Working with Data

> *Datasets are the single most valuable (and most expensive) outputs of our research. If an evaluation cost one million dollars, shouldn't we value the underlying dataset at one million dollars as well?*
> *—Bhavani Prathap Kasina, former Associate Director Research, J-PAL South Asia, current Regional Director, Asia and Latin America, IPA.*

It is critical to ensure data safety and security. Data must be backed up and protected at all times, with appropriate procedures undertaken to ensure compliance with IRB and other research protocols. Having ensured the security of the data, following appropriate data cleaning and analysis protocols is essential for accurate results, and for ensuring a transparent, replicable data analysis. The broad categories of data management dealt with in this section are:

1. Data Security, Backup, and Storage
2. Data Cleaning
3. Data Analysis

Data Security, Backup, and Storage

There are a number of questions that need to be answered when it comes to data storage and security:

1. Where should we store, and how should we back up, our data? Should it be stored in the cloud? If not, why not?
2. Does all data need to be password protected? What constitutes

a good password? What are appropriate protocols for transmitting passwords between team members?

3. What constitutes personally-identifying information, or other sensitive data that needs to be encrypted?
4. If the data is to be stored in a non-secure location, how does one go about removing identifiers? Can de-identified data be shared with parties that are not on the IRB approval for dealing with identified data?
5. How often should the data be backed up and in how many locations should backups be made?
6. How can we transfer data, both internally and externally? What constitutes a “secure channel” for transferring data?

For one classification of what constitutes sensitive information, consult the Harvard Data Classification Table.

J-PAL recommends the use of VeraCrypt, an updated version of TrueCrypt, to read and create encrypted storage volumes used for storing sensitive data. However, the recent audit of VeraCrypt found issues with Windows XP, which are outside of the scope for VeraCrypt to address. Anyone still running Windows XP is strongly encouraged to upgrade their computers to Windows 7 or Windows 10.

As a resource to the community, we have modified the truecrypt Stata command to work with VeraCrypt. The updated package can be found here. Please note that this package is in beta form—if you come across any problems, please submit an issue here. Once finalized, we will make the package available via Stata SSC archive. We have also developed a guide to installing and using VeraCrypt software.

BoxCryptor is another option widely used for encryption of data stored in the Cloud (using Services such as Box, Dropbox, Google Drive, etc.)

BitLocker is a service that encrypts all data stored on the Windows operating system.

Data Cleaning

Even data from the best designed surveys typically require some preparatory and cleaning work before it is ready for analysis. Some of the questions that we will need to tackle with incoming data include:

- How should data be named, recoded, and labeled?
- Are there standard best practices for preparing data before analysis? What standard checks should be run on unique identifiers and variable values?
- How do we deal with conflicts in the data?
- How should we deal with missing data?
- What is the best way to check for logical consistency in answers and to verify answers in the data against survey options?

Data Analysis

When conducting analysis of the data from our project, we are often interested in the relationship between two variables. As examples, we can use analysis to: test theories, understand relationships between variables, predict outcomes, and run simulations. Analysis of impact can range from the basic—such as testing whether there is a statistically significant difference in outcomes for individuals in the treatment vs. the control group—to the more complex—such as using the data to look at heterogeneous impacts, estimate parameters of a structural model, etc.

- For a broad overview of the theory behind causal inference in randomized evaluations, refer to the

Randomization Toolkit.

- Mastering 'Metrics, a textbook aimed at undergraduates by Joshua Angrist and Jörn-Steffen Pischke, is a good guide to the econometrics behind drawing causal inference in various study designs, including randomized evaluations.
- Colin Cameron and Praveen Trivedi have a comprehensive guide to conducting microeconometric analysis using the software Stata.
- Chuck Huber and David Drukker have a series of posts on the Stata blog on using the program's treatment effects command, "teffect."
- Christopher Baum's slides (2013) on using Stata for data management and reproducible research contain many valuable tips and tools.
- For further technical resources on using software to manage and analyze data, refer to the Software Tools section.

Please note that the practical research resources referenced here were curated for specific research and training needs and are made available for informational purposes only. Please email us for more information.

Published on *The Abdul Latif Jameel Poverty Action Lab*
(https://www.povertyactionlab.org)