How to Borrow Information from Unlinked Data? A Relative Density Approach for Predicting Unobserved Distributions with an Application to Wage Inequality Research

Siwei Cheng

Department of Sociology

New York University

How to Borrow Information from Unlinked Data? A Relative Density Approach for Predicting Unobserved Distributions with an Application to Wage Inequality Research

Abstract

One of the most important developments in the current era of social sciences is the growing availability and diversity of data, big and small. Social scientists increasingly combine information from multiple datasets in their research. While conducting statistical analyses with linked data is relatively straightforward, borrowing information across unlinked data can be much more challenging due to the absence of unit-to-unit linkages. This paper proposes a new methodological approach for borrowing information across *unlinked* surveys to predict unobserved distributions. The gist of the proposed approach lies in the idea of using the *relative density* between the observed and unobserved distributions in the reference data to characterize the difference between the two distributions and borrow that information to the base data. The key benefit of this approach is that it relaxes the conditions of comparable sample representativeness and comparable measurement across data, and instead relies on the assumption that the relative density between the observed and unobserved distributions is the same or very similar across data. It also has the additional advantage of allowing the researcher to borrow information about the entire distribution, rather than a few summary statistics. The approach also comes with a method for incorporating and quantifying the uncertainty in its output. We illustrate the formulation of this approach, demonstrate with simulation examples, and then apply it to address the problem of employment selection in wage inequality research.

Introduction

One of the most important developments in the social sciences of the current era is the growing availability and diversity of data. Social scientists draw inference about social processes not only by using a single dataset, but also, and increasingly so, by combining information from multiple datasets and taking advantage of their differential strengths. There are two types of setups for combining information across data. The first type is *linked data* — that is, datasets in which a unit-level identifier is available for mapping observations in one dataset to their records in another dataset. For example, micro-data from household surveys can be linked to individuals' administrative tax and benefit records using the Social Security Number (SSN) or a protected identification key (PIK) (Abowd et al., 2006; Grusky et al., 2015). The second type is *unlinked data* — that is, data sets that do not share a common identifier which allows for unit-to-unit linkage. For example, while the major national surveys in the United States, such as the Current Population Survey and the Panel study of Income Dynamics, contain useful information about various aspects of individuals and households, it is difficult to link individuals in one survey to another because of the lack of publicly available individual-level identifiers across these surveys.

Conducting statistical analysis with linked data is usually not much different from analyzing a single data set, because the linkage keys allow the researcher to directly merge two data matrices. Yet, borrowing information across unlinked data can be much more challenging. While aggregate-level measures, such as mean and variance, within subgroups of population can be borrowed across the data, two conditions are usually required. The first is *comparable sample representativeness*: the data should be representative of the same or similar population. For example, we can borrow information on wage distributions from the PSID to CPS, assuming that they are both representative of the general population after appropriate weights are applied, but we cannot always directly borrow distribution of wages from a single firm to all firms because they are not representative of the same population. The second is *comparable measurement*: the key variable whose distribution we want to predict needs to be measured consistently across data. For example, if we would like to directly borrow income distribution across two data sets, we need to assume that the income variable is measured consistently in these two data. The requirement of these two conditions for borrowing variable statistics across unlinked data substantially limits the possible data whose information we can combine across.

This paper proposes a new methodological approach for borrowing information across unlinked data to predict unobserved distributions. Specifically, this approach can be used in scenarios where some variable of interest is unobserved for part of the sample in the one data set but this unobserved distribution can be observed or measured using a proxy in an external, unlinked data set. The gist of the proposed approach lies in the idea of using the *relative density* between the observed and unobserved distributions in the reference data (i.e. the data set with a smaller sample but containing information on the unobserved distribution) to characterize the *difference* between the two distributions and borrow that information into the base data (i.e. the main data set with a larger sample but containing no information on the unobserved distribution). The methodological approach for constructing the relative density builds on the statistical literature on relative distribution methods, also known as grade transformation (Ćwik and Mielniczuk, 1989; Handcock and Morris, 2006). The key benefit of this approach is that it relaxes the conditions of comparable sample representativeness and comparable measurement across data, and instead relies on a different assumption: that the relative density between the observed and unobserved distributions is the same or similar between datasets. One additional advantage of the relative density approach is that it allows the researcher to borrow information about the shape of the entire distribution, rather than a few summary statistics.

We start with describing our key motivating example — the problem of employment selection, also known as the sample selection bias — in wage inequality research. Employment selection problem happens when the wages are observed for those who are working but unobserved for those who are not working, and that the observed distribution differs systematically from the unobserved distribution. We follow up with a discussion on the strengths and limitations of previous approaches. Then, we introduce our analytic approach using both mathematical formulation and simulation examples. Finally, we apply the approach to the empirical analysis of wage inequality over the past two decades in the United States, focusing specifically on the ways in which accounting for the unobserved wage distribution among the nonworkers will affect our conclusions about wage inequality.

References

Abowd, J., M. Stinson, and G. Benedetto (2006). Final report to the social security administration on the sipp/ssa/irs public use file project.

Ćwik, J. and J. Mielniczuk (1989). Estimating density ratio with application to discriminant analysis. *Communications in Statistics-Theory and Methods 18*(8), 3057–3069.

Grusky, D. B., T. M. Smeeding, and C. M. Snipp (2015). A new infrastructure for monitoring social mobility in the united states. *The ANNALS of the American Academy of Political and Social Science 657*(1), 63–82.

Handcock, M. S. and M. Morris (2006). *Relative distribution methods in the social sciences.* Springer Science & Business Media.