How to Find Your Dataset

Bolun Zhang

4/6/2021

Main Approaches to Find Your Dataset

Usually in quantitative research in the social science, you either construct your own data set, or you will use a data set that someone else built. Since the rising of the computational social science and big data in the social science, the first way is becoming more popular, but the second way is still the mainstream. This handout will focus on the second approach to the find the data set you need.

There will be three main ways to find the data set you need:

1. Through previous research on the same topic: what kind of data are they using?
2. Focusing on your own research question: what kind of variable can you use to answer your research question?
3. Commonly used data sets in the social science and other ways to find them.

The following part will cover these three approach respectively.

Data from Previous Research

Usually, a topic that you are interested in has already been covered by some previous research. One way to get your data is to first check whether you can access the same data set that other people have already used.

This is also related to the replication of previous research. The importance of replication has been emphasized by several important scholars in the social science, and more researchers now offer the data and codes for replications online. There is usually a link listed in the paper or the web page of it in the journal.

You can also search research on the following website, to see whether there are papers around your interested topics that offers replications data. The pitfall of this approach is that sometimes the authors only offer recoded data due to some regulation rules.

- https://dataverse.harvard.edu/

Evan if you are not using the same data, it is still useful to read scripts that authors provide. There is usually more information about operationalization in the codes, and it is a good way to learn coding.

You can also reach out to the author directly. Most authors will be happy about the fact that someone else other than the editors and reviewers is interested in their works.

Focusing on the Research Question

The other way is to cut in through what variable can help you answer your research question. You can search related variables on ICPSR through the description or related questions in the survey.

Usually, you can first search for variables that can be used as dependent variables or key independent variable. This will result in a short list. Among data sets on the short list, you can further search for other variables.

- https://www.icpsr.umich.edu/web/pages/
  - ICPSR includes a great variety of datasets that covers not only the U.S, but also other societies.
  - Its searching function is really easy to use.

General and Domain Specific Datasets

GSS and its family

GSS usually covers a great number of variables, though many of them are year specific. It is still valuable to revisit this data set.

Also, GSS's mode is very influential. You can find counterpart data set in other countries. It is possible to carry out certain comparative studies. A few examples are listed.

- GSS: https://gss.norc.org/
- KGSS: https://www.icpsr.umich.edu/web/ICPSR/series/288
- CGSS: https://dss.princeton.edu/catalog/resource825

PSID

The Panel Study of Income Dynamics (PSID) is a very influential panel data set to study social stratification and social mobility. There is a restricted version where geo-location data is provided. You need to have a authorization to utilize the restricted version.

- https://psidonline.isr.umich.edu/

PSID is very difficult to begin with, but a really useful data set in the social science.

NCES data sets

NCES provides a series of data about the education field in the U.S. These data sets can be grouped into two parts: individual level, and school and above.

- https://nces.ed.gov/
- NELS, ELS, HSLS is a series of education attainment data sets that are frequently used in education sociology.
- CCD and PSS are data about public school and private school. CCD also has a school district and county level data.

Other data set about the U.S.

IPUMS provides a easier way than American Fact Finder to access census data and American Community Survey data and many other data products about the United States. You can check their website for more details.

- https://ipums.org/

Other Specific Areas

Usually, in the handbook of a sub-area in sociology, there will be a chapter that introduces some data set that one can use. For example, in *Methodological Practices in Social Movement Research*, there is a chapter about protest event analysis. Almost every important data set is listed there. So referring to handbooks, though a old fashioned way, is still helpful

Other way to find your data

You can also search related data on the following website.

- https://datasetsearch.research.google.com/
- https://www.kaggle.com/datasets
- https://github.com/

Technical Issues

You can refer to Intermediate Importing Data in R: Chapter 5 Importing data from statistical software packages.

- https://campus.datacamp.com/courses/intermediate-importing-data-in-r/importing-data-from-statistical-software-packages?ex=1
- https://strengejacke.github.io/sjlabelled/reference/read_spss.html

If you have further question, please let me know.