Abadie's Semiparametric Difference-in-Difference Estimator

Kenneth Houngbedji*

*Agence Française de Développement

Stata Users Group meeting, July 2017 - Paris

Outline

1. Motivation
2. Framework
3. Example
4. Limitations

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.
    - Randomization was not possible,

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.
  - Randomization was not possible,
  - Selection into treatment depends on covariates which determine also the treatment outcome

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.
  - Randomization was not possible,
  - Selection into treatment depends on covariates which determine also the treatment outcome
  - Conditional exogeneity is not plausible.

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.
    - Randomization was not possible,
    - Selection into treatment depends on covariates which determine also the treatment outcome
    - Conditional exogeneity is not plausible.
- Abadie (2005) proposes an estimator to estimate average effect of treatment on the treated.

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.
  - Randomization was not possible,
  - Selection into treatment depends on covariates which determine also the treatment outcome
  - Conditional exogeneity is not plausible.
- Abadie (2005) proposes an estimator to estimate average effect of treatment on the treated.
  - When data are available before and after treatment for treated and non treated observations

Motivation

- Researchers are sometimes interested in studying the impact of reform or intervention using non experimental data.
  - Randomization was not possible,
  - Selection into treatment depends on covariates which determine also the treatment outcome
  - Conditional exogeneity is not plausible.
- Abadie (2005) proposes an estimator to estimate average effect of treatment on the treated.
  - When data are available before and after treatment for treated and non treated observations
  - Conditional parallel trend assumption is plausible.

Semiparametric difference-in-difference estimator

- The estimator proceeds in three steps.

Semiparametric difference-in-difference estimator

- The estimator proceeds in three steps.
  - First, compute change of outcomes over time for each observation;

Semiparametric difference-in-difference estimator

- The estimator proceeds in three steps.
  - First, compute change of outcomes over time for each observation;
  - Second, estimate the probability to be treated for each observation and use it to weight each observation;

Semiparametric difference-in-difference estimator

- The estimator proceeds in three steps.
    - First, compute change of outcomes over time for each observation;
    - Second, estimate the probability to be treated for each observation and use it to weight each observation;
    - Last, compare weighted change over time across treated and non-treated groups.

Semiparametric difference-in-difference estimator

- The estimator proceeds in three steps.
  - First, compute change of outcomes over time for each observation;
  - Second, estimate the probability to be treated for each observation and use it to weight each observation;
  - Last, compare weighted change over time across treated and non-treated groups.
- Inference takes also into account that the propensity score is estimated.

Semiparametric difference-in-difference estimator

- The estimator proceeds in three steps.
    - First, compute change of outcomes over time for each observation;
    - Second, estimate the probability to be treated for each observation and use it to weight each observation;
    - Last, compare weighted change over time across treated and non-treated groups.
- Inference takes also into account that the propensity score is estimated.
- Heterogeneity of treatment effect can also be investigated.

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.
- Each subject has two potential outcomes : $(\mathbf{y}_{1t}\ ,\ \mathbf{y}_{0t})$.

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.
- Each subject has two potential outcomes : $(\mathbf{y}_{1t}\ ,\ \mathbf{y}_{0t})$.
  - $\mathbf{y}_{1t}$ is the value of $\mathbf{y}$ if the subject receives the treatment by time t;

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.
- Each subject has two potential outcomes : $(\mathbf{y}_{1t}\ ,\ \mathbf{y}_{0t})$.
    - $\mathbf{y}_{1t}$ is the value of $\mathbf{y}$ if the subject receives the treatment by time t;
    - $\mathbf{y}_{0t}$ is the value of $\mathbf{y}$ had the participant not received the treatment at time t;

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.

- Each subject has two potential outcomes : $(\mathbf{y}_{1t}\ ,\ \mathbf{y}_{0t})$.

  - $\mathbf{y}_{1t}$ is the value of $\mathbf{y}$ if the subject receives the treatment by time t;
  - $\mathbf{y}_{0t}$ is the value of $\mathbf{y}$ had the participant not received the treatment at time t;

- $\mathbf{d}_t$ is equal to 1 when a participant is treated by time t and 0 otherwise.

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.
- Each subject has two potential outcomes : $(\mathbf{y}_{1t}\ ,\ \mathbf{y}_{0t})$.
  - $\mathbf{y}_{1t}$ is the value of $\mathbf{y}$ if the subject receives the treatment by time t;
  - $\mathbf{y}_{0t}$ is the value of $\mathbf{y}$ had the participant not received the treatment at time t;
- $\mathbf{d}_t$ is equal to 1 when a participant is treated by time t and 0 otherwise.
- At baseline b no one is treated.

Notations

- We want to estimate the causal effect of a treatment on a variable of interest $\mathbf{y}$ at some time t.
- Each subject has two potential outcomes : $(\mathbf{y}_{1t}\ ,\ \mathbf{y}_{0t})$.
  - $\mathbf{y}_{1t}$ is the value of $\mathbf{y}$ if the subject receives the treatment by time t;
  - $\mathbf{y}_{0t}$ is the value of $\mathbf{y}$ had the participant not received the treatment at time t;
- $\mathbf{d}_t$ is equal to 1 when a participant is treated by time t and 0 otherwise.
- At baseline b no one is treated.
- $\mathbf{x}_b$ is a vector of covariates measured at baseline.

The estimator

The average treatment effect on the treated (ATET) is:

$$\text{ATET} \equiv \mathbb{E}\Big(\mathbf{y}_{1t} - \mathbf{y}_{0t} \mid \mathbf{d}_t = 1\Big) \tag{1}$$

The estimator

The average treatment effect on the treated (ATET) is:

$$\text{ATET} \equiv \mathbb{E}\Big(\mathbf{y}_{1t} - \mathbf{y}_{0t} \mid \mathbf{d}_t = 1\Big) \tag{1}$$

Key assumptions:

$$\mathbb{E}\Big(\mathbf{y}_{0t} - \mathbf{y}_{0b}\Big|\, \mathbf{d}_t = 1\,,\ \mathbf{x}_b\Big) = \mathbb{E}\Big(\mathbf{y}_{0t} - \mathbf{y}_{0b}\Big|\, \mathbf{d}_t = 0\,,\ \mathbf{x}_b\Big). \tag{2}$$

$$\mathbb{P}\left(\mathbf{d}_t = 1\right) > 0 \text{ and } \pi\left(\mathbf{x}_b\right) < 1. \tag{3}$$

The estimator

The average treatment effect on the treated (ATET) is:

$$\text{ATET} \equiv \mathbb{E}\Big(\mathbf{y}_{1t} - \mathbf{y}_{0t} \mid \mathbf{d}_t = 1\Big) \tag{1}$$

Key assumptions:

$$\mathbb{E}\Big(\mathbf{y}_{0t} - \mathbf{y}_{0b} \Big| \, \mathbf{d}_t = 1\,,\ \mathbf{x}_b\Big) = \mathbb{E}\Big(\mathbf{y}_{0t} - \mathbf{y}_{0b} \Big| \, \mathbf{d}_t = 0\,,\ \mathbf{x}_b\Big). \tag{2}$$

$$\mathbb{P}\left(\mathbf{d}_t = 1\right) > 0 \text{ and } \pi\left(\mathbf{x}_b\right) < 1. \tag{3}$$

The semiparametric difference-in-difference estimator is the sample analog of:

$$\mathbb{E}\left(\frac{\mathbf{y}_t - \mathbf{y}_b}{\mathbb{P}\left(\mathbf{d}_t = 1\right)} \times \frac{\mathbf{d}_t - \pi\left(\mathbf{x}_b\right)}{1 - \pi\left(\mathbf{x}_b\right)}\right). \tag{4}$$

Estimating the propensity score

- Abadie (2005) suggests to approximate the propensity score $\pi(\mathbf{x}_b)$ semiparametrically using a polynomial series of the predictors.

Estimating the propensity score

- Abadie (2005) suggests to approximate the propensity score $\pi\left(\mathbf{x}_b\right)$ semiparametrically using a polynomial series of the predictors.

- We can either use a linear probability specification or a series logit estimator (SLE) (see Hirano et al., 2003) .

Estimating the propensity score

- Abadie (2005) suggests to approximate the propensity score $\pi\left(\mathbf{x}_{b}\right)$ semiparametrically using a polynomial series of the predictors.

- We can either use a linear probability specification or a series logit estimator (SLE) (see Hirano et al., 2003) .

  - The approximation of $\pi\left(\mathbf{x}_{b}\right)$ produced by the linear probability model can be written as follows:

$$\hat{\pi}\left(\mathbf{x}_{b}\right) = \hat{\gamma}_{0} + \hat{\gamma}_{1} \times \mathbf{x}_1 + \sum_{i=1}^{\mathrm{k}} \hat{\gamma}_{2i} \times \mathbf{x}_2^{i} \tag{5}$$

Estimating the propensity score

- Abadie (2005) suggests to approximate the propensity score $\pi\left(\mathbf{x}_b\right)$ semiparametrically using a polynomial series of the predictors.

- We can either use a linear probability specification or a series logit estimator (SLE) (see Hirano et al., 2003) .

  - The approximation of $\pi\left(\mathbf{x}_b\right)$ produced by the linear probability model can be written as follows:

$$\hat{\pi}\left(\mathbf{x}_b\right) = \hat{\gamma}_0 + \hat{\gamma}_1 \times \mathbf{x}_1 + \sum_{i=1}^{\mathrm{k}} \hat{\gamma}_{2i} \times \mathbf{x}_2^i \tag{5}$$

  - The approximation of $\pi\left(\mathbf{x}_b\right)$ produced by a series logit estimator will be as follows:

$$\hat{\pi}\left(\mathbf{x}_b\right) = \Lambda\left(\hat{\gamma}_0 + \hat{\gamma}_1 \times \mathbf{x}_1 + \sum_{k=1}^{\mathrm{K}} \hat{\gamma}_{2k} \times \mathbf{x}_2^k\right) \tag{6}$$

Stata command absdid

- The basic syntax for the command `absdid` is:

Stata command absdid

- The basic syntax for the command `absdid` is:

`absdid` *depvar [if] [in]* , tvar(*varname*) xvar(*varlist*) order(#) sle

Stata command absdid

- The basic syntax for the command `absdid` is:

  `absdid` *depvar [if] [in]* , tvar(*varname*) xvar(*varlist*) order(#) sle

- Additional options includes:

Stata command absdid

- The basic syntax for the command `absdid` is:

  `absdid` *depvar [if] [in]* , tvar(*varname*) xvar(*varlist*) order(#) sle

- Additional options includes:
  - yxvar(*varlist*): list of variables to explore heterogeneity of treatment effect.

Stata command absdid

- The basic syntax for the command `absdid` is:

  `absdid` *depvar [if] [in]* , tvar(*varname*) xvar(*varlist*) order(#) sle

- Additional options includes:
  - yxvar(*varlist*): list of variables to explore heterogeneity of treatment effect.
  - csinf(#) to drop observations of which the propensity score is less than the value provided as `csinf`. The default is `csinf(0)`.

Stata command absdid

- The basic syntax for the command `absdid` is:

`absdid` *depvar [if] [in]* , tvar(*varname*) xvar(*varlist*) order(#) sle

- Additional options includes:
  - yxvar(*varlist*): list of variables to explore heterogeneity of treatment effect.
  - csinf(#) to drop observations of which the propensity score is less than the value provided as `csinf`. The default is `csinf(0)`.
  - csup(#) to drop observations of which the propensity score is greater than the value provided as `csup`. The default is `csup(1)`.

Average Effect of Land Certificates on Labour Supply

Outcomes	Mean	ATET
Labour supply of male adults	135.540	-12.042***
	(7.758)	(7.917)
- Pre-planting	22.196	-9.513***
	(1.384)	(2.401)
- Planting	14.404	-0.164
	(1.104)	(1.149)
- Weeding	18.053	-1.972
	(1.257)	(1.788)
- Harvest	18.842	0.475
	(1.227)	(1.489)
- Threshing	15.193	-2.318*
	(1.001)	(1.290)
Number of households	161	591

Average effect across different groups

	Mean	(1)	(2)	(3)
Outcome: Labor supply by male adults				
Constant	22.196	-9.513***	3.927	6.648
	(1.384)	(2.401)	(8.060)	(10.526)
- Distance to plot (mins)			0.252	0.261
			(0.278)	(0.284)
- Number of plots at baseline				-2.382**
				(1.104)
Number of households	161	591	591	591

Testing parallel trend assumption

Outcomes	ATET in 2004		
	Mean	(SDID)	(DID)
Labor supply	119.789	3.113	-27.843***
	(6.881)	(7.531)	(6.977)
- Women	38.857	2.673	-7.490***
	(2.436)	(2.761)	(2.390)
- Men	80.932	0.439	-20.353***
	(4.827)	(5.781)	(5.101)
Number of households	161	591	669

Limitations

- The semiparametric difference-in-difference approach is mostly suited for longitudinal surveys with a baseline and follow-up rounds.

Limitations

- The semiparametric difference-in-difference approach is mostly suited for longitudinal surveys with a baseline and follow-up rounds.

- However, it is possible to modify extend it to include repeated cross section data.

Limitations

- The semiparametric difference-in-difference approach is mostly suited for longitudinal surveys with a baseline and follow-up rounds.
- However, it is possible to modify extend it to include repeated cross section data.
- For a set of control variables, the estimates vary with

Limitations

- The semiparametric difference-in-difference approach is mostly suited for longitudinal surveys with a baseline and follow-up rounds.
- However, it is possible to modify extend it to include repeated cross section data.
- For a set of control variables, the estimates vary with
    - the type of approximation used;

Limitations

- The semiparametric difference-in-difference approach is mostly suited for longitudinal surveys with a baseline and follow-up rounds.
- However, it is possible to modify extend it to include repeated cross section data.
- For a set of control variables, the estimates vary with
  - the type of approximation used;
  - the order of the polynomial approximation used.

Thanks for your attention.

References I

Abadie, A. (2005, 01). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1 – 19.

Hirano, K., G. W. Imbens, and G. Ridder (2003, 07). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161 – 1189.