1 Using Paradata to Assess and Analyze Survey Contact Attempts

Background

Survey researchers collect paradata that provide important metrics for assessing multiple dimensions of the project, including response rates by potential participants and interviewers' adherence to fieldwork protocols. Non-response is determined by collecting information on contact attempts (i.e., people unavailable, unwilling or ineligible to participate in the survey) by interviewers in the field. Those data are inserted into industry-standard formulae for calculating response rates. In order to standardize and monitor this process, LAPOP uses specialized software and an add-on module to capture location and time data while interviewers attempt to secure face-to-face interviews in the field. In theory, these data that track the interviewers' paths can be used to assess the interviewers' adherence to protocols in the field.

Objective

The objective of this project is to conduct an assessment and descriptive analysis of LAPOP's "attempts dataset" to help us understand what information it can provide about interviewer adherence to fieldwork protocols. Following protocol, each full response is preceded by some number *N* of attempts, with variables encoding the outcome of each attempt. In particular, we want to know when interviewers record the attempt at the threshold of the house, and record the attempt one at a time, or do they insert many attempts at once to approximate the number of attempts that they made in a given day (i.e., a "burst" of attempts). We call this interviewer behavior "delayed reporting." Additionally, we want to identify "underrepo›rting." In this scenario the interviewers have unusually high success rates (i.e. those who always, or too often, report getting successful interviews at the first attempt).

Process

The team will first need to become familiar with the attempts dataset from the 2018/19 AmericasBarometer (we will provide data for Ecuador, El Salvador, and Mexico to begin with), and reshape it for the purpose of analyzing it. The attempts database contains, for each attempt, the system captured right before the attempts module and then right before each attempt (to make sure the interviewer is attempting in different households). The dataset also has a time measure. We recommend starting with just one country's attempts data, until the team has gained sufficient project knowledge to expand. In the second stage, the team will analyze the data in order to distinguish between regular patterns (in accord with protocols) and "burst" patterns (delayed reporting) or underreporting patterns of recording attempts. We are interested in understanding how we can assess this data – for example, we would like to know if all interviewers record “bursts”, or is the behavior isolated to subsets of the team within a country and more, or less common across different country teams and are bursts more frequent in higher crime regions? If time permits, we are also interested in understanding whether the data can be used to assess the extent to which the interviewers’ paths follow protocols for the selection of homes. Finally, we aim to identify “underreporting” by looking at the distributions and standard deviations of attempts and finding those individuals who are closer to the left tail of the distribution (1 attempt for each successful interview).

Resources

- Point of contact: Dr. Daniel Montalvo (LAPOP’s Director of Survey Research Operations)
    - Can provide ideas, and data, for linking the contacts attempt database to measures of crime and insecurity
- Attempts module datasets (with neighborhood indicator)
- Response datasets (with neighborhood indicator)
- Fieldwork protocol
- R code for calculating response rates (as supplemental information about data structure)

Deliverables

1. GitHub repository that contains all the files used for the project and that is organized to promote replication of all results.
2. Short overview report with descriptive analysis of the data. The report should help us better understand the data both at aggregate and fine-grained units of analysis.
3. Short presentation (20 min) of your results
4. Some of the questions we are interested in understanding better are:
    a. Are different interviewers doing different things? Where is the variation in interviewer behavior? Does the data reveal that interviewer behavior is random or are there patterns?
    b. Are interviewers adhering to the protocol?
    c. Do any other features in the data, such as crime and insecurity, trend with interview patterns? I.e., are places where crime is perceived as a threat do interviewers seem create more ‘attempts’ bursts?
    d. Time permitting, teams could search for and include external data, such as police reporting on crime levels, in their analysis
    e. We will ask each team to begin with one primary country and move on to a list of secondary countries for their consideration, if time allows.

2 Uncovering Profiles of Central Americans with Emigration Intentions

Background

The number of Central Americans seeking to leave their countries has increased in recent times. This upward trend has motivated those influencing policy, and those researching the topic, to try to better understand what factors are pushing individuals to emigrate from countries such as El Salvador, Guatemala, and Honduras. Key suspects include crime, crime-related insecurity, economic vulnerability, lack of job opportunities, poor governance, and food insecurity. A standard practice is to “horse-race” proxies for these factors via regression analysis, but what if some of these factors tend to cluster together in certain individuals? Perhaps crime victimization is less relevant on its own than it is when it is combined with perceptions of poor governance (e.g., impunity, corruption), and perhaps that joint condition matters more for women than men. Standard regression approaches are not well-suited to undercover the varying profiles of those with intentions to emigrate.

Objective

The objective of this project is to analyze survey data on intentions to emigrate out of El Salvador, Guatemala, and Honduras, to assess and describe how various potential predictors of emigration intentions cluster together. We want to understand what alternatives to regression analysis we could use to answer this question, and what types of answers we get conditional on different approaches (e.g., cluster analysis, factor analysis).

Process

The team will first need to become familiar with the topic by reading a few short LAPOP reports and briefs that provide background in terms of likely predictors and the use of the AmericasBarometer survey dataset to assess emigration intentions. The team will then use the AmericasBarometer datasets for El Salvador, Guatemala, and Honduras to perform factor and cluster analyses and report on the findings from their analysis, comparing across different approaches and comparing to answers provided in recent LAPOP reports that use descriptive and regression approaches. We recommend beginning with one country before expanding, so that the team can take stock of what it learns in its initial study of one country to develop and test out an approach to analyzing data from the other two countries. If time permits, we’d be interested in replicating the final approach with data from earlier AmericasBarometer surveys, or other countries (e.g., Mexico), to see if the patterns are similar or different across time and/or place.

Resources

- Point of contact: Claire Evans (PhD candidate and LAPOP Research Assistant)
- Response datasets (subscriber level) with Q14(F) included
  - Q14 asks an individual whether they have an intention to leave their country to work, study, or live abroad in the next three years
  - 2018/19 contains a follow-up question (Q14F) in some countries, asking respondents to Q14 how likely they are to emigrate--in those cases the overall likelihood to emigrate is of most interest
- Reports containing prior analyses of these questions
  - Mexico
  - Honduras
  - Guatemala
  - El Salvador
  - Ecuador

Deliverables

1. GitHub repository that contains all the files used for the project and that is organized to promote replication of all results.
2. Short (10 pages or less) report with descriptive analysis of the data. The report should help us better understand the data both at aggregate and fine-grained units of analysis. Some of the questions we are interested in better understanding are:
   a. Explore the data to help us better understand the motivation for emigration among the interviewees.
   b. Provide a visual narrative that describes interviewees' intention to emigrate.
   c. Examine patterns using linear modeling and compare linear modeling approaches to machine learning approaches. These can be unsupervised methods (such as clustering) or other methods the team deems useful.
   d. Develop a model to best predict which interviewees are likely to emigrate and/or present a 'profile discovery' analysis so we can better understand who is most likely to emigrate.
3. Short presentation (20 min) of your results
4. Students will be assigned one country's data to begin with and a list of other countries to explore, if time allows.

3 Developing a Prototype for a Data Playground (Online Data Analysis Dashboard)

Background

The LAPOP lab directs and processes dozens of survey projects every two years, and makes datasets publicly available. The raw datasets are available in different formats, for skilled users. However, non-skilled users – such as students, journalists, and individuals working in international development – often want to query the datasets. They need an interface that is user-friendly and permits them to generate simple tabulations of the data: at a minimum, univariate and bivariate tables that show frequencies, proportions, measures. LAPOP's current method is increasingly out-dated and clunky: it is an interactive system hosted by the University of Costa Rica. Because it is run by an external party, it is cumbersome for LAPOP to administer and update. LAPOP has been working with the Data Science Institute on a new "dashboard" or "playground" for data queries and visualizations, but that project is not yet complete.

Objective

The objective of this project is to develop a proposal for a dashboard/playground through which external parties can query the LAPOP AmericasBarometer dataset via a point-and-click approach. The resulting output must accurately represent the survey, for example in including appropriate weights and sample design adjustments in the calculations, and easy to interpret.

Process

The team will first assess the existing interactive portal for data analysis available on LAPOP's website, as well as data query portals on other survey project websites (e.g., the ArabBarometer), to assess the state of things. The team will then work to develop a dashboard prototype for a new interactive portal for querying the AmericasBarometer dataset, to produce tabulations that reflect the data accurately, including with respect to the appropriate weights and sample design adjustments.

Resources

- Point of contact: Maita Schade (LAPOP Statistician and Head of Online Division)
- LAPOP survey datasets
- Previous "System for Online Data Analysis": http://lapop.ccp.ucr.ac.cr/en
- Readings on the analysis of weighted survey data

Deliverables

1. Provide a review of existing dashboards of survey projects online
    a. what do others do, what works and what doesn't?
    b. Analyze other dashboards' design and delivery
        i. https://www.arabbarometer.org/survey-data/data-analysis-tool/
        ii. https://afrobarometer.org/online-data-analysis/analyse-online
        iii. http://www.latinobarometro.org/latOnline.jsp
        iv. Any others you can find!
2. Explore LAPOP's data and make suggestions for which visualizations might be the most interesting to showcase online for our audience.
3. Short presentation (20 min) of your results
4. Develop ideas for an online dashboard through the development of a Shiny app allowing for interactive analysis of AmericasBarometer data from different angles.