Exam 1: The Bullet or the Ballot?

36-402, Spring 2015

Due at 11:59 pm on Wednesday, 4 March 2015

Many people assume that violence, while perhaps dangerous or evil, is more *effective* politically than non-violence. In this exam, we will examine whether, in fact, non-violent political movements are more or less likely to achieve their goals than violent ones. Moreover, we will look at the conditions which make non-violence more or less likely to succeed.

Our data set, gathered by political scientists who have studied exactly these questions, is `navc.csv` on the class website. The units of analysis here are political movements or campaigns. For each movement, the data records:

- The name of the movement (`campaign`);
- The country the movement was in (`country`);
- The peak year of the movement's activity (`year`);
- Whether the movement fully achieved its aims (1.0), achieved partial success (0.5), or failed (0) (`outcome`);
- An indicator variable (`nonviol`), 1 for non-violent movements and 0 for others;
- A quantitative measure of how democratic the government of the country was, from -10 for very un-democratic governments to a possible maximum of +10 (`democracy`);
- An indicator for the government being under international sanctions (`sanctions`);
- An indicator for whether the government received aid from other governments to help deal with the movement (`aid`);
- An indicator for the movement's receiving aid from foreign governments (`support`);
- An indicator for the government's using violence to repress the movement (`viol.repress`);
- An indicator for whether substantial portions of the security (military and police) forces of the government sided with the movement (`defect`);
- The duration of the movement, in days (`duration`).

Specific analytic issues you must address In general, are non-violent movements more likely to be successful than violent ones? Does violent repression by the government make movements more or less likely to be successful, and is there a difference in this effect between movements which are themselves violent and non-violent? Similarly, what is the effect of foreign aid to the government and to the movement? Do non-violent movements become more likely to succeed as the government becomes more democratic? Does the difference in probability of success between violent and non-violent movements vary with how democratic the government is? All of these should be answered with reference to the results in your model (or models).

Models Use a generalized additive model with a logistic link function; smooth all continuous predictor variables, and include all categorical variables, except campaign and country names, as your default. (Departures from this should be carefully justified.) Be sure to include the year as a predictor variable, and explain the interpretation of your estimated effects for the year. Some of the analytic issues above may be most easily addressed through including interaction terms, or through fitting different models on subsets of the data; describe any such variations, and the reasons for your choices.

Note 1: Before fitting a model with a logistic link function, you will need to re-code partial successes as either successes or failures. Explain which one you chose, and briefly justify your decision.

Note 2: The analysis could also be done with kernel models, and doing so would receive full credit, but computations may take too long. (This could however avoid needing to re-code partial successes.)

Inferential Statistics and Model Assessment You may not assume that R's default standard errors or p-values on estimated regression coefficients 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 probabilities of success. Be sure to describe your procedure and explain why you chose it, that is, why it is appropriate to answer the questions at hand.

Format

Your main report should be a humanly-readable document of at most 10 single-spaced pages, including figures. It should have the following sections:

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;
- 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.

Numerical results Numerical quantities should be written out to appropriate precision, i.e., neither more nor fewer significant digits than appropriate.

Code *All* statistical results must be supported by appropriate code, or they will receive no credit. ("Show your work.") The ideal would be to use R Markdown, or knitr+LaTeX, to embed all computations in a humanly readable document, and submit both the knitted version and the source[1] As a second best, it is acceptable to submit a PDF document containing all text and figures, and a separate .R file, containing all supporting computations, clearly labeled via the comments so that it is easy to see which claims or results go with which pieces of code.

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.

[1] See examples at `http://yihui.name/knitr/demos/`, and the useful chunk options like `echo` at `http://yihui.name/knitr/options/`.

Pictures (5) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text.

Code (15) 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 (if applicable).

Modeling (15) 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.

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 reference to the results of specific tests.

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.