Homework 1: The Secret of the Primitive Accumulation?

36-402, Spring 2017, Section A

Due at 11:59 pm on Wednesday, 25 January

Background

Modern economic growth begins with the industrial revolution in Britain and the rest of northwestern Europe around 1800. But before that, those countries were already some of the most prosperous parts of the world, which is part of why the industrial revolution happened there. Explaining that pre-industrial growth, sometimes called "primitive accumulation", is therefore an important problem in economics and world history.

In this assignment, we will look at a model, and the accompanying data, for one of the leading contemporary theories of this growth. The theory, in brief, is that the key factor driving this early growth was trade across the Atlantic, involving New World plantations (for, e.g., sugar), slaves, and imperial ventures in Africa and Asia. Countries which could participate in these Atlantic trades grew richer than others. The theory goes on that countries which started with relatively free institutions that protected private property were able to take more advantage of the Atlantic trade than countries with more absolutist institutions (and that merchants growing rich from the trade encouraged even less absolutist institutions later on).

The scholars who have elaborated this theory have gathered a data set of the relevant variables, which is available as RAJ.csv on the class website. This also contains a number of variables which are intended as controls, because they are important to rival theories. Each row in the data represents a particular country in a particular year. The countries are limited to Europe and Asia. As is usual for real data, some variables are missing for some rows.

1. `country`
2. `year`
3. `urbanization` An estimate of the fraction of the population living in cities and towns.
4. `population` The total population of the country (in thousands of people)
5. `coastToArea` The ratio of the country's Atlantic coast-line (in miles) to its total land area (in square miles).

6. `execConstr` A rating of how constrained the executive branch of the country's government was, on a seven-point ordinal scale from 1 (least constrained) to 7 (most constrained).
7. `initialConstr` Rating of how constrained the executive branch of the country's government was "initially", averaging ratings for 1400 and 1500
8. `atlTrade` An index of the volume of the trade carried over the Atlantic, across *all* countries.
9. `westernEurope` An indicator for whether the country is part of western Europe.
10. `easternEurope` An indicator for whether the country is part of eastern Europe.
11. `wars` The number of wars the country engaged in, per year, over the period.
12. `protestant` Whether the country's inhabitants are primarily Protestant Christians[1].
13. `roman` Whether the country was part of the Roman Empire.
14. `gdppc` An estimate of per-capita GDP, in current dollars. (This is known to be very imprecise.)

Urbanization is being used here as a proxy for the over-all level of economic development. The favored model of the scholars who proposed this is that the level of urbanization of country i in year t is

$$u_{it} = d_t + \delta_i + \alpha_t W_i M_t + \beta A_t P_i + \gamma_t C_i M_t + \eta A_t P_i C_i + \epsilon_{it} \quad (1)$$

where d_t and δ_i are "fixed effects" for year and country (respectively); W_i is an indicator for country i being in western Europe; A_t is the index of Atlantic trade in year t; P_i is country i's potential for Atlantic trade, measured by its coastline-to-area ratio; C_i is the country's "initial" level constraint on the executive, in 1400–1500; M_t is an indicator variable, 1 if the year is $\geq$ 1500 and 0 earlier; and ϵ_{it} combines noise and measurement error.

Data-Analysis Problems

1. (5) Load the data as a data frame. Create a column in the data frame to represent the M_t variable. Show that this column has been correctly calculated.
2. Estimate the model.
   (a) (2) Explain how your R formula corresponds to Eq. 1. (This is one question where including R code in your written answer is OK.)

[1] This changed for several countries in the study over the period, and it's not clear what year was used to set this variable.

(b) (3) Report point estimates and 95% confidence intervals for each coefficient, to reasonable precision.

(c) (1) For some ways of estimating the model, you may find that some coefficients are NA. Explain why this is OK. (If you can estimate the model in a way which doesn't lead to any NA coefficients, this is a free point for you.)

(d) (3) Describe the assumptions which went in to your calculation of the confidence intervals.

(e) (3) Describe, in words suitable for a historian who has taken a course like 225/226, but *not* 401, the interpretation of the η coefficient.

3. Consider two countries which are otherwise identical, except for their institutions score, which is 1.5 for one country and 2.5 for another country.

   (a) (3) Assume that both countries have the median level of potential for Atlantic trade. Calculate and plot the predicted difference in their level of urbanization for each year, along with 95% confidence levels.

   (b) (1) Assume that both countries' potential for Atlantic trade is at the 25th percentile. Re-do the plot from the previous problem.

   (c) (1) Assume that both countries' potential for Atlantic trade is at the 75th percentile. Re-do the plot.

   (d) (5) Qualitatively, according to the model, do "good" (high-scoring) institutions seem to contribute to economic development? Do they contribute more as time goes on? Does the answer depend on the level of potential for Atlantic trade? How much uncertainty is there around this answer?

4. (5) Plot fitted/predicted values of urbanization for each data point against the actual values. Describe the pattern you see, and whether it indicates any problem with the model.

5. (6) Plot residuals and squared residuals against all of the continuous predictors. Describe any patterns you find. Do the residuals look random, with zero mean and constant variance?

6. (3) Plot the distribution of residuals. How Gaussian is it?

7. (5) Create a box plot of the distribution of residuals by country. Do all countries have the same distribution? Should they?

8. (5) Calculate the leverage for each data point, and plot the distribution of leverages by country. Comment on any patterns you find.

9. (5) Calculate the Cook's distance for each data point, and plot their distribution by country. Comment on any patterns you find.

10. (5) Re-estimate the model with $\alpha_t W_i M_t$ replaced by $\alpha_t W_i$, and $\gamma_t C_i M_t$ replaced by $\gamma_t C_i$. What has happened to the coefficients? Can you explain why, mathematically, this happened? Can you think of a substantive (economic or historical) reason for requiring these interactions?

11. (5) Over-all, how sound do the statistical assumptions of the original (Eq. 1) model seem? How reliable does that model seem?

Theory Problems

These are all from the exercises at the end of Chapter 1 of the textbook.

1. (5) Exercise 7.

2. (6) Exercise 8.

3. (6) Exercise 9.

4. (7) Exercise 10.

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.