Stat 133 Final Project: Part 1 (Due: 7/25)

July 14, 2014

Your final project will use data from a Dialect Survey conducted by Bert Vaux. The questions and results can be found at http://www4.uwm.edu/FLL/linguistics/dialect/index.html. In particular, you will be working with questions that focus on lexical (as opposed to phonetic) differences. Included in this folder, you will find the file: LINGDATA.TXT which contains the variables ID, CITY, STATE, ZIP, Q50-Q121 (with a few of the questions removed), LAT, and LONG.

Cleaning

As with many applied projects, the data you are recieving needs to be cleaned before you can begin your analysis. Your first task will be dealing with missing responses. Currently, the Q50-Q121 variables range from 0 to the maximum number of responses for a given question. A value of 0 indicates that a respondent omitted that question when taking the survey. Please deal with omitted responses in the following way:

1. Create a subset of the data where all respondents (rows) that omitted *every* question have been removed.

   (a) In the DATA-CLEANING.R file, create the variable <n.no.response> indicating the number of observations you removed

2. With your subset, find the number of omitted responses for each remaining observation. Create a histogram of this data with the title "Number Omitted Questions" and x-axis label "questions omitted".

3. Again with your subset, find the value of the 99^{th} percentile for number of omitted questions. Create a variable for this value called <non.response.cutoff>. Remove all observations from your subset that omitted more than <non.response.cutoff> questions.

4. Save the remaining observations (after your two rounds of removals) in a file named "ling-data-clean.data".

Reformatting

Unfortunately, the data is not in the proper format for you to perform your analysis. Currently, individual questions are stored as variables and responses are given by an integer values. For your analysis, it will be convenient to create a variable for each question-response combination and store responses as a binary value (1 indicating the respondent chose a particular question-response pair, 0 indicating they did not). For

example consider a two question survey where Q1 has three possible responses and Q2 has two possible responses. The data you currently have might look something like this:

	Q1	Q2
respondent1	2	1
respondent2	3	1
respondent3	0	2

You will need to convert this data to look like:

	Q1.1	Q1.2	Q1.3	Q2.1	Q2.2
respondent1	0	1	0	1	0
respondent2	0	0	1	1	0
respondent3	0	0	0	0	1

In the REFORMATTING.R file please create the function "makeBinary". Your function should take a vector of response values for each question and a vector (of equal length) giving the possible number of responses to each question. Your funtion should return a binary vector that converts responses as previously described.

Use your "makeBinary" function to reformat the data. Save your reformatted responses as a dataframe in a file named "binary-ling-data.data". Your dataframe should contain the following variables with the *same names* and in the *same order:*

ID, CITY, STATE, ZIP, lat, long, Q050.1, Q050.2, ..., Q121.6, Q121.7

with the questions not included in LINGDATA.TXT removed from the above numbering. We have included the first 10 reformatted observations for you to compare with.

Extras

You will need to complete and submit the above tasks by the 25^{th}. However, these are far from a complete examination of the data. The following should provide you with a sense of the types of questions you should start to ask when doing applied research. We *strongly* recommend that you consider them (along with other similar questions) and place your findings in your final report due at the end of the semester:

- Do some questions appear to be omitted more frequently than others? Does this seem to be at all related to geography?
- Do the observations you removed appear to be sampled uniformly with respect to geography, or do certain regions have a disproportionately large number of removed observations?