample of 28 countries for which we have source data. Comment on the differences, especially on the model with latitude. Why are the standard errors in the first stage regression so close to those from the unrestricted sample despite the dramatic reduction of sample size?

(d) The other problem Albuoy discusses is the different sources of data used even in the Curtin's database. Some are from soldiers located in barracks, some from soldiers on military campaign, some from enslaved laborers. Albuoy classified these using variables *campaign* and *slave*. Control for these variables in the models above, both on the sample of 64 and 28 countries. Comment.

Joint classroom discussion: AJR (2012) reply

Of course, AJR had to reply to Albuoy's critique. Read the introduction to AJR (2012).

(1) What is the main argument for why the original AJR paper included the imputed data from other countries to countries where no data were available? Section II.C can help.

(2) Why do AJR believe that the use of the data that was available in Europe during the 19th century and for example used for insurance purposes could have been reflected the decisions about the location and type of colonial settlements? Introductory section can help.

(3) Why do AJR dismiss Albuoy's argument of non-comparability of campaign and non-campaign data? Can this argument be consistent with Albuoy's results in which the campaign and slave data turn out to be significant in the first stage regressions? See section I.A.

Group work: AJR (2012)

(1) Up until now we have assumed a linear relationship between mortality and foundations of institutions. Maybe there is an upper threshold above which the mortality rate is just too much to take to set up neo-Europes. Further, maybe the rates in the mortality data also reflect extremes rather than averages. See section II.A on the choice of a cap on mortality rates.

(a) What are the reasons for capping mortality at 250 (and other rates)? Read section II.A.

(b) Now use the variable on absolute levels, cap the mortality rates at 250 and create a new log mortality variable using the censored variable. Estimate the model $GDP_i = \alpha + \beta\ \widehat{avgexpropriation}_i + \varepsilon$ (where average expropriation

[6]You will be getting slightly different results from the Albuoy paper as they also use Huber-White robust standard errors and some mortality data are updated. This makes no substantial changes, though. Using `ivreg 2sls` you cannot use both options.

rate (1985-1995) is instrumented using the capped log mortality rate) using IV regression on both the original sample of 64 countries and the Albuoy subsample of 28 countries. You can then include absolute distance from equator. Compare to the regressions to original Albuoy results.[7]

(c) Further, AJR note that the Albuoy (non-)results are driven by an outlier, Gambia. Look at the figures you produced for the original AJR replication and locate Gambia. Why is Gambia a weird observation? See section II.D, and footnote 46 in particular.

(d) Repeat the previous step estimation and drop Gambia (GMB) from the sample. Comment.

Are you convinced by the original AJR story?

[7]Note: clustering should still be done on the uncapped log mortality variable.