Exam 2: Big Sky Country

36-402, Advanced Data Analysis

Due at 11:59 pm on Thursday, 21 April 2016

Instructions

Please read the problem background carefully, before beginning the data analysis. Adequate data analysis here *will* require you to go beyond what you know from linear regression, and use methods from this class. You will be graded not just on the technical correctness of your results, but also on the soundness of the reasoning you use to get to the results, and the clarity with which you communicate both your reasons and your results.

Allowed and prohibit resources You can use your notes, the textbooks, and anything other printed or electronic reference (with exceptions noted below), if it is properly acknowledged. However, **all your work must be your own**. You may not, under any circumstances, discuss the exam with anyone other than the professors and the teaching assistants. This prohibition specifically includes classmates, friends, relatives, strangers, and people online. You may freely use solutions provided *this semester* for previous homeworks and exams, with acknowledgment. You may not read, copy, "consult", "study" or otherwise have anything to do with solutions for this class in previous years. You also may not post this exam or any portion of it to any online forum.

Using prohibited resources, or any form of collaboration, is not just cheating but easily detected cheating. If you find that you have broken these rules, please contact the professors as soon as possible to arrange for an oral examination. Otherwise, being detected in cheating will result in formal academic disciplinary action under the university's policy on academic integrity.

If you have any questions about what is and is not allowed, please ask the professors.

Background

All over the world, people invoke historical episodes and experiences as they try to make sense of political events, and to try to bring people around to their point of view on political matters. Inevitably, these uses of history are highly selective (nobody can pay attention to everything), and often highly "motivated" (people pay more attention to examples that reinforce what they already like), but they can also be consequential (people sometimes change their minds because of historical examples). Cultural anthropologists are interested in studying how popular historical memory interacts with widely-shared values, and how *differences* in the way the past is perceived, within a single population, relate to differences in values.

The data for this exam come from a survey of attitudes towards historical episodes, political values, and human rights, conducted in a small formerly-Communist country, which we may call "Paristan", by anthropologists, studying how the citizens of the new republic made sense of the transition to capitalism and democracy[1]. The survey was conducted in two waves, in 1998 and 2003; the subjects surveyed were different each time, as was the sample size.

Data

The surveys were conducted as interviews, with four sets of variables extracted from recordings of the interviews. One set are demographic variables about the survey subjects (Table 4). The other three variables all concerned whether the subjects *mentioned* certain topics or ideas, divided into attitudes about the past (Table 1), general political values (Table 2), and attitudes specifically about human rights (Table 3).

The scientists who conducted the survey are interested in how the distributions of all three sets of variables have changed between 1998 and 2003. They are also interested in testing the idea that general values (variables in Table 2) "mediate between" attitudes towards the past (Table 1) and attitudes on human rights (Table 3). In particular, they want to know if the changes in attitudes about human rights between 1998 and 2003 can be accounted for by changes in attitudes about historical episodes, while the *relationships* between attitudes towards the past and values, and between values and human rights, did not change.

Because the same subjects were not re-surveyed in both waves, the demographics of the two samples are different, and it is also possible that the differences in attitudes between the two waves (if any) can be explained by changing demographics. (E.g., perhaps young uneducated male members of the majority ethnic group all have pretty similar attitudes, and there were more in the 2003 sample.)

[1]The investigators have kindly given permission for the data to be used in this class, but some identifying details are disguised since it is not yet published, hence the pseudonym.

postsocialist	Post-socialist years
intl.stds	International standards of democracy
socialist	Socialist era
monarchy	Monarchy (through early 20th century)
feudal	Feudal era

Table 1: Variables recording mentions of historical eras; coded 0 if not mentioned, 1 if mentioned.

freedom.oppression	Freedom from oppression
personal.dignity	Personal dignity
selfdetermination	National self-determination
national.dignity	National dignity and acceptance in the international community

Table 2: Variables mentioning political values; coded 0 if not mentioned, 1 if mentioned

hr.personal.dignity	Human rights bring personal dignity
hr.equality	HR bring equality
hr.political.freedom	HR bring political freedom
hr.participation	HR bring citizen participation in government
hr.econ.freedom	HR bring economic freedoms
hr.socioeconomics	HR brings socioeconomic rights
hr.selfdetermination	HR brings self-determination
hr.natl.respect	HR brings respect for the nation
hr.violated	HR are violated or cause problems
hr.support	Government should support HR
hr.democracy	HR are linked to democracy
hr.mentioned	HR mentioned in any way

Table 3: Variables recording attitudes towards human rights; coded 0 if not mentioned, 1 if mentioned

`location`	0 if provinces, 1 if national capital region
`gender`	0 female, 1 male
`residence`	0 rural, 1 urban
`age`	1, 17–26
	2, 27–39
	3, 40–54
	4, 55+
`education`	0, < high school
	1, secondary school
	2, technical college
	3, university+
`occupation`	0, unemployed
	1, student (working age)
	2, pensioner
	3 government worker
	4, NGO worker
	5, private sector
	6, farmer or herder
`ethnicity`	0, minority A
	1, minority B
	2, other minorities
	3, majority
	4, NA

Table 4: Demographic variables and their codes. Note that `age` and `education` are ordinal variables, but `occupation` and `ethnicity` are just categorical.

Scientific Conjectures

The researchers who gathered the data theorized that:

- Responses to the questions about historical attitudes should follow a cluster or mixture-model distribution, with a limited number of clusters.
- Responses to the questions about general political values should follow their own mixture model.
- Responses to questions about human rights should follow a third mixture model.
- Which cluster a person falls into for general political values should be the sole cause of the clusters they fall into for historical attitudes and attitudes towards human rights. That is, the researchers think that there are no other systematic causes of attitudes regarding history and human rights.
- Which cluster someone falls into for general political values should be caused by their pre-existing demographic variables.

Problems

For problems 1 through 5, you may simply drop rows which contain NAs in the relevant variables.

1. (5) Fit a mixture model for the "attitudes towards the past" variables, including determining the optimal number of mixture components or clusters. Report the conditional distributions of the observables for each cluster, along with the cluster proportions, accompanied by appropriate measures of uncertainty.
2. (5) Similarly, fit a mixture model for the "general political values" variables.
3. (5) Similarly, fit a mixture model for the "human rights" variables.
4. The researchers' theory, described above, implies some conditional independence relations among the latent and observable variables.
   (a) (5) Explore the extent to which the demographic variables can predict membership in the clusters estimated for problem 2, and summarize your conclusions.
   (b) (5) Explore whether estimated membership in the clusters for problems 1 and 3 are independent given the estimated cluster membership from problem 2, and summarize your conclusions.

(c) (5) Explore whether estimated membership in the clusters for problem 1 is conditionally independent of demographic variables, given estimated membership in the clusters for problem 2. Summarize your conclusions.

5. (a) (3) Draw a graphical model representing the researchers' theory. Does it imply that the *estimated* cluster memberships used in problem 4 should follow the same conditional independence relations as the *actual* latent variables?

   (b) (2) Is the graphical model for the researchers' theory the only graphical model which predicts this pattern of conditional independences among the observed and estimated (not latent) variables? If so, explain why; if not, present another graphical model with the same implications.

6. (5) Some of the demographic variables contain missing values. Does the presence of these missing values contain any information about the opinion variables? Can cases with missing values simply be dropped when demographic variables are relevant? If not, explain and implement an appropriate method for dealing with the missing values.

   Hints: Things you can explore here include: the distributions of other variables for data points with and without NAs in the demographic variables; formal (χ^2) tests of independence; the over-all distribution of p-values from such tests.

7. (5) Write an assessment of the researchers' theory about the causal structure of these variables. Refer to specific results obtained as answers to the questions above, and be sure to note any relevant points of uncertainty.

Rubric

Words (5) 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 (10) 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 used, and used appropriately.

Inference (10) 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.

Conclusions (10) 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.

Some Hints

In no particular order:

1. This is not a regression assignment, though there may be some parts of some problems where regression is useful.

2. You may use either `poLCA` or `multimixEM` (from `mixtools`) to estimate mixture models. Note that the former might require re-coding some variables, and the latter requires its data to be a matrix, not a data frame. Also, you might have to tweak some of the default settings in either package to get good estimates on this data.

   — If you find another package for fitting categorical mixture models which you like better, you are welcome to use it.

3. Two variables can be perfectly dependent even though their correlation coefficient is exactly zero. Two variables can be perfectly dependent even though their linear regression coefficient is exactly zero. Significantly non-zero correlation or linear regression coefficients are evidence of dependence, but not the other way around. If you want to establish *in*dependence, you need to examine the full distributions. (Several ways of doing this are discussed at various points in the textbook.)

4. Gaussian mixture models will not be useful here at all. Neither will factor analysis[2].

Extra credit (5) Up to five 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.

[2]There are forms of factor analysis where discrete observables descend from continuous latent variables. You are welcome to try them out here.