Replication repository for "Model averaging and double machine learning"

Authors: Achim Ahrens, Christian B. Hansen, Mark E. Schaffer, Thomas Wiemann

Working paper: arxiv

Introductory example

Figures: Figure 1 (`example1.png`, `example2.png`)

Relevant code folder: `sim_SJ`

Simulation code:

1. `euler_SJ.sh` starts multiple batch jobs that run instances of the simulation script.
2. The simulation script is in `sim_SJ.do`. It generates simulated samples and performs the DDML estimations. Note that the data-generating process (DGPs) `dgp0==2` and `dgp0==5` are relevant for generating the above figures.
3. `output_for_applied.do` generates the output figures.

The `log` folder keeps log files of the simulations. The `out` folder contains multiple `dta` files with point estimates and standard errors per seed and per DGP. `output` contains the above output figures.

Calibrated simulation (Section 4.1+4.2)

Figures:

- Figure 5 (`sim_small_linear.png` and `sim_small_nonlinear.png`)

Tables:

- Table 1 (`sim_output_folds2.tex`)
- Table 2 (`sweights_nnls1_*.tex`)

Appendix Tables:

- Table A.1 (`mspe_*.tex`)
- Table A.2 (`sim_output_folds2_se.tex`)
- Table A.3-A.4 (`sweights_*.tex`)
- Table A.5 (`timing_long.tex`)
- Table B.3 (`sim_output_dgp0_*folds_withse.tex`)
- Table B.4 (`sim_output_dgp1_*folds_withse.tex`)
- Table B.5 (`sim_small_cover_wide_folds*.tex`)

Data: `Data/PVW_data.dta`

Relevant code folder: `sim_Advantages`

Simulation code:

1. `euler_large.sh` (for N=99150) and `euler_small.sh` (for N<=9915) start multiple batch jobs that run instances of the simulation script.
2. The simulation script is in `sim_Adv.do`.
3. `sim_Adv_output.do` creates all output figures with the exception of Table A.5, which is created in `sim_Adv_output_time.do`.

DDML and Stacking in Very Small Samples (Section 4.1)

Figures:

- Figure 4 (`pdsa_bbias.png`, `pdsb_bbias.png`)

Tables:

- Table 3 (`sim_WZ/bias_olspds_f10.tex`, `sim_WZ_linear/bias_ddml_*.tex`)
- Table 4 (`sim_WZ_linear/ssw_*folds10.tex`)

Appendix Tables:

- Table B.1 (`sim_WZ/fullsample_*.tex`)
- Table B.2 (`sim_WZ/ssw_*folds10.tex`)

Relevant code folders: `sim_WZ`, `sim_WZ_linear`

Data: `restatw.dat`, `data_spec1.mat`, `data_spec2.mat`. These files are taken from Wüthrich & Zhu (2023, ReStat) and combined into `data_401k_final.dta`; see `simWZ_prepare.R` and `simWZ_prepare.do`.

Code:

1. `sim_WZ_linear/euler_WZ.sh` (for DDML with linear candidate learners) and `sim_WZ/euler_WZ.sh` (for DDML with full set of candidate learners) start multiple batch jobs that run instances of the estimation scripts.
2. `sim_WZ_linear/simWZ.do` (for DDML with linear candidate learners) and `sim_WZ/simWZ.do` (for DDML with full set of candidate learners) draw bootstrap samples and perform the DDML estimations.
3. `sim_WZ_linear/sim_WZ_output.do` (for DDML with linear candidate learners) and `sim_WZ/sim_WZ_output.do` (for DDML with full set of candidate learners) create the output tables & figures.

Gender gap in citations (Section 5.1)

Relevant code folder: `scopus_cites`

Data: The data was kindly shared with us by Advani, Ash, Cai & Rasul (2021). Due to restrictions, we are not able to share the data. Please contact the authors for data access.

Tables and Figures:

- Table 5 (`scopus_cites/results_log70.png`)

- Figure 6 (`weights_log.tex`, `mspe_log_joined.tex`)

Appendix Figures and Tables:

- Figure C.1 (`results_log60.png`, `results_log90.png`)
- Table C.1 (`results_tab.tex`)

Code:

1. `pull_data.R` extracts sample from raw original data.
2. `data_processing.R` loads the raw data, and uses the Namsor API to predict gender from author names.
3. `save_BERT_features.R` extracts BERT embeddings.
4. `generate_dfm.R` creates word count matrices.
5. `data_prep.R` does further data processing and generates the predictor matrices.
6. `run_ddml.R` performs the DDML estimations.
7. Auxiliary files: `mdl_keras.R` (neural net learner), `pdslasso.R` (PDS-lasso), `rlasso2.R` (plugin lasso used by pdslasso), `ddml_auxiliary.R` (for creating outputs)
8. `Scopus_output.R` creates the output files (also estimates OLS + PDS-lasso).

For computation on a PBS or slurm-based computing cluster, we rely on the scriptflow makefile `sflow.py`. Replication using scriptflow requires adjusting the computing cluster's account details (here).

Gender gap in wages (Section 5.2)

Relevant code folder: `GWG`

Data: `gender_gap_ML_processed.dta`. The data file is prepared in `Preprocess/gender_gap_ML_v7_prep.do`.

Tables and Figures:

- Figure 7 (`GWG/all_estimates.png`)

Appendix Tables:

- Table D.1 (`Interactive_mse_weights.tex`)
- Table D.2 (`GWG/CLS_weights_onlypooled.tex`)
- Table D.3 (`Single-best_weights.tex`)
- Table D.4 (`regression_results_1.tex`)
- Table D.5 (`regression_results_2.tex`)

Code:

1. `run_euler.sh` starts multiple batch jobs that run instances of the estimation script.
2. `GWG.do` is the estimation script.
3. `Output.R` and `Output_weights_mse.R` create the output files.

Data sets

1. `PVW_data.dta`

Variables:

```
nifa            float   %9.0g                 Net non-401(k) financial assets
net_tfa         float   %9.0g                 Net total financial assets
tw              float   %9.0g                 Total wealth
age             byte    %9.0g                 Age of the head of the household
inc             float   %9.0g                 Household income
fsize           byte    %9.0g                 Household size
educ            byte    %9.0g                 Years of education of the head of the
household
db              byte    %9.0g                 Defined benefit pension status indicator
marr            byte    %9.0g                 Married indicator
twoearn         byte    %9.0g                 Two-earner status indicator
e401            byte    %9.0g                 401(k) eligibility
p401            byte    %9.0g                 401(k) participation
pira            byte    %9.0g                 IRA participation indicator
hown            byte    %9.0g                 House ownership indicator
```

Number of observations and summary statistics:

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        nifa |      9,915    13928.64    54904.88          0    1430298
     net_tfa |      9,915    18051.53     63522.5    -502302    1536798
          tw |      9,915    63816.85    111529.7    -502302    2029910
         age |      9,915    41.06021     10.3445         25         64
         inc |      9,915    37200.62    24774.29      -2652     242124
-------------+---------------------------------------------------------
       fsize |      9,915     2.86586    1.538937          1         13
        educ |      9,915    13.20625    2.810382          1         18
          db |      9,915    .2710035    .4445003          0          1
        marr |      9,915    .6048411    .4889094          0          1
     twoearn |      9,915    .3808371    .4856171          0          1
-------------+---------------------------------------------------------
        e401 |      9,915    .3713565    .4831919          0          1
        p401 |      9,915    .2616238     .439541          0          1
        pira |      9,915    .2421583    .4284112          0          1
        hown |      9,915    .6351992    .4813985          0          1
```

2. `restatw.dta`

Variables: The main variables of interest are the same as in `PVW_data.dta`. For a detailed description:

> Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Supplement to "Program evaluation and causal inference with high-dimensional data". Econometrica Supplemental Materials. URL: https://www.econometricsociety.org/publications/econometrica/2017/01/01/program-evaluation and-causal-inference-high-dimensional-data (last accessed 09/06/2021)

Number of observations and summary statistics:

Variable	Obs	Mean	Std. dev.	Min	Max
wgt	9,915	4603.056	2666.726	.1294955	36424.79
ira	9,915	3462.872	9648.028	0	100000
a401	9,915	3990.459	12825.84	0	153000
hval	9,915	63595.87	73679.37	0	300000
hmort	9,915	30022.51	40056.88	0	150000
hequity	9,915	33573.36	51830.77	-40000	300000
nifa	9,915	13928.64	54904.88	0	1430298
net_nifa	9,915	10414.15	56028.86	-502302	1430298
tfa	9,915	21566.03	62565.04	0	1536798
net_tfa	9,915	18051.53	63522.5	-502302	1536798
tfa_he	9,915	51624.9	93253.39	-502302	1687115
tw	9,915	63816.85	111529.7	-502302	2029910
age	9,915	41.06021	10.3445	25	64
inc	9,915	37200.62	24774.29	-2652	242124
fsize	9,915	2.86586	1.538937	1	13
educ	9,915	13.20625	2.810382	1	18
db	9,915	.2710035	.4445003	0	1
marr	9,915	.6048411	.4889094	0	1
male	9,915	.2059506	.4044149	0	1
twoearn	9,915	.3808371	.4856171	0	1
dum91	9,915	1	0	1	1
e401	9,915	.3713565	.4831919	0	1
i1	9,915	.0641452	.2450238	0	1
i2	9,915	.19647	.397348	0	1
i3	9,915	.209178	.4067422	0	1
i4	9,915	.1726677	.3779788	0	1
i5	9,915	.1214322	.3266453	0	1
i6	9,915	.1585477	.3652722	0	1
i7	9,915	.0773575	.2671714	0	1
p401	9,915	.2616238	.439541	0	1
pira	9,915	.2421583	.4284112	0	1
hown	9,915	.6351992	.4813985	0	1
a1	9,915	.1440242	.351132	0	1
a2	9,915	.2096823	.4071024	0	1

```
          a3 |      9,915    .2974281    .4571496          0          1
-------------+---------------------------------------------------------
          a4 |      9,915    .2154312    .4111419          0          1
          a5 |      9,915    .1334342    .3400605          0          1
        nohs |      9,915    .1282905    .3344298          0          1
          hs |      9,915    .3768028    .4846093          0          1
       smcol |      9,915    .2444781     .429799          0          1
-------------+---------------------------------------------------------
         col |      9,915    .2504286    .4332817          0          1
           a |      9,915    2.984569    1.238896          1          5
        icat |      9,915    3.875441    1.735169          1          7
        ecat |      9,915    2.617045    .9969147          1          4
          f1 |      9,915    .2269289    .4188674          0          1
-------------+---------------------------------------------------------
          f2 |      9,915    .2346949    .4238294          0          1
          f3 |      9,915    .1954614    .3965755          0          1
          f4 |      9,915    .2083712    .4061641          0          1
          f5 |      9,915    .1170953      .32155          0          1
          f6 |      9,915    .0174483    .1309412          0          1
-------------+---------------------------------------------------------
           f |      9,915    2.806354    1.392442          1          6
        zhat |      9,915    .3713663    .1941871   .0278493   .7856908
      wntfa1 |      9,915    .3713716    .3865303          0          1
      wntfa2 |      9,915    .3713713    .3867707          0          1
      wntfa3 |      9,915     .371386    .3871684          0          1
-------------+---------------------------------------------------------
      wntfa4 |      9,915     .371388    .3876389          0          1
      wntfa5 |      9,915    .3713817    .3879718          0          1
      wntfa6 |      9,915    .3713582    .3881998          0          1
      wntfa7 |      9,915    .3707906    .3891362          0          1
  wnet_nifa1 |      9,915    .3713681    .3865905          0          1
-------------+---------------------------------------------------------
  wnet_nifa2 |      9,915    .3713804    .3868005          0          1
  wnet_nifa3 |      9,915    .3713957    .3871623          0          1
  wnet_nifa4 |      9,915    .3713865    .3876176          0          1
  wnet_nifa5 |      9,915    .3713763    .3878892          0          1
  wnet_nifa6 |      9,915     .371365    .3881237          0          1
-------------+---------------------------------------------------------
  wnet_nifa7 |      9,915     .371413    .3888467          0          1
        wtw1 |      9,915    .3713698    .3864794          0          1
        wtw2 |      9,915    .3713698    .3867207          0          1
        wtw3 |      9,915    .3713432     .387168          0          1
        wtw4 |      9,915    .3713428    .3873409          0          1
-------------+---------------------------------------------------------
        wtw5 |      9,915    .3713434    .3879759          0          1
        wtw6 |      9,915    .3713618    .3884121          0          1
        wtw7 |      9,915    .3713464    .3887333          0          1
    net_n401 |      9,915    13877.02    59604.84    -502302    1467798
  wnet_n4014 |      9,915     .371391    .3876098          0          1
```

3. `gender_gap_ML_processed.dta`

Variables:

```
Name: age_r
Description: Person resolved age from BQ and QC check (derived)

Name: gender_r
Description: Person resolved gender from BQ and QC check (derived)

Name: b_q01a
Description: Education - Highest qualification - Level

Name: b_q01b
Description: Education - Highest qualification - Area of study

Name: d_q06c
Description: Current work - Part of a larger organisation

Name: d_q08a
Description: Current work - Managing other employees

Name: d_q09
Description: Current work - Type of contract

Name: d_q10
Description: Current work - Hours/week

Name: d_q10_t1
Description: Hours per week at this job or business - range of hours (Trend-IALS/ALL)

Name: d_q14
Description: Current work - Job satisfaction

Name: i_q04b
Description: About yourself - Learning strategies - Relate new ideas into real life

Name: i_q04d
Description: About yourself - Learning strategies - Like learning new things

Name: i_q04h
Description: About yourself - Learning strategies - Attribute something new

Name: i_q04j
Description: About yourself - Learning strategies - Get to the bottom of difficult
things

Name: i_q04l
```

Description: About yourself - Learning strategies - Figure out how different ideas fit togeth

Name: i_q04m
Description: About yourself - Learning strategies - Looking for additional info

Name: i_q05f
Description: About yourself - Cultural engagement - Voluntary work for non-profit organisatio

Name: i_q06a
Description: About yourself - Political efficacy - No influence on the government

Name: i_q07a
Description: About yourself - Social trust - Trust only few people

Name: i_q07b
Description: About yourself - Social trust - Other people take advantage of you

Name: i_q08
Description: About yourself - Health - State

Name: j_q02a
Description: Background - Living with spouse or partner

Name: j_q03b
Description: Background - Number of children

Name: j_q03d1
Description: Background - Age of the youngest child

Name: j_q03d1_c
Description: Background - Age of the youngest child (categorised, 4 categories)

Name: j_q04c1_c
Description: Background - Age of immigration (categorised, 9 categories)

Name: j_q06b
Description: Background - Mother/female guardian - Highest level of education

Name: j_q07b
Description: Background - Father/male guardian - Highest level of education

Name: yrsqual
Description: Highest level of education obtained imputed into years of education (derived)

Name: pared
Description: Highest of mother or father’s level of education (derived)

```
Name: impar
Description: Parents’ immigration status (derived)

Name: imgen
Description: First and second generation immigrants (derived)

Name: leavedu
Description: Respondent’s age when leaving formal education (derived)

Name: nfehrs
Description: Number of hours of participation in non-formal education (derived)

Name: pvlit1
Description: Literacy scale score - Plausible value 1

Name: pvnum1
Description: Numeracy scale score - Plausible value 1

Name: nfe12jr
Description: Participated in non-formal education for job-related reasons in 12 months
preced

Name: nfe12njr
Description: Participated in non-formal education for non job-related reasons in 12
months pr

Name: lnearn
Description: Natural logarithm of hourly earnings

Name: new_reg_tl2
Description: Geographical region - Respondent (OECD TL2) (coded)

Name: new_isic1c
Description: Industry classification of respondents job at 1-digit level (ISIC rev 4),
curre

Name: new_isco1c
Description: Occupational classification of respondents job at 1-digit level (ISCO
2008), cu
```

Descriptives:

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       age_r |      4,889    39.77153    12.15181         16         65
    gender_r |      4,889    .5825322    .4931918          0          1
      b_q01a |      4,889    9.131724     5.33235          1         16
      b_q01b |      4,889    3.593168    2.273892          0          9
```

```
      d_q06c |      4,876      1.3226    .4675192          1          2
-------------+---------------------------------------------------------
      d_q08a |      4,889    1.638781     .480403          1          2
       d_q09 |      4,877    1.466475    1.114982          1          6
       d_q10 |      4,888    34.73957    12.91088          1        125
    d_q10_t1 |      4,888    2.059124    .6700212          1          6
       d_q14 |      4,888    2.022095    .9091777          1          5
-------------+---------------------------------------------------------
      i_q04b |      4,873    3.160887    .9574126          1          5
      i_q04d |      4,889    3.902434    .8845442          1          5
      i_q04h |      4,879    3.614675    .8720513          1          5
      i_q04j |      4,889    3.873185    .9090577          1          5
      i_q04l |      4,884    3.604832    .9500451          1          5
-------------+---------------------------------------------------------
      i_q04m |      4,888    3.994476    .8853734          1          5
      i_q05f |      4,889    1.611577    1.017179          1          5
      i_q06a |      4,875    2.648615     1.18473          1          5
      i_q07a |      4,885    2.265711    1.117311          1          5
      i_q07b |      4,883    2.195576    1.006556          1          5
-------------+---------------------------------------------------------
       i_q08 |      4,887    2.195416    .9673738          1          5
      j_q02a |      4,889    1.086725    .6686253          0          2
      j_q03b |      4,889    1.329924    1.353515          0         11
     j_q03d1 |      4,889    7.633054    10.64093         -1         48
   j_q03d1_c |      4,889    1.352424    1.686556          0          4
-------------+---------------------------------------------------------
   j_q04c1_c |      4,889    .5125793    1.660992          0          9
      j_q06b |      4,889     1.51442    .8341663          0          3
      j_q07b |      4,889    1.603395    .8784546          0          3
     yrsqual |      4,889    12.38085     4.35737         -1         16
       pared |      4,889     1.51892    1.237416         -1          3
-------------+---------------------------------------------------------
       impar |      4,868     2.69166     .668985          1          3
       imgen |      4,889     2.57026    .9252013          0          3
     leavedu |      4,889    19.47044    13.16482         -1         63
      nfehrs |      4,889     60.4821    183.9337         -1       1920
      pvlit1 |      4,889    280.8468     44.3119   111.9843   419.1831
-------------+---------------------------------------------------------
      pvnum1 |      4,889     270.921      49.283   50.56118   445.1982
     nfe12jr |      4,889    .5755778    .4943055          0          1
    nfe12njr |      4,889    .0490898    .2160776          0          1
      lnearn |      4,889    2.768354     .579042  -4.871062   6.785444
 new_reg_tl2 |      4,887    7.732965    3.381178          1         11
-------------+---------------------------------------------------------
  new_isic1c |      4,889    12.08652    5.480295          1         22
  new_isco1c |      4,889    5.318265    2.441491          1         11
      tenure |      4,887    8.437078    8.868766          0         53
 immig_years |      4,889     .696666    6.684973         -1         58
  yrsqual_na |      4,889     .072612    .2595251          0          1
```

```
-------------+--------------------------------------------------------
  leavedu_na |      4,889    .2082225    .4060784          0          1
immig_year~a |      4,889     .894866    .3067573          0          1
  j_q03d1_na |      4,889    .3800368    .4854452          0          1
    pared_na |      4,889    .1444058    .3515367          0          1
   nfehrs_na |      4,889    .3771732    .4847284          0          1
```

References

Kaspar Wüthrich, Ying Zhu; Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis. The Review of Economics and Statistics 2023; 105 (4): 982–997. https://doi.org/10.1162/rest_a_01128

Advani, A., Ash, E., Cai, D., & Rasul, I. (2021). Race-related research in economics and other social sciences. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3846227