Data Analysis Exam 1: "The sound of gunfire, off in the distance"

36-402, Spring 2024

Due at 6:00 pm on Thursday, 29 February 2024

This is a take-home data analysis exam. Please read this whole document carefully before beginning to work.

The rules on allowed resources and collaboration are stricter than for homework; please refer to the syllabus and the course policies. If you are unsure what is allowed, ask the professor.

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 analyze the data, not force the analysis to match a pre-conceived idea.

Writing Instructions

Please submit *one* file to Gradescope: the knitted PDF of your report.

Write up your work as a scientific report of **at most 10 pages**, including figures and tables. Nothing beyond the tenth page will be read[1]. Use the following outline, unless you have strong reasons to deviate from it:

1. Introduction describing the scientific problem and the data set, possibly including *relevant* summary statistics or exploratory graphs.
2. Models with subsections
   (a) Specifying the model (or models) you estimated, and explaining why you chose those specifications rather than others;
   (b) Giving the relevant estimated coefficients and/or functions (possibly in visual form), along with suitable measures of uncertainty;
   (c) 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.

[1] Do not try to game this: fonts should be no smaller than 9 points, margins should be reasonable, etc. Hide your code (`echo=FALSE`), unless the code is the clearest and shortest way to convey an idea. (It rarely is.)

3. RESULTS answering the analytical questions quantitatively, and with suitable measures of uncertainty, referring back to your estimated model.

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. The reader should be able to assess your arguments using *just* your PDF report, without having to read your code.

Research Problem and Data

Since the end of the Second World War, civil wars within countries have become much more common than wars between states. Understanding the circumstances which make them more likely is thus a problem of considerable importance for both social science and for preventing human misery. Two leading theories suggest different situations which can make a country vulnerable to civil wars.

1. One theory argues that civil wars are easier to start and maintain in countries whose economies are heavily dependent on commodity exports, where rebels can seize, and sell, some part of the commodity production.
2. Another theory is that civil wars tend to start in countries where there are strong ethnic divisions, and one ethnic group dominates the government and economy.

This data-analysis exam will look at the quantitative evidence in favor of (or against) these theories.

Our data `http://www.stat.cmu.edu/~cshalizi/uADA/24/exams/1/ch.csv`, comes from an influential study of the causes of civil wars. Every row of the data represents a combination of a country and of a five year interval — the first row is `Afghanistan, 1960`, really meaning Afghanistan, 1960–1965. The variables are:

- The country name;
- The year;
- An indicator for whether a civil war *began* during that period — the code of `NA` means an on-going civil war, while 0 denotes continuing peace;
- Exports, really a measure of how dependent the country's economy is on *commodity* exports;
- Secondary school enrollment rate for males, as a percentage[2];

[2] I have been unable to find an explanation anywhere of why this rate is greater than 100 for some data points.

- Annual growth rate of GDP;
- An index of the geographic concentration of the country's population (which would be 1 if the entire population lives in one city, and 0 if it evenly spread across the territory);
- The number of months since the country's last war or the end of World War II, whichever is more recent[3];
- The natural logarithm of the country's population;
- An index of social "fractionalization", which tries to measure how much the country is divided along ethnic and/or religious lines;
- An index of ethnic dominance, which tries to measure how much one ethnic group runs affairs in the country.

Specific Questions and Issues

You should estimate a model which predicts the outbreak (not the continuation or the ending) of civil war. All other variables except country and year are potentially usable as predictors. You should assess whether, within this model, your estimates (and their uncertainties) support or undermine the two theories of the origins of civil war discussed above. Specifically, you need to assess not only which variables predict the origin of civil wars, but also how important they are compared to other variables, and what each theory says about which variables should matter. You should also carefully examine how well your the model fits the data, particularly considering outliers (especially if they are also influential points) and the pattern of residuals.

Inferential Statistics and Model Assessment Don't presume that R's default standard errors, p-values or confidence intervals 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 the particular problem you are solving.*

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 binary 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.

[3]This appears to count only civil and not foreign wars.

Rubric

As usual, this describes the ideal.

Words (15) 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 (10) All numerical results or summaries are reported to justified precision (neither more nor less), and with suitable measures of uncertainty attached when applicable. All numbers reported are either generated by the code reproducibly, or derived by explicit mathematical calculations.

Pictures (10) 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 (as appropriate), and sit near the relevant pieces of text. All figures and tables are generated reproducibly by the code.

Modeling (15) 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 (15) 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 (20) 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 analytical methods are unusually insightful, or where the analysis extends the required analytical questions in an interesting way.