Exam 1: Imperialism, the Earliest Stage of Capitalism?

36-402, Spring 2016

Due at 11:59 pm on Thursday, 3 March 2016

This is a take-home data analysis exam. You are allowed to use your notes, the textbook, and any other printed or electronic references. **You are under no circumstances allowed to consult with any person other than the professors and the teaching assistants.** You are expected to comply with the CMU policy on academic integrity. Unauthorized help will result in failing the exam, and possibly more severe disciplinary action. If you are unsure what is allowed, ask the professors.

Please submit two files to Blackboard: one is the PDF of your report; the other is the .Rmd (or .Rnw) file which produced it.

While there are wrong answers, there are many possible right answers. Any data analysis decisions or conclusions that you make should be justified and explained. Your job is to correctly ananlyze the data, not force the analysis to match a pre-conceived idea.

Writing Instructions

Write up your work as a scientific report of **at most 10 pages**. Nothing beyond the tenth page will be read[1] A suggested (not mandatory) outline:

- Introduction describing the scientific problem and the data set, possibly including *relevant* summary statistics or exploratory graphs.
- Models with subsections
  - Describing the specification of the model (or models) you estimated, and explaining why you decided to use those specifications rather than others;
  - Giving the relevant estimated coefficients and/or functions (possibly in visual form), along with suitable measures of uncertainty;

[1] Do not try to game this: fonts should be no smaller than 9 points, margins should be reasonable, graphs and tables should be embedded in the report and count against the length. You will find it a good idea to hide your code (`echo=FALSE`), except in the rare situations where a line of code is the clearest and shortest way to convey an idea.

  - Checking the goodness of fit of the model, including a description of the test procedures you used, why you chose those ways of checking the model, what the results were, and what they told you about the ability of the model to describe the data set.

- Results answering the analytical questions quantitatively, and with suitable measures of uncertainty, with reference to your estimated model or models.

You may assume that the reader has a general familiarity with the contents of 401, and with the models and methods we have covered so far in the course, but will need to be reminded of any details. The reader should not be assumed to have any prior familiarity with the data set.

Research Problem and Data

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 usual, 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[2].
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

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

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; and ϵ_{it} combines noise and measurement error.

Specific Questions and Issues

You should estimate the baseline linear model (Eq. 1). You should also assess whether, within this model, your estimates (and their uncertainties) support or undermine the theory. You should also carefully examine how well the model fits

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

the data, particularly considering outliers (especially if they are also influential points) and the pattern of residuals.

You should also see whether it is possible to devise another model which predicts the data better than Eq. 1, and whether such a model is actually better.

Finally, you should use your preferred, estimated model to assess how much evidence there is in favor of the theory that early-modern economic growth in Europe was driven by cross-Atlantic trade, and the interaction of that trade with initial institutions.

Inferential Statistics and Model Assessment You may not assume that R's default standard errors or p-values for regression models can be trusted. Uncertainty should be assessed using suitable bootstrap or simulation procedures. (Be sure to explain why you used the procedure you did.) If you need to compare two models in terms of predictive accuracy, this should not be done through R's default significance tests or R^2's, but through either a suitable bootstrap or cross-validation (again, explain the reasoning behind your choices). Exceptions will be made if you can successfully argue that the default calculations are reliable *for this problem*.

Model checking The answers you give to the substantive analytical questions rest on your estimated model, so you need to include some assessment of the model's goodness of fit. The exact way in which you do this is left up to your initiative; it may help to remember that the model is predicting a quantitative outcome. Be sure to describe your procedure and explain why you chose it, that is, why it is appropriate to answer the questions at hand.

Rubric

As usual, this describes the ideal.

Words (10) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (5) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (5) All figures and tables shown are relevant to the argument for the ultimate conclusions. Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text.

Code (10) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. All code runs, and the Markdown file knits.

Modeling (20) Regression model specifications are described clearly and in appropriate detail. There are clear explanations of how estimating the model helps to answer the analytical questions, and rationales for all modeling choices. If multiple models are compared, they are all clearly described, along with the rationale for considering multiple models, and the reasons for selecting one model over another, or for using multiple models simultaneously. Models beyond those covered in 401 are seriously considered, and, if not ultimately used, are rejected for sound, data-driven reasons.

Inference (20) The actual estimation of model parameters or estimated functions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty.

Checking (15) The goodness-of-fit of the model is actively probed by means of tests suitable to that class of model. The tests chosen are described, along with the rationale for using those tests. The execution of the tests is technically correct, and the results of the checks are clearly described. The extent to which the results of the model assessment build or undermine confidence in the conclusions is laid out clearly, with references to specific pieces of evidence.

Conclusions (15) The substantive, analytical questions are all answered as precisely as the data and the model allow. The chain of reasoning from estimation results about the model, or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.

Extra credit (10) Up to ten points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.