Applications of the AIPW Estimator in Causal Inferences

Di Liu

StataCorp

Table of Contents

Evolution of the AIPW estimator

We will talk about the **AIPW**-style estimator (Robins and Rotnitzky 1995) in causal inferences.

- Estimating **ATE** and **ATET** for cross-sectional data:
  - Low-dimensional/parametric settings (Robins and Rotnitzky 1995)
  - High-dimensional/semiparametric settings (Farrell 2015 and Chernozhukov et al. 2018)
- Difference-in-differences for panel data:
  - Homogeneous **ATET** (Sant'Anna and Zhao 2020)
  - Heterogeneous **ATET** (Callaway and Sant'Anna 2021)
- Heterogeneous treatment effects (Semenova and Chernozhukov 2021, Knaus 2022, and Kennedy 2023)

The **AIPW** estimators in Stata

- Estimating **ATE** and **ATET** for cross-sectional data:
  - Low-dimensional/parametric settings (`teffects aipw`)
  - High-dimensional/semiparametric settings (`telasso`)
- Difference-in-differences for panel data:
  - Homogeneous **ATET** (user-written `drdid`)
  - Heterogeneous **ATET** (`xthdidregress` and `hdidregress`)
- Heterogeneous treatment effects (I will show some examples)

Table of Contents

Example: 401(k) eligibility effects

We want to know the average treatment effects (ATE) of the 401(k) eligibility on the personal net financial assets (Chernozhukov et al. 2018):

$$\textbf{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

where

- Treatment is 401(k) eligibility status
- Outcome is the personal net financial assets
- $Y(1) \equiv$ potential outcome if being eligible for 401(k)
- $Y(0) \equiv$ potential outcome if being not eligible for 401(k)

Fundamental missing data problem: only one of $Y(1)$ or $Y(0)$ is observed for each individual.

Key assumptions to identify the ATE

- **Conditional independence**: Conditional on a set of control variables, the potential outcomes are independent of the treatment assignment.
  $\Longrightarrow$ We can use the observed outcome in the treated group as a proxy to estimate the treated potential outcome in the control group, and vice versa.
  $\Longrightarrow$ Use $E[Y|treat = 1, X]$ to estimate $E[Y(1)|treat = 0, X]$
- **Overlap**: There is always a positive probability that any given unit is treated or untreated.
  $\Longrightarrow$ We can always find similar units (same value of X) in both treated and control groups.
- **I.I.D**: identically independent distributed observations.
  $\Longrightarrow$ Unit i does not interfere with unit j ($\forall i \neq j$)

The model in a potential-outcome framework

The model is

$$y = g(\tau, \mathbf{x}) + u, \quad \mathbb{E}[u|\mathbf{x}, \tau] = 0$$
$$\tau = m(\mathbf{x}) + v, \quad \mathbb{E}[v|\mathbf{x}, \tau] = 0$$

where

- y is the observed outcome
- τ is the treatment status (1 treated, 0 untreated)
- $g(1, \mathbf{x}) \equiv \mathbb{E}[Y(1)|\mathbf{x}]$ and $g(0, \mathbf{x}) \equiv \mathbb{E}[Y(0)|\mathbf{x}]$
- $m(\mathbf{x}) \equiv \Pr[\tau = 1|\mathbf{x}]$ (propensity score)

$$\textbf{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1)|\mathbf{x}] - \mathbb{E}[Y(0)|\mathbf{x}]] = \mathbb{E}[g(1, \mathbf{x}) - g(0, \mathbf{x})]$$

The AIPW (Robins and Rotnitzky 1995) estimator

$$\textbf{ATE} = \mathbb{E}\left[Y(1, \mathbf{x})_{AIPW} - Y(0, \mathbf{x})_{AIPW}\right]$$

where

$$Y(1, \mathbf{x})_{AIPW} = g(1, \mathbf{x}) + \frac{\tau(y - g(1, \mathbf{x}))}{m(\mathbf{x})}$$

$$Y(0, \mathbf{x})_{AIPW} = g(0, \mathbf{x}) + \frac{(1 - \tau)(y - g(0, \mathbf{x}))}{1 - m(\mathbf{x})}$$

Notice that

$$\textbf{ATE} = \mathbb{E}[g(1, \mathbf{x}) - g(0, \mathbf{x})]$$

The red terms are Agumented terms using the Inverse of Probability Weighting; thus **AIPW** was born.

Example: 401(k) eligibility

```
. webuse assets
(Excerpt from Chernozhukov and Hansen (2004))

. describe

Contains data from https://www.stata-press.com/data/r18/assets.dta
 Observations:         9,913                  Excerpt from Chernozhukov and
                                                Hansen (2004)
    Variables:            10                  15 Jun 2022 14:15
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
assets          float   %9.0g                 Net total financial assets
age             byte    %9.0g                 Age
income          float   %9.0g                 Household income
educ            byte    %9.0g                 Years of education
pension         byte    %16.0g     lbpen      Pension benefits
married         byte    %11.0g     lbmar      Marital status
twoearn         byte    %9.0g      lbyes      Two-earner household
e401k           byte    %12.0g     lbe401     401(k) eligibility
ira             byte    %9.0g      lbyes      IRA participation
ownhome         byte    %9.0g      lbyes      Homeowner
-------------------------------------------------------------------------------
Sorted by: e401k
```

Outcome: `assets`   Treatment: `e401k`

teffects aipw

```
. egen incomecat = cut(income), group(5)
. global controls educ age i.(pension married twoearn ira ownhome  incomecat)
. teffects aipw (assets $controls) (e401k $controls)
Iteration 0:   EE criterion =  2.445e-21
Iteration 1:   EE criterion =  1.154e-23

Treatment-effects estimation                    Number of obs     =      9,913
Estimator      : augmented IPW
Outcome model  : linear by ML
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
      assets | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATE          |
       e401k |
   (Eligible |
          vs |
 Not elig..) |   8019.463   1152.038     6.96   0.000      5761.51    10277.42
-------------+----------------------------------------------------------------
POmean       |
       e401k |
 Not eligi.. |   13930.46    817.613    17.04   0.000     12327.97    15532.96
------------------------------------------------------------------------------
```

The double robustness

The **AIPW** estimator is **doubly robust**: only one of the treatment or outcome model needs to be correctly specified for consistent estimation of **ATE**.

Suppose that only the treatment model is correctly specified. Let $\widehat{g}(\tau, \mathbf{x})$ be an incorrect outcome model.

$$\mathbb{E}[Y(1,\mathbf{x})_{AIPW}|\mathbf{x}] = \widehat{g}(1,\mathbf{x}) + \mathbb{E}\left[\frac{\tau(y-\widehat{g}(1,\mathbf{x}))}{m(\mathbf{x})}\Big|\mathbf{x}\right]$$

Then $\mathbb{E}\left[\frac{\tau(y-\widehat{g}(1,\mathbf{x}))}{m(\mathbf{x})}\Big|\mathbf{x}\right]$ is

$$\Pr[\tau=1|\mathbf{x}] * \mathbb{E}\left[\frac{y-\widehat{g}(1,\mathbf{x})}{m(\mathbf{x})}\Big|\mathbf{x},\tau=1\right] + \Pr[\tau=0|\mathbf{x}] * 0$$

$$= m(\mathbf{x})\,\mathbb{E}\left[\frac{y-\widehat{g}(1,\mathbf{x})}{m(\mathbf{x})}\Big|\mathbf{x},\tau=1\right] = \mathbb{E}[y|\mathbf{x},\tau=1] - \widehat{g}(1,\mathbf{x})$$

The double robustness (continued)

$$
\begin{aligned}
\mathbb{E}[Y(1,\mathbf{x})_{AIPW}|\mathbf{x}] &= \widehat{g}(1,\mathbf{x}) + \mathbb{E}[y|\mathbf{x},\tau=1] - \widehat{g}(1,\mathbf{x}) \\
&= \mathbb{E}[y|\mathbf{x},\tau=1] \\
&= \mathbb{E}[Y(1)|\mathbf{x},\tau=1] \\
&= \mathbb{E}[Y(1)|\mathbf{x}]
\end{aligned}
$$

where the last equality comes from the assumption of conditional independence. Similarly, $\mathbb{E}[Y(0,\mathbf{x})_{AIPW}|\mathbf{x}] = \mathbb{E}[Y(0)|\mathbf{x}]$.
Thus,

$$
\mathbb{E}[Y(1,\mathbf{x})_{AIPW} - Y(1,\mathbf{x})_{AIPW}] = \mathbb{E}[\mathbb{E}[Y(1) - Y(0)|\mathbf{x}]] = \mathbb{E}[Y(1) - Y(0)]
$$

even if the outcome model is incorrectly specified.

Table of Contents

More vs. fewer variables

We want to estimate the treatment effects of 401(k) eligibility on financial assets, but we have the following dilemma:

- On the one hand, we think a simple specification may not be adequate to control for the related confounders. So we need more variables or flexible models.

  $\Longrightarrow$

  - Adding interactions among variables as controls.
  - Generating B-splines of continuous variables as controls.
  - There are many raw variables.

- On the other hand, flexible models decrease the power to learn about the treatment effects. So we need fewer variables or simple models. $\Longrightarrow$ The model may not converge!

Set controls

```
. //---- orthogonal polynomial ----//
.
. orthpoly age, degree(6) generate(_orth_age*)
. orthpoly income, degree(8) generate(_orth_inc*)
. orthpoly educ, degree(4) generate(_orth_educ*)
.
. //---- define controls ---------//
.
. global cvars _orth*
. global fvars pension married twoearn ira ownhome
. global controls2 $cvars i.($fvars) c.($cvars)#i.($fvars) ///
>          i.($fvars)#i.($fvars)
```

There are 248 controls and 9913 observations.

Include all the controls?

```
. cap noi teffects aipw (assets $controls2) (e401k $controls2)
treatment model has 5 observations completely determined; the model, as
specified, is not identified
```

- Including too many controls will violate the overlap assumption!
- In practice, to avoid conflicts, researchers usually do some sort of model selection, but they conduct inference as if there is no model selection or assuming the selected model is correct!
  - It is mostly dangerous! Very! (Leeb and Pötscher 2005, 2008)

Conflits between the C.I. and overlap assumptions

- **Conditional independence**: $\mathbb{E}(y(\tau)|\mathbf{x}, \tau) = \mathbb{E}(y(\tau)|\mathbf{x})$. Dependent on a set of control variables, the potential outcome is independent of the treatment assignment.
- **Overlap**: $m_0(\mathbf{z}) > 0$. There is always a positive probability that any given unit is treated or untreated.

Conflicts

- The more covariates we have, the easier the CI assumption is satisfied.
- Certain specific values of covariates may not be observed in some treatment groups, which means the violation of the overlap assumption.

Honestly solve the conflicts

- We need to select variables that matter to outcome and treatment. We only need some of them!

- The inference should be robust to model-selection mistakes. We admit that we made the model selection and that we may select the wrong variables. $\Longrightarrow$ **Neyman orthognality**
  A Neyman orthogonal moment condition is defined as

$$\mathbb{E}\left[\psi(W;\theta_0,\eta_0)\right] = 0$$
$$D_0[\eta-\eta_0] = 0$$

  where

$$D_r[\eta-\eta_0] = \partial_r \left\{\mathbb{E}\left[\psi(W;\theta_0,\eta_0+(\eta-\eta_0)r)\right]\right\}$$

  for all $r \in [0,1)$. When D_r is evaluated at $r = 0$, we denote it as $D_0[\eta-\eta_0]$

Treatment effects + lassos

$$\textbf{ATE} = \mathbb{E}\left[Y(1,\mathbf{x})_{AIPW} - Y(0,\mathbf{x})_{AIPW}\right]$$

where

$$Y(1,\mathbf{x})_{AIPW} = g(1,\mathbf{x}) + \frac{\tau(y - g(1,\mathbf{x}))}{m(\mathbf{x})}$$

$$Y(0,\mathbf{x})_{AIPW} = g(0,\mathbf{x}) + \frac{(1-\tau)(y - g(0,\mathbf{x}))}{1 - m(\mathbf{x})}$$

- We use lasso-type techniques to predict $g(1,\mathbf{x})$, $g(0,\mathbf{x})$, and $m(\mathbf{x})$.
- It is just a version of `teffects aipw` with lassos.
- It is doubly robust, i.e., either the outcome or treatment model can be misspecified.
- It is Neyman orthogonal; it is robust to model-selection mistakes (Not RA or IPW estimators).

telasso

```
. telasso (assets $controls2) (e401k $controls2)

Estimating lasso for outcome assets if e401k = 0 using plugin method ...
Estimating lasso for outcome assets if e401k = 1 using plugin method ...
Estimating lasso for treatment e401k using plugin method ...
Estimating ATE ...

Treatment-effects lasso estimation    Number of observations      =      9,913
Outcome model:   linear               Number of controls          =        248
Treatment model: logit                Number of selected controls =         29

------------------------------------------------------------------------------
             |               Robust
      assets | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATE          |
       e401k |
   (Eligible |
          vs |
 Not elig..) |   8408.417   1259.405     6.68   0.000     5940.029    10876.81
-------------+----------------------------------------------------------------
POmean       |
       e401k |
 Not eligi.. |   13958.04   874.6395    15.96   0.000     12243.78    15672.31
------------------------------------------------------------------------------
```

On average, being eligible for a 401(k) will increase financial assets by $8408.

Table of Contents

Double machine learning

Double machine learning means cross-fitting + resampling.

Why do we need it?

- Cross-fitting relaxes the requirements in the sparsity assumption.
  - **Without cross-fitting**, the sparsity assumption requires

    $$s_g^2 + s_m^2 \ll N$$

    where s_g and s_m are the number of actual terms in the outcome and treatment models, respectively.
  - **With cross-fitting**, the sparsity assumption requires

    $$s_g * s_m \ll N$$

- Resampling reduces the randomness in cross-fitting.

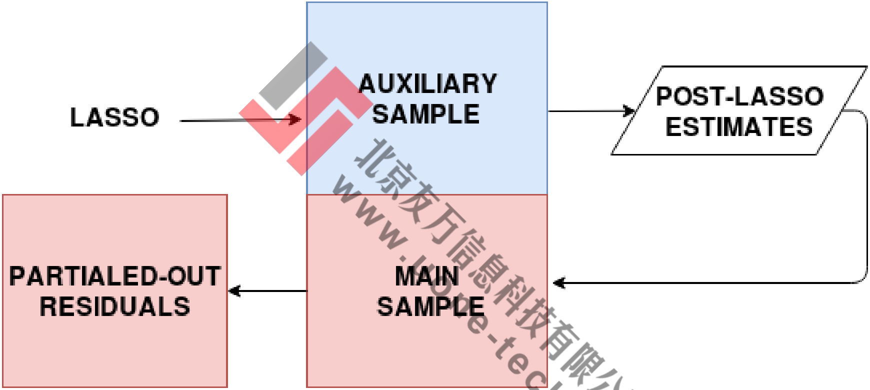
Basic idea of double machine learning

$$ATE = \mathbb{E}\left(g(1,\mathbf{x}) + \frac{\tau\left(y - g(1,\mathbf{x})\right)}{m(\mathbf{z})}\right) - \mathbb{E}\left(g(0,\mathbf{x}) + \frac{(1-\tau)\left(y - g(0,\mathbf{x})\right)}{1 - m(\mathbf{z})}\right)$$

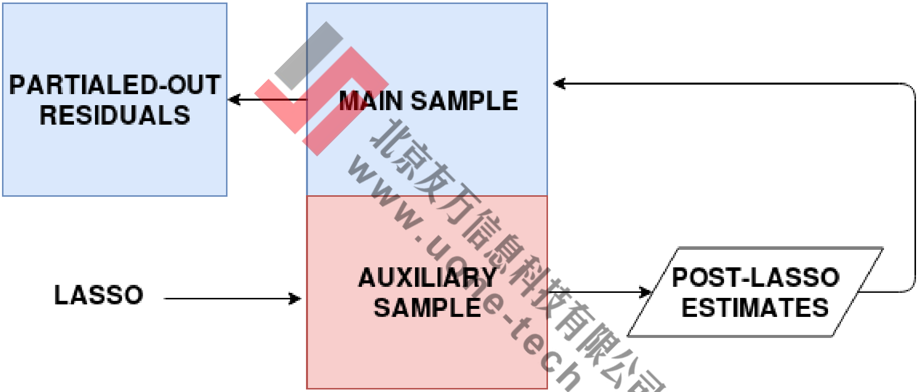
Basic idea

1. Split sample into auxiliary part and main part;
2. All the machine-learning techniques are applied to the auxiliary sample;
3. All the post-lasso residuals are obtained from the main sample;
4. **Switch the role of auxiliary sample and main sample**, and do steps 2 and 3 again;
5. Solve the moment equation using the full sample.

2-fold cross-fitting (I)



2-fold cross-fitting (II)



Cross-fitting

```
. telasso (assets $controls2) (e401k $controls2), xfolds(5) rseed(123)
Cross-fit fold 1 of 5 ...
Estimating lasso for outcome assets if e401k = 0 using plugin method ...
Estimating lasso for outcome assets if e401k = 1 using plugin method ...
Estimating lasso for treatment e401k using plugin method ...
(... output omitted ...)
Treatment-effects lasso estimation    Number of observations         =      9,913
                                      Number of controls             =        248
                                      Number of selected controls    =         43
Outcome model:   linear               Number of folds in cross-fit   =          5
Treatment model: logit                Number of resamples            =          1
```

assets	Coefficient	Robust std. err.	z	P>|z|	[95% conf. interval]	
ATE						
e401k (Eligible vs Not elig..)	8244.876	1521.009	5.42	0.000	5263.754	11226
POmean						
e401k Not eligi..	14271.34	921.0897	15.49	0.000	12466.03	16076.64

Cross-fitting + resampling

```
. telasso (assets $controls2) (e401k $controls2), xfolds(5) resample(3) rseed(1
> 23)

Resample 1 of 3 ...
Cross-fit fold 1 of 5 ...
Estimating lasso for outcome assets if e401k = 0 using plugin method ...
Estimating lasso for outcome assets if e401k = 1 using plugin method ...
Estimating lasso for treatment e401k using plugin method ...
(... output omitted ...)

Treatment-effects lasso estimation    Number of observations       =     9,913
                                      Number of controls           =       248
                                      Number of selected controls  =        47
Outcome model:   linear               Number of folds in cross-fit =         5
Treatment model: logit                Number of resamples          =         3

------------------------------------------------------------------------------
             |               Robust
      assets | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATE          |
       e401k |
   (Eligible |
          vs |
 Not elig..) |    8132.74   1434.918     5.67   0.000     5320.353    10945.13
-------------+----------------------------------------------------------------
POmean       |
       e401k |
 Not eligi.. |   14175.17   907.9799    15.61   0.000     12395.56    15954.78
------------------------------------------------------------------------------
```

Table of Contents

Heterogeneous treatment effects

- So far, we focus on measuring the **ATE**, but a single mean is not good enough to summarize the treatment effects.
- We want to understand the driving mechanism underlying the treatment effects. $\Longrightarrow$ Who is benefitting more or less?

For example, we want to know how the treatment effects of 401(k) eligibility vary with education or income categories.

Another look at the AIPW estimator

$$\Gamma(\mathbf{x}) \equiv Y(1, \mathbf{x})_{AIPW} - Y(0, \mathbf{x})_{AIPW} = \mathbb{E}[\textit{treatment effects}|\mathbf{x}]$$

$$\textbf{ATE} = \mathbb{E}\left[\Gamma(\mathbf{x})\right]$$

$$\textbf{ATET} = \mathbb{E}\left[\Gamma(\mathbf{x})\middle|\tau = 1\right]$$

Then, the ATE over the subgroups $G = g$ is just

$$\mathbb{E}\left[\Gamma(\mathbf{x})\middle|G = g\right]$$

Similarly, the ATE over a specific value of continuous variable $Z = z$ is

$$\mathbb{E}\left[\Gamma(\mathbf{x})\middle|Z = z\right]$$

Estimating strategies

Group ATE

$$\mathbb{E}\left[\Gamma(\mathbf{x})\middle|G=g\right]$$

1. We already have an estimate of $\Gamma(\mathbf{x})$ after `teffects aipw` or `telasso` $\Longrightarrow$ use `predict ..., te` to construct $\Gamma(\mathbf{x})$.
2. Run `regress` $\Gamma(\mathbf{x})$ `i.G`

ATE over a continuous variable

$$\mathbb{E}\left[\Gamma(\mathbf{x})\middle|Z=z\right]$$

1. Run `npregress series` $\Gamma(\mathbf{x})$ `Z`.

See discussions in Semenova and Chernozhukov (2021), Knaus (2022), and Kennedy (2023).

Example: Treatment effects for each income group

```
. // ---- fit model ----//
. qui teffects aipw (assets $controls) (e401k $controls)
.
. // ---- predict treatment effects ---- //
. predict myte, te
.
. // ---- income group ---- //
. table incomecat, stat(min income) stat(max income)      ///
>         stat(median income) nototal
```

	Minimum value	Maximum value	Median
incomecat			
0	0	17196	12240
1	17214	26523	21735
2	26526	37275	31482
3	37296	53841	44379
4	53844	242124	69612

Example: Treatment effects for each income group

```
. regress myte ibn.incomecat, noconstant
```

Source	SS	df	MS
Model	1.1208e+12	5	2.2416e+11
Residual	1.3020e+14	9,908	1.3141e+10
Total	1.3132e+14	9,913	1.3247e+10

Number of obs	=	9,913
F(5, 9908)	=	17.06
Prob > F	=	0.0000
R-squared	=	0.0085
Adj R-squared	=	0.0080
Root MSE	=	1.1e+05

myte	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
incomecat						
0	3748.291	2575.567	1.46	0.146	-1300.345	8796.927
1	1035.475	2573.619	0.40	0.687	-4009.343	6080.293
2	5509.986	2574.918	2.14	0.032	462.6239	10557.35
3	8749.087	2574.268	3.40	0.001	3702.997	13795.18
4	21052.43	2574.268	8.18	0.000	16006.34	26098.51

```
.
. test 4.incomecat = 3.incomecat = 2.incomecat, mtest(bonferroni)
 ( 1)  - 3.incomecat + 4.incomecat = 0
 ( 2)  - 2.incomecat + 4.incomecat = 0
```

	F(df,9908)	df	p > F
(1)	11.42	1	0.0015*
(2)	18.22	1	0.0000*
All	10.14	2	0.0000

* Bonferroni-adjusted *p*-values

Example: Treatment effects over education

```
. npregress series myte educ, knots(3)
warning: you have entered variable educ as continuous but it only has 18
        distinct values.  The estimates may differ substantially if you
        inadvertently include a discrete variable as continuous

Computing approximating function

Computing average derivatives
Cubic B-spline estimation                   Number of obs      =        9,913
                                            Number of knots    =            3
```

myte	Effect	Robust std. err.	z	P>|z|	[95% conf.	interval]
educ	2693.11	1388.459	1.94	0.052	-28.22017	5414.441

Note: Effect estimates are averages of derivatives.

The marginal effect of education (in years) on the 401(k) eligibility treatment effects is $415.

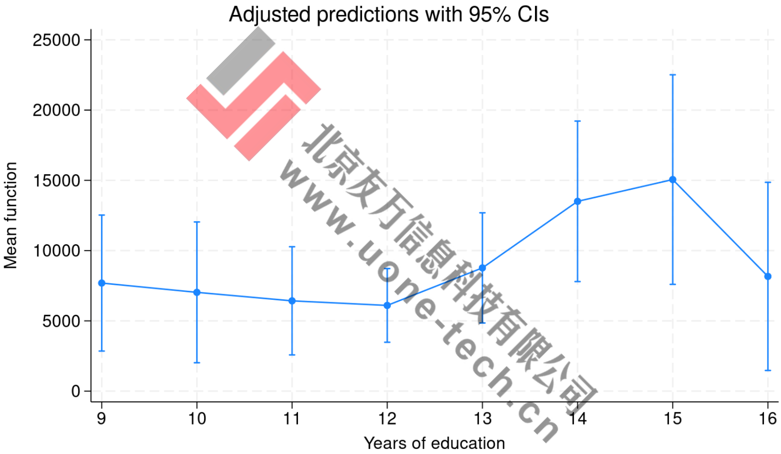
Example: Treatment effects over education

```
. margins, at(educ = (9(1)16))

Adjusted predictions                                   Number of obs = 9,913
Model VCE: Robust

Expression: Mean function, predict()
1._at: educ =  9
2._at: educ = 10
3._at: educ = 11
4._at: educ = 12
5._at: educ = 13
6._at: educ = 14
7._at: educ = 15
8._at: educ = 16
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	7691.175	2469.007	3.12	0.002	2852.011	12530.34
2	7029.716	2555.966	2.75	0.006	2020.115	12039.32
3	6426.178	1964.316	3.27	0.001	2576.19	10276.17
4	6100.159	1337.229	4.56	0.000	3479.238	8721.08
5	8770.296	2000.363	4.38	0.000	4849.656	12690.94
6	13506.69	2914.037	4.64	0.000	7795.283	19218.1
7	15056.14	3805.146	3.96	0.000	7598.191	22514.09
8	8165.443	3415.943	2.39	0.017	1470.317	14860.57

Example: Treatment effects over education

```
. marginsplot
Variables that uniquely identify margins: educ
```



Example: Linear projection of treatment effects

```
. regress myte educ age income i.(married ownhome twoearn)

      Source |       SS           df       MS      Number of obs   =     9,913
-------------+----------------------------------   F(6, 9906)      =      5.54
       Model |  4.3743e+11         6  7.2904e+10   Prob > F        =    0.0000
    Residual |  1.3025e+14     9,906  1.3148e+10   R-squared       =    0.0033
-------------+----------------------------------   Adj R-squared   =    0.0027
       Total |  1.3068e+14     9,912  1.3185e+10   Root MSE        =   1.1e+05

------------------------------------------------------------------------------
        myte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |  -160.0135   459.9507    -0.35   0.728     -1061.61    741.5834
         age |   257.0527   119.0901     2.16   0.031     23.61187    490.4934
      income |   .2175988   .0589338     3.69   0.000     .1020766    .3331211
             |
     married |
    Married  |   -3021.45   3203.746    -0.94   0.346    -9301.445    3258.545
             |
     ownhome |
        Yes  |   3750.313   2695.386     1.39   0.164    -1533.193    9033.818
             |
     twoearn |
        Yes  |   100.0405   3194.365     0.03   0.975    -6161.566    6361.647
       _cons |  -9110.624    8088.33    -1.13   0.260     -24965.4    6744.149
------------------------------------------------------------------------------
```

Summary

- AIPW estimator in the classical settings (`teffects aipw`).
- High-dimensional controls (`telasso`).
- Use AIPW scores to estimate the heterogeneous treatment effects. (Note: In the ideal case, we can construct the AIPW scores using cross-fitting. It would require some programming.)
- In the heterogeneous DID settings, AIPW also plays an important role. (See `xthdidregress` and `hdidregress` from last year's talk.)

References

Callaway, B., and P. H. Sant'Anna. 2021. Difference-in-Differences with multiple time periods. *Journal of Econometrics* 225: 200–230.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21: C1–C68.

Farrell, M. H. 2015. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189: 1–23.

Kennedy, E. H. 2023. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17: 3008–3049.

Knaus, M. C. 2022. Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal* 25: 602–627.

Leeb, H., and B. M. Pötscher. 2005. Model selection and inference: Facts and fiction. *Econometric Theory* 21(1): 21–59.

———. 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* 142(1): 201–211.

Robins, J. M., and A. Rotnitzky. 1995. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association* 90: 122–129.

Sant'Anna, P. H., and J. Zhao. 2020. Doubly robust difference-in-differences estimators. *Journal of Econometrics* 219: 101–122.

Semenova, V., and V. Chernozhukov. 2021. Debiased machine learning of conditional average treatment effects and other causal functions. *Econometrics Journal* 24: 264–289.

Table of Contents

Proofs for Neyman orthognality and double robustness of the AIPW ATE estimator

Di Liu
StataCorp

Contents

0.1 Proof for ATE score is Neyman orthogonal

We need to prove the moment condition is zero at true parameters, and also this moment condition is robust to machine learning mistakes.

Step 1: we need to prove $\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0$

Proof.

$$\begin{aligned}\mathbb{E}[\psi(W;\theta_0,\eta_0)] =& \mathbb{E}\left[(g_0(1,X)-g_0(0,X))\right] + \mathbb{E}\left[\frac{D\left(Y-g_0(1,X)\right)}{m_0(X)}\right] \\ &- \mathbb{E}\left[\frac{(1-D)\left(Y-g_0(0,X)\right)}{1-m_0(X)}\right] - \theta_0\end{aligned}$$

Where the second and third term are zero. The second term is

$$\begin{aligned}\mathbb{E}\left[\frac{D\left(Y-g_0(1,X)\right)}{m_0(X)}\right] =& Pr(D=0)*0 + Pr(D=1)\,\mathbb{E}\left[\frac{D\left(Y-g_0(1,X)\right)}{m_0(X)}\middle| D=1\right] \\ =& Pr(D=1)\,\mathbb{E}\left[\mathbb{E}\left(\frac{D\left(Y-g_0(1,X)\right)}{m_0(X)}\middle| D=1,X\right)\right] \\ =& Pr(D=1)\,\mathbb{E}\left[\frac{D}{m_0(X)}\,\mathbb{E}\left(Y-g_0(1,X)\middle| D=1,X\right)\right]\end{aligned}$$

Notice $\mathbb{E}\left(Y-g_0(1,X)\middle| D=1,X\right) = 0$, so $\mathbb{E}\left[\frac{D(Y-g_0(1,X))}{m_0(X)}\right] = 0$.

The third term is

$$
\begin{aligned}
\mathbb{E}\left[\frac{(1-D)\,(Y-g_0(0,X))}{1-m_0(X)}\right] =& Pr(D=0)\,\mathbb{E}\left[\frac{1\,(Y-g_0(0,X))}{1-m_0(X)}\middle| D=0\right] + Pr(D=1)*0 \\
=& Pr(D=0)\,\mathbb{E}\left[\mathbb{E}\left(\frac{1\,(Y-g_0(0,X))}{1-m_0(X)}\middle| D=0,X\right)\right] \\
=& \mathbb{E}\left[\frac{1}{1-m_0(X)}\,\mathbb{E}\left(Y-g_0(0,X)\middle| D=0,X\right)\right]
\end{aligned}
$$

Notice that $\mathbb{E}\left(Y-g_0(0,X)\middle| D=0,X\right)=0$, so $\mathbb{E}\left[\frac{(1-D)(Y-g_0(0,X))}{1-m_0(X)}\right]=0$.

By the definition of $\theta_0=\mathbb{E}\left[g_0(1,X)-g_0(0,X)\right]$, so $\mathbb{E}\left[\psi(W;\theta_0,\eta_0)\right]=0$. □

Step 2: we need to prove $D_0[\eta-\eta_0]=0$

Proof.

$$
\begin{aligned}
\mathbb{E}[\psi\,(W;\theta,\eta_0+(\eta-\eta_0)\gamma)] =& \mathbb{E}[(g_0(1,X)+(g(1,X)-g_0(1,X))\gamma)] \\
&-\mathbb{E}[(g_0(0,X)+(g(0,X)-g_0(0,X))\gamma)] \\
&+\mathbb{E}\left[\frac{D\,(Y-(g_0(1,X)+(g(1,X)-g_0(1,X))\gamma))}{(m_0(X)+(m(X)-m_0(X))\gamma)}\right] \\
&-\mathbb{E}\left[\frac{(1-D)\,(Y-(g_0(0,X)+(g(0,X)-g_0(0,X))\gamma))}{1-(m_0(X)+(m(X)-m_0(X))\gamma)}\right] \\
&-\theta
\end{aligned}
$$

Under some regularity conditions, the derivative and expectation operator are interchangeable. So $D_0[\eta-\eta_0]$ is

$$
\begin{aligned}
D_0[\eta-\eta_0] =& \partial_\gamma\left\{\mathbb{E}[\psi\,(W;\theta,\eta_0+(\eta-\eta_0)\gamma)]\right\}\Big|_{\gamma=0} \\
=& \mathbb{E}[(g(1,X)-g_0(1,X))]-E[(g(0,X)-g_0(0,X))] \\
&-\mathbb{E}\left[\frac{D(g(1,X)-g_0(1,X))}{m_0(X)}\right] \\
&-\mathbb{E}\left[\frac{D(Y-g_0(1,X))(m(X)-m_0(X))}{m_0(X)^2}\right] \\
&+\mathbb{E}\left[\frac{(1-D)(g(0,X)-g_0(0,X))}{1-m_0(X)}\right] \\
&-\mathbb{E}\left[\frac{(1-D)(Y-g_0(0,X))(m(X)-m_0(X))}{(1-m_0(X))^2}\right]
\end{aligned}
$$

Notice that

$$\begin{aligned}
\mathbb{E}\left[\frac{D(g(1,X)-g_0(1,X))}{m_0(X)}\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{D(g(1,X)-g_0(1,X))}{m_0(X)}\middle| X\right]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}(D|X)\frac{(g(1,X)-g_0(1,X))}{m_0(X)}\right\} \\
&= \mathbb{E}\left\{m_0(X)\frac{(g(1,X)-g_0(1,X))}{m_0(X)}\right\} \\
&= \mathbb{E}[(g(1,X)-g_0(1,X))]
\end{aligned}$$

similarly

$$\mathbb{E}\left[\frac{(1-D)(g(0,X)-g_0(0,X))}{1-m_0(X)}\right] = \mathbb{E}[(g(0,X)-g_0(0,X))]$$

Now

$$\begin{aligned}
&\mathbb{E}\left[\frac{D(Y-g_0(1,X))(m(X)-m_0(X))}{m_0(X)^2}\right] \\
&= Pr(D=0)*0 + Pr(D=1)\,\mathbb{E}\left[\frac{D(Y-g_0(1,X))(m(X)-m_0(X))}{m_0(X)^2}\middle| D=1\right] \\
&= Pr(D=1)\,\mathbb{E}\left\{\mathbb{E}\left[\frac{D(Y-g_0(1,X))(m(X)-m_0(X))}{m_0(X)^2}\middle| D=1,X\right]\right\} \\
&= Pr(D=1)\,\mathbb{E}\left\{\frac{D(m(X)-m_0(X))}{m_0(X)^2}\,\mathbb{E}\left[Y-g_0(1,X)\middle| D=1,X\right]\right\}
\end{aligned}$$

But $\mathbb{E}\left[Y-g_0(1,X)\middle| D=1,X\right]=0$, so $\mathbb{E}\left[\frac{D(Y-g_0(1,X))(m(X)-m_0(X))}{m_0(X)^2}\right]=0$.

Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(1-D)(Y-g_0(0,X))(m(X)-m_0(X))}{(1-m_0(X))^2}\right] \\
&= Pr(D=1)*0 + Pr(D=0)\,\mathbb{E}\left[\frac{(1-D)(Y-g_0(0,X))(m(X)-m_0(X))}{(1-m_0(X))^2}\middle| D=0\right] \\
&= Pr(D=0)\,\mathbb{E}\left\{\mathbb{E}\left[\frac{(1-D)(Y-g_0(0,X))(m(X)-m_0(X))}{(1-m_0(X))^2}\middle| D=0,X\right]\right\} \\
&= Pr(D=0)\,\mathbb{E}\left\{\frac{(1-D)(m(X)-m_0(X))}{(1-m_0(X))^2}\,\mathbb{E}\left[Y-g_0(0,X)|D=0,X\right]\right\}
\end{aligned}$$

But $\mathbb{E}\left[Y-g_0(0,X)\middle| D=0,X\right]=0$, so $\mathbb{E}\left[\frac{(1-D)(Y-g_0(0,X))(m(X)-m_0(X))}{(1-m_0(X))^2}\right]=0$

So indeed, $D_0[\eta-\eta_0]=0$ $\square$

0.2 Unconfoundness and overlap assumptions

Assumption 1. *Unconfoundness assumption: Conditional on X, the treatment assignment mechanism is independent of the potential outcome. A weaker version of this assumption is the conditional mean independence. Which is*

$$\mathbb{E}(y_0|X, D) = E(y_0|X) \tag{1}$$
$$\mathbb{E}(y_1|X, D) = \mathbb{E}(y_1|X) \tag{2}$$

That is $g_0(0, X) = E(y_0|X)$ and $g_1(1, X) = E(y_1|X)$.

Assumption 2. *Overlap assumption: $0 < Pr(D|X) < 1$.*

These two assumptions are needed for identification of our estimators.

- The unconfoundness assumption allows us to use $\mathbb{E}(y|X, D = 0)$ to replace $\mathbb{E}(y_0|X)$, and use $\mathbb{E}(y|X, D = 1)$ to replace $\mathbb{E}(y_1|X)$. This means we can use the observed outcome to learn the conditional mean of the potential outcome.
- The overlap assumption allows $\theta = \mathbb{E}(\mathbb{E}(y_1|X) - \mathbb{E}(y_0|X))$

The observed outcome y can be written as $y = y_0 + D(y_1 - y_0)$.

$$\begin{aligned}
\mathbb{E}(y|X, D) &= \mathbb{E}(y_0 + D(y_1 - y_0)|X, D) \\
&= \mathbb{E}(y_0|X, D) + D[\mathbb{E}(y_1|X, D) - \mathbb{E}(y_0|X, D)] \\
&= \mathbb{E}(y_0|X) + D[\mathbb{E}(y_1|X) - \mathbb{E}(y_0|X)]
\end{aligned}$$

where the third equality comes from the unconfoundness assumptions. If D = 1, $\mathbb{E}(y|X, D = 1) = \mathbb{E}(y_1|X)$; if D = 0, $\mathbb{E}(y|X, D = 0) = \mathbb{E}(y_0|X)$.

Notice that in order to compute ATE or ATET, we need $g_0(1, X) = \mathbb{E}(y_1|X)$. By unconfoundness assumption, we can use the observed outcome variable moment $\mathbb{E}(y|X, D = 1)$ to get $\mathbb{E}(y_1|X)$.

The ATE is an expecation over population, so the overlap assumption guarantees that $\theta = \mathbb{E}(\mathbb{E}(y|X, D = 1) - \mathbb{E}(y|X, D = 0))$ is identifiable.

0.3 Proof for ATE estimator is doubly robust

Proof.

$$
\begin{aligned}
\theta_0 = &\left[\mathbb{E}(g_0(1,X)) + \mathbb{E}\left(\frac{D(Y-g_0(1,X))}{m_0(X)}\right)\right] \\
&- \left[\mathbb{E}(g_0(0,X)) + \mathbb{E}\left(\frac{(1-D)(Y-g_0(0,X))}{1-m_0(X)}\right)\right]
\end{aligned}
$$

Let's consider two scenarios. First, assume that the outcome model is correctly specified, so $g_0(0,X) = E(Y|X, D=0)$ and $g_0(1,X) = E(Y|X, D=1)$. Then the second term and and the fourth term are zero. They have already been proved in the proof of Neyman orthogonality in 0.1. So θ_0 is indeed ATE.

Second, assume that the only the propensity score model is correctly specified, so $\mathbb{E}(D|X) = m_0(X)$.

$$
\begin{aligned}
\mathbb{E}\left(\frac{D(Y-g_0(1,X))}{m_0(X)}\right) &= \Pr(D=1)\,\mathbb{E}\left[\mathbb{E}\left(\frac{(Y-g_0(1,X))}{m_0(X)}\middle| X, D=1\right)\right] \\
&= \Pr(D=1)\,\mathbb{E}\left[\frac{1}{m_0(X)}(\mathbb{E}(Y|X,D=1) - g_0(1,X))\right] \\
&= \mathbb{E}\left[\frac{D}{m_0(X)}(\mathbb{E}(Y_1|X) - g_0(1,X))\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(D|X)}{m_0(X)}(\mathbb{E}(Y_1|X) - g_0(1,X))\right] \\
&= \mathbb{E}(Y_1) - \mathbb{E}(g_0(1,X))
\end{aligned}
$$

Similarly, we can prove that $\mathbb{E}\left(\frac{(1-D)(Y-g_0(0,X))}{1-m_0(X)}\right) = \mathbb{E}(Y_0) - E(g_0(0,X))$. So again $\theta_0 = \mathbb{E}(Y_1) - E(Y_0)$. □

The above proof also sheds light on how to compute the potential outcome. To preserve the double robustness, we need to compute $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_0)$ by inverse probability adjustment. Specifically,

$$
\begin{aligned}
\mathbb{E}(Y_1) &= \mathbb{E}(g_0(1,X)) + \mathbb{E}\left(\frac{D(Y-g_0(1,X))}{m_0(X)}\right) \\
\mathbb{E}(Y_0) &= \mathbb{E}(g_0(0,X)) + \mathbb{E}\left(\frac{(1-D)(Y-g_0(0,X))}{1-m_0(X)}\right)
\end{aligned}
$$