# Template README and Guidance

INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see [Endorsers](https://social-science-data-editors.github.io/template_README/Endorsers.html). It is available as [Markdown/txt](https://github.com/social-science-data-editors/template_README/blob/master/template-README.html), [Word](https://social-science-data-editors.github.io/template_README/templates/README.docx), [LaTeX](https://social-science-data-editors.github.io/template_README/templates/README.tex), and [PDF](https://social-science-data-editors.github.io/template_README/templates/README.pdf). In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with > INSTRUCTIONS). Please ensure that a PDF is submitted in addition to the chosen native format.

## Overview

INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.

This replication package contains the data and the code to generate Table 1 in the paper.

## Data Availability and Provenance Statements

### Statement about Rights

* x I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
* x I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the [LICENSE.txt](https://social-science-data-editors.github.io/template_README/LICENSE.txt) file.

### Summary of Availability

* x All data **are** publicly available.
* ☐ Some data **cannot be made** publicly available.
* ☐ **No data can be made** publicly available.

### Details on each Data Source

### Data from Decennial Census 2000

This paper uses publicly available data from the 2000 Census of Population and Housing (US Census Bureau, 2006). We use the PUMS 5 percent sample. It can be downloaded from ICPSR. Registration is required, but free and without conditions.

* For this paper, we only use Alaska data. Users can just download the Alaska ASCII file by clicking on the download link to the right of the file “DS2 Alaska”. We need the file called “13568-0002-Data.txt”.

## Dataset list

|  |  |  |  |
| --- | --- | --- | --- |
| Data file | Source | Notes | Provided |
| data/ICPSR\_13568/DS0002/13568-0002-Data.txt | US Census Bureau (2006) | Public-use | Yes |
| data/derived/regression\_input.dta | All listed | Combines multiple data sources, serves as input for Table 2, 3 and Figure 5. | Yes |
|  |  |  |  |

## Computational requirements

INSTRUCTIONS: In general, the specific computer code used to generate the results in the article will be within the repository that also contains this README. However, other computational requirements - shared libraries or code packages, required software, specific computing hardware - may be important, and is always useful, for the goal of replication. Some example text follows.

INSTRUCTIONS: We strongly suggest providing setup scripts that install/set up the environment. Sample scripts for [Stata](https://github.com/gslab-econ/template/blob/master/config/config_stata.do), [R](https://github.com/labordynamicsinstitute/paper-template/blob/master/programs/global-libraries.R), [Julia](https://github.com/labordynamicsinstitute/paper-template/blob/master/programs/packages.jl) are easy to set up and implement. Specific software may have more sophisticated tools: [Python](https://pip.pypa.io/en/stable/user_guide/#ensuring-repeatability), [Julia](https://julia.quantecon.org/more_julia/tools_editors.html#Package-Environments).

### Software Requirements

INSTRUCTIONS: List all of the software requirements, up to and including any operating system requirements, for the entire set of code. It is suggested to distribute most dependencies together with the replication package if allowed, in particular if sourced from unversioned code repositories, Github repos, and personal webpages. In all cases, list the version *you* used.

* Stata (code was last run with version 15)
  + estout (as of 2018-05-12)
  + rdrobust (as of 2019-01-05)
  + the program “0\_setup.do” will install all dependencies locally, and should be run once.
* Python 3.6.4
  + pandas 0.24.2
  + numpy 1.16.4
  + the file “requirements.txt” lists these dependencies, please run “pip install -r requirements.txt” as the first step. See <https://pip.pypa.io/en/stable/user_guide/#ensuring-repeatability> for further instructions on creating and using the “requirements.txt” file.
* Intel Fortran Compiler version 20200104
* Matlab (code was run with Matlab Release 2018a)
* R 3.4.3
  + tidyr (0.8.3)
  + rdrobust (0.99.4)
  + the file “0\_setup.R” will install all dependencies (latest version), and should be run once prior to running other programs.

Portions of the code use bash scripting, which may require Linux.

Portions of the code use Powershell scripting, which may require Windows 10 or higher.

### Controlled Randomness

INSTRUCTIONS: Some estimation code uses random numbers, almost always provided by pseudorandom number generators (PRNGs). For reproducibility purposes, these should be provided with a deterministic seed, so that the sequence of numbers provided is the same for the original author and any replicators. While this is not always possible, it is a requirement by many journals’ policies. The seed should be set once, and not use a time-stamp. If using parallel processing, special care needs to be taken. If using multiple programs in sequence, care must be taken on how to call these programs, ideally from a main program, so that the sequence is not altered.

* ☐ Random seed is set at line \_\_\_\_\_ of program \_\_\_\_\_\_

### Memory and Runtime Requirements

INSTRUCTIONS: Memory and compute-time requirements may also be relevant or even critical. Some example text follows. It may be useful to break this out by Table/Figure/section of processing. For instance, some estimation routines might run for weeks, but data prep and creating figures might only take a few minutes.

#### Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

* ☐ <10 minutes
* ☐ 10-60 minutes
* ☐ 1-2 hours
* ☐ 2-8 hours
* ☐ 8-24 hours
* ☐ 1-3 days
* ☐ 3-14 days
* ☐ > 14 days
* ☐ Not feasible to run on a desktop machine, as described below.

#### Details

The code was last run on a **4-core Intel-based laptop with MacOS version 10.14.4**.

Portions of the code were last run on a **32-core Intel server with 1024 GB of RAM, 12 TB of fast local storage**. Computation took 734 hours.

Portions of the code were last run on a **12-node AWS R3 cluster, consuming 20,000 core-hours**.

INSTRUCTIONS: Identifiying hardware and OS can be obtained through a variety of ways: Some of these details can be found as follows:

* (Windows) by right-clicking on “This PC” in File Explorer and choosing “Properties”
* (Mac) Apple-menu > “About this Mac”
* (Linux) see code in [tools/linux-system-info.sh](https://github.com/AEADataEditor/replication-template/blob/master/tools/linux-system-info.sh)`

## Description of programs/code

INSTRUCTIONS: Give a high-level overview of the program files and their purpose. Remove redundant/ obsolete files from the Replication archive.

* Programs in programs/01\_dataprep will extract and reformat all datasets referenced above. The file programs/01\_dataprep/main.do will run them all.
* Programs in programs/02\_analysis generate all tables and figures in the main body of the article. The program programs/02\_analysis/main.do will run them all. Each program called from main.do identifies the table or figure it creates (e.g., 05\_table5.do). Output files are called appropriate names (table5.tex, figure12.png) and should be easy to correlate with the manuscript.
* Programs in programs/03\_appendix will generate all tables and figures in the online appendix. The program programs/03\_appendix/main-appendix.do will run them all.
* Ado files have been stored in programs/ado and the main.do files set the ADO directories appropriately.
* The program programs/00\_setup.do will populate the programs/ado directory with updated ado packages, but for purposes of exact reproduction, this is not needed. The file programs/00\_setup.log identifies the versions as they were last updated.
* The program programs/config.do contains parameters used by all programs, including a random seed. Note that the random seed is set once for each of the two sequences (in 02\_analysis and 03\_appendix). If running in any order other than the one outlined below, your results may differ.

### (Optional, but recommended) License for Code

INSTRUCTIONS: Most journal repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a LICENSE.txt file, separately from the README, possibly combined with the license for any data provided. Some code may be subject to inherited license requirements, i.e., the original code author may allow for redistribution only if the code is licensed under specific rules - authors should check with their sources. For instance, some code authors require that their article describing the econometrics of the package be cited. Licensing can be complex. Some non-legal guidance may be found [here](https://social-science-data-editors.github.io/guidance/Licensing_guidance.html).

The code is licensed under a MIT/BSD/GPL [choose one!] license. See [LICENSE.txt](https://social-science-data-editors.github.io/template_README/LICENSE.txt) for details.

## Instructions to Replicators

* Run programs/00\_setup\_stata.do at least once (it is also called from main.do)
* Run programs/main.do to run all steps in sequence.

### Details

* programs/00\_setup.do: will create all output directories, install needed ado packages.
* programs/01\_dataprep:
  + These programs were last run at various times in 2018.
  + Order does not matter, all programs can be run in parallel, if needed.
  + A programs/01\_dataprep/main.do will run them all in sequence, which should take about 2 hours.
* programs/02\_analysis/main.do.
  + If running programs individually, note that ORDER IS IMPORTANT.
  + The programs were last run top to bottom on July 4, 2019.
* programs/03\_appendix/main-appendix.do. The programs were last run top to bottom on July 4, 2019.
* Figure 1: The figure can be reproduced using the data provided in the folder “2\_data/data\_map”, and ArcGIS Desktop (Version 10.7.1) by following these (manual) instructions:
  + Create a new map document in ArcGIS ArcMap, browse to the folder “2\_data/data\_map” in the “Catalog”, with files “provinceborders.shp”, “lakes.shp”, and “cities.shp”.
  + Drop the files listed above onto the new map, creating three separate layers. Order them with “lakes” in the top layer and “cities” in the bottom layer.
  + Right-click on the cities file, in properties choose the variable “health”… (more details)

## List of tables and programs

The provided code reproduces:

* ☐ All numbers provided in text in the paper
* **x** All tables and figures in the paper
* ☐ Selected tables and figures in the paper, as explained and justified below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure/Table # | Program | Line Number | Output file | Note |
| Table 1 | 02\_table1.do |  | results/.csv |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

## References

United States Census Bureau. Census of Population and Housing, 2000 [United States]:  Public Use Microdata Sample:  5-Percent Sample    . [distributor], 2006-01-12. https://doi.org/10.3886/ICPSR13568.v1