

Open Research Data in Economics

Velichka Dimitrova

Open Economics Working Group, Open Knowledge
Foundation, London, United Kingdom

Just a few decades ago, particularly in the 1970s and 1980s, empirical work in economics lacked credibility: modifications to functional form, sample size or controls could change the findings and conclusions. Edward Leamer (1983) criticised the fragility of econometric results, saying that to draw inferences from data described in econometric texts, it was necessary “to make whimsical assumptions”. For a long time nobody trusted the results of econometric papers.

Since then, better research designs, experiments or good quasi-experiments has lead to a credibility revolution in economics (Angrist and Pischke 2010) and “taking the “con” out

How to cite this book chapter:

Dimitrova, V. 2014. Open Research Data in Economics. In: Moore, S. A. (ed.) *Issues in Open Research Data*. Pp. 141–150. London: Ubiquity Press. DOI: <http://dx.doi.org/10.5334/ban.i>

of econometrics”. Leamer’s judgement of the empirical work of his time – that “hardly anyone takes anyone else’s data analysis seriously” seems to be less justified today, largely due to quality research designs. Miguel et al. (2014) argue that these changes have been particularly pronounced in development economics with a large number of randomised trials in recent years.

Parallel to this trend, new opportunities of gathering and processing data has made some researchers enthusiastic about the opportunities to create novel research designs, to analyse large and granular datasets, allowing for better measurements of economic effects and outcomes, etc. (Einav and Levin 2013).

Data itself has become “the new oil” or “a new asset class” (Schwab et al. 2011). In many sub-disciplines of economics, the surging number of empirical papers has attested that greater availability of micro data has permitted “rigorous empirical analyses of questions that cannot be answered purely based on theory” (Raj and Finkelstein 2012).

The hype about all the opportunities which data creates often pays less need to the questions of access to data, reproducible research and transparency. Who if not economists understand the value generated by having open access to knowledge and data as well as the benefits of knowledge as a public good?¹.

Making economics research data and code available serves to enable scholarly enquiry and debate and to ensure that the results of economics research can be reproduced and verified. This is the

¹ Having the properties of non-rivalrousness and non-excludability, knowledge could be considered a public good or at least an “impure public good” as returns to some forms of knowledge can be appropriated to some extent (Stiglitz 1999).

rationale behind the Open Economics Principles², a Statement on the Openness of Data and Code – <http://openeconomics.net/principles/>. The purpose of the Principles is to provide some basic guidelines on why, how and when data in economic research should be open.

The first Principle is to have “**open data by default**” where *“data in its different stages and formats, program code, experimental instructions and metadata – all of the evidence used by economists to support underlying claims – should be open as per the Open Definition³, free for anyone to use, reuse and redistribute”*. Having open data by default sets a gold standard for research in economics, where any researcher would have to abide by this principle where possible. Some empirical economists do provide access to their data and code on their websites and actively encourage their research to be replicated (where Joshua Angrist’s data archive⁴ is a leading example), yet there are still relatively few who do so.

Whilst many initiatives exist in the field of the natural sciences, social scientists and economists have been more hesitant about opening up data and code. Economists work with diverse and often sensitive data. Original empirical work depends on having unique datasets with individuals, households or firms as observation units. Such data may contain sensitive information or may be subject to confidentiality agreements. The researchers may also not own the data they work with.

² The Open Economics Principles were created by The Open Economics Working Group of the Open Knowledge Foundation. The statement was brought forward by an Advisory Panel (<http://openeconomics.net/advisory-panel/>) of economics professors, funders and practitioners with the support of the Alfred P. Sloan Foundation.

³ <http://opendefinition.org/>

⁴ <http://economics.mit.edu/faculty/angrist/data1/data>

Therefore, the second Principle recognises that *“there are often cases where for reasons of **privacy, national security and commercial confidentiality** the full data cannot be made openly available. In such cases researchers should share analysis under the least restrictive terms consistent with legal requirements, and abiding by the research ethics and guidelines of their community”*. Researchers would still be encouraged to open up non-sensitive data, summary data, metadata and code where applicable, as legal agreements may often allow for some degree of sharing.

Privacy and confidentiality are, however, not the only reasons for not opening up data and code. Access to quality and high-frequency data is often not free and requires significant investment of research resources. Gathering particular novel datasets requires a resource investment and researchers may not be willing to share data until they have exhausted all returns associated with their investment.

For that reason, the third Principle attempts to summarise the need to offer a reward associated with sharing as it deals with **reward structures and data citations** – *“recognizing the importance of data and code to the discipline, reward structures should be established in order to recognise these scholarly contributions with appropriate credit and citation in an acknowledgement that producing data and code with the documentation that make them reusable by others requires a significant commitment of time and resources”*.

The Principles also draw attention to the efforts of data curators which often are under-appreciated, but who have a major role in supporting researchers in gathering, documenting, storing and sharing research data.

Further rewards associated with the sharing of data and come may become more common in the future. Data citations are seen

as a way to reward the efforts of researchers in producing data and making it easier for others to find and access datasets. If researchers begin to cite data the same way they cite articles and books, it would allow for tracking the data's impact, verifying and re-using it as well as acknowledging the contribution of the data producers⁵.

Nevertheless, datasets in economics are most frequently related to one or more academic papers. The data and code serve to enable the verification of empirical results. Thus, the fourth Principle deals with the **data availability**: *“Investigators should share their data by the time of publication of initial results of analyses of the data, except in compelling circumstances. Data relevant to public policy should be shared as quickly and widely as possible. Funders, journals and their editorial boards should put in place and enforce data availability policies requiring data, code and any other relevant information to be made openly available as soon as possible and at latest upon publication.”*

Recognising that data and code should be made available, economics journals have put in place data availability policies. The American Economic Review, which could be seen as setting the tone for the policy of other journals⁶, requires the authors of accepted empirical papers to provide prior to publication all necessary data and computation necessary for replication and promises to make it available on the AER website⁷. Accordingly, the majority of the more recent AER articles have their datasets available online.

⁵ See the DataCite project for details: <https://www.datacite.org/>

⁶ The project EdaWaX evaluated the data availability of economics journals: <http://openeconomics.net/resources/data-policies-of-economic-journals/>

⁷ The data availability policy of the American Economic Review: <http://www.aeaweb.org/aer/data.php>

In fact, the availability of raw data related to a paper is not a new issue. In what became to be regarded as the first referee report of an article submitted to *Econometrica*, Ragner Frisch commented on the work of Henry Schulz in October 1932:

“I would also suggest that you include a table giving the raw data you have used. ... I think the publishing of the raw data is very important in order to stimulate criticism and control” (Bjerkholt 2013).

Another emerging area is the pre-registration of economics and social science studies, especially where experiments are involved. For instance if the researchers are running a randomised controlled trial, they would have to state in their trial protocol what kind of outcomes they would like to observe. The more outcomes we look at, the more probable it is that there would be some indicator with a significant effect size. Stating *ex-ante* what the purpose of the trial is and what outcomes will be observed sets out a transparent research process.

The American Economic Association (AEA) launched in 2013 a registry for randomized controlled trials in economics (<https://www.socialscienceregistry.org/>) “to address the growing number of requests for registration by funders and peer reviewers, make access to results easier and more transparent, and help solve the problem of publication bias by providing a single place where all trials are registered in advance of their start”⁸. Pre-registration would help improve the quality of randomized experiments and tackle the selective presentation of results, the inadequate documentation of hypothesis testing and data mining.

⁸ See short announcement at <http://openeconomics.net/2013/07/04/the-aea-registry-for-randomized-controlled-trials/>

Funders have also established data management and sharing plans where researchers are required to outline their approach to gathering, storing and disseminating their research data. However, many funders have to face the trade-off between giving more research funding and setting aside a pot for supporting the documentation of research. In line with these developments the U.S. government released a policy memorandum⁹, promising specific funding for making federally-funded research freely available to the public, giving specific attention to digital data.

The Economic and Social Research Council in the UK also requires data management and data sharing plans from all grant applicants and recognises that data sharing and re-use are “becoming increasingly important”¹⁰. Funders of research are aware that having research and its underlying data out in the open has the potential of multiply the impact of the original project, thus making better use of the research resources.

The fifth Principle refers to the **openness of publicly funded data**: *“publicly funded research work that generates or uses data should ensure that the data is open, free to use, reuse and redistribute under an open license – and specifically, it should not be kept unavailable or sold under a proprietary license. Funding agencies and organizations disbursing public funds have a central role to play and should establish policies and mandates that support these principles, including appropriate costs for long-term data availability in the funding of research and the evaluation of such policies, and independent funding for systematic evaluation of open data policies and use.”*

⁹ http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

¹⁰ ESRC Research Data Policy, September 2010 – http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

As publicly funded research is done in the public interest, it should be also open for the public to access, as the greatest benefit would be realised when data and code are made open and publicly available. The analysis done by economists and social scientists is also often used to inform policy-making and serves as evidence for government interventions or de-regulation. Public engagement and trust are some of the underlying reasons for making economics research data and code openly available.

Economists like Reinhart and Rogoff¹¹ as well as Piketty¹² who have come under scrutiny with regard to their research methodology and data have had publish corrections or respond to criticisms. Where economic research results are adopted as recommendations in policy-making, it is essential that the methodology and data underlying these results can be reviewed and scrutinised. A lot of the economics evidence base may remain undiscovered or unused if not published in the proper way.

Therefore, the sixth Principle deals with **usability and discoverability**: *“as simply making data available may not be sufficient for reusing it, data publishers and repository managers should endeavour to also make the data usable and discoverable by others for example: documentation, the use of standard code lists, etc., all help make data more interoperable and reusable and submission of the data to standard registries and of common metadata enable greater discoverability”*.

Better systems and frameworks have emerged to encourage and enable the sharing of data and code. Projects like the Open Science Framework (<https://osf.io/>) provide platforms to researchers for

¹¹ <http://www.ft.com/cms/s/0/433778c4-b7e8-11e2-9f1a-00144feabdc0.html#axzz3DjfB0D2P>

¹² http://www.nytimes.com/2014/05/30/upshot/thomas-piketty-responds-to-criticism-of-his-data.html?_r=0&abt=0002&abg=0

storing and sharing their data throughout the research lifecycle, with the aim to increase productivity of academics and the efficiency of sharing. Web-hosting services with revision controls systems may be a model for collaboration projects also in the social sciences where researchers would be able to share their code and work more effectively.

Further tools exist for economic researchers to share their research data, e.g. projects like DataVerse at Harvard (<http://thedata.org/>) offer online repositories for research data. It is generally not the lack of available tools, which hinders openness of economic data.

There are many potential benefits for sharing data: it enhances the visibility and the impact of one's research: it allows for the scrutiny of research findings, promotes new uses of the data and avoids unnecessary costs for duplicate research. The revolution of credibility in econometrics needs to embrace open data in order to realise its full potential.

References

- Angrist, Joshua D., and Jörn-Steffen Pischke. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24.2 (2010): 3–30.
- Bjerkholt, Olav. *Promoting econometrics through Econometrica* 1933–37. No. 28/2013. Memorandum, Department of Economics, University of Oslo, (2013).
- Chetty Raj and Finkelstein Amy "Program Report: The Changing Focus of Public Economics Research, 1980-2010" NBER Reporter (2012).
- Einav, Liran, and Jonathan D. Levin. "The data revolution and economic analysis" National Bureau of Economic Research (2013).

- Leamer, Edward E. "Let's Take the Con out of Econometrics." *The American Economic Review* 73.1 (1983): 31–43.
- Miguel, E., et al. "Promoting Transparency in Social Science Research" *Science* (2014): 30–31.
- Schwab, K., et al. "Personal Data: The Emergence of a New Asset Class." World Economic Forum Report. (2011).
- Stiglitz, Joseph E. "Knowledge as a global public good." *Global public goods: International cooperation in the 21st century* 308 (1999): 308–25.