Final Exam

36-402, Section A, Spring 2019

Due at 6 pm on Thursday, 9 May 2019

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

Writing Instructions

Please submit two files to Canvas: one is the PDF (or HTML) of your report; the other is the .Rmd (or .Rnw) file which produced it. Both will be graded.

The exam requires you to answer a sequence of questions. Points are as indicated for each question. In addition, you will be graded on a number of aspects of the quality of your writing and presentation of your results. The quality rubric will account for 40% of your exam grade.

Strikes

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, `http://www.stat.cmu.edu/~cshalizi/uADA/19/exams/2/strikes.csv`, kindly provided by a distinguished specialist in the field, contains information about the incidence of strikes, and several variables which are plausibly related to that, for 18 developed (OECD) countries during 1951–1985:

- Country name (not to be used unless indicated otherwise)
- Year (also not to be used unless indicated otherwise)
- Strike volume, defined as "days [of work] lost due to industrial disputes per 1000 wage salary earners"
- Unemployment rate (percentage)
- Inflation rate (consumer prices, percentage)
- "parliamentary representation of social democratic and labor parties". (For the United States, this is the fraction of Senate seats held by the Democratic Party.)
- A measure of the centralization of the leadership in that country's union movement, on a scale of 0 to 1.
- Union density, the fraction of salary earners belonging to a union (only available from 1960).

Note that some variables are missing (`NA`) for some cases. You will need to handle NAs in some sensible way, and explain what that way is.

1. (10) Use `pc()` from `pcalg` to obtain a graph, assuming all relations between variables are linear. Report the causal parents (if any) and children (if any) of every variable. If the algorithm is unable to orient one or more of the edges, report this, and in later parts of this problem, consider all the graphs which result from different possible orientations.
2. (12) Linearly model each variable as a function of its parents. Report the coefficients (to reasonable precision), the standard deviation of the regression noise (ditto), and 95% confidence intervals for all of these, as determined by bootstrapping the residuals.
3. You should find that there is no edge between strike volume and union density (neither is the parent of the other), but that there is at least one directed path linking them (either density is an ancestor of strike volume, or the other way around).

[1] Or it used to be, anyway.

(a) (8) Find the expected change in the descendant from a one-standard-deviation increase in the ancestor above its mean value.

(b) (5) Linearly regress the descendant on all the other variables, including the ancestor. According to this regression, what is the expected change in the descendant, when the ancestor increases one SD above its mean value and all other variables are at their mean values?

4. Check the linearity assumption for each variable which has a parent. (Putting in interactions and/or quadratic terms is inadequate and will result in only partial credit at best.) *Hint:* Chapter 9.

   (a) (5) Describe your method, and why it should work.

   (b) (5) Report the p-value for each case, to reasonable precision.

   (c) (5) What is your over-all judgment about whether it is reasonable to model each endogenous variable as linearly related to its parents? If you need more information than just p-values to reach a decision, describe it.

5. (10) Discuss the over-all adequacy of the model, on both statistical grounds (goodness-of-fit, appropriateness of modeling assumptions, etc.) and substantive, scientific ones (whether it makes sense, given what is known about the processes involved).

Rubric

As usual, this describes the ideal.

Words (10) The text is laid out cleanly, with clear divisions and transitions between questions. 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.

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 other resources is explicitly acknowledged and cited in the comments. Functions or procedures not taken from the notes or from well-established packages have accompanying tests which check whether the code does what it is supposed to. All code runs, and the Markdown file knits.

Extra credit (5) Write a one-page summary of your findings. As much as possible, use ordinary language (as opposed to mathematical formulas, computer code, or statistical jargon). You may find it useful to adopt the perspective of trying to give advice to a policy-maker who would like to know what actions to take to reduce (or increase) strikes.