PRDN Data Practices – Advice for Researchers

Work in the PRDN may be slightly different from work that you have previously done on your own or departmental machines. Here are some of the important components to keep in mind when analyzing data on the PRDN.

User folders

Each researcher is equipped with their own personal folder on the PRDN. Each researcher's folder is named with their Duke NetID and no one else on the project has access to that personal folder. We strongly suggest saving all syntax from all statistics programs in one's personal folder. If you are working on multiple projects using a single data file, we recommend a subfolder for each of the projects. Contact your PRDN administrator if you would like to share syntax between researchers. It is possible to set up a project based user folder or transfer syntax like output is transferred.

Output

One of the PRDN's greatest strengths is one of the most challenging aspects for researchers to get used to. All analyses with the "Researcher" access level in the PRDN are done on the virtual machine and the results can not be directly downloaded to your personal machine. Therefore, the following processes are in place to download output results to one's own computer, where modifications and printing can be conducted. Each project handles these output processes somewhat differently, so contact your project level personnel who have Manager access in the PRDN.

Typically, each project has a folder named *file transfer* that is available to all researchers on the project. Researchers that need results and visualization from their analyses for papers, presentation, or further analyses can save and move that output from their own user folder to the file transfer folder (the output data must comply with the data provider's requirements). Contact the project manager for the data and let them know the output is ready to be moved. They will check it and move it to a less restrictive server or email the file attachment or some other process.

The PRDN terms of use clearly define the non-disclosure of information. Any identifiable data found to be outputted may lead to the termination of use of the PRDN.

Creation of Analysis Files

For the majority of projects, we do not recommend the creation of permanent personal analysis datasets. Creating datasets for use may be convenient and what you have always done, but can create problems in a shared server environment. First, changes and improvements to the parent dataset are not integrated into permanent analysis files created before the improvements. Second, new permanent files take up space and memory and depending on the size of the original file or the number of users that create their own files, this can grow exponentially. Finally, analysis files may include many cases and variables not being analyzed as part of the project, which can lead to analysis inefficiencies, especially with larger datasets.

In all major statistical software packages, temporary datasets can be constructed that:

- Include some or all cases
- Include some or all variables
- Merge data from across multiple datasets

Do not read in all variables in a datafile unless you are analyzing all of the variables. This will slow down the processing of your analysis. Rather, think mindfully about what variables you need and which ones you do not in order to successfully complete your project. Remember, you can always modify your code to include other variables as needed.

If you have any questions about how to implement these strategies, contact the PRDN administrators. We are always looking for ways to improve the PRDN while maintaining the high level of security inherent in the design. If you have any suggestions on improving any of these issues, contact the SSRI data security team at ssri-datasecurity@duke.edu