Replication Material for "Prior Work Experience and Entrepreneurship: The Careers of Young Entrepreneurs"

Nicolas Gendron-Carrier *

McGill University

November 4, 2024

*Department of Economics, McGill University, 855 Sherbrooke St. West, Montreal, QC, Canada, H3A2T7. e-mail: nicolas.gendron-carrier@mcgill.ca. website: https://www.nicolasgendroncarrier.com.

1 Overview

The replication package described here contains all the code used to generate results for "Prior Work Experience and Entrepreneurship: The Careers of Young Entrepreneurs" by Gendron-Carrier published in the *Journal of Labor Economics*. The data source is a custom-built version of the Canadian Employer-Employee Dynamics Database (CEEDD) (Statistics Canada, 2018). The CEEDD incorporates Canadian administrative tax records on firms, workers, and business owners in each year 2001-2012. Statistics Canada employees at the Canadian Centre for Data Development and Economic Research (CDER) prepared the CEEDD data files used in this project. These files are not publicly available. Access can be granted to interested researchers by following the procedure described in Section 2. Researchers who are granted access will be able to replicate all the results reported in the paper using the code provided in the replication package. The data can only be accessed from one of Statistics Canada's Research Data Centres (RDC).

2 Data Availability and Provenance Statements

Researchers can apply to get access to the data used in this study by visiting:

https://www.statcan.gc.ca/en/microdata/data-centres.

Please cite the following data product in your application: Custom Extraction - Canadian Employer Employee Dynamics Database (CEEDD) 5805, Product ID MAD-8006-204529.3.

Statistics Canada is committed to Open Science, working to advance open science in four key areas: FAIR (findable, accessible, interoperable, reusable) and open data, open publications, open communication, and open code. Replicability is an essential part of the FAIR data principles and aligned with the Transparency principle of the Open Science approach. For further information about Statistics Canada's Open Science Action Plan, please refer to this https://www.statcan.gc.ca/en/about/osap.

2.1 Statement about Rights

The author of the manuscript has legitimate access to and permission to use the data used in this manuscript.

2.2 Summary of Availability

The data cannot be made publicly available.

2.3 Details on Data Source

a. Canadian Employer-Employee Dynamics Database (CEEDD). The data used in this project is a custom-built version of the CEEDD (Statistics Canada, 2018). All files are located in the "DATA" folder of the replication archive located on the Statistics Canada server (restricted access).

The CEEDD files provided include:

- Information about individuals and business owners.
  - Data Files:
    (i) "Individuals_'year'_rl.dta" (2001-2012)
    (ii) "SE_all_'year'_rl.dta" (2001-2012)
  - Brief Description: (i) Matched employer-employee data for 2001-2012 and (ii) information about business owners.
- Information about Canadian postal codes.
  - Data Files:
    (i) "PC_masterfile.dta"
    (ii) "geogcor.dta"
  - Brief Description: (i) List of all postal codes in Canada and (ii) associated geographical information.
- Information about firms.
  - Data Files:
    (i) "Enterprise_'year'_rl.dta" (2001-2012)
    (ii) "Entid_bn_'year'.dta" (2001-2012)
    (iii) "annualized_gifi_by_ent_final.dta"
    (iv) "Connexions_'i'_'j'.dta" (2001-2002 to 2011-2012)
  - Brief Description: (i) Firm-level characteristics, (ii) business identifiers, (iii) additional balance sheet information, and (iv) year-to-year firm connections used to construct longitudinal firm identifiers.

3 Computational requirements

The code was run on a desktop computer operating Windows. The code uses Stata 17. It was last run on a machine with 64GB of RAM. Total run time exceeds 21 days. The code requires the installation of the following Stata packages:

- FTOOLS (Correia, 2016)
- REGHDFE (Correia, 2014)
- ESTOUT (Jann, 2004)

4 Description of programs and associated output

The replication package contains the code used to organize the CEEDD data, construct the estimation samples, and generate all the results presented in the paper. The "readme_run.do" file calls all other Stata programs in the archive and serves as a table of contents. Below, I describe the other dofiles and associated output in the order in which the programs should be run.

- "gc_1_create_firm_panel.do" creates the firm-level panel data.
- "gc_2_clean_individual_datasets.do" prepares the cross-sectional individual-level datasets.
- "gc_3_create_individual_panel.do" creates the individual-level panel data.
- "gc_4_akm.do" and "gc_4_firm_info.do" are used to classify firms. Associated output:
  - Table A1: Firm-level outcomes by firm class. The log file for "gc_4_firm_info.do" contains the statistics that appear in this table.
- "gc_5_create_estimation_sample_men_start25_occ_v'j'.do" ($j = 1, 2$) prepares the estimation samples.
- "gc_6_descriptive_analysis_men.do" generates descriptive statistics for the paper. Associated output:
  - Table 1 uses results from
    - "gc_date_ss_rowtransit30_men_start25_occ_v2_def6.xlsx"
    - "gc_date_ss_coltransit30_men_start25_occ_v2_def6.xlsx"
  - Table 2 (except the second and third rows) uses results from

    - * "gc_date_ss_all_occupations_men_start25_occ_v2_def6.xlsx"
  - Table 3 and Table A4 uses results from
    - * The log file for "gc_6_descriptive_analysis_men.do". The results are generated in "Step 6" of the code.
  - Table A2 uses results from
    - * "gc_date_ss_lifecycle_men_start25_occ_v2_def6.xlsx"
  - Figure 1 uses
    - * "gc_date_ss_lny_dist_legend_men_rl.gph" from "Step 8" in the code
  - Figure 2 uses
    - * "gc_date_stylized_lifecycle_cohort_legend_rl.gph" from "Step 9" in the code
  - Figure C1 and Table 2 (second row) uses results from
    - * "gc_date_ss_lifecycle_lny_men_start25_occ_v2_def6.xlsx"
- "gc_7_reduced_form_men.do" produces reduced-form evidence on the returns to prior work experience in entrepreneurship. Associated output:
  - Table A3 uses results from
    - * "gc_date_OLS_men_start25_occ_v2_def6_rl.txt"
- "gc_8_structural_1st_stage_main.do" estimates the first stage parameters of the main model with five types and three types. Associated output:
  - Tables C3-C7 use results from
    - * "gc_structural_y11_a25_v2_d6_q5_m3.txt"
    - * "gc_structural_y12_a25_v2_d6_q5_m3.txt"
    - * "gc_structural_y13_a25_v2_d6_q5_m3.txt"
    - * "gc_structural_y30_a25_v2_d6_q5_m3.txt"
    - * "gc_structural_y20_a25_v2_d6_q5_m3.txt"
  - Table 4, Figure 3, Figure 4, Figure C2, and Figure C3 are produced using the numbers reported in tables C3-C7.
  - Tables D3-D7 use results from
    - * "gc_structural_y11_a25_v2_d6_q3_m3.txt"
    - * "gc_structural_y12_a25_v2_d6_q3_m3.txt"
    - * "gc_structural_y13_a25_v2_d6_q3_m3.txt"

        * "gc_structural_y30_a25_v2_d6_q3_m3.txt"
        * "gc_structural_y20_a25_v2_d6_q3_m3.txt"

- "gc_8_structural_1st_stage_basic.do" estimates the first stage parameters of the simpler model with homogeneous firms and homogeneous returns to experience across types. Associated output:
    - Tables D12-D14 use results from
        * "gc_structural_y10_B_a25_v2_d6_q5_m2.txt"
        * "gc_structural_y30_B_a25_v2_d6_q5_m2.txt"
        * "gc_structural_y20_B_a25_v2_d6_q5_m2.txt"

- "gc_8_structural_1st_stage_alt.do" estimates the first stage parameters of the main model with five types and using an alternative definition of entrepreneurship. Associated output:
    - Tables D19-D23 use results from
        * "gc_structural_y11_a25_v1_d6_q5_m3.txt"
        * "gc_structural_y12_a25_v1_d6_q5_m3.txt"
        * "gc_structural_y13_a25_v1_d6_q5_m3.txt"
        * "gc_structural_y30_a25_v1_d6_q5_m3.txt"
        * "gc_structural_y20_a25_v1_d6_q5_m3.txt"

- "gc_9_structural_2nd_stage_main.do" estimates the second stage parameters of the main model with five types and three types. Associated output:
    - Table C1 and Table C2 use results from
        * "gc_structural_util_a25_v2_d6_q5_m3.txt"
    - Figure 5 uses the numbers reported in Table C1 and Table C2.
    - Table D1 and Table D2 use results from
        * "gc_structural_util_a25_v2_d6_q3_m3.txt"

- "gc_9_structural_2nd_stage_basic.do" estimates the second stage parameters of the simpler model with homogeneous firms and homogeneous returns to experience across types. Associated output:
    - Table D10 and Table D11 use results from

        * "gc_structural_util_B_a25_v2_d6_q5_m2.txt"

- "gc_9_structural_2nd_stage_alt.do" estimates the second stage parameters of the main model with five types and using an alternative definition of entrepreneurship. Associated output:
  - Table D17 and Table D18 use results from
    * "gc_structural_util_a25_v1_d6_q5_m3.txt"

- "gc_10_structural_fit_main.do" performs full simulations and assess model fit for the main model with five and three unobservable types. Associated output:
  - Table C8 Panel C uses results from
    * "gc_fit_lifecycle_a25_v2_d6_q5_m3.xlsx"
  - Table C9 Panel B uses results from
    * "gc_fit_rowtransit_a25_v2_d6_q5_m3.xlsx"
    * "gc_fit_coltransit_a25_v2_d6_q5_m3.xlsx"
  - Figure C1 and Table 2 (third row) uses results from
    * "gc_fit_lifecycle_lny_a25_v2_d6_q5_m3.xlsx"
  - Table D8 Panel B uses results from
    * "gc_fit_lifecycle_a25_v2_d6_q3_m3.xlsx"
  - Table D9 Panel B uses results from
    * "gc_fit_rowtransit_a25_v2_d6_q3_m3.xlsx"
    * "gc_fit_coltransit_a25_v2_d6_q3_m3.xlsx"

- "gc_10_structural_fit_basic.do" performs full simulations and assess model fit for the simpler model with homogeneous firms and homogeneous returns to experience across types. Associated output:
  - Table D15 Panel B uses results from
    * "gc_fit_B_lifecycle_a25_v2_d6_q5_m2.xlsx"
  - Table D16 Panel B uses results from
    * "gc_fit_B_rowtransit_a25_v2_d6_q5_m2.xlsx"
    * "gc_fit_B_coltransit_a25_v2_d6_q5_m2.xlsx"

- "gc_10_structural_fit_alt.do" performs full simulations and assess model fit for the main model with five types and using an alternative definition of entrepreneurship. Associated output:
  - Table D24 Panel B uses results from
    - "gc_fit_lifecycle_a25_v1_d6_q5_m3.xlsx"
  - Table D25 Panel B uses results from
    - "gc_fit_rowtransit_a25_v1_d6_q5_m3.xlsx"
    - "gc_fit_coltransit_a25_v1_d6_q5_m3.xlsx"
- "gc_10_structural_fit_27on.do" performs full simulations for the main model with five types, starting simulations at age 27.
- "gc_11_structural_expost.do" performs a post-estimation, ex-post analysis of types and more model fit. Associated output:
  - Table C8 Panel B uses results from
    - "gc_Xp_lifecycle_a25_v2_d6_q5_m3.xlsx"
  - Table C11 uses results from
    - "gc_Xp_rowtransit_2012_a25_v2_d6_q5_m3.xlsx"
    - "gc_Xp_rowtransit_2012rl_a25_v2_d6_q5_m3.xlsx"
    - "gc_Xp_coltransit_2012_a25_v2_d6_q5_m3.xlsx"
    - "gc_Xp_coltransit_2012rl_a25_v2_d6_q5_m3.xlsx"
  - Table C10 uses results from
    - "gc_Xp_lifecycle_2012_a25_v2_d6_q5_m3.xlsx"
    - "gc_Xp_lifecycle_2012rl_a25_v2_d6_q5_m3.xlsx"
  - Table A5 uses results from
    - "gc_Xp_SS_types_a25_v2_d6_q5_m3.xlsx"
  - Table 5 uses results from
    - "gc_Xp_bytype_rl_a25_v2_d6_q5_m3.xlsx"
  - Table C12 uses results from
    - "gc_Xp_bytype_rl_a25_v2_d6_q5_m3.xlsx"
    - "gc_Xp_bytype_a25_v2_d6_q5_m3.xlsx"
  - Table C13 uses results from

        * "gc_Xp_bytype26_rl_a25_v2_d6_q5_m3.xlsx"
        * "gc_Xp_bytype26_a25_v2_d6_q5_m3.xlsx"
    - Table C14 uses results from
        * "gc_Xp_bytype29_rl_a25_v2_d6_q5_m3.xlsx"
        * "gc_Xp_bytype29_a25_v2_d6_q5_m3.xlsx"
    - Table C15 uses results from
        * "gc_Xp_bytype32_rl_a25_v2_d6_q5_m3.xlsx"
        * "gc_Xp_bytype32_a25_v2_d6_q5_m3.xlsx"
    - Table C16 uses results from
        * "gc_Xp_bytype35_rl_a25_v2_d6_q5_m3.xlsx"
        * "gc_Xp_bytype35_a25_v2_d6_q5_m3.xlsx"
    - Table A6 uses results from
        * "gc_Xp_bytypeNAICS_rl_a25_v2_d6_q5_m3.xlsx"

- The policy simulations are done in:
    - "gc_12_structural_counterfactuals_p2.do"
    - "gc_12_structural_counterfactuals_p3.do"
    - "gc_12_structural_counterfactuals_p2_27on.do"
    - "gc_12_structural_counterfactuals_p3_27on.do"

- Table 6 and Table C17 use results generated by:
    - "gc_13_counterfactuals_save.do"
    - "gc_13_counterfactuals_save_27on.do"

5 Instructions to replicators

To replicate all the results reported in the paper:

1. Follow the steps outlined in Section 2 to get access to the replication package located on the Statistics Canada server.

2. Create the following folders inside your main project folder within the RDC:

    (a) "dofiles"

(b) "DATA"

(c) "DATA\gc_esamples" (used to save estimation samples)

(d) "DATA\gc_ster" (used to save other generated output)

(e) "ResultsFolder" (used to save the results that appear in the paper)

3. Copy all provided CEEDD data files inside the "DATA" folder.

4. Copy all Stata dofiles provided in this replication package inside the "dofiles" folder.

5. Install the following Stata packages inside the RDC:

   - FTOOLS (Correia, 2016)
   - REGHDFE (Correia, 2014)
   - ESTOUT (Jann, 2004)

6. Amend the absolute paths at the top of each dofile to match your environment within the RDC. Then, execute the Stata dofile "dofiles\readme_run.do." The "readme_run.do" file calls all other Stata programs in the "dofiles" folder and serves as a table of contents.

References

Correia, S. (2014): "REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects," Statistical Software Components, Boston College Department of Economics.

——— (2016): "FTOOLS: Stata module to provide alternatives to common Stata commands optimized for large datasets," Statistical Software Components, Boston College Department of Economics.

Jann, B. (2004): "ESTOUT: Stata module to make regression tables," Statistical Software Components, Boston College Department of Economics.

Statistics Canada (2018): "Canadian Employer-Employee Dynamics Database (CEEDD), 2001-2012," Record Number 5228. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5228. Accessed: November 27, 2018.