Public Opinion Question Database Documentation

December 23, 2019

Contents

Project Description

The purpose of this Github repository is to allow researchers to merge numerous public opinion datasets from a variety of sources into a single dataset. The repository houses all of the R code needed to load, parse, import, and merge files, as well as to identify questions and analyze the data.

Summary

The files in the Github respository originated as part of a larger study on abortion attitudes and opinions conducted from 2017-2019. A portion of the project, conducted at the University of Michigan, involved creating a single file that comprehensively archived survey questions asked on abortion attitudes over time. For this part of the project, data were collected from representative surveys or polls that asked a question about abortion. The research team combed data repositories for all publicly available data, relied on restricted use data, and included proprietary data where it could be made availble. Datasets were in a variety of formats, including fwf, dat, data, sav, por, portable, rdata, tab, rda, sps, and dta. This Github respository archives the code associated with combining and organizing the datasets.

—-DOES THIS INCLUDE TOPILINE ONLY FILES? CAN I CALL THESE DATASETS???

Data

The data associated with this project are restricted access; data were gathered from collections and sources that require membership and can only be made available directly from the organizations responsible for archiving the information. Data come from a variety of archives including Dataverse, the Roper Center, ICPSR, RAND, Quinnipiac, and the Pew Center, and sources such as GSS, ANES, ABC News-Washington Post Polls, etc.

Universe

500 studies, over 5 million respondents, WHAT SHOULD I PUT ABOUT THE SIZE/SCOPE OF THE STUDY?

Citation

Pasek, J. and Lippman, J. R. 2019 Public Opinion on Abortion Question Archive. Ann Arbor, MI, GitHub repository.

—-NOTE: FOR DOI https://help.github.com/en/github/creating-cloning-and-archiving-repositories/referencing-and-citing-content

Funding

Basic Concepts

Data Directory

Top level data directory for all surveys. Includes folders from each survey organization which contain every survey that the organization has run on the topic of interest. Within each individual survey folder, there are typically several files—often the data files in several formats and related survey documentation.

Keyword Files

Metadata file in csv format providing study information at the dataset level. Provided by Roper.

Parser Files

Files manually created to read dat format from Roper. Question-level CSV file that lists question by topic, question wording and response categories, and contains other metadata, such as survey begin and end dates, N, sample type, survey organization, etc.

Topic Files

Metadata file in csv format providing information at the question level. Provided by Roper. Catalogues each question by topic, and contains other metadata, such as survey begin and end dates, N, survey organization, survey sponsor, sample type, etc.

Topline File

For each survey question, percentage of respondents selecting each answer category. Usually csv, pdf, txt, rtf, doc, docx. Roper toplines are csv files, usually called “report.csv”. Roper is the only source of toplines in the abortion study. Roper csv topline files are set up so that each case is a response category with variables related to the question and the percentage of respondents selecting the response.

User Guide

This section provides an overview of all of the R files in the repository. To use the R files, the user begins with a top level data directory that includes folders for each survey containing several files, often with data files in several formats and related survey documentation.

01 Get All File Info

This R file (01_GetAllFileInfo.r) identifies all the files located within the data directory to be included in the user's project and gives the user information about what is being imported. Specifically, this R routine produces a csv and rdata file containing information about all of the files located within the data directory, looking within each survey folder and identifying files by extension type. The routine searches the data directory for data except for "donotread." It deals with duplicate names by assuming that they are from different filing schemes. This information is then saved to a folder called "Summary Information" as a csv file (StudyFiles.csv) and also produces an rdata file (located in EachFileExports/01_FileInformation.rdata).It can, for example, tell the user how many csv, doc, dta, or other file types are located in the data directory.

02 Load and Parse

The purpose of the 02_LoadAndParse.r file is to catalogue the data by producing a single csv and rdata file that contain the metadata from each survey or topline. This includes study ID, start and end date for the study, country where data collection took place, study sponsor, survey organization, type of sample, N, mode, notes, study title, and other information. For the abortion study, topic files and toplines were used that were created by Roper. These types of files are typically available for any study, and can be exported. For the abortion study, some metadata for surveys from other sources were manually recorded, and these are also imported at this step. The routine also produces error files for toplines and parsers, indicating duplicate question numbers, missing metadata, parsers or toplines that cannot be used or have parsing failures. Error files list the studies or toplines with problems, and these can be investigated for correction or follow-up.

03 Import Datasets

The 03_ImportDatasets.r file imports all of the data from data files including rdata, sps, spss and portable (por) files, stata, and tab and saves them collectively based on file type. It also parses and then imports all data from files originally from dat format (i.e., the Roper files), linking the parser files with the dat files and exporting the numeric data. For all of the file formats, errors are flagged where necessary, and troubleshooting files are produced in csv format that list all of the files that cannot be imported, categorized by problem type. Note that the parser files will work on any kind of fixed-width data.

04 Copy Data to Dropbox And Merge

This file (04_CopyDataToDropboxAndMerge.r) builds an rdata file that provides the labels for the question wording and response options for the datasets (or for each dataset????????????) from the previous step. It does this by first extracting and adding the wording and response options using the parser files and adding them to the previously parsed datasets, and provides error files that list datasets where the number of response options in the parser file and the number of values in the data do not match. It also lists errors from the process of building the parsed datasets in a csv file. Next, this R file then combines all the datasets that were imported using the previous file (i.e., 03_ImportDatasets.r), listing errors in csv files that deal with duplicate level names, corrupt data imports, and empty data frames. Finally, it identifies question wording and response options for all of the variables based on the labels provided in the original sources of data.

05 Question Identifications

The Question Identification R file (05_QuestionIdentifications.r) adds the qestion wording and response options from the topline files to the rdata file in the previous step so that there is a single, comprehensive rdata file of question wording and response options. It does this primarily by extracting the question wording and response options from all of the topline files. The topline files, from Roper, are arranged such that each response category is a separate case. Only toplines in csv format are imported. It also produces a csv troubleshooting file that lists all of the toplines where there are duplicate question names so that the user can go back and fix these in the original topline files.

06 Wording Flags

The purpose of this routine is to find the variables and recode and relabel them. The R file goes through all of the question and response options and converts as many of the labels as possible to a single standard. To do this, it first calls 99_TranslateResponses.r which orders demographic variables and cleans label abbreviations and misspellings. It looks at the variables sex, race, hispanic origin, state, region, education, religion (denomonation, fundamentalism, attendance frequency), weight, partisanship, party strength, presidential approval, age, marital status, and–for the abortion study–abortion questions.

The routine writes three types of csv files for each variable that is being incorporated into the final dataset: CheckTranslate files, CheckWords files, and CheckDrops files. The check translate files show how well the routine did at relabeling the answer categories. The user can examine any of the csv files in CheckTranslate to compare the original categories with the relabeled answer categories. The column in this csv file, labeled "unmatched" shows any unusual labels that can't be relabled by the routine. The "n" column shows how many times the labeling scheme from the "original" column appears in the datasets. The csv files in the CheckDrops folder show variables that have been dropped from the final dataset because they are unmatched after translation; since they can't be appropriately recoded or relabeled they are dropped.

The routine produces an rdata file with cleaned question and response options located at EachFileExports/06_SelectedQuestionInfo.rdata. Please note that the 06_WordingFlags.r file takes a long time to run.

—writes "CheckWords/RaceWords.csv" question wording - how often each type appears?

07 Abort Classes

07_AbortClasses.r pulls in the raw question-level data from the previous step (06 Wording Flags) and categorizes it by question and response type. It excludes questions on abortion that are not relevant to abortion attitudes (e.g., recognizing the position of candidate's positions on abortion). For the remaining questions, it merges them with the metadata from the second r file (i.e., 02_MetaInformation.rdata) so that more information can be appended to the question categorizations.

The routine then writes three types of item banks as csv files: Item Bank by Wordings, Item Banks by Question Subtype, and Item Bank By Question Type.The first of these lists every question asked on abortion with the full question wording, the responses, the number of times asked, when it was last asked, the total N, the survey organization, and sponsor, the data type, and provides the classification of the question by type and subtype (e.g., Attitude Strength and Importance). The latter two files collapse the Item Bank by Wording file into question subtype and type. Overall, there are 9 subtypes (Importance, Access, Maternal Circumstances, Morality, Pregnancy Circumstances, Pregnancy Term, Regulation, Pro-Choice/Pro-Life, and

NA). The Item Bank by Question Type tells us, for example, that Importance questions were asked 1,342 times of 1,786,826 respondents between 1955 and 2018 (when the datasets were incorporated into the project), and provides us with further metadata related to the survey organizations and sponsors that asked that type of question.

The routine writes 07-01_UncategorizedAbortionQs.csv, a troubleshooting file, to list the questions (and number of times they were asked) that appear in the NA category of the Item Bank files. Finally, the routine writes the rdata file for the categorizations, EachFileExports/07-01_AbortionVariablesCategorized.rdata.

08 Demog Classes

08_DemogClasses.r pulls in the cleaned questions and response options from the demographic variables in step 06 Wording Flags, and categorizes variables from that step by type. It also further cleans the demographic variables.

09 Demog Imputations

The file 09_DemogImputations.r imports the abortion variables as individual level data. It puts the responses from factor variables in order based on their ordinality, using the first factor from principal components analysis to order the responses. It also builds several study level datasets. Finally, the routine uses Amelia to multiply impute missing data.

10 Combine Datasets ReDo

The purpose of 10_CombineDatasetsReDo.r is to build a dataset where each case is an answer to an abortion attitude question. Each case contains the “respID” (the respondents answer), the study ID (“StudyID”), and a “UniqueID” for the abortion attitude question. In this construction of the dataset, each person may be represented more than once (for example, if they answer more than one question about abortion attitudes).

The routine also cleans up and adds on the imputed data. It limits the imported data to either directional (e.g., more or less support) or strength measures (e.g., strongly agree to strongly disagree), and produces troubleshooting files for variables that did not import correctly. Next, it figures out what answers rarely get chosen, limiting the dataset to questions for which >80% give a moderate probability answer (of 5% or more). (I DON’T REALLY KNOW WHAT THIS MEANS).

Next, the routine identifies the items where the responses have known ordinality and direction (and more than two response options). This serves as a training dataset for items with unknown direction (that is, items with unknown direction have more than two response options and are known to be ordinal, but have unknown direction). It creates csv files listing the uncategorized response options and the untrained items, and writes an rdata file “10_IndividualMasterPreReversals.rdata.”

11 Standardize Metrics

The 11_StandardizeMetrics.r file takes the ordinal directional questions and strength questions (from the training dataset) recodes them from 0 to 1 and standardizes them. Next, it runs a prediction on the standardized data and imputes demographics for cases. Then, it predicts probabilities with the imputed demographics on the standardized versions in the training set. After that, it takes the predicted probabilities for the trained and untrained data, identifying the variables to be reversed. To do this, it correlates the predicted probability with imputed demographics with the untrained data within each variable to see whether they are positive or negative with the predicted probability. Then, the data are recoded, the regression is rerun, and the data are tuned. The data are then rebuilt at study level To assess correlations between items by type. The routine also produces a figures “CorrelationsBetweenVariableTypes.pdf” that shows these correlations. Finally, it writes a file, “10-02_FullDatasetWithReversals.rdata” in which all ordinal variables are coded in the correct ordinal direction and each case is an individual’s abortion measure answer.

THESE ARE THE INDIVIDUALS ANSWERS TO ABORTION QUESTIONS AND NOT THE TOTAL ANSWERS PER CATEGORY, YES?

12 Analyses

The 12_Analyses.r file produces plots to examine positions on abortion (i.e., directional differences) and differences in the strength of position on abortion by various demographics, including age, sex, and party identification. It produces several pdf files of multiple plots by question type (directional or strength) and demographics over time.

—Loop Back Recode????

13 Topline Analyses

This routine, like 11_StandardizeMetrics.r reverse codes variables in the topline files so that all variables run in the same ordinal direction. It merges similar variables across studies and produces figures of the distribution of responses to different types of abortion questions over time.

97 Temp Foundation Code

—No idea what this does

98 Wording Tests

–Checks the label abbreviations. How is this different from 06 check translate files?

99 Translate Responses

Called by 06_WordingFlags.r, this file orders demographic variables and cleans label abbreviations and misspellings.See the section on Wording Flags for more information.