CAnD3 Reproducibility Report

Reproducer: Samuel Nemeroff

Original Author: Nicole Antunes Rezende

Project Research Question:

How does the region of origin influence the educational attainment of immigrants, controlling for sex, province of residence, and age at immigration?

Summary:

While the authors program was incredibly easy to follow, and the crosstable was almost perfectly reproducible, the regression outputs vary significantly.

Replicators Computing Environment:

Software:

Windows 11 Pro Operating System

Version 10.0.22631 Build 22631

RStudio 2024.04.2+764 "Chocolate Cosmos" Release (e4392fc9ddc21961fd1d0efd47484b43f07a4177, 2024-06-05) for windows

Hardware:

Processor: AMD Ryzen 9 7950X3D 16-Core Processor 5.7GHz

Installed RAM: 64 GB DDR5 6400mhz

Graphics Processing Unit: Nvidia RTX 4070 FE, 12gb GDDR6X

Data Sources:

1. 2016 Canadian Census

Replication Steps:

1. Download 2016 Census PUMF
2. Isolate variables identified by initial author
3. Download packages readr, dplyr, and tibble.
4. Followed recoding and labelling instructions outlined in INSTRUCTIONS_Nicole_CAnD3_RRWM.pdf.
5. Executed code.

Figure or Table	Relevant Program	Replicated (Comments)
Output_crosstable1_code_NicoleAR	Project CODE_Nicole_CAnD3_RRWM.R	Yes – additional responses of "missing" in originregion were included in reproduced code, with difficulty in removing them. However, all numbers for valid originregion outputs were identical.
Output_regression_code_NicoleAR	Project CODE_Nicole_CAnD3_RRWM.R	No – significant variation across regression results.

Discrepancies:
Minor discrepancy in the crosstable, but major discrepancies in regression. Several discrepancies arose, including the inclusion of the not applicable/not available answers in my own regression output, which, coupled with the inclusion of responses of not applicable/not available for originregion, may account for the lack of similarity.

Classification:
Incomplete reproducibility

Additional notes:
As mentioned above, my code seems to be including missing answers in the regression and crosstabels. In my own code, I included values of 88 and 99 and classified them as "missing or not applicable" for AGEIMM. I also included 88 in originregion, giving it an NA value. After some investigation, I believe this to be a result of me using recoding functions included in the DPLYR package, while Nicole's was written with different functions. I have not yet had the chance to try to amend the code after looking at hers, as my replication attempt was done before looking at her code. However, when running her code, downloaded from the GitHub repository, I was able to get identical results.