README File for "Crowd-out, Education, and Employer Contributions to Workplace Pensions: Evidence from Canadian Tax Records"

Derek Messacar

1 Datasets

Proprietary datasets from Statistics Canada were used in this study, to which access by external researchers is restricted. Specifically, the following datasets were used:

- Longitudinal Administrative Databank (LAD)
- Longitudinal Worker File (LWF)
- Census of Canada

The LAD and Census of Canada files are available at several Reseach Data Centres (RDC) located on university campuses across Canada. The Canadian Research Data Centre Network (CRDCN) is an initiative by Statistics Canada, the Social Sciences and Humanities Research Council (SSHRC), and university consortia to facilitate data access and strengthen Canada's social research capacity. The LWF is available at the Canadian Centre for Data Development and Economic Research (CDER), a section of Statistics Canada established to provide researchers with secure access to business and microdata for analytical research.

For more information about these Centres, including how researchers may apply for access to data, please visit:

- CRDCN: https://crdcn.org
- CEDR: http://www.statcan.gc.ca/eng/cder/index

2 Replication Files

This section describes the replication files for the analyses based on the LAD and Census-Tax Linkages datasets, respectively. The STATA statistical software for data analysis was used exclusively in this study, running on the Windows operating system.

2.1 LAD

From the raw LAD data, I extract the following variables for the years 1990-2010 inclusive and construct the (unbalanced) panel dataset:

- *LAD Variables:* `lin__i, wgt2_i, sxco_i, yob__i, yod__i, flag_1991-flag2010, immfli, fin__i, age__i, age__p, fcmp_i, indfli, psco_i, prco_i, t4e__i, t4e__p, txi__i, txi__p, aftici, afticp, sei__i, sei__p, dues_i, dues_p, eins_i, eins_p, tirc_i, tirc_p, nftxci, nftxcp, nptxci, nptxcp, sop4ai, sop4ap, oasp_i, oasp_p, oaspri, oasprp, cqpp_i, cqpp_p, clcppi, clcppp, cqppdi, cqppdp, t4rp_i, t4rp_p, rrspci, rrspcp, mdexci, mdexcp, disdni, disdnp, disdoi, disdop, t4rspi, t4rspp, tpajai, tpajap, clkgli, clkglp, invi_i, invi_p, totdni, totdnp, kid1_i-kid6_i, dsbcqi, dsbcqp, saspyi, saspyp, rsppii, rrspdi, rrspli, disdti, disdtp, hbpshi, hbpshp, naic1i, t4cnti, oasfli, espa_i, espa_p, espadi, espadp, rdsp_i, rdsp_p`

I construct a variable indicating the tax year (named `year`). The names and descriptions of these variables are provided in the LAD Dictionary, which is currently available at:

- *LAD Dictionary: http://www23.statcan.gc.ca/imdb-bmdi/document/4107_D9_T15_V1-eng.htm*

This dataset is saved as "LAD.dta" in a subdirectory called "LAD\Data\".

To prepare the panel dataset for analytical use, the following .do files are run. They are to be saved in a subdirectory called "LAD\Data_Cleaning\".

- *LAD_Data_Cleaning.do:* Sample restrictions, selection of relevant variables, data cleaning, and creation of relevant derived variables
- *LAD_Commands_filtering.do:* Additional sample selection and cleaning, which is called to execute within the "Data_cleaning" .do file

The following .do files are run to produce the empirical results. They are to be saved in a subdirectory called "LAD\Analysis\".

- *LAD_1_Summary_Statistics.do:* Columns (1) and (3) of Table 1
- *LAD_2_Summary_Statistics_Unionized.do:* Columns (2) and (4) of Table 1
- *LAD_3_Summary_Statistics_RPPMembers.do:* Descriptive statistics in Section 5.3 comparing workplace pension members with and without voluntary retirement saving
- *LAD_4_RKD_Graphs.do:* Figure 1
- *LAD_5_RKD_Primary.do:* Estimates in Panels (a), (b) and (c) of Figure 1 and in columns (1), (4) and (5) of Table 2
- *LAD_6_RKD_FixedEffects.do:* Column (2) of Table 2
- *LAD_7_RKD_DD.do:* Estimates in Panel (d) of Figure 1 and in column (3) of Table 2
- *LAD_8_RKD_Unconditional.do:* Columns (6) and (7) of Table 2
- *LAD_9_RKD_NetContributions.do:* Columns (1) and (4) of Table 3
- *LAD_10_RKD_WDHistory.do:* Columns (2) and (3) of Table 3
- *LAD_11_RKD_ContHistory.do:* Estimates in Panel (a) of Figure 2
- *LAD_12_RKD_ContRoom.do:* Estimates in Panel (b) of Figure 2
- *LAD_13_RKD_Bandwidth.do:* Columns (1), (2) and (3) of Table A-1
- *LAD_14_RKD_Polynomial.do:* Column (4) of Table A-1
- *LAD_15_Histograms.do:* Figure A-1
- *LAD_16_RKD_Permutation_Narrow.do:* Estimates for the note about the permutation test of Ganong and Jäger (2016) in the Online Appendix, using the narrow region.
- *LAD_17_RKD_Permutation_Moderate.do:* Estimates for the note about the permutation test of Ganong and Jäger (2016) in the Online Appendix, using the moderate region.
- *LAD_18_RKD_Permutation_Wide.do:* Estimates for the note about the permutation test of Ganong and Jäger (2016) in the Online Appendix, using the wide region.
- *LAD_19_RKD_IV.do:* Estimation of the primary crowd-out analysis using a two-stage least squares approach, as described in endnote #3.

2.2 Census-Tax Linkages

The raw dataset comprises T1 and T4 tax records, and 1991 and 2006 Census of Canada responses. The linkage was carried out by Wilkins et al. (2008). I extract the following variables for the years 1990-2010 inclusive and construct the (unbalanced) panel dataset based on the tax records:

- *T1 variables:*
    - `casenum`: A unique identifier for individuals (renamed `id`)
    - `year`: Calendar year
    - `age`: Age is calculated from the date of birth on the T1 files
    - `sex`: The sex of the individual
    - `prov_t1`: The province or territory of residence of the individiual on December 31 of the tax year (renamed `province`)
    - `marst`: The marital status of each individual
    - `t4earn_individual`: Total employment income received by an individual in a given year, summed across all paid jobs (renamed `empinc`)
    - `t1_union_dues`: Union dues (renamed `dues`)
    - `t1_empe_padj_amt`: Amount of the employee's pension adjustment for the registered pension plan or a deferred profit sharing plan (renamed `penadj`)
    - `rrsp_contrib`: The amount contributed to registered retirement savings plans by the individual in a given year (renamed `rspcont`)
    - `t4rsp_inc`: The amount withdrawn from from registered retirement savings plans by the individual in a given year (renamed `rspwd`)
    - `disability_dedn`: Disability tax deduction amount for self (renamed `disability`)
    - `tot_inc_calc`: Total income before tax (renamed `totinc`)
    - `oas_ben`: Old Age Security pension income (renamed `oasinc`)
- *T4 variables:*
    - `casenum`: A unique identifier for individuals (renamed `id`)
    - `year`: Calendar year
    - `can_long_naics1`: The 4-digit North American Industrial Classification that is assigned to the business enterprise at the national level (renamed `naics`)
    - `earnflag1`: For each individual, earnings from different jobs are compared and the job from which the highest earnings were received is identified

Note that the variables are renamed, as indicated in the listings, to match the naming convention used in the LAD. The T4 information for the highest-paying job is used. The following variables from the 1991 and 2006 Censuses of Canada are merged into the panel tax dataset:

- *1991 Census of Canada:*
    - `hlos`: Highest level of schooling
    - `value`: Value of dwelling
    - `birth_pl`: Province or territory of birth, if born in Canada
    - `compw1`: Composite weight

- *2006 Census of Canada:*

    - `hcdd`: Defined equivalently as hlos
    - `value`: Defined above
    - `pob`: Defined equivalently as birth_pl
    - `compw1`: Defined above

I construct a variable indicating whether the match occurred to the 2006 Census of Canada data (named `census06`). The compulsory schooling variables are available as additional materials from Oreopoulos (2006), published on the *American Economic Review* website. This dataset is saved as "CensusTax.dta" in a subdirectory called "Census_Tax_Linkage\Data\".

To prepare the panel dataset for analytical use, the following .do files are run. They are to be saved in a subdirectory called "Census_Tax_Linkage\Data_Cleaning\".

- *CTL_Commands_Education.do:* Sample restrictions, selection of relevant variables, data cleaning, and creation of relevant derived variables for the returns-to-education analysis of Section 7.
- *CTL_Commands_Crowdout.do:* Sample restrictions, selection of relevant variables, data cleaning, and creation of relevant derived variables for the analysis of crowd-out by education of Section 6.

The following .do files are run to produce the empirical results. They are to be saved in a subdirectory called "Census_Tax_Linkage\Analysis\".

- *CTL_1_Education_Regressions.do:* Table 5 (run before *CTL_Commands_Crowdout.do*)
- *CTL_2_ Education_Graphs.do:* Figure 4
- *CTL_3_Crowdout_Regressions.do:* Table 4 and descriptive statistics in Section 6 comparing the characteristics of workers in the RKD sample by level of education
- *CTL_4_Crowdout_Graphs.do:* Figure 3
- *CTL_5_Optimal_Bandwidth.do:* Estimates for the note about the optimal bandwidth test of Calonico et al. (2014) in the Online Appendix.

References

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *Stata Journal*, 14(4):909–946.

Ganong, P. and Jäger, S. (2016). A permutation test for the regression kink design. Unpublished Manuscript.

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review*, 96(1):152–175.

Wilkins, R, Tjepkema, M., Mustard, C., and Choinière, R. (2008). The Canadian census mortality follow-up study, 1991 through 2001. *Health Reports*, 19(3): 25–43. Statistics Canada Catalogue no. 82-003-XPE.