Oshrat Hochman (GESIS Leibniz Institue for the Social Sciences)

Open science: Open Methodology

Topic 4 – Getting the statistics out

Keywords: Open Methodology European Microdata Documentation

Introduction

The concept of open science, has gained immense importance in the last few years throughout the world and particularly in Europe, where the European Commission presented the vision that by 2020, "all European researchers will be able to deposit, access and analyze European scientific data through a European Open Science Cloud." When we think about open science we often think about three main domains we wish to make accessible: access to scientific publications, access to scientific data, and access to scientific source code, or in other words computer syntax, programs and packages. Going back to the foundations of empirical science it is clear that we need to gain access to scientific publications in order to learn what questions they ask and what answers they provide. We then also need access to the observed data in order to replicate the findings and see whether the answer provided is sustainable or not, and to software that can allow us to analyze it.

Methods / Problem statement

However, having all this access will get us only half of the way. The rest of the way, we will often times have to figure out on our own, using incomplete information regarding the analysis that led to the reported findings and to the conclusion derived from them. This is why we need to speak about "open methodology".For example, a researcher claims that the level of education determines the social status of individuals. We can all agree that this claim is highly plausible, but should we engage in testing the empirical validity of this claim we are immediately challenged by two questions? What is "level of education" and what is "social status"? These questions may appear simple but they are crucial for the empirical test: in order to test the validity of this claim one would have to test it using the exact same definition of level of education, and of social class. Otherwise, any finding we derive that disagrees with the researchers claim would be dismissed as associated with the replication procedure.

Results / Proposed solution

Some of the European microdata from official statistics present a unique challenge in this regard namely one associated with the ex-ante Output Harmonization: Eurostat regulations define mandatory variables, called "target variables". However these target variables are not collected using the same formulation or by means of the same questionnaire across different countries. In fact, in some countries the target variables are not collected through questionnaires at all but rather delivered from existing register data. The variation in the information used to construct the target variables may reduce the ability of empirical research to replicate existing results for a specific country using such target variables. A second example of the need for better methodological documentation is associated with derived variables like disposable household income, where again, users get access to an end result of some statistical procedure without knowing how this result was achieved, what assumptions it makes regarding the different components and so on.

Conclusions

Documentation of the national data, and the statistical procedure used to transform this data into the international target or derived variables in English (i.e. open methodology) will not solve the problem of

replication to the full, but will insure that users are aware of potential reasons for their inability to replicate exiting findings.