

Estimating IATEs in Stata using cate

Giovanni Cerulli
CNR-IRCrES

2025 Stata Users Group, Italy
Milan, September 25, 2025

Why CATE Matters

In **causal inference** and **program evaluation**, the Conditional Average Treatment Effect (**CATE**) matters for a series of reasons:

- **Goes beyond the ATE:** the Average Treatment Effect is often too aggregated and can hide crucial differences across individuals or groups.
- **Captures heterogeneity:** CATE shows how treatment effects vary with observable characteristics, revealing “who benefits most” and “who benefits least.”
- **Improves decision-making:** policies and interventions can be tailored to the right subpopulations, increasing effectiveness and efficiency.
- **Supports fairness:** identifying differential effects helps avoid policies that disproportionately advantage or disadvantage certain groups.
- **Enhances external validity:** by linking effects to covariates, results are more generalizable to new populations and contexts.

Setup and Model Fitting

Given a binary treatment variable, $D \in \{0, 1\}$, two potential outcomes $Y(1)$, $Y(0)$, and a set of (exogenous) covariates X .

Conditional Average Treatment Effect (CATE)

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

- Captures how the causal effect varies *with covariates* X .
- **ATE** = $\mathbb{E}[\tau(X)]$ averages CATE over the population.
- **IATE**: predicted effect at an individual's X_i (model-based).

Identification: When is CATE well-defined from data?

Key assumptions (observational data)

- ① **Unconfoundedness** (selection on observables): $\{Y(1), Y(0)\} \perp D \mid X$.
- ② **Overlap** (positivity): $0 < e(X) = \mathbb{P}(D = 1 \mid X) < 1$ a.s.

Estimation view (sketch)

- **Nuisance functions**: outcome regressions $\mu_d(x) = \mathbb{E}[Y \mid D = d, X = x]$ and propensity score $e(x)$.
- Two CATE estimation approaches:
 - ① **Partialing-out** estimator
 - ② **Augmented inverse-probability weighting** estimator (doubly robust)
- **Cross-fitting** for reducing overfitting bias with ML learners.

From CATE to IATEs and GATEs

- **CATE** function $\tau(x)$: maps covariates to treatment effect.
- **IATEs**: $\hat{\tau}(X_i)$ — individualized predictions from the estimated CATE.
- **GATEs**: group averages of IATEs, e.g. by income category:

$$\text{GATE}(g) = \mathbb{E}[\hat{\tau}(X) \mid G = g]$$

- Why useful:
 - Detect heterogeneity $\Rightarrow$ who benefits most/least.
 - Inform targeting/policy design and fairness considerations.

CATE Estimation with Cross-Fitting and ML (**Partialling-out**)

- **Step 1. Nuisance estimation (cross-fitted)**

$$\hat{\mu}_d(x) = \mathbb{E}[Y \mid D = d, X = x] \quad (\text{linear lasso})$$

$$\hat{e}(x) = \Pr(D = 1 \mid X = x) \quad (\text{logit lasso})$$

- **Step 2. Residualization (partialling-out)**

$$\tilde{Y}_i = Y_i - \hat{\mu}(X_i), \quad \tilde{D}_i = D_i - \hat{e}(X_i)$$

- **Step 3. CATE estimation (heterogeneity learner)**

$$\tilde{Y}_i = \tau(X_i) \tilde{D}_i + \varepsilon_i, \quad \hat{\tau}(x) \approx \text{Random Forest}(x)$$

- **Formula:**

$$\hat{\tau}(x) = \frac{\mathbb{E}[\tilde{Y} \cdot \tilde{D} \mid X = x]}{\mathbb{E}[\tilde{D}^2 \mid X = x]}$$

Partialing-out estimator

```
cate po (ovar catevarlist) (tvar) [if] [in] [, options]
```

Augmented inverse-probability weighting estimator

```
cate aipw (ovar catevarlist) (tvar) [if] [in] [, options]
```

Definitions:

- `ovar`: continuous outcome of interest.
- `catevarlist`: covariates for the CATE model (conditioning variables).
 - May include factor variables; see [U] 11.4.3 Factor variables.
- `tvar`: binary variable representing the treatment indicator.

Stata code

```
. use https://www.stata-press.com/data/r19/assets3
. global catecovars age educ i.(incomecat pension married twoearn ira ownhome)
. cate po (assets $catecovars) (e401k), rseed(12345671)

* IATEs visualization and tests
. categraph histogram
. estat heterogeneity

* GATEs by income (reuse estimated CATE)
. cate, group(incomecat) reestimate
. categraph gateplot
. estat gatetest
```


Dataset and Research Questions

Goal: Estimate the effect of 401(k) eligibility (e401k) on net financial assets (asset).

Key questions:

- Are treatment effects on net wealth **heterogeneous**?
- How do effects vary across prespecified groups (e.g., incomecat)?

Data: 1990 *Survey of Income and Program Participation (SIPP)*

Sample: households, information on head of household

- incomecat, age, educ
- pension, married
- ira, ownhome, twoearn

Approach:

- With `teffects` $\Rightarrow$ average effect (ATE).
- With `cate` $\Rightarrow$ **individualized average treatment effects (IATEs)**.

Variables in the Dataset

Variable	Description	Type
asset	Net financial assets (outcome)	Continuous
e401k	401(k) eligibility (treatment)	Binary
incomecat	Household income category	Categorical
age	Age of household head	Continuous
educ	Years of education	Continuous
pension	Receives pension benefit	Binary
married	Marital status	Binary
ira	Has IRA account	Binary
ownhome	Owns a home	Binary
twoearn	Two earners in household	Binary

Note: Covariates are stored in a global macro `$catecovars` for use in estimation.

Application in Stata

- Load the dataset and define covariates:

Stata code

```
. use https://www.stata-press.com/data/r19/assets3  
. global catecovars age educ i.(incomecat pension married twoearn ira ownhome)
```

- Fit the model with partialing-out estimator, outcome assets, treatment e401k.

Stata code

```
. cate po (assets $catecovars) (e401k), rseed(12345671)
```

- Cross-fitting with lasso models:
 - Outcome: linear lasso
 - Treatment: logit lasso
 - CATE model: random forest

Results and Interpretation

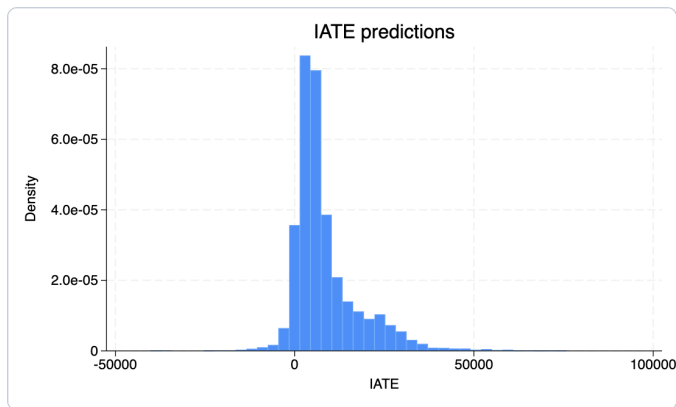
- Output (summary):
 - Observations: 9,913
 - Estimator: Partialing out, 10-fold cross-fit
 - ATE (e401k eligible vs not): **\$7,937** higher assets (SE = 1153, $p < 0.001$)
 - POmean (not eligible): **\$14,016**
- Process:
 - Cross-fitting lasso → estimate outcome/treatment models
 - Random forest → estimate IATE function
 - AIPW scores → compute ATE
- Beyond the ATE: cate also estimates **IATEs**.
 - Use `categraph` histogram to visualize distribution of IATEs.

Distribution of IATEs

- Histogram of predicted IATEs:

Stata code

```
. categraph histogram
```



Testing for Heterogeneity

- Visual inspection $\neq$ statistical evidence.
- Use `estat heterogeneity` to formally test:

Stata code

```
. estat heterogeneity
```

- H_0 : Treatment effects are homogeneous
- Test result: $\chi^2(1) = 4.11, p = 0.0427$
- $\rightarrow$ Reject H_0 : evidence of heterogeneous effects

IATEs by Education Level

- Explore how IATEs vary with educ, fixing other covariates at their means/base levels.

Stata code

```
. categraph iateplot educ
```

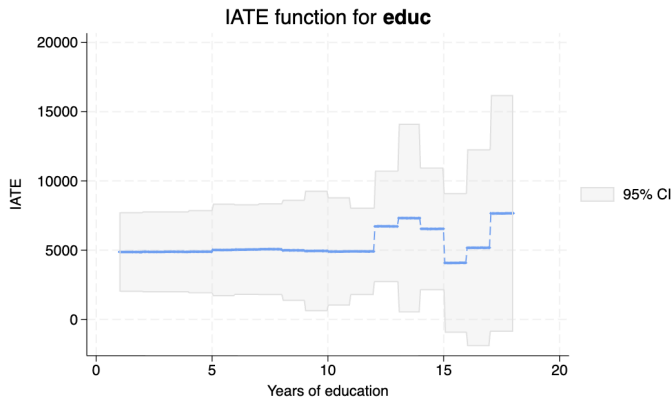
Note: IATE estimated at fixed values of covariates other than educ.

Variable	Statistic	Value	Type
age	mean	41.05891	continuous
incomecat	base	0	factor
ira	base	0	factor
married	base	0	factor
ownhome	base	0	factor
pension	base	0	factor
twoearn	base	0	factor

- By default: age fixed at mean (41.1), other factors at base categories.

Interpreting IATEs by Education

`categraph iateplot` also shows **95% pointwise confidence intervals**. Findings: Below 10 years of education, effects appear constant; Between 12–15 years, effects are larger while holding other covariates fixed; Beyond 15 years, variability increases and CIs widen, making inference less precise.



Estimating GATEs over Prespecified Groups

- Goal: characterize heterogeneity by estimating group average treatment effects (**GATEs**).
- Example: income categories (`incomecat`).

Stata code

```
. table incomecat, stat(min income) stat(max income) stat(median income) nototal
```

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

Estimating GATEs (using IATEs)

- Reuse the estimated IATE function to compute GATEs:

Stata code

```
. cate, group(incomecat) reestimate
```

Estimating GATE ...

Conditional average treatment effects	Number of observations	= 9,913
Estimator: Partialing out	Number of folds in cross-fit	= 10
Outcome model: Linear lasso	Number of outcome controls	= 17
Treatment model: Logit lasso	Number of treatment controls	= 17
CATE model: Random forest	Number of CATE variables	= 17

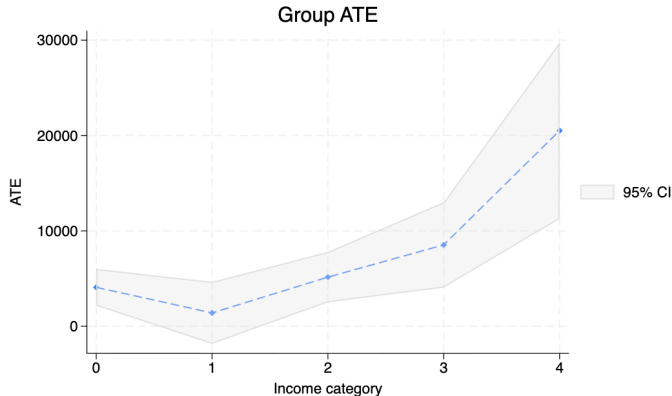
		Robust				
assets	Coefficient	std. err.	z	P> z	[95% conf. interval]	
GATE						
incomecat						
0	4087.014	987.7124	4.14	0.000	2151.133	6022.895
1	1399.398	1663.193	0.84	0.400	-1860.4	4659.196
2	5154.329	1349.842	3.82	0.000	2508.688	7799.97
3	8532.238	2287.664	3.73	0.000	4048.499	13015.98
4	20510.94	4723.741	4.34	0.000	11252.58	29769.3
ATE						
e401k (Eligible vs Not eligi..)	7937.182	1153.017	6.88	0.000	5677.309	10197.05
POMean						
e401k Not eligible	14016.38	833.4423	16.82	0.000	12382.87	15649.9

GATE Plot and Group Heterogeneity Test

- Visualize GATEs:

Stata code

```
. categraph gateplot
```



Stata code

```
. estat gatetest
```

```
. estat gatetest
```

Group treatment-effects heterogeneity test

H0: Group average treatment effects are homogeneous

```
( 1)  [GATE]0bn.incomecat - [GATE]1.incomecat = 0
( 2)  [GATE]0bn.incomecat - [GATE]2.incomecat = 0
( 3)  [GATE]0bn.incomecat - [GATE]3.incomecat = 0
( 4)  [GATE]0bn.incomecat - [GATE]4.incomecat = 0
```

```
      chi2(4) =   18.44
Prob > chi2 = 0.0010
```

- Read more about CATE in **[CAUSAL] cate** in the *Stata Causal Inference and Treatment-Effects Estimation Reference Manual*.
- Learn more about Stata's **causal inference features**.
- Explore new features in **Stata 19**, in particular those related to causal inference.