Description of data and programs used in "Dynamic Incentives in Retirement Earnings-Replacement Benefits"

1. Data access

We cannot make the datasets used in the project publicly available. The *social security records main analysis sample* and the *firms sample* administrative datasets were provided by Banco de Previsión Social (BPS) to Instituto de Economía (IECON), Universidad de la República, under a data sharing agreement (contact: Andrés Dean, andres.dean@fcea.edu.uy, +598 24131007). The *retirement sample* administrative dataset was provided by BPS to Departamento de Economía, Facultad de Ciencias Sociales (DECON), Universidad de la República, also under a data sharing agreement (contact: Ianina Rossi, ianina.rossi@cienciassociales.edu.uy, +598 24180938). Researchers can request these administrative datasets to BPS (https://www.bps.gub.uy/10429/banco-de-prevision-social.html).

The household survey data files that we use were produced by IECON. These are standardized versions (to make some variables compatible across years) of the survey Encuesta Continua de Hogares (ECH) for the years 1996 to 2016, which are produced by Instituto Nacional de Estadística and publicly available at https://www.ine.gub.uy/encuesta-continua-de-hogares1. The standardized datafiles are proprietary data of IECON but can be requested by contacting IECON (http://www.iecon.ccee.edu.uy/, contact: Martín Leites, martin.leites@fcea.edu.uy, +598 24131007).

2. Raw data files

Online Appendix B provides details about the data sources and samples.

- BPS main analysis sample administrative records: work histories between April 1996 and March 2016 of a random sample of 299,883 workers who were registered in social security at least one month during the period. These confidential data were provided by Uruguay's social security administration, Banco de Previsión Social (BPS), to IECON-UDELAR. It includes the following raw files:
    - *cabezales.txt*
    - *personas.txt*
    - *lineas_1.txt* and *lineas_2.txt*
- BPS retirement sample: The retirement sample is a dataset that contains retirement records for a sample of retired workers receiving retirement benefits for at least one month between December 2001 and February 2016, matched to the workers' employment records between April 1996 and March 2015. These confidential data were provided by Uruguay's social security administration, Banco de Previsión Social (BPS), to DECON-UDELAR. It includes the following raw files:
    - *personas_2017.dta*
    - *puestos_2017.dta*
    - *empresas_2017.dta*
    - *pasividades_2017.dta*
    - *pasivos_2017.dta*
    - *activos_2017.dta*

    - *prestaciones_activos_2017.dta*
- BPS firms sample: dataset provided by BPS that contains the information reported to the social security administration of a random sample of firms operating between the years 1996 and 2013 (firms sample). The data consist of the full history of administrative records between 1996 and 2013 for the firms sampled. These confidential data were provided by Uruguay's social security administration, Banco de Previsión Social (BPS), to IECON-UDELAR. It includes the following raw files:
    - *cabezales_C.txt*
    - *lineas_C_1996.txt* to *lineas_C_2013.txt*
- Household survey (ECH): annual labor force survey Encuesta Continua de Hogares (ECH), conducted by the Instituto Nacional de Estadística (INE). We use standardized datafiles that are proprietary data of IECON-UDELAR, which include the following files:
    - *p96.dta* to *p99.dta* for the years 1996 to 1999
    - *p0.dta* to *p16.dta* for the years 2000 to 2016
- Minimum contribution base (MCB) levels: we constructed a spreadsheet of the minimum contribution base from historical values of the standard contribution base (BFC) available from BPS's website https://www.bps.gub.uy/bps/valoreshistoricos.jsp?contentid=5479 (the MCB corresponds to 11 BFC). We provide these values in the following datafile:
    - *FICTO UNIPERSONALES.csv*

3. Code

Below we describe all the programs used in setting up the data and conducting the analysis – all code is made available as part of this package. The code was run using Stata MP version 15.1.

3.1 Programs for data setup

Programs used for data set up for the tables and figures presented in the paper

- **prepare0.do** reads and cleans the raw administrative data from BPS for the *main analysis sample*. This program produces the dataset *admindata.dta* which contains the full records for all the workers in the dataset, with one observation per individual, firm, and time (month-year). It calls the intermediate do-file **ipc.do** to add information on consumer price index.
- **prepare1_figure1c.do** reads *admindata.dta*, conducts some sample restrictions, creates new variables, adds information on the minimum contribution base, and saves the dataset *mainsample_age50.dta*. This dataset contains one observation per individual and time (month-year). It also produces the plot of age-earnings profiles around age 50 presented in Figure 1, panel C in the paper.
- **prepare2.do** reads *mainsample_age50.dta*. It prepares the data and creates variables for the descriptives and event studies centered around age 50, saving the data file *mainsample_age50analysis.dta*.
- **prepare3.do** reads and cleans the raw administrative data from BPS for the *retirement sample*. This program produces the dataset *retirementsample.dta* which contains the full records for all the workers in the dataset, with one observation per individual and time (month-year). It calls the intermediate do-file **ipc.do** to add information on consumer price index.

- **prepare4.do** reads *retirementsample.dta*, conducts some sample restrictions, creates variables for the estimation of the retirement age model, collapses the variables at the individual level, and saves the dataset *retirmodeldata_retirsample.dta*.
- **prepare5.do** reads *admindata.dta*, conducts some sample restrictions, creates variables for the retirement age model prediction, collapses the variables at the individual level, and saves the dataset *retirmodeldata_mainsample.dta*.
- **retirmodel-tableE1.do** estimates the retirement model using the Retirement Sample (using *retirmodeldata_retirsample.dta)*, and computes retirement probability predictions in the Retirement Sample and the Main Analysis Sample (using *retirmodeldata_mainsample.dta*). The predicted probabilities for the main sample are saved in the datafile *mainsample_predictions.dta*.
- **prepare6.do** uses sample selection from *mainsample_age50analysis.dta* and data from *admindata.dta*, adds predicted retirement from *mainsample_predictions.dta*, and saves *mainsample_medbcw.dta* (for main analysis) and *mainsample_expbcw.dta* (for robustness checks in Appendix E, Figure E.3).
- **prepare7.do** reads and cleans the raw household survey data. This program produces the dataset *ech_clean.dta* with the household survey sample for employed and self-employed workers ages 45 to 57. It calls the intermediate do-files **ipc.do** to add information on consumer price index, and **bfc.do** to add information on the administrative base of social contributions used for the computation of social security contributions and gross earnings. It creates variables for the descriptives and event studies centered around age 50.

Additional programs used for data set up for the tables and figures presented in the appendix

- **prepare8.do** reads *retirementsample.dta*. It conducts additional sample restrictions to create an analysis sample of employed and self-employed workers in the retirement sample, and creates variables for event studies centered around the start of the benefit calculation window (10 years before observed retirement), saving the data file *retirsample_bcw.dta*.
- **prepare9.do** reads *admindata.dta,* it merges it with the main analysis sample, and creates a balanced panel by imputing zeros in periods without reported earnings, to conduct the analysis of sample attrition presented in Appendix D.4, Figure D.3. It saves the dataset *extmargin_medbcw.dta*.
- The programs **prepare10.do** to **prepare13.do** prepare a dataset to conduct the placebo analysis presented in Appendix D.6:
    - **prepare10.do** starts from the dataset *admindata.dta* and conducts sample restrictions to create the sample of public sector workers used in the placebo analysis. It creates new variables, adds information on the minimum contribution base, and saves the dataset *data_placebo.dta*.
    - **prepare11.do** starts from the dataset *admindata.dta* and creates a working dataset for the placebo sample to use in the prediction of retirement based on retirement prediction model. It saves the dataset *retirmodeldata_placebo.dta*.
    - **prepare12.do** estimates the retirement model using the Retirement Sample (identically to retirmodel-tableE1.do, using *retirmodeldata_retirsample.dta)*, and computes retirement probability predictions in the placebo sample (using *retirmodeldata_placebo.dta*). The predicted probabilities are saved in the datafile *placebo_predictions.dta*.

    - **prepare13.do** uses the sample selection and data from *data_placebo.dta*, adds predicted retirement from *placebo_predictions.dta*, and saves *placebo_medbcw.dta* (to be used in the placebo analysis in Appendix D.6, Figure D.5).
- **prepare14.do** reads and cleans the raw administrative data from BPS for the *firm sample*. It conducts sample restrictions to create the small firm sample of firms and creates variables for event studies centered around the first time a worker turns age 50. It calls the intermediate do-file **ipc.do** to add information on consumer price index. It saves the data file *firmsample_age50.dta,* which has one observation per firm and time (month-year), and is used to conduct the analysis presented in Appendix D.9, Table D.8.

3.2 Programs to produce figures and tables

Programs that produce figures and tables presented in the paper

- **figure1ab.do** reads *retirementsample.dta* and creates the plots figure1A and figure1B, shown in Figure 1, Panels A and B. Figure 1A plots the histogram of discrete age in years at the time of retirement. Figure 1B plots the empirical CDF of continuous age (in months) at retirement.
- **prepare1_figure1c.do** plots average monthly earnings reported to social security of self-employed and employed workers at ages 45 to 57, shown in Figure 1, Panel C. This is done within the same program that cleans the administrative data (see description above).
- **figure2.do** reads *mainsample_medbcw.dta* and estimates the main event studies of reported earnings around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the three plots presented in Figure 2.
- **figure3.do** reads *ech_clean.dta* and estimates the event studies of earnings and hours of work around the age 50, using the household survey samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the three plots presented in Figure 3.
- **table1_admin.do** reads *mainsample_age50analysis.dta* and produces the parts of Table 1 that use administrative data (columns 1-4). The table must be assembled manually with the output produced by this program and by table1_survey.do.
- **table1_survey.do** reads *ech_clean.dta* and produces the parts of Table 1 that use household survey data (columns 5-8). The table must be assembled manually with the output produced by this program and by table1_admin.do.
- **table2.do** reads *mainsample_medbcw.dta* and estimates semi-parametric event study models of reported earnings around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, with different sets of control variables, and produces Table 2.

Programs that produce figures and tables presented in the appendix

- **figureC1.do** reads *mainsample_medbcw.dta* and estimates semi-parametric event study models of self-employed reported earnings for four subsamples: for the full sample of self-employed workers, the sub-sample of self-employed workers who report above the minimum contribution base at baseline, the sub-sample of frequent self-employment reporters (reporting density of at least 70%), and the sub-sample of self-employed workers in the two-pillar cohorts. It then plots the coefficients and confidence intervals and produces Appendix Figure C.1.

- **figureC2.do** reads *mainsample_medbcw.dta* and estimates an event study of the probability of reporting the minimum contribution base for self-employed workers, and plots the estimated coefficients and confidence intervals in Appendix Figure C.2.
- **figureC3.do** reads *mainsample_medbcw.dta* and estimates an event study of self-employed earnings for the sub-sample of frequent self-employment reporters (reporting density of at least 70%), and plots the estimated coefficients and confidence intervals in Appendix Figure C.3.
- **figureC4.do** reads *mainsample_medbcw.dta* and estimates event studies of employee reported earnings for the sub-samples of workers of firms with 1 to 4 employees (Appendix Figure C4, Panel A), 5 to 9 employees (Panel B), 10 to 49 employees (Panel C), and 50 or more employees (Panel D).
- **figureC5.do** reads *mainsample_medbcw.dta* and estimates an event study of employee reported earnings for the sub-sample of small-firm employees with strong employment attachment (reporting density of at least 70% and at least 1 year of job tenure). It plots the estimated coefficients and confidence intervals in Appendix Figure C.5.
- **figureC6.do** reads *mainsample_medbcw.dta* and estimates semi-parametric event study models of employee reported earnings for three subsamples: for the full sample of small firm workers, the sub-sample of high employment attachment small-firm employees (reporting density of at least 70% and at least 1 year job tenure), and the sub-sample of high employment attachment small-firm employees that stay on the same job for the sample period. It then plots the coefficients and confidence intervals and produces Appendix Figure C.6.
- **figureC7.do** reads *mainsample_medbcw.dta* and estimates event studies of the likelihood of changing jobs around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the three plots presented in Appendix Figure C.7.
- **figureC8.do** reads *mainsample_medbcw.dta* and estimates event studies of reported hours worked and implied wage per hour around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the six plots presented in Appendix Figure C.8.
- **figureC9.do** reads *mainsample_medbcw.dta* and estimates semi-parametric event study models of employee reported hours of work and implied wage per hour, for small firm workers, by industry. It then plots the coefficients and confidence intervals and produces the two plots presented in Appendix Figure C.9.
- **figureD1.do** reads *mainsample_age50analysis.dta* and estimates event studies of reported earnings around the age 50, using the main analysis samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the three plots presented in Appendix Figure D.1.
- **figureD2.do** reads *mainsample_medbcw.dta* and estimates event studies with individual fixed effects of reported earnings around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the three plots presented in Appendix Figure D.2.
- **figureD3.do** reads the balanced dataset *extmargin_medbcw.dta* and conducts difference-in-differences event studies of the likelihood of being in sample around the predicted start of the benefit calculation window, for the main analysis samples of self-employed and small-firm

employees using large-firm employees as a control group. It plots the estimated coefficients and confidence intervals in the two plots presented in Appendix Figure D.3.

- **figureD4.do** reads *mainsample_medbcw.dta* and estimates difference-in-differences event studies of reported earnings around the predicted start of the benefit calculation window, for the main analysis samples of self-employed and small-firm employees using large-firm employees as a control group. It plots the estimated coefficients and confidence intervals in the two plots presented in Appendix Figure D.4.
- **figureD5.do** reads *placebo_medbcw.dta* and estimates two specifications of an event study of reported earnings for the placebo sample of publicly owned firms, without and with individual fixed effects, and plots the estimated coefficients and confidence intervals in the two plots of Appendix Figure D.5.
- **figureD6.do** reads *mainsample_medbcw.dta* and estimates an event study of self-employed reported earnings as a ratio of the minimum contribution base, around the predicted start of the benefit calculation window, using the main analysis sample of self-employed workers. It plots the estimated coefficients and confidence intervals in Appendix Figure D.6.
- **figureD7.do** reads *ech_clean.dta* and estimates the event studies of net earnings around the age 50, using the household survey samples of self-employed, small-firm employees and large-firm employees, and plots the estimated coefficients and confidence intervals in the three plots presented in Appendix Figure D.7.
- **figureE1.do** reads *mainsample_predictions.dta* and plots the histogram of the predicted retirement age in the main sample, producing Appendix Figure E.1.
- **figureE2.do** reads *mainsample_medbcw.dta* and estimates event studies of reported earnings around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, using the predicted probability of retirement at age 60 as weights. It plots the estimated coefficients and confidence intervals in the three plots presented in Appendix Figure E.2.
- **figureE3.do** reads *mainsample_expbcw.dta* and estimates event studies of reported earnings around an alternative prediction of the start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, with and without individual fixed effects. It plots the estimated coefficients and confidence intervals in the six plots presented in Appendix Figure E.3.
- **figureE4.do** reads *retirsample_bcw.dta* and estimates event studies of reported earnings around the start of the benefit calculation window, measured as 10 years before observed retirement, for the samples of self-employed, small-firm employees and large-firm employees in the retirement sample. It plots the estimated coefficients and confidence intervals in the three plots presented in Appendix Figure E.4.
- **tableB3_admin.do** reads *mainsample_age50analysis.dta* and produces the rows of Appendix Table B.3 that use administrative data. The table must be assembled manually with the output produced by this program and by tableB3_survey.do.
- **tableB3_survey.do** reads *ech_clean.dta* and produces the rows of Appendix Table B.3 that use household survey data. The table must be assembled manually with the output produced by this program and by tableB3_admin.do.
- **tableC1.do** reads *mainsample_medbcw.dta* and estimates event study models of self-employed reported earnings, for the sub-sample of self-employed workers reporting the minimum contribution base at baseline, on one hand, and reporting more than the minimum contribution base at baseline, on the other, with different sets of control variables, producing Appendix Table C.1.

- **tableC2.do** reads *mainsample_medbcw.dta* and estimates three groups of semi-parametric event study models of self-employed reported earnings (for the sub-sample of frequent self-employment reporters (reporting density of at least 70%), for all self-employed workers, interacting event time dummies with an indicator for transition cohorts, and for all self-employed workers, interacting event time dummies with an indicator for having employees at baseline), with different sets of control variables, producing Appendix Table C.2.
- **tableC3.do** reads *mainsample_medbcw.dta* and estimates semi-parametric event study models of employee reported earnings for the sub-sample of high employment attachment small-firm employees (reporting density of at least 70% and at least 1 year job tenure), and the sub-sample of high employment attachment small-firm employees that stay on the same job for the sample period, with different sets of control variables. It produces Appendix Table C.3.
- **tableC4.do** reads *mainsample_medbcw.dta* and produces Appendix Table C.4, which shows summary statistics of reported hours worked, for all employed workers in the main analysis sample, by industry.
- **tableD1.do** reads *mainsample_age50analysis.dta* and estimates semi-parametric event study models and parametric models with linear age trends of reported earnings around age 50, using the main analysis samples of self-employed, small-firm employees and large-firm employees, with different sets of control variables, and produces Appendix Table D.1.
- **tableD2.do** reads *mainsample_medbcw.dta* and estimates parametric models with linear event-time trends of reported earnings around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, small-firm employees and large-firm employees, with different sets of control variables, and produces Appendix Table D.2.
- **tableD3.do** reads *mainsample_medbcw.dta* and estimates dynamic difference-in-differences models of reported earnings around the start of the benefit calculation window, for the samples of self-employed and small-firm employees, using large-firm employees as a control group, with different sets of control variables. It produces Appendix Table D.3.
- **tableD4.do** reads *mainsample_medbcw.dta* and estimates a semi-parametric event study model of reported earnings as a ratio of the minimum contribution base, around the predicted start of the benefit calculation window, using the main analysis samples of self-employed, with different sets of control variables, and produces Appendix Table D.4.
- **tableD5.do** reads *ech_clean.dta* and estimates semi-parametric event study models and parametric models with linear age trends of earnings and hours of work around age 50, using the household survey sample of self-employed workers, with different sets of control variables, and produces Appendix Table D.5.
- **tableD6.do** reads *ech_clean.dta* and estimates semi-parametric event study models and parametric models with linear age trends of earnings and hours of work around age 50, using the household survey sample of small-firm employees, with different sets of control variables, and produces Appendix Table D.6.
- **tableD7.do** reads *ech_clean.dta* and estimates semi-parametric event study models and parametric models with linear age trends of underreporting of earnings around age 50, using the household survey samples of self-employed, small-firm employees and large-firm employees, and produces Appendix Table D.7.
- **tableD8.do** reads *firmsample_age50.dta* and estimates semi-parametric event study models and parametric models with linear trends for the *firm sample* around the time when a worker turns age 50. It produces Appendix Table D.8.
- **retirmodel-tableE1.do** estimates the retirement model using the Retirement Sample (using *retirmodeldata_retirsample.dta)*, and computes retirement probability predictions in the

Retirement Sample and the Main Analysis Sample (using *retirmodeldata_mainsample.dta*). These predicted probabilities are saved in the datafiles *retirsample_predictions.dta* and *mainsample_predictions.dta*, respectively. It also tabulates the results of the model, producing Appendix Table E.1.

- **tableE2.do** reads *retirsample_predictions.dta* and tabulates observed and predicted probabilities of the retirement model and produces Appendix Table E.2.
- **tableF1.do** performs back-of-the-envelope calculations of costs under different scenarios and produces Appendix Table F.1.
- **tableF2.do** performs back-of-the-envelope calculations of costs under different scenarios with an alternative measure of the estimated increase in earnings and produces Appendix Table F.2.