Consumption Partial Insurance in the Presence of Tail Income Risk

INSTRUCTIONS FOR DATA REPLICATION*

Anisha Ghosh Alexandros Theloudis

December 10, 2025

This `README` file guides the reader through the replication of the results in the paper 'Consumption Partial Insurance in the Presence of Tail Income Risk' (Ghosh and Theloudis, 2025). The code in this replication package constructs the analysis sample and estimation results from publicly available data, as we describe below.

The code has three main parts. One main Stata script assembles the data, selects the main analysis sample, and carries out bootstrap resampling. Subsequently, one Matlab script carries out the estimation of the income and consumption processes. Finally, an R script performs the welfare calculations. The code runs in approximately 15 hours on a modern MacBook Pro, almost equally split among the three parts.

1 Data Availability

1.1 Statement about Rights

We certify that as authors of the manuscript, we have legitimate access to and permission to use the data used in this manuscript. All data are freely available online.

1.2 Details on Data Sources

Our two main data sources are the Panel Study of Income Dynamics (Institute for Social Research, 2021) and the Consumer Expenditure Survey (Bureau of Labor Statistics, 2019a).

In addition to those, we also use data on prices (Bureau of Labor Statistics, 2019b) and the federal minimum wage (Department of Labor, 2019), the data provided by Blundell et al. (2008) as part of their replication package, and tables with estimates of income moments from Guvenen et al. (2021).

Below we describe how the user can get access to these data.

*Ghosh: Desautels Faculty of Management, McGill University; Theloudis (corresponding author): Department of Econometrics & Operations Research, Tilburg University, email: a.theloudis@gmail.com.

Table 1: Raw data

	Source	Data	File	Is file provided?
1.	PSID	Family File `YYYY`	`fYYYYER.dta`	No, but available online
2.	PSID	Individual File Cross-year	`i1968_2019.dta`	No, but available online
3.	CEX	Interview Data `yyyy`	`fmliyyyy1x.dta`	No, but available online
			`fmliyyyyq.dta`	No, but available online
			`memiyyyy1x.dta`	No, but available online
			`memiyyyyq.dta`	No, but available online
4.	CPI	Consumer Price Index:		
		Total for United States	`BLS_CPIU_1931_2019.xlsx`	Yes
		Childcare & Schooling	`ChildcareSchoolingPrices.xlsx`	Yes
		Food	`FoodPrices.xlsx`	Yes
		Fuel & Utilities	`FuelUtilitiesPrices.xlsx`	Yes
		Medical Services & Drugs	`MedicalPrices.xlsx`	Yes
		Transportation	`TransportationPrices.xlsx`	Yes
5.	MW	Federal Minimum Wage	`StLouisFed_Federal_Min_Wage.csv`	Yes
6.	BPP	Replication Data	`data.dta`	Yes
			`cexall.dta`	Yes
			`finprice.dta`	Yes
			`natpr.dta`	Yes
			`tax9192.dta`	Yes
7.	GKOS	Empirical estimates	`gkos_2021_moments.xlsx`	Yes

Notes: For PSID, see Institute for Social Research (2021); '`YYYY`' corresponds to the odd years from 1997 to 2019. For CEX, see Bureau of Labor Statistics (2019a); '`yyyy`' corresponds to all years in 1997–2020, '`q`' corresponds to quarters 2-4. For CPI, see Bureau of Labor Statistics (2019b). For MW, see Department of Labor (2019). For BPP, see Blundell et al. (2008). For GKOS, see Guvenen et al. (2021).

1.2.1 Access to PSID household data (items 1-2 in table 1)

The paper uses household data from the Panel Study of Income Dynamics (Institute for Social Research, 2021). Specifically, we use:

- the Main Study Family Files in waves 1997-2019 (item 1 in table 1; files available biennially), which contain information on household income, consumption, demographics;
- the Main Study Individual File Cross-year (item 2 in table 1), which allows us to assemble households in a panel.

To access the data, please register at https://simba.isr.umich.edu/U/login.aspx, accept the conditions of use, and confirm your registration through an email.

The raw data is subject to a redistribution restriction, but can be freely downloaded from https://simba.isr.umich.edu/data/data.aspx. This is the public data section of the PSID website, providing custom files to users.

For the Main Study Family Files, please go to *Data Center* > *File*, then choose *PSID Family-level* > *PSID Main Family Data* from the breakdown menu. For survey year 1997,

click on the relevant part in the menu (1997: 6,747 observations 2,263 variables), select all variables in the text box, and add them to cart. In the cart, select the format 'ASCII Data with Stata Statements' and download the output as a zip folder. The folder contains a large text file (the data; named after the specific job assigned to this task by the PSID server) and a do file (named similarly) that reads the data. Edit this do file to correctly read the location of the text file in your computer and run it. Save the output data as `f1997.dta` in a directory of your choice; `PSIDdir` is set inside the main Stata script and should point to this directory (more on this below). Please repeat this process for all other years until 2019, while remaining logged in to your account.

For the Main Study Individual File Cross-year, please go to *Packaged Data > Main and Supplemental Studies*, then choose *Main Study > Family Files* from the breakdown menu, and download the 'Cross-year Individual: 1968-2023' file.[1] The zip folder contains a large text file (the data) and a do file that reads the data. Edit this do file to correctly read the location of the text file in your computer and run it. Save the output data as `i1968_2019.dta` within the `PSIDdir` directory mentioned above.

1.2.2 Access to CEX household data (item 3 in table 1)

The paper uses interview data from the Consumer Expenditure Survey in years 1997-2020 (Bureau of Labor Statistics, 2019a, data available annually). Specifically, we use:

- the quarterly expenditure files, which contain information on household expenditures (file `fmliyyyy1x.dta` for quarter 1, files `fmliyyyyq.dta` for quarters 2-4);
- the quarterly salary files, which contain information on income (file `memiyyyy1x.dta` for quarter 1, files `memiyyyyq.dta` for quarters 2-4).

The raw data is subject to a redistribution restriction, but can be freely downloaded from the BLS website at https://www.bls.gov/cex/pumd_data.htm. This is the public use microdata site, providing zip files separately for each survey year.

Please scroll down and select the 'Stata' tab of the page; then download the zip folders that contain the interview data in years 1997 to 2020. For instance, for year 1997, the download saves the zip `intrvw97` in your computer. Please unzip this file, extract the folder `intrvw97`, and save it in a directory of your choice; `CEXdir` is set inside the main Stata script and should point to this directory (more on this below). You may discard all other folders in the zip. Please repeat the process for all other years until 2020.[2]

1.2.3 Access to other data (items 4-7 in table 1)

We use data on prices and the current price index to carry out the demand estimation in the CEX and convert all monetary amounts to 2018 dollars. The data, originally provided

[1] This file name changes when the PSID releases a new wave. This does not affect the replication of our code; please follow the same steps even if the period of the cross-year individual file is extended beyond 2023.

[2] In the end of this process, please copy the 2020Q1 files from the 2019 into the 2020 folder (and consistent with how other Q1 files are named, call them `201x`) because as per the CEX website "The 2020 PUMD package for the Interview Survey only contains Q2-Q4 of 2020 and Q1 of 2021. If you need Q1 2020, please download the 2019 data package".

by the Bureau of Labor Statistics (Bureau of Labor Statistics, 2019b), are downloaded from the BLS at https://www.bls.gov/cpi/data.htm.

To obtain the price data, please select the 'All Urban Consumers current series' database. Using the multi-screen option, please select the options for U.S. city average, current base, SA0 All Items, monthly, time period 1919-2019 including annual averages, and download. Repeat similarly for the prices of SAF1 Food, SEEB Tuition, other school fees, and childcare, SAH2 Fuels and utilities, SAM2 Medical care services, and SAS4 Transportation services in the period 1997-2019 (or a longer horizon). A copy of the data is provided as part of this archive; the downloaded files have been renamed according to table 1 (item 4).

We use data on the federal minimum hourly wage for non-farm workers to carry out our sample selection. The data are provided by the Department of Labor through the Federal Reserve Economic Data repository at https://fred.stlouisfed.org/. Please search the repository for 'FEDMINNFRWG'; when the data show on screen as a graph, please select the period 1938-10-01 to 2019-06-01 and download. A copy of the data is provided as part of this archive; the downloaded file has been renamed according to table 1 (item 5).

We use the replication data from Blundell et al. (2008) to investigate how the frequency of observations (annual versus biennial) may affect the estimates of partial insurance. The data are available through the American Economic Association's data repository at https://www.openicpsr.org/openicpsr/project/113270/version/V1/view. A copy of the relevant data files is provided as part of this archive (item 6).

We finally use estimates of income moments in the administrative data of Guvenen et al. (2021). The data are on Fatih Guvenen's website at https://www.fatihguvenen.com/published-papers. Please locate the paper on the page and download the file 'Data Moments for Men'. A copy of this file is provided as part of this archive (item 7).

2 Contents of Replication Package

2.1 Contents

The contents of our replication package are shown in table 2. The first item in the table, `Data/psid_imputed.dta`, is processed data after we assemble the PSID and CEX samples, deflate all monetary amount using the CPI, carry out sample selection, estimate the demand for food, and impute consumption from the CEX into the PSID sample. We provide this file so that users who do not wish to download the raw data (section 1.2) may still replicate the later parts of our code, most importantly the estimation of the income and consumption processes. More details on this data file follow in section 3.2 below.

2.2 Description of Code

The code has three main parts. The first part of the code consists of eight Stata scripts (items 2-9 in table 2):

1. The file `GT00_Shell.do` sets the directories, installs the necessary packages, and calls all other Stata scripts. The user will only need to interact with this file in order to replicate all our results (detailed instructions in section 3.2 below).

Table 2: Contents of replication package

	Name	Type	Notes
1.	`Data/psid_imputed.dta`	data file	*see text*
2.	`Code/Stata/GT00_Shell.do`	Stata script	master file Stata
3.	`Code/Stata/GT01_Extract_PSID.do`	Stata script	
4.	`Code/Stata/GT02_Selection_Variables.do`	Stata script	
5.	`Code/Stata/GT03_Impute_CEX.do`	Stata script	
6.	`Code/Stata/GT04_First_Stage.do`	Stata script	
7.	`Code/Stata/GT05_Export_GMM.do`	Stata script	
8.	`Code/Stata/GT06_Bootstrap.do`	Stata script	
9.	`Code/Stata/GT07_BPP.do`	Stata script	
10.	`Code/Matlab/GT_Shell.m`	Matlab script	master file Matlab
11.	`Code/Matlab`	directory	contains Matlab functions & scripts
12.	`Code/R/GT_Welfare_Costs.R`	R script	master file R
13.	`BLS_CPIU_1931_2019.xlsx`	data file	prices
14.	`ChildcareSchoolingPrices.xlsx`	data file	prices
15.	`FoodPrices.xlsx`	data file	prices
16.	`FuelUtilitiesPrices.xlsx`	data file	prices
17.	`MedicalPrices.xlsx`	data file	prices
18.	`TransportationPrices.xlsx`	data file	prices
19.	`StLouisFed_Federal_Min_Wage.csv`	data file	min wage
20.	`gkos_2021_moments.xlsx`	data file	estimates from Guvenen et al. (2021)
21.	`BPP`	directory	data from Blundell et al. (2008)

2. The file `GT01_Extract_PSID.do` extracts the raw PSID data, assembles the panel, and merges it with the price data in `BLS_CPIU_1931_2019.xlsx` and the minimum wage data in `StLouisFed_Federal_Min_Wage.csv`.

3. The file `GT02_Selection_Variables.do` carries out the sample selection, cleans and harmonizes variables, deflates all monetary amounts, designs appropriate subsamples, and produces the table of descriptive statistics in the appendix (table C.1).

4. The file `GT03_Impute_CEX.do` extracts the raw CEX data, assembles quarterly expenditure into annual equivalent amounts, carries out the sample selection, conducts the demand estimation in the CEX using the good-specific prices (items 14-18 in table 2), and imputes consumption into the PSID. It produces the table for the comparison of sample means (table C.4) and the figures for the moments of log consumption in the PSID and CEX in the appendix (figure C.1).

5. The file `GT04_First_Stage.do` runs the first-stage estimation for income and consumption, and creates a table for the sources of income shocks (table C.3).

6. The file `GT05_Export_GMM` exports the data in a format that is suitable for the structural estimation in Matlab; it also estimates the variance of income measurement error.

7. The file `GT06_Bootstrap.do` carries out block resampling with replacement and repeats the first-stage estimation and the estimation of income measurement error in each sample; it exports the data in a format that is suitable for the structural estimation.

8. The file `GT07_BPP.do` assembles income and consumption growth at annual and biennial rates from the original data of Blundell et al. (2008); it exports the data in a format that is suitable for the structural estimation.

The second part of the code consists of one main Matlab script, `GT_Shell.m` (item 10 in table 2), which calls several other scripts and functions within `Code/Matlab`.

`GT_Shell.m` reads the data exported from the previous part of the code, estimates the empirical moments of income and consumption, estimates the income process and the linear and quadratic consumption functions, replicates BPP using income and consumption growth at annual and biennial rates, conducts sensitivity analyses, and exports all other tables in the paper and the appendix that are not mentioned above. Appendix table C.2 specifically requires the empirical estimates in `gkos_2021_moments.xlsx`.

The third and final part of the code consists of one R script, `GT_Welfare_Costs.R` (item 12 in table 2), which carries out the welfare calculations of appendix D and creates all five tables in that appendix.

The tables and figures that are created in the various parts of the code are respectively saved in folders `Tables` and `Figures`, generated by `GT00_Shell.do`.

3 Instructions to Replicators

3.1 Software and Hardware Requirements

The code requires Stata MP 18. Earlier versions of Stata may run the code, but we have not checked for backward compatibility. We make use of commands that are part of the default distribution of Stata, as well as commands available in packages `estout`, `catplot`, `coefplot`, `dataout`, `ivreg2`, and the graphics scheme `lean1`. Lines 59-64 in `GT00_Shell.do` install these dependencies directly from the Internet. A user who has not previously installed these packages must have access to the Internet in order to run the code.

The code runs in Matlab 2024a; recent earlier versions will not cause problems but we have not exhaustively checked backward compatibility. The code uses the optimization, econometrics, and statistics and machine learning toolboxes.

The code requires the differential evolution optimization library (DEoptim) in R.

The code runs in approximately 15 hours on a modern MacBook Pro, almost equally split among the three parts (Stata, Matlab, R). The raw data, including the bootstrap samples, require about 4GB disk space. The disc space taken up by the output is negligible.

3.2 Instructions

Follow these instructions to run our code:

- Download the raw data as per section 1.2 herein; ensure that the directories `PSIDdir` and `CEXdir` contain a copy of all files 1-3 of table 1. All other files are provided as part of the replication package.
- Edit `GT00_Shell.do` for the default directories in lines 34-43 and run it.
- Edit `GT_Shell.m` for the default directories in lines 41 & 44 and run it.
- Edit `GT_Welfare_Costs.R` for the default directory in line 35. Run it, and that's it!

Note 1: Users who do not wish to get access to the raw data should comment out lines 67-69 in `GT00_Shell.do`. The code will then run without processing the raw data; instead it will make use of the file `Data/psid_imputed.dta`, which we make available as part of this replication package. This pathway will create most output shown in the paper/appendix, except for tables C.1, C.4, and figure C.1.

Note 2: The Stata script uses results that are gradually stored in your system's RAM memory, so we do not recommend interrupting the code and resuming it later. Ideally, `GT00_Shell.do` should be allowed to run once, without interruptions.

References

Blundell, R., L. Pistaferri, and I. Preston (2008). Consumption inequality and partial insurance. *American Economic Review 98*(5), 1887–1921.

Bureau of Labor Statistics (1913–2019b). Consumer price index. Monthly CPI data, historical and current; includes CPI-U, CPI-W, and supplemental files.

Bureau of Labor Statistics (1999–2019a). Consumer expenditure survey. Public-use microdata files.

Department of Labor (1938–2019). Federal minimum hourly wage for nonfarm workers. Monthly series; original data from U.S. Department of Labor, compiled via FRED.

Ghosh, A. and A. Theloudis (2025). Consumption partial insurance in the presence of tail income risk. *Journal of Political Economy Macroeconomics*. Accepted manuscript.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2021). What do data on millions of U.S. workers reveal about lifecycle earnings dynamics? *Econometrica 89*(5), 2303–2339.

Institute for Social Research (2021). Panel study of income dynamics, public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.