

Overview

This is a replication package for “**Conditional Choice Probability Estimation with an Imperfectly Measured Latent State**” (Yujung Hwang, Quantitative Economics), containing R code files for Monte Carlo simulation results (Bus engine replacement/Schelling’s segregation), and STATA/R code files for empirical example estimation section (Model of labor supply and mental health).

Description of package content

1. **Data Availability and Provenance Statements**
2. **Computational requirements**
3. **Software Requirements**
4. **Memory, Runtime, Storage Requirements**
5. **Instructions to Replicators**
6. **List of Tables and Programs**
7. **Contact Information**

Data Availability and Provenance Statements

All Monte Carlo simulation results do not require any data.

The estimation section results used the following data, the newest version of which is distributed by UK Data Archive under End User License. I do not have the right to redistribute this data, so I only provide the URL, where they can be downloaded. You may download the End User License data immediately after you register with the UK Data Service (<https://www.data-archive.ac.uk/>) and agree with their terms of use.

The Harmonised BHPS/UKHLS data I used (15th edition) has been taken down from their website. To access this old version, you should contact the UK Data Archive helpdesk by filling out the form at <https://beta.ukdataservice.ac.uk/help> The UKDA team typically try to reply within five business days.

- University of Essex, Institute for Social and Economic Research. (2022). Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 15th Edition. UK Data Service. SN:6614, DOI: <http://doi.org/10.5255/UKDA-SN-6614-16>

I used Wright (2020) STATA code files to create a cumulative year of experience variable in BHPS/UKHLS data. These code files are distributed under [CC-By Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/), which allows redistribution. Therefore, I copied the downloaded code files in the replication folder. Note that I made minor modifications to the original code files (e.g., updating file directories and file names).

- Wright, L. (2020). Producing working-life histories in the BHPS and UKHLS 2017-2020. [Data Collection]. <http://doi.org/10.5255/UKDA-SN-854327>

I used the UK Consumer Price Index downloaded from the ONS website and the CPI data file is included in the replication package.

- Office for National Statistics (ONS). *Consumer price inflation time series*. UK: ONS. Available at: <https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/consumerpriceindices> (accessed: 17-08-2022.).

Computational requirements

All codes were run on either a Dell mobile workstation New XPS 15 9510 model (16 cores/64GB Ram/11th Gen Intel(R) Core(TM) i9-11900H @ 2.50GHz 2.50 GHz), or a Johns Hopkins university computation slurm system server (Rockfish, “shared” or “parallel” partition, <https://www.arch.jhu.edu/guide/>, or RIT HPC, “cpu” partition). The CPU chips used in these servers are as follows:

- Rockfish “shared” partition: Intel(R) Xeon(R) Gold 6448Y CPU @ 2.10 GHz
- Rockfish “parallel” partition: Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz
- RIT HPC “cpu” partition: Intel(R) Xeon(R) Platinum 8480+ CPU @ 2.00 GHz

Software Requirements

STATA Packages Used

I used STATA MP Version 17. Install “sutex2” by typing

```
ssc install sutex2
```

Monte Carlo Simulations / Estimation / R codes

Conda environment used for Monte Carlo simulations and Estimation was saved as DDCM_Estimation.yml file. To recreate the same environment (named “R442_DDCM”), move the “DDCM_Estimation.yml” to your working directory.

First, load the anaconda3 module. I used Rockfish and RIT HPC servers with the following anaconda modules.

```
module load anaconda3/2024.02-1
```

or

```
module load anaconda3/2023.09
```

Next, create the conda environment using the .yml file

```
conda env create -f DDCM_Estimation.yml
```

To activate the environment, type

```
conda activate R442_DDCM
```

or

```
source activate R442_DDCM
```

In addition to packages specified in DDCM_Estimation.yml, I used the following R packages on Dell mobile workstation.

```
install.packages("rstudioapi")
```

```
install.packages("devtools")
```

```
devtools::install_github("thomasp85/patchwork")
```

```
install.packages("expm")
```

```
install.packages("vtable", repos= "https://cloud.r-project.org/")
```

To list the packages specified in DDCM_Estimation.yml file again,

```
install.packages(c("readstata13", "nloptr", "tidyverse", "pracma", "Rcpp", "mlogit", "kableExtra", "tidyr", "readr", "optimPara
```

```
l1el", "xlsx", "foreach", "xtable", "foreign", "forcats", "nnet",  
"iterators", "purrr", "dplyr", "lubridate", "Hmisc", "plm", "sos",  
"doParallel"), repos = "https://cran.r-project.org")
```

To see the exact package version numbers, read the “DDCM_Estimation.yml” file. **You can open the file in a Notepad program.**

Memory, Runtime, Storage Requirements

Quite a few codes require substantial computational resources.

For example, I have run Monte Carlo simulations 200 times each for 160 set ups (which differs by a model, sample size, proxy informativeness parameter). Evaluating one function takes <5 minutes, but repeating this estimation 32000 times takes quite a long time. So I ran these codes in parallel in server.

The main estimation code “estimateBHPS_UKHLS_DDCM_250410.R” can be run using 16GB Ram, a single CPU core. I have set the starting value close to the final estimate, so the code can finish within 6 hours on Rockfish “shared” partition.

The standard errors are computed through bootstrapping, and this requires much computational resources. I manually ran 10 bootstrapping codes, each of which repeats 10 estimations. Due to memory constraint on server, I could not run one bootstrap code that conducts 100 estimations simultaneously. Each bootstrapping code was run on the Rockfish “parallel” partition, with 48 cores / 192 GB Ram, and was completed within 58 hours, when all ten codes were run simultaneously.

Note: In the replication package, the master script has been revised to call the estimation and bootstrapping codes **sequentially**. The run times reported for master scripts 8, 16, and 29 reflect a setup in which the computationally intensive routines were executed on a separate server; in that configuration, the master script primarily aggregates results and generates tables/figures.

Num ber	Type	Code	Server	Partiti on	CP U	Mem ory	Time
1	Monte Carlo Segregation	MonteCarloThreeProxyEveryYearV4Final.R	RIT HPC	cpu	100	100GB	2.25 hours
2	Monte Carlo Segregation	MonteCarloRotatingModuleV4Final.R	RIT HPC	cpu	100	100GB	2.75 hours
3	Monte Carlo Segregation	MonteCarloNoProxyV4Final.R	RIT HPC	cpu	100	100GB	0.75 hours
4	Monte Carlo Segregation	MonteCarloTwoProxyEveryYearV4Final.R	RIT HPC	cpu	100	100GB	2.5 hours
5	Monte Carlo Segregation	MonteCarloOneProxyEveryYearV4Final.R	RIT HPC	cpu	100	100GB	2.75 hours
6	Monte Carlo Segregation	MonteCarloThreeOneProxyV4Final.R	RIT HPC	cpu	100	100GB	2.5 hours
7	Monte Carlo Segregation	MonteCarloOneProxyEveryYearNoMEV4Final.R	RIT HPC	cpu	100	100GB	0.5 hour
8	Monte Carlo Segregation	masterCodeV4SegFinal.R	Dell mobile workstation		1	<1GB	<1min
9	Monte Carlo BusEngine	MonteCarloBusThreeProxyEveryYearV4Final.R	RIT HPC	cpu	100	100GB	2.25 hours
10	Monte Carlo BusEngine	MonteCarloBusRotatingModuleV4Final.R	RIT HPC	cpu	100	100GB	4 hours
11	Monte Carlo BusEngine	MonteCarloBusNoProxyV4Final.R	RIT HPC	cpu	100	100GB	3 hours
12	Monte Carlo BusEngine	MonteCarloBusTwoProxyEveryYearV4Final.R	RIT HPC	cpu	100	100GB	3.5 hours
13	Monte Carlo BusEngine	MonteCarloBusOneProxyEveryYearV4Final.R	RIT HPC	cpu	100	100GB	7.5 hours
14	Monte Carlo BusEngine	MonteCarloBusThreeOneProxyV4Final.R	RIT HPC	cpu	100	100GB	6 hours
15	Monte Carlo BusEngine	MonteCarloOneProxyEveryYearNoMEV4Final.R	RIT HPC	cpu	100	100GB	1.25 hour
16	Monte Carlo BusEngine	masterCodeV4Final.R	Dell mobile workstation		1	<1GB	<1min
17	Estimation Data Cleaning	master_DDCM_241202.do	Dell mobile workstation		1	<4GB	<0.5hour
18	Estimation	estimateBHPS_UKHLS_DDCM_250410.R	Rockfish	shared	4	3GB	6 hours
19	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part1.R	Rockfish	parallel	48	192GB	24-58 hours
20	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part2.R	Rockfish	parallel	48	192GB	24-58 hours
21	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part3.R	Rockfish	parallel	48	192GB	24-58 hours
22	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part4.R	Rockfish	parallel	48	192GB	24-58 hours

23	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part5.R	Rockfish	parallel	48	192GB	24-58 hours
24	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part6.R	Rockfish	parallel	48	192GB	24-58 hours
25	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part7.R	Rockfish	parallel	48	192GB	24-58 hours
26	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part8.R	Rockfish	parallel	48	192GB	24-58 hours
27	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part9.R	Rockfish	parallel	48	192GB	24-58 hours
28	Estimation SE Bootstrapping	bootstrapBHPS_UKHLS_DDCM_250410_part10.R	Rockfish	parallel	48	192GB	24-58 hours
29	Estimation	master_DDCM_250410.R	Dell mobile Workstation		1	<1GB	<5 mins

Instructions to Replicators

Monte Carlo Simulation / Comparison with Hotz and Miller (1993)

I conducted 200 simulations for 160 different survey designs that differ by the number and frequency of proxies, proxy informativeness. Each simulation takes only <5 minutes on a personal laptop, but because of large volumes of simulations, I conducted most simulations in parallel on university server. In replication package, I modified the master code file so that all files can be executed sequentially.

1. Move the entire “MonteCarlo” folder to your project area, either personal laptop or university server file directory.
2. Create the same virtual environment using “DDCM_Estimation.yml” file.
3. Run `./MonteCarlo/BusEngine/R/masterCodeV4Final.R` and `./MonteCarlo/Segregation/R/masterCodeV4SegFinal.R` to generate all results. The two master code files use the “rstudioapi” package to update file paths based on the directory in which the code is saved.

```
library(rstudioapi)
rootdir <-
dirname(dirname(rstudioapi::getSourceEditorContext
())$ path))
```

```

codedir <- paste0(rootdir, "/R")
excelldir <- paste0(rootdir, "/output/excel")
graphdir <- paste0(rootdir, "/output/graph")

output_excelldir <- excelldir
output_graphdir <- graphdir

```

To save time, you may execute the MonteCarlo* jobs in parallel, in any order, on a server. **In that case, you should update the file directories in each code file.** After they finish, copy all generated output files into ./output/excel/ **before** running the downstream scripts that produce the graphs.

Estimation

The estimation codes consist of STATA code file for data cleaning and R code files for estimation. To replicate the estimation results, follow the below steps.

- 1) Download BHPS/UKHLS data from <http://doi.org/10.5255/UKDA-SN-6614-16>
Note that I used 15th edition. The later release might include small changes in data.
Put the raw BHPS/UKHLS data (.dta files) in subfolder, “./Estimation/rawdata”
- 2) Update the directories in “./Estimation/do/work-life-histories-ukhls-bhps-master/Do Files/Launch Programme.do” and run the code. The code will create an intermediate dataset “Merged Dataset.dta”, containing cumulative years of experience variable. Save the “Merged Dataset.dta” in ./Estimation/data folder
- 3) Update the directories in “./Estimation/do/master_DDCM_241202.do” and run the code. It will create an intermediate data for estimation, “cIn_BHPS_UKHLS_DDCM_241202.csv” in data folder and descriptive statistics for Table 6.1.
- 4) Create the same virtual environment for R program using “DDCM_Estimation.yml” file and install additional R packages listed in “Software Requirements” section. They are R packages, “rstudioapi”, “devtools”, “patchwork”, “expm”, “vtable”.
- 5) Update the directories in “./Estimation/R/master_DDCM_250410.R” and run the code to generate all results.

Originally, estimateBHPS_UKHLS_DDCM_250410.R and bootstrapBHPS_UKHLS_DDCM_250410_part(1–10).R were executed **in parallel** on a computation server to save time. In this replication package, the master script has been modified to run all components **serially**.

- a. If you want to run only subsets of the 10 bootstrap code files, you should set “numbootfile” parameter equal to the number of bootstrap files you want to run in **line 48** of “master_DDCM_250410.R”. Note that this parameter is also used in “combineBHPS_UKHLS_DDCM_250410.R”, which aggregates the bootstrap estimates. See “combineBHPS_UKHLS_DDCM_250410.R” for more details.
- b. If you have enough computational resources, you may run 10 bootstrap files in parallel in any order on a computation server. In that case, update the file directories in **line 21-24** in each bootstrap code file. You may need to write job submission batch files specific to your local system.

List of Tables and Programs

Number	Graph /Table	Output File Name	Code	Line	Comments
1	Table 6.1	sampleCharBHPS_UKHLS_241202.tex	describe_BHPS_UKHLS_241202.do	31-33	
2	Table 6.1	sampleCountBHPS_UKHLS_241202.xlsx	describe_BHPS_UKHLS_241202.do	40-42	
3	Table 6.3	flowutil_parest_nr_100_250410.csv	combineBHPS_UKHLS_DDCM_250410.R	56	call estimate files prepared by "estimateBHPS_UKHLS_DDCM_250410.R" and "bootstrapBHPS_UKHLS_DDCM_250410_partX.R"
4	Table 6.3	marriage_parest_nr_100_250410.csv	combineBHPS_UKHLS_DDCM_250410.R	95	
5	Table 6.3	unobs_tran_parest_nr_100_250410.csv	combineBHPS_UKHLS_DDCM_250410.R	67	
6	Table 6.3	wage_parest_nr_100_250410.csv	combineBHPS_UKHLS_DDCM_250410.R	86	
7	Figure 6.2	mentalhealthtypeshare_250410.png	functionBHPS_UKHLS_DDCM_250410.R		called by estimateBHPS_UKHLS_DDCM_250410.R
8	Figure 6.3	modelfit_loginc_age_250410.png	functionBHPS_UKHLS_DDCM_250410.R	824	called by estimateBHPS_UKHLS_DDCM_250410.R

9	Figure 6.3	modelfit_loginc_exp_250410.png	functionBHPS_UKHLS_DCM_250410.R	858	called by estimateBHPS_UKHLS_DCM_250410.R
10	Figure 6.3	emp_fit_age_250410.png	combineBHPS_UKHLS_DCM_250410.R	214	call estimate files prepared by "estimateBHPS_UKHLS_DDCM_250410.R" and "bootstrapBHPS_UKHLS_DDCM_250410_partX.R"
11	Figure 6.3	emp_fit_exp_250410.png	combineBHPS_UKHLS_DCM_250410.R	232	call estimate files prepared by "estimateBHPS_UKHLS_DDCM_250410.R" and "bootstrapBHPS_UKHLS_DDCM_250410_partX.R"
12	Figure 6.4	mentalhealthpred_age30_250410.png	functionBHPS_UKHLS_DCM_250410.R	1454	called by estimateBHPS_UKHLS_DCM_250410.R
13	Figure 6.4	mentalhealthpred6yr_age30_250410.png	functionBHPS_UKHLS_DCM_250410.R	1481	called by estimateBHPS_UKHLS_DCM_250410.R
14	Figure 6.5	counter_empprob_250410.png	functionBHPS_UKHLS_DCM_250410.R	1876	called by estimateBHPS_UKHLS_DCM_250410.R
15	Figure 6.5	counter_mentalhealth_250410.png	functionBHPS_UKHLS_DCM_250410.R	1915	called by estimateBHPS_UKHLS_DCM_250410.R
16	Figure 6.5	counter_logInc_250410.png	functionBHPS_UKHLS_DCM_250410.R	1890	called by estimateBHPS_UKHLS_DCM_250410.R
17	Table B.1	prxcorr_250410.tex	estimateBHPS_UKHLS_DCM_250410.R	57	
18	Table B.2	testRankCond_241202.tex	testNumComp_BHPS_UKHLS_DDCM_241202.R	142-143	
19	Table B.3	M_param_est_250410.xlsx	functionBHPS_UKHLS_DCM_250410.R	872-888	called by estimateBHPS_UKHLS_DCM_250410.R
20		M_param_SE_250410.xlsx	combineBHPS_UKHLS_DCM_250410.R	113	M_param_est_250410.xlsx shows estimates and M_param_SE_250410.xlsx shows SE estimates
21	Figure D.2	BusNoProxyV4Final_label.png	plotMonteCarloBusV4Final.R	1165	call estimate files prepared by "MonteCarloBusNoProxyV4Final.R"
22	Figure D.3	BusThreeProxyEveryYearV4Final_label.png	plotMonteCarloBusV4Final.R	367	call estimate files prepared by "MonteCarloBusThreeProxyEveryYearV4Final.R"
23	Figure D.4	BusRotatingModuleV4Final_label.png	plotMonteCarloBusV4Final.R	546	call estimate files prepared by

					"MonteCarloBusRotatingModuleV4Final.R"
24	Figure D.5	BusTwoProxyEveryYearV4Final_label.png	plotMonteCarloBusV4Final.R	722	call estimate files prepared by "MonteCarloBusTwoProxyEveryYearV4Final.R"
25	Figure D.6	BusOneProxyEveryYearV4Final_label.png	plotMonteCarloBusV4Final.R	899	call estimate files prepared by "MonteCarloBusOneProxyEveryYearV4Final.R"
26	Figure D.7	BusThreeOneV4Final_label.png	plotMonteCarloBusV4Final.R	1078	call estimate files prepared by "MonteCarloBusThreeOneProxyV4Final.R"
27	Figure D.9	SegNoProxyV4Final_label.png	plotMonteCarloV4Final.R	965	call estimate files prepared by "MonteCarloNoProxyV4Final.R"
28	Figure D.10	SegThreeProxyEveryYearV4Final_label.png	plotMonteCarloV4Final.R	328	call estimate files prepared by "MonteCarloThreeProxyEveryYearV4Final.R"
29	Figure D.11	SegRotatingModuleV4Final_label.png	plotMonteCarloV4Final.R	470	call estimate files prepared by "MonteCarloRotatingModuleV4Final.R"
30	Figure D.12	SegTwoProxyEveryYearV4Final_label.png	plotMonteCarloV4Final.R	613	call estimate files prepared by "MonteCarloTwoProxyEveryYearV4Final.R"
31	Figure D.13	SegOneProxyEveryYearV4Final_label.png	plotMonteCarloV4Final.R	755	call estimate files prepared by "MonteCarloOneProxyEveryYearV4Final.R"
32	Figure D.14	SegThreeOneProxyEveryYearV4Final_label.png	plotMonteCarloV4Final.R	894	call estimate files prepared by "MonteCarloThreeOneProxyV4Final.R"
33	Figure D.15	BusOneProxyEveryYearNoMEV4Final_label.png	plotAppendixMonteCarloBusNoMEV4Final.R	254	call estimate files prepared by "MonteCarloBusOneProxyEveryYearNoMEV4Final.R"
34	Figure D.16	SegOneProxyEveryYearNoMEV4Final_label.png	plotAppendixMonteCarloNoMEV4Final.R	215	call estimate files prepared by "MonteCarloOneProxyEveryYearNoMEV4Final.R"

Reference

University of Essex, Institute for Social and Economic Research. (2022). Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 15th Edition. UK Data Service. SN:6614, DOI: <http://doi.org/10.5255/UKDA-SN-6614-16>

Wright, L. (2020). Producing working-life histories in the BHPS and UKHLS 2017-2020. [Data Collection]. <http://doi.org/10.5255/UKDA-SN-854327>

Contact Information

Dr. Yujung Hwang (yujungghwang@gmail.com, Johns Hopkins University)