Directorate General Economics, Statistics and Research

01.09.2023

3. Output from BELab: Standards for Output Control

BELab operational guides

Statistics Department. BELab Unit

ÍNDICE

1 Introduction

This document is part of BELab's operating manuals and sets out a series of rules[1], recommendations and best practices to ensure that the work carried out by external researchers using BELab microdata is disseminated securely.

2 Output Control Principles

2.1 Principle of reproductibility

The results provided for review and extraction should be easily reproducible. Section 4 of the document '2. Staying at BELab: Working Guidelines' **outlines the recommendations to follow for compliance with these standards**

1. **Program code reproductibility:** All results must be generated by a code file named 'main_YYMMDD' that can be executed without errors and must contain the necessary code to reproduce the files to be extracted. This program can begin with loading the original data provided by BELab or from a file of pre-processed data. The program must consistently execute the same steps and produce precisely the same results as those previously presented for review.

2. **Software Description:** Any computer software used to generate results must be clearly described at the beginning of the 'main_YYMMDD' file (name and version number). Along with the analysis software version number, the names of all packages (e.g., R, Python, Octave) used should also be included.

3. **Data Description:** All BELab datasets used to generate results must be clearly described at the beginning of the 'main_YYMMDD' file (DOI if available, variables, and year). Any externally sourced datasets used must also be described.

4. **Reproducibility of Publishable Output:** The files reproduced in the review process must be exactly identical to those generated by the researcher. In case of any variation in the reproduction of the files, they will not be extracted.

2.2 Anonymisation principles

The following rules aim to facilitate compliance with data confidentiality regulations. **External researchers are responsible at all times** for ensuring that their results meet the criteria ensuring complete anonymization.

1. **Non-extraction of identifiers:** Identifiers cannot be included in the results to be extracted or in the codes that generate them.

2. **Non-extraction of microdata:** No result may contain microdata. This entails refraining from extracting subsets of data, as well as tables, graphs, codes, or log

[1] BELab reserves the right to modify, supplement, or expand these standards during the Output Control process, if deemed necessary.

files that contain microdata themselves. Consequently, the extraction of minimum and maximum values is not permitted.

3. **Minimum number of observations:** All the results to be extracted should be based on at least three different observations. This applies both to aggregate results (averages, medians, etc.) and to charts and tables (at least three observations per cell/information node). The simplest way to demonstrate compliance with this criterion is to always generate the frequency table associated with each result.

4. **Degrees of freedom:** Regression models must be calculated with at least ten observations and must also have at least ten degrees of freedom.

5. **Dominance Rule (%p):** It is necessary to ensure that the largest observation does not exceed 85% of the total weight of the analyzed value or any other weighting used.

   Example: For calculating total sales in a specific sector for a particular year, let's consider only 3 companies. The total volume is 100 million euros, with the following composition: 90 million from the largest company, and 5 million from each of the others. In this case, the largest company is potentially identifiable due to its contribution to the total value.

6. **Confidentiality in multiple tables, control of differences:** If the results are calculated based on a G population, but are subsequently recalculated for an X subset of G, the rules explained above must be met for observations of the difference. Otherwise, the individual observations could be identified on the basis of the differentiation.

   Example: we have a table with all the firms in a given sector and another with the firms in that sector that exceed an X volume of sales. We would have to create a third table with the firms that do not reach such X volume and check that the confidentiality criteria are met in that table; otherwise, the firms could be identified by differentiation.

7. **Dichotomous (0-1) Categorical Variables (Dummies):** When calculating averages of these variables, there should be a minimum of three observations for each category (three observations with 0 and three with 1).

8. **Treatment of zeros and missing values:** Zeros are permitted in regressions and descriptive statistical analysis, provided they do not represent missing values in dichotomous and categorical variables. In descriptive statistics, missing values will not be taken into account for determining the number of different observations used. If missing values are imputed, the number of imputed and observed observations should be reported.

2.3 Principle of Verifiability

The Output Control process involves significant time and effort from the BELab Team. To optimize the utilization of Data Lab resources and minimize wait times from data extraction requests, external researchers must adhere to the following guidelines:

1. **Log file:** The log function (log file) must be activated for each program code. Logging should begin before describing the research project's content and before the first line of calculation code.

2. **Code Order and Structure:** Code should be visually structured in a clear manner, so that individual code blocks (header, individual analytical stages, etc.) are visually distinct. Indentation should be used for loops. Lengthy programs or analytical steps should be divided into smaller code files, for instance ('0_master.do', '1_data_preparation.do', '2_descriptive_analysis.do', and '3_regressions.do')

3. **Code Comments:** The code should include sufficient comments to make it understandable even to individuals not familiar with the project within a reasonable time.

4. **Clear Variable Names:** All variables names should be as informative as possible and used consistently. Variable labels and brief descriptions should be provided for all user-generated data as well as for all externally sourced data. If variables (categorical) are created or modified, corresponding value labels need to be assigned to these values.

5. **Specification of Anonymization Compliance:** The researcher should include code justifying compliance with the anonymization requirements described above. For instance, providing frequency tables, model descriptions, or any other element that demonstrates adherence to the requested output standards.

6. **Re-evaluation of Output:** If requesting the extraction of code that has been previously reviewed but with minor changes, these changes should be specifically noted in the new request. Whenever possible, researchers should only present for evaluation the portions of the program that they have modified.

2.4 Principle of reasonable use of resources

As a general rule, elements whose extraction from BELab is requested are only to be directly used in a publication. For this reason, visiting researchers must respect the principle of reasonable use of resources, especially at the time of deciding which results they wish to extract. The number of elements to be presented must be consistent with what is normally expected in the sphere of an empirical scientific article.

In general, visiting researchers must take into account the following rules:

1. **Exploratory data analysis cannot be part of the output to be extracted.** Only analyses that may be published directly can be presented for review. The task of

selecting results worthy of publication is part of the work to be conducted by researchers in BELab.

2. **Maximum number of lines in the output.** BELab does not establish, a priori, a maximum number of code lines to be reviewed during the Output Control process, but it reserves the possibility of doing so if researchers do not use the Laboratory's resources reasonably. This entails being prudent in the quantity of output requested, program use, execution times, use of sessions, etc.

2.5 Principle of responsibility

Visiting researchers are responsible for ensuring compliance with all the principles and rules set out in this document. Failure to comply with these rules will entail BELab's refusal to deliver the calculation results. Researchers must respect the following rules:

1. **Checking all calculation results for the purpose of publication:** Before researchers ask BELab to review an output they wish to extract, they must first check that the Output Control principles have been applied. Once that has been done, they will ask the BELab team to review their data. They will place the output requested in the /Out/Output folder of their project and indicate, as detailed in these guidelines, the elements required to carry out their review.

2. **Functioning of the code:** If the program code contains syntax or other errors, BELab will leave the codes uncorrected and ask the researchers to correct them.

3. **Output control format:** Program results and codes will only be accepted for output control if they are editable and are presented as unformatted text or .csv files. Charts must have a read-only (static) format and be presented in .jpeg or .png format.

3 Publication Control Principles

The following guidelines aim to assist researchers in more easily complying with the publication control standards ('publication control').

1. **Copy of Publications:** It is the responsibility of researchers to provide a copy of the published works that they produce and that contain research results from the analyses conducted during their stay at BELab.

2. **Referencing sources:** The researcher commits to mentioning the ultimate data source in any publication resulting from this study, as indicated in the respective guide of each database.

3. **Referencing charts and tables:** All charts and tables should be referenced as follows: "Source: BELab. Banco de España Data Laboratory, <name of the set of microdata used from the BELab catalogue (if appropriate, with the common abbreviation)>, <period during which the microdata were used>, own calculations."

4. **Specification of type of data access:** Each publication must specify the type of access the researcher had to the data, e.g. in-person access from a dataroom (indicate whether Madrid or Barcelona), remote access or mixed access.

5. **Specification of datasets used, use of DOI:** All datasets used in the research project must be cited, indicating the name and, where appropriate, the DOI.