# Creating Replication Packages

Jason Thomas

IPR Community & Collaboration Workshop

Feb 15th, 2023

# Introduction: Replication Packages

- I am here to tell you to eat your vegetables 🍅 🥕 🫑
- To help keep a balance of ideal and practical, we have
  - **Son**: starting grad school next Fall

    > "Every project should be started with the intention of making it replicable! This is how we make improve the quality of science and thus improve the world."

  - **Ja**: been in the game a while

    > "I've got 18 meetings, 3 classes to teach, and mouths to feed. I slept 4 hours last night and am only standing because of caffeine. You want me to do what now?!?"

# Introduction: Game plan

- Please feel free to take sides and interrupt with additional points in the spirit of either Son or Ja
- Game Plan
  - review some of the general expectations of journals and editors
  - walk through a few examples of replication packages
  - introduce a few tools and best practices for making replication packages
- http://github.com/buckipr/IPRCC_Replication

# Guidance

# Guidance: Sources

- Demographic Research
- American Journal of Political Science (PDF)
    - includes info for qualitative studies
- American Economic Association (jounals: AER, JEL, JEP )
    - FAQ & README template for social science from (AEA)
- Social Science Data Editors
- Open Science Framework

# Guidance: Perspectives

> "Before even looking at your data, you should sculpt a beautiful README guide, describing how the files are organized. Then create your file system and a master do-file. As you populate your project with well-commented code, update your README file with succinct descriptions of what each file does.
>
> Remember, the person who will replicate your work the most is YOU, so be kind to your future self and stay true to this process."
>
> --- Son

# Guidance: Another Perspectives

> "The analysis is already done, and now I have to make a replication packages to get this editor off my back. I didn't know I was submitting to this blasted journal, so what do I need to do so they will publish my stuff?
>
> Also, I'm using restricted data, so why do I even need to do this?!?"
>
> --- Ja

We've got you covered Ja, here is a quick run down of what you will likely need to do.

# Guidance: General Requirements

- Some journals indicate *where* to store your replications package (or provide recommendations)
- `README` file
  - general instructions: how/where to get the data; how/when to run your programs/scripts
  - include the name of each file included in the package with a description of what is/does (description of the organization of the folders and files is also useful)
  - version of software and libraries used to produce the results

# Guidance: General Reqs (cont.)

- Code (e.g., do-file or R scripts)
  - "extensive comments" or code should be "well commented"
  - indications of how commands are related to figures or tables in the manuscript

    > "The following commands recode variables X and Y in preparation of the logistic regression model. The following commands create Figure 1 in the article."

  - OK to assume familiarity with the program/language
  - a *master file* may be recommended: (a script/program that runs all of the other scripts/programs in the appropriate order)

# Guidance: General Reqs (cont.)

- Data & Codebook
  - original source data (or a description of where and how to obtain them)
  - clear and useful variable names; (unique case ids *may also be required*)
  - ok to include links or point to official documentation (but still need to describe any new variables created)
- Restricted Data
  - there will likely be a discussion with the editor about how to proceed
  - some mention that the data need to be submitted to the editorial staff and reviewed by a 3rd party replicator
  - still need to include a description of how to obtain the data (and possible of variables used)

# Data Sharing

> "It took 5 months to make my analytic data set! And now you want me to hand it over?
>
> Also, I created new measures with fancy stats, and want to use them for another paper."
>
> --- Ja

- It may be possible to make arrangements with the editor about when data are released
- Include a clear statement (in README or separate file) that anyone using the data or code should cite you as the author (& discuss with editor)

# Examples

- Van Hood & Bachmeier (Dem Res 29-1)
- Ambugo & Yahirun (Dem Res 34-8)
- Saccardo & Serra-Garcia AER 113-2
- Clark Demography 2019
- Hauer Open Sci
- Hauer NC

# General Advice

- Project organization is key!
    - pick a strategy/layout and stick to it so it because automatic
    - treat README file as a source for text used in your article
- Best Coding Practices (reuse/recycle!)
    - templates for scripts are key!
    - e.g. opening comment block that prints software version & date, and sets up working environment (e.g. file structure)
    - write comments so you can paste them into your manuscript (e.g., I need to recode this variable because of the outliers)

# General Advice (cont.)

- Automate table/figure construction
  - there are many good tools for this...copy and paste are not among them!
- Master scripts really are a good idea
  - set up your project so you run it on your office computer, your laptop, and the desktop at home
- Son: "What about Accessibility?!?"
  - cost and barriers to tools

# Useful Tools & Resources

- **GitHub**
    - popular site for making code (and smaller data sets) accessible
    - compatible with markdown (useful for README files)
    - also provides version control
- Version Control Software
    - PHD by Jorge Cham

# Useful Tools & Resources (cont)

- Open Science Framework -- hosting projects and making them available
    - web-based project (and file management)
    - useful for collaboration and sharing/updating files among team members
    - can share unpublished/in-progress/submitted work as well
- Other sites for storing data
    - ICPSR
    - Zenodo.org
    - DataCite
    - Harvard Dataverse Network

# Useful Tools & Resources (cont)

Dynamic Documents and related tools

- **R**
    - R Markdown -- combine your results and manuscript into a single document
    - stargazer
    - kableExtra
    - furniture

# Useful Tools & Resources (cont)

Dynamic Documents and related tools

- **Stata**
  - Markstat for Stata
  - Tools starting with Stata 15
  - Tables in Stata 17
  - estout and tabstat
- StatTag

# Recap

- Write once then reuse (e.g., README file; comments in code)
- Creating and using master files is a great habit
- Automate as much as possible
- Learning new tools is an investment that will make you more efficient in the future (and improve science 😄)

# Thanks!

# Journal-Specific Requirements

# PNAS

- make materials, data, and associated protocols available (code and scripts)
- available in a public repository (exceptions must be noted: e.g. legal/ethical restrictions on sharing; logistical/size of files)
- General instructions for accessing data & reproducing results (README file)
- Statement for data sharing plans (to be included in article)

# Amer Jour of Pol Sci

- Source dataset (only variables used in analysis)
  - meaningful variable names (recommended)
  - unique case id (required)
  - for complicated procedures (e.g., Bayesian simulation or bootstrap replications) only the final analytic dataset need be included (but need to include code to reproduce replication data sets)
  - codebook (for each data file)

# Amer Jour of Pol Sci (cont)

- Source dataset (cont.)

- *restricted data*: need explicit approval from Editor (upon initial submission); no need to make data available, but provide instructions for accessing data and formatting and variable definition information
  - *data for future research*: authors can request the data be retained for a "limited amount of time" (need permission from Editor) with a statement of when they will be available; **Must make data available to Journal staff and contractor for verifying replicability**

# Amer Jour of Pol Sci (cont.)

- Code for running relevant software (for analysis in article and any supplementary/supporting analysis)
  - use of extensive comments (can assume familiarity with statistical software)
  - "The following commands recode variables X and Y in preparation of the logistic regression model"; "The following commands create Figure 1 in the article"
  - include version of software
- Instructions (ReadMe File -- actual file)
  - names for all the files in the package (data, code, results) and brief description; grouping of files into headings and/or folders is recommended for larger projects (many files);
- instructions for Qualitative Work as well

# American Econ Assoc

- Data and programs should be archived in the AEA Data and Code Repository
    - *Exceptions*: author must retain all materials and provide reasonable assistance to requests and clarification and replication
- "the data set(s)" (just analytic?) & sufficient information to access the source data file(s)
- code for preparing the data and producing the results
- "description sufficient to allow the programs to be run"
- If applicable (e.g., original data collection):
    - survey instrument and interviewer instructions
    - code for experiment or survey collection mechanism
    - instructions and details on subject selection

# American Econ Assoc (cont.)

- README file (PDF, text, markdown)
    - data availability statement (how, where, and under what conditions can the original source data be obtained); includes requirements for restricted data (when possible, restricted data should be provided to editor and third- party replicator)
    - list of all files and their purpose
    - instructions for how the replication should be conducted
- codebooks (ok to reference publicly available versions)
- code ("a master script is strongly encouraged")

# Demographic Research

- (welcomes the submission of replication papers that seek to replicate other published work)
- computer code ("program files, spreadsheets producing tables or figures from summary data")
  - well commented -- indicate the purpose of certain commands and formulas, and how they relate to the results in article
  - version of software used to generate the results
- package everything into a single zip file, accompanied by a readme file that lists and describes each file
- "if possible, small sample dataset from which interested readers could generate illustrative results"
  - if data cannot be provided, then a supplementary document with meta-information (where and how to obtain the files, which variables were used with basic description, how many records)