Final Exam: The Union Makes Us Strong

36-402, Advanced Data Analysis

Due at 10 am on Monday, 9 May 2011

Instructions

You will be sent a data set (CSV format) by e-mail to your Andrew account. Each data set is slightly different. Work only with your own. It should have 625 rows and 8 columns. If you have not received a data set, or cannot open it, or it has the wrong format, contact Prof. Shalizi by 9 am on Wednesday, 27 April. If you do not do so, the presumption will be that you have received and can read your data.

Your work for this exam will be a formal report. You will be graded equally on the technical accuracy with which you employ your chosen statistical tools; the quality of the reasons you give for selecting those tools (and not others) and for supporting your conclusions; and the skill with which you use words, graphs and numbers to communicate. Be clear, be concise, and use your own words.

Divide your report into four marked sections: introduction and data summary; methods; results; conclusions. You may sub-divide these, but you must have these four sections; they will be weighted equally.

1. *Introduction* Describe the data and the problem. This can be brief, but should be comprehensible to someone who has *not* read this assignment. Include any exploratory analyses you do to guide your choice of methods.

2. *Methods* Describe your methods. Give enough detail that someone who had taken 401 but not 402 would understand. Explain why your methods are suitable to the problem — how *that* statistical analysis answers *this* non-statistical question. Be explicit about the assumptions your methods make about the data-generating process, and how (if at all) the assumptions can be checked.

3. *Results* Give the results of your statistical analysis. As appropriate, include checks on model assumptions, and on the quality of fits to the data. Describe the results in words, accompanied by numbers and/or graphs as needed — raw or minimally edited R output is unacceptable.

4. *Conclusions* Relate your statistical results to substantive, scientific conclusions. Discuss the statistical significance of your results, their scientific or practical significance, and the uncertainty in your conclusions. As far

as possible, be quantitative about your uncertainty. If there are assumptions your have been unable to check, or sources of uncertainty you are unable to quantify, be explicit about them, and discuss how much they might compromise your conclusions. End with a statement of the strongest conclusions that your data and analyses can support.

All figures should go together after the main text. R may go in an appendix.

Your report should be no more than 12 pages (excluding figures and appendix). Text after page 12 may or may not get graded. There is no minimum length, but anything less than 4 pages is probably too short.

Turn in a hard-copy of the write-up to Prof. Shalizi, either in his office (Baker Hall 229C) or in his mailbox in the statistics department (Baker Hall 232). Include a signed copy of the last page of this exam as a cover sheet. Do not submit your write-up electronically.

Turn in your code by uploading a plain text file to Blackboard. Name the file `andrewID-3.R`, where `andrewID` is your actual Andrew username. Make sure the file can be loaded into R and run; files in other formats (in particular, Word) will not be graded. Include code which will allow us to reproduce all your figures and analyses; this is part of showing your work.

Turn in your work by 10 am on Monday, 9 May. If you have not been able to finish, turn in whatever you have done, for partial credit. Late exams will get no credit.

Background

Finding the factors which control the frequency and severity of strikes by organized workers is an important problem in economics, sociology and political science[1]. Our data set, kindly provided by Prof. Bruce Western of Harvard University[2], contains information about the volume of strikes, and several variables which are plausibly related to this, for 18 developed (OECD) countries during 1951–1985:

1. Country name
2. Year
3. Strike volume, defined as "days lost due to industrial disputes per 1000 wage salary earners"
4. Unemployment rate (percentage)
5. Inflation rate (consumer prices, percentage)
6. "parliamentary representation of social democratic and labor parties"[3]

[1] Or it used to be, anyway.

[2] Whom you should *not* bother with questions.

[3] For the United States, this appears to be the fraction of Congressional seats held by Democrats.

7. Prof. Western's measure of the centralization of the leadership in that country's union movement.

8. Union density, the fraction of salary earners belonging to a union (only available from 1960).

Union centralization is constant for each country over the observation period; the others vary year to year within each country.

The question of interest is, **What determines the volume of strikes?** Specifically, **If you wished to minimize the volume of strikes, which variables would you want to control, and to what values would you want to move them?** (There may be more multiple, qualitatively distinct situations which lead to a low volume of strikes.)

Specific points you may want to consider:

- Some variables are missing for some cases (coded NA).
- Should some variables be transformed?
- What causal structure does your analysis assume?
- Can each year be treated as an independent case? If not, which variables in year t depend on which variables in year $t-1$?
- Which variables might be causal descendants of strike volume? What should be done with them?
- Which variables have interactive and/or non-linear effects on strike volume?
- Should dummy variables be introduced for individual countries, or groups of countries? If so, should they be interacted with any other variables?
- Do any countries or years appear to be outliers? If so, what should be done with them?
- Should the year be included as a variable? If so, how?
- Can you really say anything about causation with this data? If not, what kind of predictions can you make?

Some of these points may be more appropriately addressed in your exploratory analyses than in your results.

36-402, Advanced Data Analysis, Spring 2011

Final Examination

I have read the university policy on cheating and plagiarism (`http://www.cmu.edu/policies/documents/Cheating.html`). I have completed this take-home examination honestly, without giving or receiving prohibited assistance to anyone.

SIGNED:

NAME: