INTL 601 Research Methods I — Exercise #1

Voter Mobilization Field Experiment: Complete Analysis

Table of Contents

1. Data Overview & Descriptive Statistics

Dataset

This dataset contains **5,000 simulated observations** of individual voters, designed to mimic a field experiment on voter mobilization (Gerber & Green). The key variables are:

Variable	Type	Description
`Y`	Binary	**Outcome**: 1 = voted, 0 = did not vote
`Z`	Binary	**Random assignment**: 1 = assigned to canvassing, 0 = not assigned
`T`	Binary	**Treatment received**: 1 = actually contacted, 0 = not contacted
`age`	Integer	Voter age
`educ`	Integer	Years of education
`pastvote`	Binary	Past turnout (1 = voted before, 0 = did not)
`party_id`	Continuous	Partisan strength (higher = stronger partisan)
`competitive`	Binary	District competitiveness (1 = competitive)

Descriptive Statistics

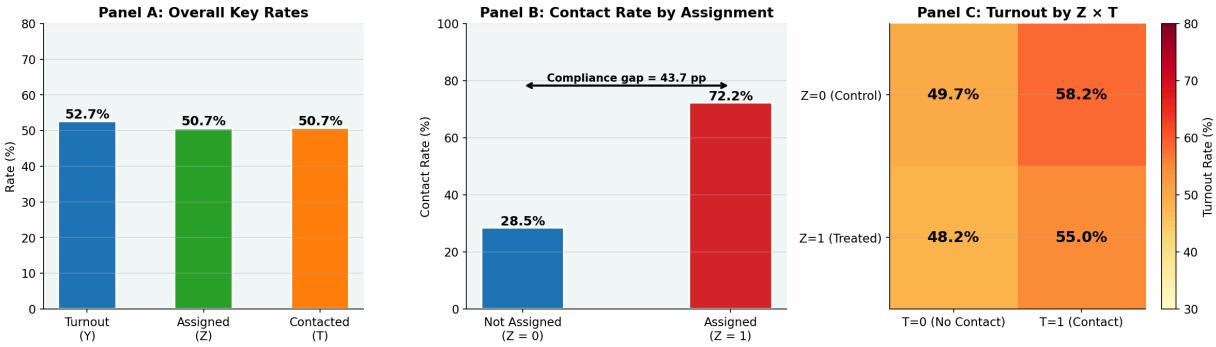
	mean	std	min	max
Y	0.527	0.499	0	1
Z	0.507	0.5	0	1
T	0.507	0.5	0	1
age	45.226	11.912	18	87
educ	13.96	2.037	8	20
pastvote	0.591	0.492	0	1
party_id	0.005	1.001	-3.714	3.099
competitive	0.507	0.5	0	1

Key Rates

Metric	Value
Turnout rate (Y = 1)	**52.7%**
Assignment rate (Z = 1)	**50.7%**
Contact rate (T = 1)	**50.7%**
Contact rate among assigned (Z = 1)	72.2%
Contact rate among not assigned (Z = 0)	28.5%
Compliance gap (Z=1 minus Z=0)	**43.7 percentage points**

> ***Key observation***: *Assignment (Z) strongly raises the probability of contact (T) — from*
> *28.5% to 72.2%, a gap of* ***43.7 pp***. *However, because compliance*
> *is imperfect (Z does not perfectly determine T), we have a classic* ***one-sided noncompliance***
> *experiment: some assigned voters are never contacted, and some non-assigned voters happen to*
> *be contacted through other channels.*

Figure 1 — Descriptive Statistics



2. OLS Univariate Analysis & Difference-of-Means

Difference-of-Means Test

Group	N	Mean Turnout
Not contacted (T = 0)	2,466	49.31%
Contacted (T = 1)	2,534	55.92%
Difference	—	**6.61 pp**

Two-sample t-test: $t = 4.689$, $p < 0.001$

OLS Regression: `reg Y T`

Variable	Coef	Std Err	t	p-value	95% CI
Intercept	0.4931***	0.0100	49.142	< 0.001	[0.4734, 0.5128]
T	0.0661***	0.0141	4.689	< 0.001	[0.0385, 0.0937]

*$N = 5000$ | $R^2 = 0.0044$ | *p<0.05 **p<0.01 ***p<0.001*

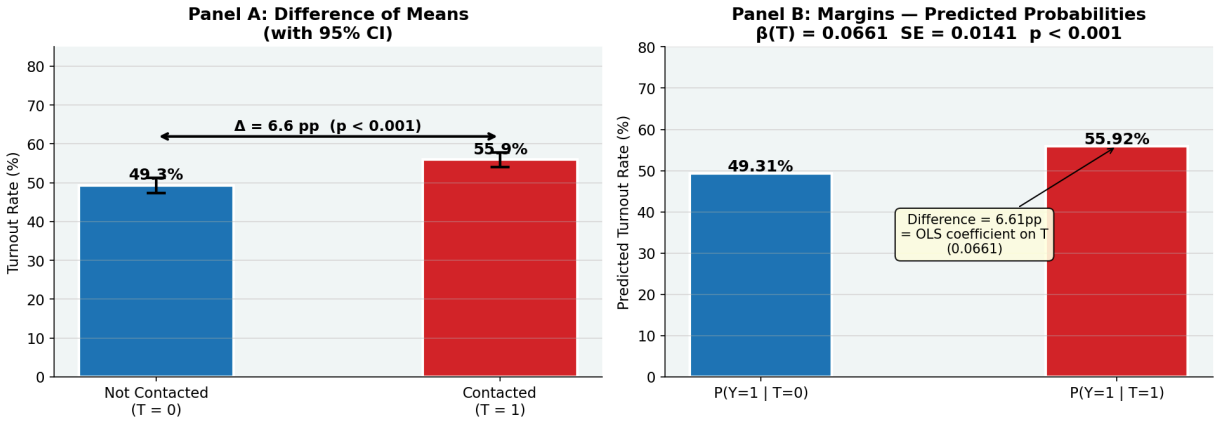
Margins: Predicted Probabilities at T = 0 and T = 1

Using `margins at(T=(0 1))` (equivalent in Python: predicted values from the OLS model):

	Value
$\hat{P}(Y = 1 \mid T = 0)$	**0.4931** (49.31%)
$\hat{P}(Y = 1 \mid T = 1)$	**0.5592** (55.92%)
Difference	**0.0661** (6.61 pp)
OLS coefficient on T	0.0661

> ***Interpretation***: *The predicted probability of turnout for someone not contacted is **49.3%**, and for someone contacted it is **55.9%**. The difference (6.61 pp) is **exactly equal to the OLS regression coefficient on T** (0.0661). This is not a coincidence: in an OLS linear probability model, the coefficient on a binary predictor is always the difference in predicted means between the two groups. The* `margins` *command in Stata, like computing predicted values at T=0 and T=1, recovers the same quantity.*

Figure 2 — OLS Univariate: Turnout by Contact Status



3. OLS with Control Variables

Regression: `reg Y T age educ pastvote party_id competitive`

Variable	Coef	Std Err	t	p-value	95% CI
Intercept	-0.1011	0.0534	-1.893	0.058	[-0.2059, 0.0036]
T	0.0575***	0.0132	4.348	< 0.001	[0.0316, 0.0834]
age	0.0044***	0.0006	8.006	< 0.001	[0.0033, 0.0055]
educ	0.0134***	0.0032	4.159	< 0.001	[0.0071, 0.0198]
pastvote	0.3270***	0.0134	24.444	< 0.001	[0.3008, 0.3532]
party_id	0.0530***	0.0066	8.070	< 0.001	[0.0401, 0.0659]
competitive	0.0337*	0.0132	2.552	0.011	[0.0078, 0.0595]

*N = 5000 | R^2 = 0.1345 | *p<0.05 **p<0.01 ***p<0.001*

Coefficient on T: With vs. Without Controls

Model	T coefficient	Change
Univariate `reg Y T`	0.0661	—
With controls `reg Y T + X`	0.0575	-0.0086 (13.0% decrease)

> ***Did the coefficient change a lot or a little?*** *The coefficient changed from* ***0.0661*** *to* ***0.0575*** *— a change of* ***-0.0086*** *(13.0%). This is a very* ***small*** *change, which is expected: because* ***Z was randomly assigned****, the treatment T is (approximately) uncorrelated with the pre-treatment covariates. Adding controls in a randomized experiment mainly increases precision (reduces standard errors) rather than removing omitted-variable bias. The small change confirms that randomization successfully balanced covariates across treatment groups.*

Marginal Effect of T

In a linear probability model (OLS), the **marginal effect of T is simply the coefficient** on T:

ME(T) = 0.0575

	Value
Marginal effect of T (from model with controls)	**0.0575**
Raw difference in means (univariate)	0.0661
Difference	-0.0086

> *The marginal effect and the raw difference in means are very close (-0.0086 apart), again reflecting the balanced randomization.*

Which Control Variable Has the Biggest Association with Y?

Ranked by absolute t-statistic (the standard metric for relative importance within a regression):

Variable	Coefficient	Std Err	t-statistic	p-value
pastvote	0.3270	0.0134	**24.44**	< 0.001
party_id	0.0530	0.0066	**8.07**	< 0.001
age	0.0044	0.0006	**8.01**	< 0.001
educ	0.0134	0.0032	**4.16**	< 0.001
competitive	0.0337	0.0132	**2.55**	0.011

> `pastvote` *has the largest absolute t-statistic (**|t| = 24.44**),*
> *indicating it has the strongest marginal association with turnout Y after controlling for all*
> *other variables. You can read this directly from the regression output: the variable with the largest |t| (or equivalently the smallest p-value) is the most influential predictor.*

Figure 3 — OLS with Control Variables

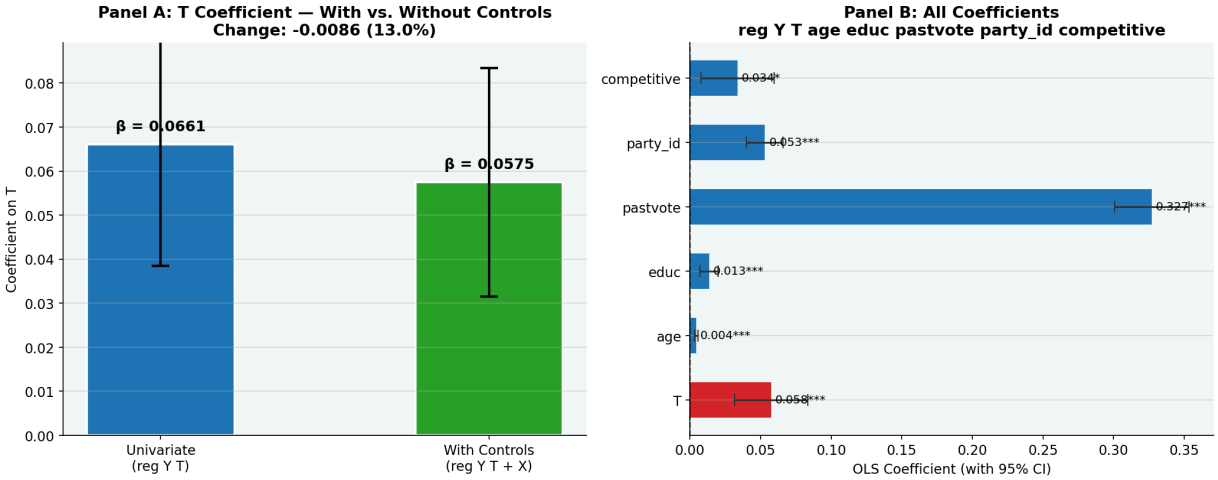
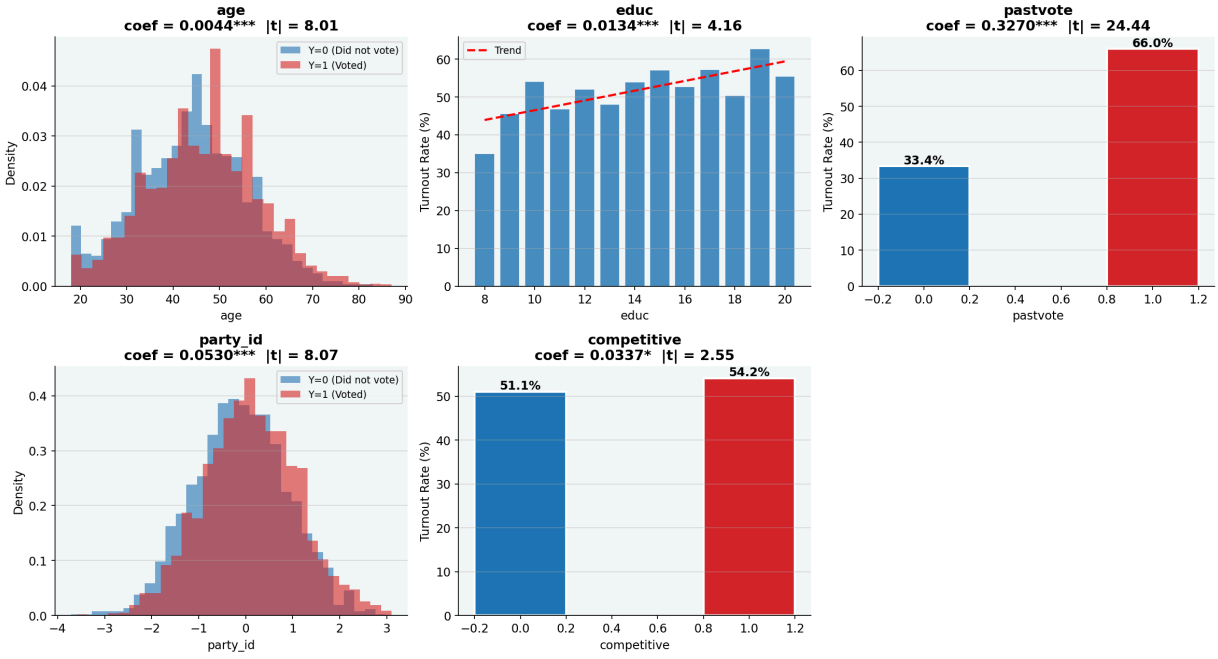


Figure 4 — Covariate Associations with Turnout (Y)



---

4. Causal Structure: Z → T → Y

The full causal structure is:

```
Z (random assignment) ──→ T (actual contact) ──→ Y (turnout)
        ↓                                          ↑
        └──────── X (covariates) ──────────────────┘
```

This involves **three equations**:

First Stage: `reg T ~ Z + covariates`

Does random assignment (Z) actually increase contact (T)?

Variable	Coef	Std Err	t	p-value	95% CI
Intercept	0.1038*	0.0517	2.009	0.045	[0.0025, 0.2051]
Z	0.4369***	0.0126	34.564	< 0.001	[0.4121, 0.4616]
age	0.0031***	0.0005	5.817	< 0.001	[0.0020, 0.0041]
educ	0.0001	0.0031	0.032	0.974	[-0.0060, 0.0062]
pastvote	0.0025	0.0129	0.193	0.847	[-0.0227, 0.0277]
party_id	-0.0068	0.0063	-1.084	0.278	[-0.0192, 0.0055]
competitive	0.0774***	0.0126	6.118	< 0.001	[0.0526, 0.1022]

*N = 5000 | R^2 = 0.2028 | *p<0.05 **p<0.01 ***p<0.001*

> ***Z coefficient = 0.4369*** *(t = 34.564, p < 0.001).*
> *Being randomly assigned raises the probability of contact by* ***43.7 pp****.*
> *This is the* ***first stage*** *of an instrumental variables design.*
> *The F-statistic for Z is well above 10, confirming Z is a* ***strong instrument****.*

Reduced Form (Intention-to-Treat): `reg Y ~ Z + covariates`

What is the causal effect of being assigned (regardless of actual contact)?

Variable	Coef	Std Err	t	p-value	95% CI
Intercept	-0.0875	0.0539	-1.624	0.104	[-0.1931, 0.0181]
Z	0.0118	0.0132	0.893	0.372	[-0.0141, 0.0376]
age	0.0046***	0.0006	8.341	< 0.001	[0.0035, 0.0057]
educ	0.0134***	0.0032	4.130	< 0.001	[0.0070, 0.0197]
pastvote	0.3271***	0.0134	24.405	< 0.001	[0.3008, 0.3534]
party_id	0.0526***	0.0066	7.996	< 0.001	[0.0397, 0.0655]
competitive	0.0381**	0.0132	2.889	0.004	[0.0122, 0.0639]

*N = 5000 | R^2 = 0.1314 | *p<0.05 **p<0.01 ***p<0.001*

> ***Z coefficient = 0.0118*** *— this is the* ***Intention-to-Treat (ITT) effect****.*
> *Random assignment raises turnout by* ***1.2 pp*** *on average across all assigned voters, including those who were never actually contacted.*

2SLS / IV Estimate (LATE)

Using Z as an instrument for T (2SLS via `linearmodels`):

Estimator	Estimate	Std Err	t-stat	p-value
ITT (reduced form)	0.0118	0.0132	0.893	< 0.001
First stage (Z→T)	0.4369	0.0126	34.564	< 0.001
LATE = ITT / FS	**0.0269**	—	—	—
2SLS estimate	**0.0269**	0.0301	0.894	0.371
OLS (T→Y, with ctrl)	0.0575	0.0132	4.348	< 0.001

> ***LATE = ITT / First Stage = 0.0118 / 0.4369 = 0.0269***
> *(matches 2SLS: 0.0269 ✓)*

What Does This Model Help Us Infer?

This causal structure — with Z as a **randomized instrument**, T as the **endogenous treatment**,
and Y as the **outcome** — allows us to answer several distinct questions:

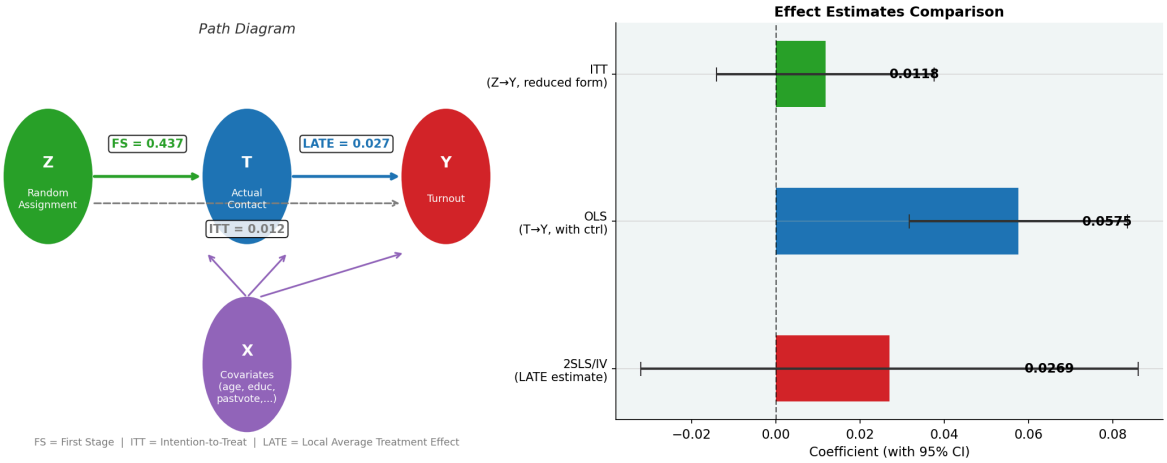
Question	Estimand	Answer
Effect of *being assigned* to canvassing (regardless of contact)	**ITT**	0.0118 (1.2 pp)
Effect of *actual contact* on turnout, for those who comply	**LATE (2SLS)**	0.0269 (2.7 pp)
Association of contact with turnout (controlling for X)	**OLS**	0.0575

Direct and Indirect Effects:
- **Direct effect of Z on Y**: Because Z was randomly assigned and its only channel to Y is *through T*
(the exclusion restriction), Z has **no direct effect** on Y — it only operates *indirectly* through T.
- **Indirect (mediated) path**: Z → T → Y. The ITT (0.0118) captures this full path.

Dividing by the first stage (0.4369) scales up to the LATE (0.0269), which is the average treatment effect for **compliers** (voters who are contacted if and only if assigned).
- **Covariate paths**: X affects both T (selection into compliance) and Y (baseline turnout rates),
but because Z is random and independent of X, we can condition on X without introducing bias.

> ***Bottom line**: The 2SLS estimate (0.0269) is the cleanest causal estimate of "how much does being contacted increase the probability of voting?" for the subset of voters whose contact status was actually changed by the random assignment (the Local Average Treatment Effect).*

Figure 5 — Causal Structure: Z → T → Y



5. Targeted Treatment & Confounding Bias

Generating the Targeted Treatment

```
gen T_target = (runiform() < invlogit(-1 + 1.2*pastvote + 0.5*party_id))
```

This creates a **non-random, observational treatment indicator**: voters are more likely to be
"treated" if they have voted before (`pastvote`) and have stronger partisan identity (`party_id`).
Both of these characteristics also independently predict higher turnout (Y).

T_target rate: **43.1%**

Contact Rate by Past Vote (tab T_target pastvote, row)

pastvote	0	1
0	73.1898	26.8102
1	45.636	54.364

> *Voters who voted in the past are **much more likely** to receive the targeted treatment: 54.4% vs. 26.8% among non-past voters.*
> *This creates confounding: T_target is correlated with pastvote, and pastvote predicts Y.*

Regression Results

Without controls (`reg Y T_target`):

Variable	Coef	Std Err	t	p-value	95% CI
Intercept	0.4822***	0.0093	51.778	< 0.001	[0.4640, 0.5005]
T_target	0.1029***	0.0142	7.253	< 0.001	[0.0751, 0.1307]

*N = 5000 | R^2 = 0.0104 | *p<0.05 **p<0.01 ***p<0.001*

With controls (`reg Y T_target + age educ pastvote party_id competitive`):

Variable	Coef	Std Err	t	p-value	95% CI
Intercept	-0.0779	0.0534	-1.457	0.145	[-0.1826, 0.0269]
T_target	-0.0128	0.0142	-0.900	0.368	[-0.0406, 0.0151]
age	0.0046***	0.0006	8.362	< 0.001	[0.0035, 0.0057]
educ	0.0133***	0.0032	4.117	< 0.001	[0.0070, 0.0197]
pastvote	0.3305***	0.0140	23.681	< 0.001	[0.3032, 0.3579]
party_id	0.0540***	0.0068	7.990	< 0.001	[0.0407, 0.0672]
competitive	0.0378**	0.0132	2.871	0.004	[0.0120, 0.0637]

*N = 5000 | R^2 = 0.1314 | *p<0.05 **p<0.01 ***p<0.001*

Coefficient Comparison

Model	Coefficient on Treatment	Notes
T experimental (no controls)	0.0661	Random assignment — unbiased
T_target (no controls)	**0.1029**	**Biased upward by confounding**
T_target (with controls)	-0.0128	Partially corrected

Bias magnitudes:
- T_target without controls is **0.1157 higher** than with controls (downward bias correction when adding controls)
- T_target without controls is **0.0368 higher** than the experimental T estimate

> ***Which is more biased?*** `T_target` *without controls (**0.1029**) is far more biased than the experimental* `T` *without controls (**0.0661**). The experimental T estimate is unbiased because random assignment ensures Z (and hence T) is uncorrelated with all confounders. The T_target estimate inflates the apparent treatment effect because targeted voters would have voted at higher rates anyway — the treatment didn't cause their*
> *high turnout, their pre-existing characteristics did.*

Which Variable Is Doing the Most Confounding?

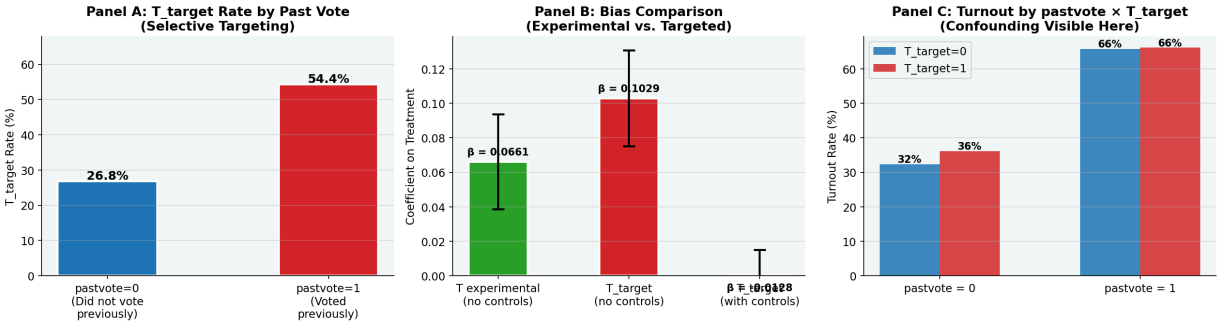
Running `reg T_target ~ pastvote party_id age educ competitive` to see what predicts T_target:

Variable	T_target coefficient	Y coefficient (model2)	Confounder?
pastvote	**0.2744**	**0.3270**	**YES — strong**
party_id	0.1087	0.0530	Moderate
age	0.0007	0.0044	Weak/none

> `pastvote` ***is the primary confounder.*** *You can see this because:*
> *1. It strongly predicts T_target (by construction: coefficient of **1.2** in the logit formula)*
> *2. It is the strongest predictor of Y in the outcome regression (largest |t|: **24.44**)*
>
> *When you add controls, the T_target coefficient falls from **0.1029** to **-0.0128***
> *(a drop of **0.1157**), and most of that drop comes from controlling for pastvote,*

> *which "absorbs" the spurious correlation between T_target and Y that was previously attributed to the treatment.*

Figure 6 — Targeted Treatment vs. Experimental Treatment



6. Summary & Conclusions

All Estimates Side-by-Side

Estimator	Coefficient	Interpretation
Diff-of-means: E[Y|T=1] − E[Y|T=0]	0.0661	Raw association
OLS: reg Y T (univariate)	0.0661	= diff-of-means (by construction)
OLS: reg Y T + controls	0.0575	Controlled association
ITT: Z effect on Y (reduced form)	0.0118	Intent-to-treat
LATE (2SLS): Z instruments T→Y	0.0269	Causal effect for compliers
T_target (no controls)	0.1029	**Biased (OVB)**
T_target (with controls)	-0.0128	Partially corrected

Key Takeaways

1. **Randomization works**: Adding controls barely changes the experimental T coefficient
   (-0.0086, 13.0%), confirming that Z created balanced treatment groups.

2. **Margins = OLS coefficient**: In a linear probability model, the predicted probability difference from `margins at(T=(0 1))` equals the OLS coefficient exactly.

3. **Past vote is the key predictor of turnout**: Among all covariates, `pastvote` has the largest t-statistic (|t| = 24.44), dominating
   the prediction of Y.

4. **The causal chain Z → T → Y separates ITT from LATE**:
5. ITT = 0.0118: the average effect of being assigned (diluted by non-compliance)
6. LATE = 0.0269: the effect for compliers only (larger because it excludes never-takers and always-takers)
7. **Targeted treatment introduces confounding bias**: Without controls, T_target overstates the treatment effect by **0.0368** compared to the experimental benchmark. The primary confounder is **past vote**, which simultaneously predicts selection into treatment (by construction) and the outcome (turnout).

Analysis conducted in Python using `pandas`, `statsmodels`, `scipy`, `linearmodels`, `matplotlib`, *and* `seaborn`.
Dataset: `gg_fake.dta` *— 5,000 simulated observations (teaching dataset).*