Comparing Experimental and Nonexperimental Methods:

What Lessons Have We Learned Four Decades after LaLonde (1986)?

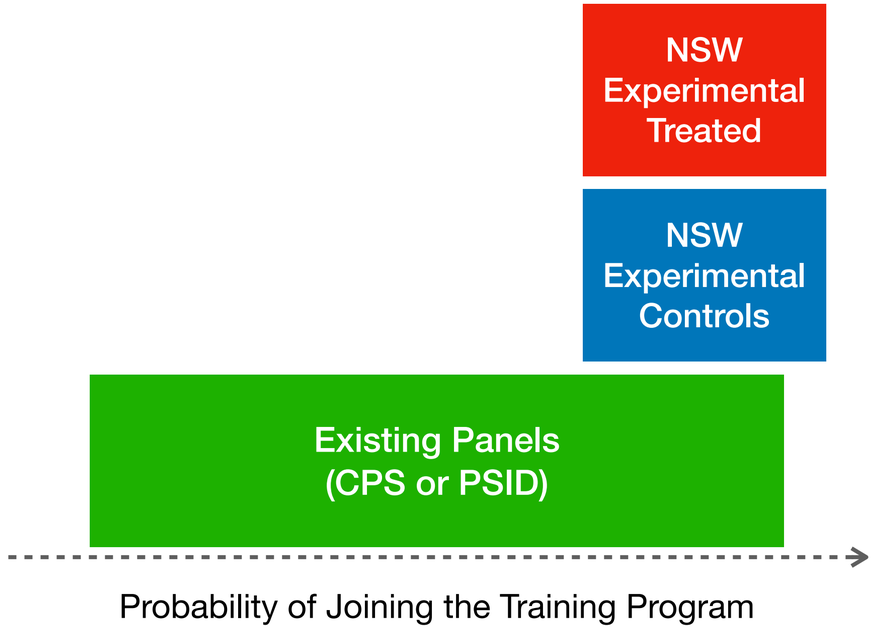
Guido Imbens and Yiqing Xu
(Stanford)

Center for Aging and Policy Studies
January 2026

LaLonde (1986)

- Lalonde (1986) analyzed data from a randomized experiment designed to evaluate the effect of a labor market program, the National Supported Work (NSW) program of mid-1970s
  - The RCT shows a substantial average effect on earnings ≈ $900 per year ($4,500 in 2026)

- LaLonde: **Could we have estimated this without an RCT?**
  - Put aside experimental control group
  - Comparison sample from public datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)



- Estimate average effect using CPS or PSID data as controls to mimic an observational study
  - Why it works? The selection process for applying to join the training program is the same between experimental treated and controls.

LaLonde's Conclusion

Abstract

> This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations.

Conclusion

> [P]olicymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors.

Why the Revisit?

- Sending shockwaves to the methods literature; kicking off the credibility revolution

- The subsequent literature in causal inference has been quite optimistic — with modern methods, we can recover experimental benchmarks using non-experimental data.

- But can we? Maybe we just got lucky.
  - It turns out, with good overlap, all modern methods can roughly recover the experimental ATE/ATT estimate using a particular non-experimental dataset (but not other datasets, and not with other causal estimands)

- Takeaways
  - Unconfoundedness is a strong assumption
  - Overlap is central (and can be improved)
  - Validating (both assumptions) is key
  - Modern estimators (e.g. AIPW) help
  - It'd be nice to check out other estimands (CATE, QTE, etc)

Roadmap

- Motivation
- **LaLonde (1986)**
- Causal Framework
- Reanalyses

Roadmap

- Motivation
- **LaLonde (1986)**
  - LaLonde's analyses
  - LaLonde's data
- Causal Framework
- Reanalyses

What LaLonde Actually Did — Regression Analyses

- Assuming zero-conditional-mean:
  - Simple linear regression: $Y_i = \delta D_i + \beta X_i + \epsilon_i$, in which X_i includes 10 baseline covariates
  - Difference-in-Differences: $\Delta Y_i = \delta D_i + \beta X_i + \epsilon_i$, in which ΔY_i is difference in income between 1975 and 1978
  - Including a lagged outcome: $\Delta Y_i = \delta D_i + Y_i^{'75} + \beta X_i + \epsilon_i$, in which $Y_i^{'75}$ is '75 earnings

Table 5

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975		NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978		Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78		Controlling for All Observed Variables and Pre-Training Earnings
		Unadjusted	Adjusted[c]	Unadjusted	Adjusted[c]	Without Age	With Age	Unadjusted	Adjusted[c]	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

What LaLonde Actually Did — Heckman's "Selection Model"

Table 6

Variables Excluded from the Earnings Equation, but Included in the Participation Equation	Comparison Group	NSW AFDC Females: Heckman Correction for Program Participation Bias, Using Estimate of Conditional Expectation of Earnings Error as Regressor in Earnings Equation — Estimate of Coefficient for: Training Dummy	NSW AFDC Females: Estimate of Expectation	NSW Males: Training Dummy	NSW Males: Estimate of Expectation
Marital Status, Residency in an SMSA, Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,129	−894	−1,333	−2,357
		(385)	(396)	(820)	(781)
	CPS-SSA-1	1,102	−606	−22	−1,437
		(323)	(480)	(584)	(449)
	NSW Controls	837	−18	899	−835
		(317)	(2376)	(840)	(2601)
Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,256	−823	–	–
		(405)	(410)		
	CPS-SSA-1	439	−979	–	–
		(333)	(481)		
	NSW Controls	–	–	–	–
Employment Status in 1976, Number of Children	*PSID*-1	1,564	−552	−1,161	−2,655
		(604)	(569)	(864)	(799)
	CPS-SSA-1	552	−902	13	−1,484
		(514)	(551)	(584)	(450)
	NSW Controls	851	147	889	−808
		(318)	(2385)	(841)	(2603)
No Exclusion Restrictions	*PSID*-1	1,747	−526	−667	−2,446
		(620)	(568)	(905)	(806)
	CPS-SSA-1	805	−908	213	−1,364
		(523)	(548)	(588)	(452)
	NSW Controls	861	284	889	−876
		(318)	(2385)	(840)	(2601)

- Identification comes from
  - Correct specification of the select model & joint normality of the error terms
  - Exclusion restriction

Reactions to Lalonde (1986)

(Credit to Josh Angrist)

- Initial reaction:
  - Heckman and Hotz (1985): specification testing gets it right
  - Heckman et al. (1997): Better data helps

- Dehejia and Wahba (1999): the propensity score method solves the problem
  - Smith and Todd (2001,2005): No, it doesn't; DW sample is special
  - DW (2005): Yes, it does
  - MHE and Kline: don't need the score; get the controls right

- Cook and Wong (2005, 2007): "well designed" observational studies (RD) come close to an RCT benchmark

- Many follow-ups in the statistical (causal inference) literature — mainly using LDW-CPS data

LaLonde's Data

- Original experiment (MDRC): about 6,600 participants; LaLonde worked with much smaller analytic samples.
- Experimental samples used by LaLonde:
  - Males: 722 individuals (297 treated, 425 controls).
  - Females: 1,158 individuals (600 treated, 585 controls).
- **Why the sample reduction?**
  - Attrition from missing post-program earnings due to incomplete follow-up interviews.
  - Random subsampling for 27- and 36-month follow-up surveys because of budget constraints.
  - Exclusion of males entering before January 1976 or still enrolled in January 1978.
- **LaLonde's ingenious design**
  - Treated experimental units combined with external comparison groups.
  - Control sources: CPS-SSA-1 (Matched CPS–SSA file) and PSID-1 (Panel Study of Income Dynamics).
  - Samples restricted to individuals under age 55 to improve comparability.

LaLonde-Dehejia-Wahba (LDW) Datasets

- Dehejia and Wahba (1999) restrict attention to the male sample, where estimates are most sensitive to functional-form specification.
- **Construction of the LDW experimental sample:**
  - Subsample of LaLonde's males with observed 1974 earnings and unemployment status.
  - Selection uses only pretreatment variables (e.g., assignment timing, prior employment), preserving orthogonality of treatment assignment.
  - Retains about 62% of the original treated male sample.
- **Nonexperimental controls:**
  - Subsets of CPS-SSA-1 and PSID-1, restricted to units with 1974 earnings and unemployment data.
  - Commonly used version combines experimental treated units with CPS-SSA-1 controls (LDW-CPS).
- **Empirical implications:**
  - LDW experimental sample shows higher pre-program unemployment and lower earnings than LaLonde's original male sample.
  - Reflects more persistent disadvantage and yields a larger estimated training effect ($1,794 vs. $886).

LaLonde's Data & LDW Datasets

Descriptive Statistics: LaLonde and Lalonde-Dehejia-Wahba Male Samples

	LaLonde Experimental		LaLonde Comparison Groups		LaLonde-Dehejia-Wahba Experimental	
	Treated	Control	CPS-SSA-1	PSID-1	Treated	Control
	(1)	(2)	(3)	(4)	(5)	(6)
Age	24.63	24.45	33.23	34.85	25.82	25.05
	(6.69)	(6.59)	(11.05)	(10.44)	(7.16)	(7.06)
Years of School	10.38	10.19	12.03	12.12	10.35	10.09
	(1.82)	(1.62)	(2.87)	(3.08)	(2.01)	(1.61)
Proportion High School Dropouts	0.73	0.81	0.30	0.31	0.71	0.83
	(0.44)	(0.39)	(0.46)	(0.46)	(0.46)	(0.37)
Proportion Married	0.17	0.16	0.71	0.87	0.19	0.15
	(0.37)	(0.36)	(0.45)	(0.34)	(0.39)	(0.36)
Proportion Black	0.80	0.80	0.07	0.25	0.84	0.83
	(0.40)	(0.40)	(0.26)	(0.43)	(0.36)	(0.38)
Proportion Hispanic	0.09	0.11	0.07	0.03	0.06	0.11
	(0.29)	(0.32)	(0.26)	(0.18)	(0.24)	(0.31)
Real Eearnings in 1975 (thousand)	3.07	3.03	13.65	19.06	1.53	1.27
	(4.87)	(5.20)	(9.27)	(13.60)	(3.22)	(3.10)
Proportion Unemployed in 1975	0.37	0.42	0.11	0.10	0.60	0.68
	(0.48)	(0.49)	(0.31)	(0.30)	(0.49)	(0.47)
Real Eearnings in 1974 (thousand)	NA	NA	14.02	19.43	2.10	2.11
			(9.57)	(13.41)	(4.89)	(5.69)
Proportion Unemployed in 1974	NA	NA	0.12	0.09	0.71	0.75
			(0.32)	(0.28)	(0.46)	(0.43)
#Observations	**297**	**425**	**15,922**	**2,490**	**185**	**260**

Roadmap

- Motivation
- LaLonde (1986)
- **Causal Framework**
- Reanalyses

Roadmap

- Motivation
- LaLonde (1986)
- **Causal Framework**
  - Identification under unconfoundedness
  - Estimation
- Reanalyses

Setup

- Units: $i = 1, \cdots, n$
- Treatment: $W_i \in \{0,1\}$
- Potential outcomes: $Y_i(w)$, where $w = 0, 1$
- Quantities of interest:
  - ATE: $\tau_{ATE} = E[Y_i(1) - Y_i(0)]$
  - ATT: $\tau_{ATT} = E[Y_i(1) - Y_i(0) \,|\, W_i = 1]$
- Question: Can we identify ATE and ATT when W_i is not randomized?
- Pre-treatment covariates: $X_i = [X_{i1}, \ldots, X_{iK}]^\top \in \mathcal{X}$
  - Predetermined and causally prior to W_i
  - Examples: sex, race, age, marital status, past income, etc.
  - X_i may be correlated with both W_i and $Y_i(w)$, thereby confounding the causal relationship

Identifying the ATT under Unconfoundedness

- **Identification Assumption**
  - $Y_i(0) \perp\!\!\!\perp W_i \mid X_i$ (Unconfoundedness)
  - $\Pr(W_i = 1 \mid X_i) < 1 - \epsilon$ (Overlap)

- **Identification Result**
  - Given unconfoundedness, in each stratum of $X = x$, we have

$$\begin{aligned}
\tau(x) &= \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] \\
&= \mathbb{E}[Y_i(1) \mid X_i = x] - \mathbb{E}[Y_i(0) \mid X_i = x] \\
\text{(by unconfoundedness \& overlap)} \qquad &= \mathbb{E}[Y_i \mid X_i = x, W_i = 1] - \mathbb{E}[Y_i \mid X_i = x, W_i = 0]
\end{aligned}$$

  - With overlap, we can average $\tau(x)$, CATT, over $\mathscr{X}$ with $W_i = 1$ to obtain the ATT.

Statistical and Causal Estimands

- Another way to think about it. Since $\mathbb{E}[Y_i \mid W_i = 1]$ is observed, the goal is to identify $\mathbb{E}[Y_i(0) \mid W_i = 1]$:

$$\begin{aligned}
\tau_{\text{ATT}} &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 1] \\
&= \mathbb{E}[Y_i \mid W_i = 1] - \int \mathbb{E}[Y_i(0) \mid X_i, W_i = 1]\, dP(X_i \mid W_i = 1) \\
&\text{(by unconfoundedness \& overlap)} \quad = \mathbb{E}[Y_i \mid W_i = 1] - \int \mathbb{E}[Y_i(0) \mid X_i, W_i = 0]\, dP(X_i \mid W_i = 1) \\
&= \mathbb{E}[Y_i \mid W_i = 1] - \int \mathbb{E}[Y_i \mid X_i, W_i = 0]\, dP(X_i \mid W_i = 1) \\
&= \mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}\{\mathbb{E}[Y_i \mid X_i, W_i = 0]\}
\end{aligned}$$

- Therefore, without unconfoundedness, we will be estimating a covariate adjusted difference

$$\mathbb{E}[Y_i \mid W_i = 1] - \int \mathbb{E}[Y \mid X_i, W_i = 0]\, dP(X_i \mid W_i = 1)$$

  but not necessary the ATT.

- In other words, all estimators may be converging to a "statistical estimand" that is uninterpretable

- Next, let's discuss how to "best" estimate $\int \mathbb{E}[Y \mid X_i, W_i = 0]\, dP(X_i \mid W_i = 1)$ using data

Estimation (for ATT)

- Subclassification / Exact matching
- Outcome modeling
  - Simple linear regression — Modeling $\mu(W, X) = \mathbb{E}[Y \mid W, X]$
  - Model-based imputation — Modeling $\mu_0(X) = \mathbb{E}[Y \mid W = 0, X]$ only, because $\mathbb{E}[Y \mid W = 1]$ is known
- Reweighting
  - Inverse probability weighting (IPW) — Resizing the control strata $\mathbb{E}[Y \mid W = 1] - \frac{1}{\pi_1} \mathbb{E}\left[(1 - W)\frac{e(X)}{1 - e(X)} Y\right]$ in which $\pi_1 = \Pr(W = 1)$ and $e(X)$ is the propensity score.
  - Entropy balancing — Find weights so that: $\omega(X) = \frac{e(X)}{1 - e(X)}$
- Augmented inverse probability weighting (AIPW)
  - Combine outcome modeling and IPW: $\mathbb{E}[Y \mid W = 1] - \frac{1}{\pi_1} \mathbb{E}\left[W \mu_0(X)\right] - \frac{1}{\pi_1} \mathbb{E}\left[(1 - W)\frac{e(X)}{1 - e(X)}\left\{ Y - \mu_0(X) \right\}\right]$,
  - "Doubly robust"
  - Special case of double/debiased matching learning (DML)

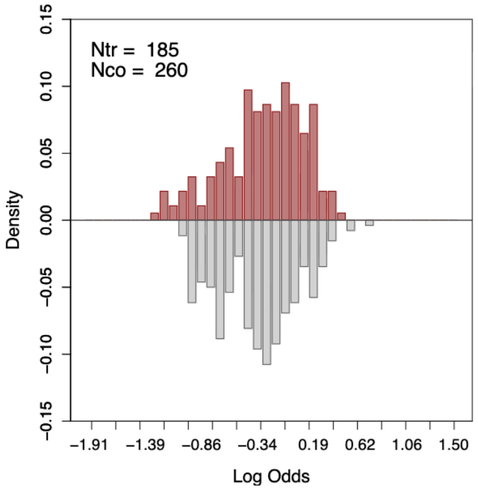
Roadmap

- Motivation
- LaLonde (1986)
- Modern Methods
- **Reanalyses**

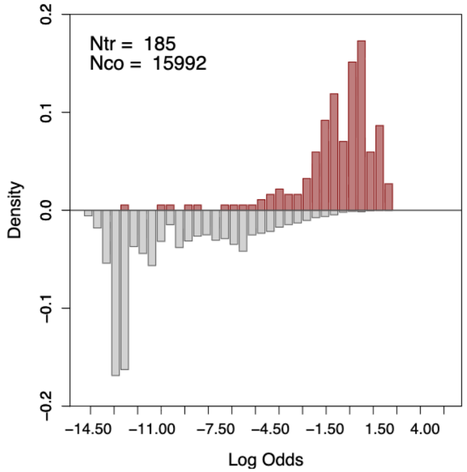
Roadmap

- Motivation
- LaLonde (1986)
- Modern Methods
- **Reanalyses**
  - The LDW data
  - Imbens, Rubin, and Sacerdote (2001)

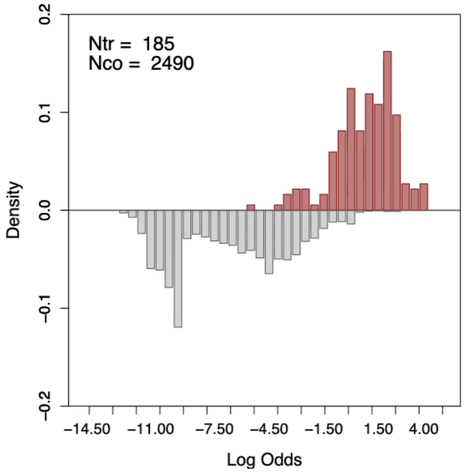
LDW Data — Overlap



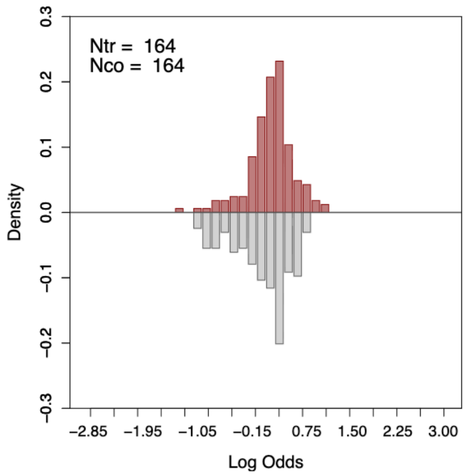
A. **LDW-Experimental**



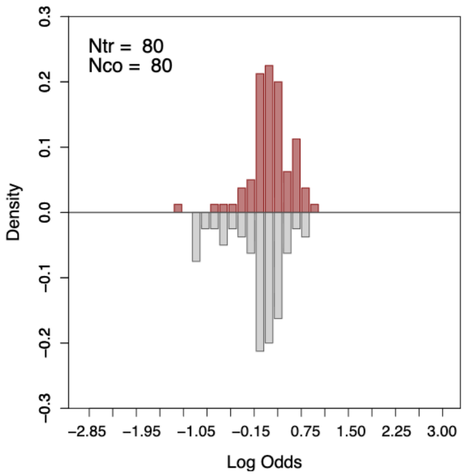
B. **LDW-CPS**



C. **LDW-PSID**

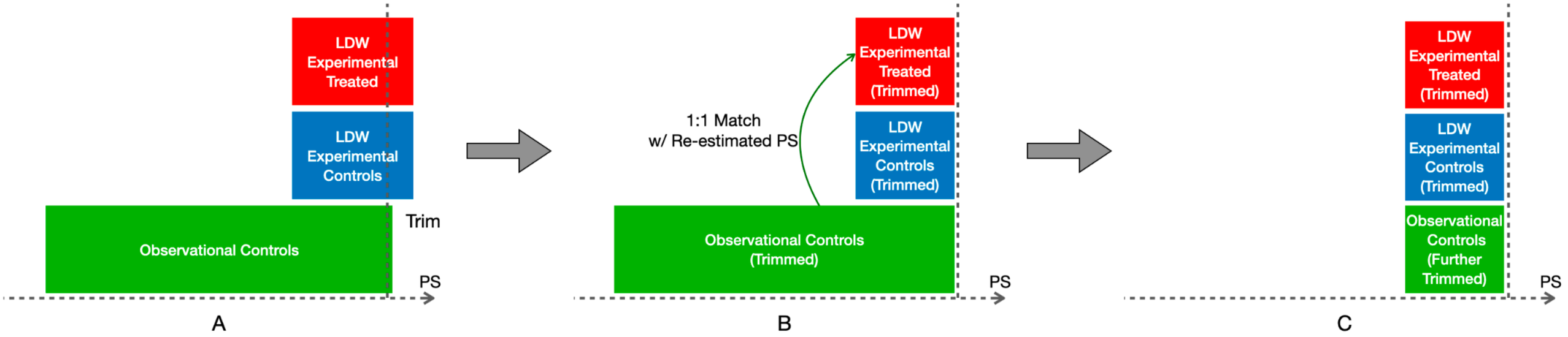


D. **Trimmed LDW-CPS**



E. **Trimmed LDW-PSID**

Trimming to Improve Overlap

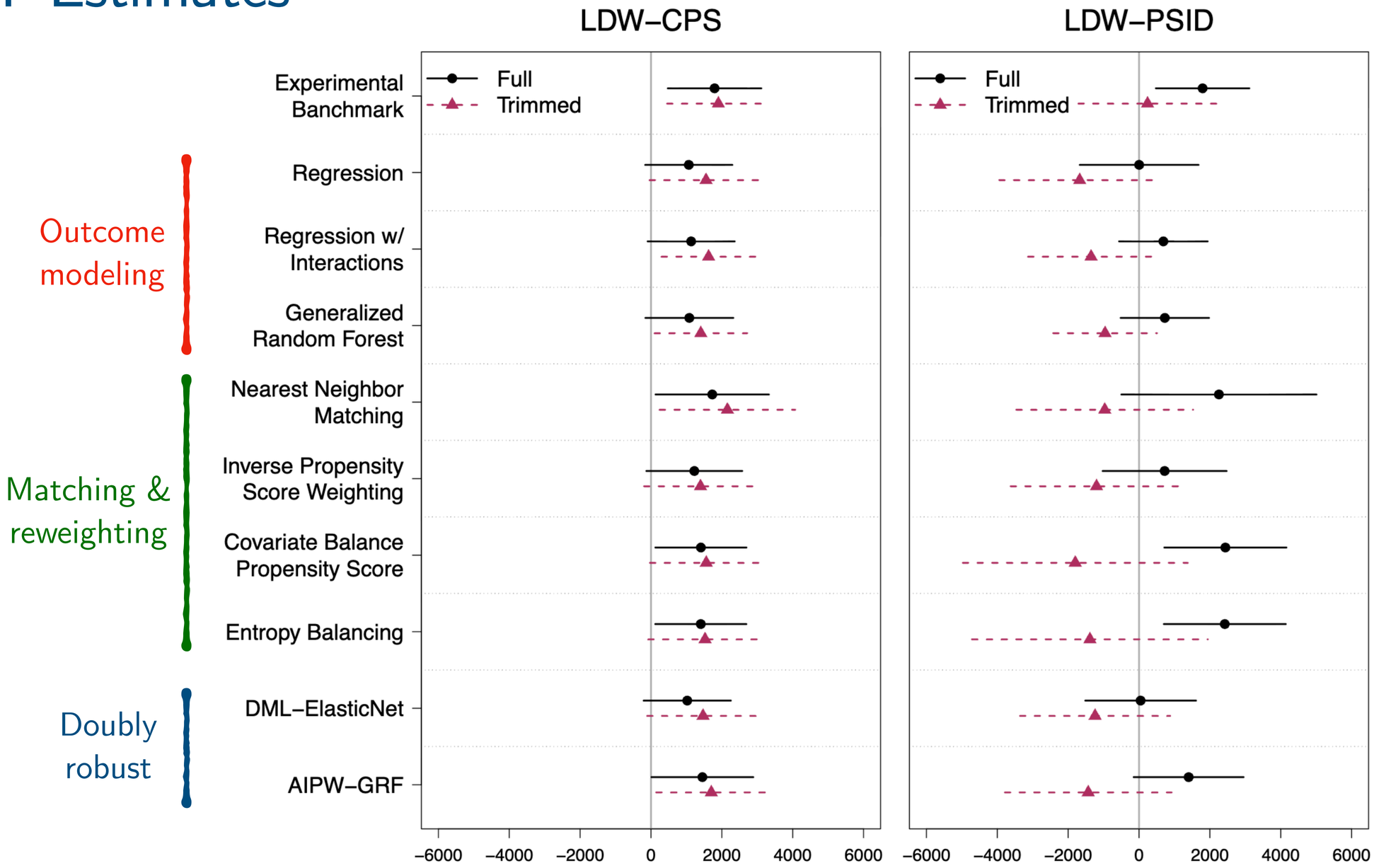


Original

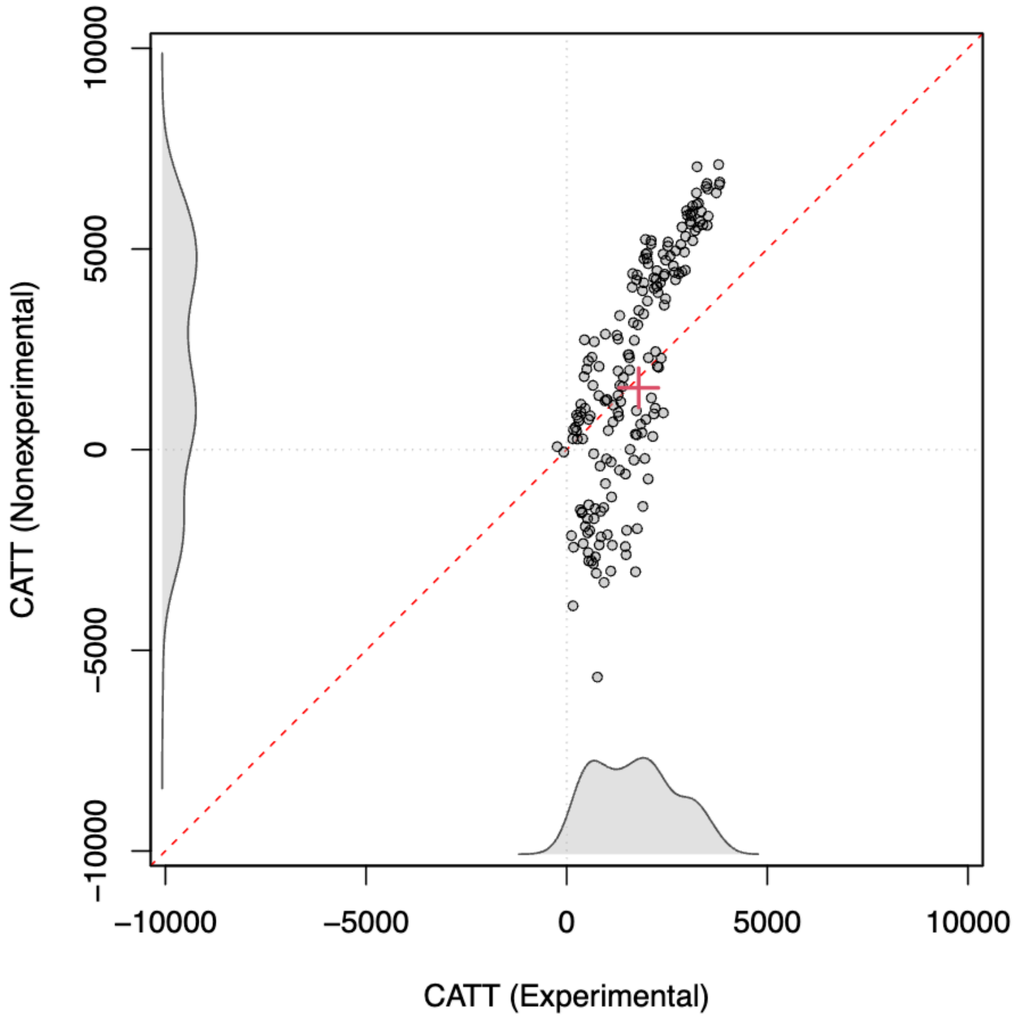
Trim experimental data
(and nonexperimental controls)

Further trim
nonexperimental controls

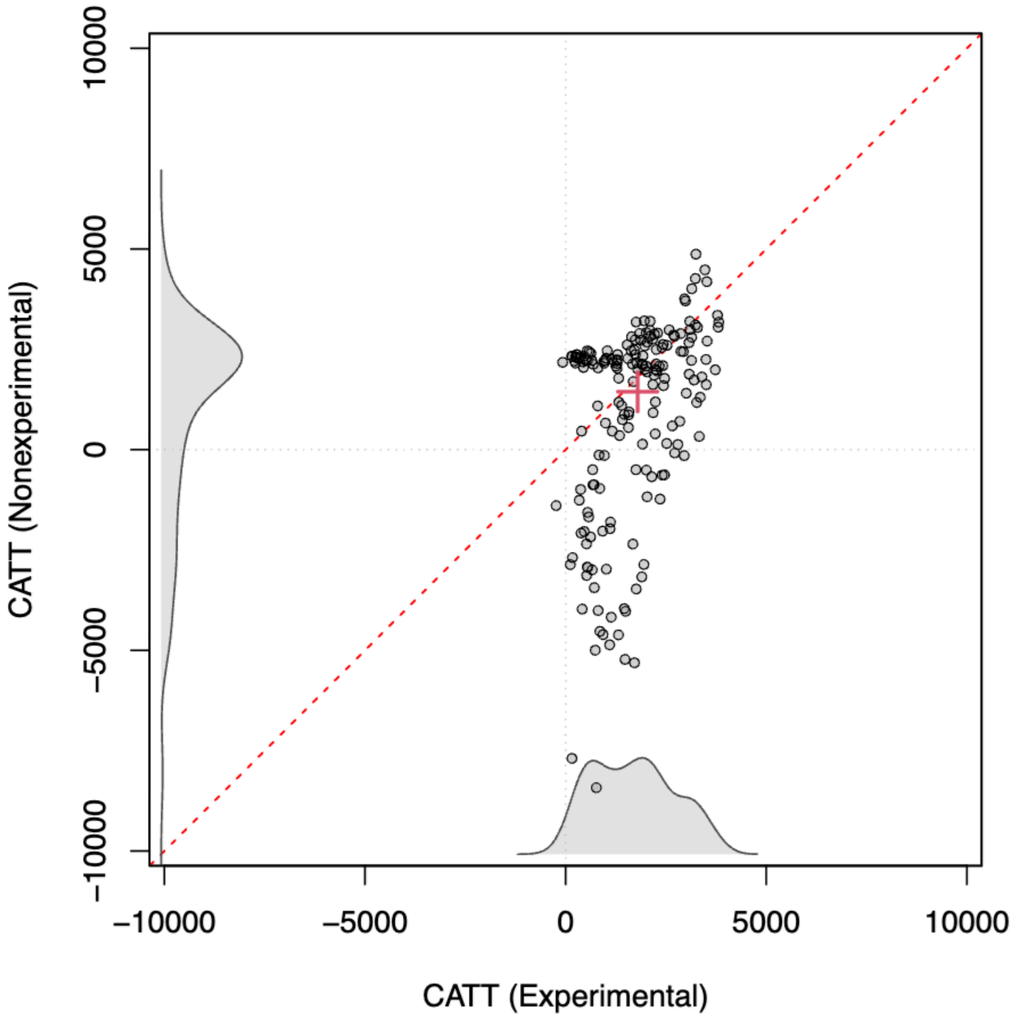
ATT Estimates



CATE Estimates (Untrimmed)

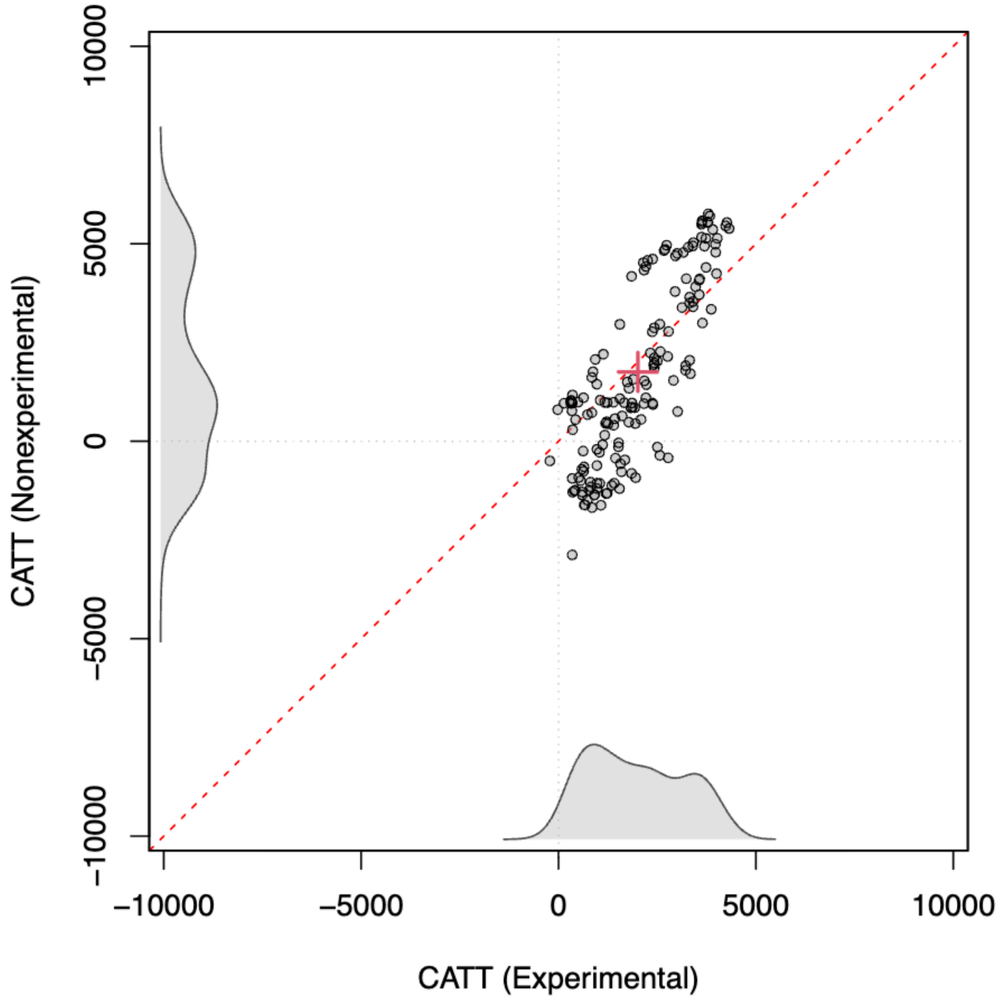


A. LDW-CPS

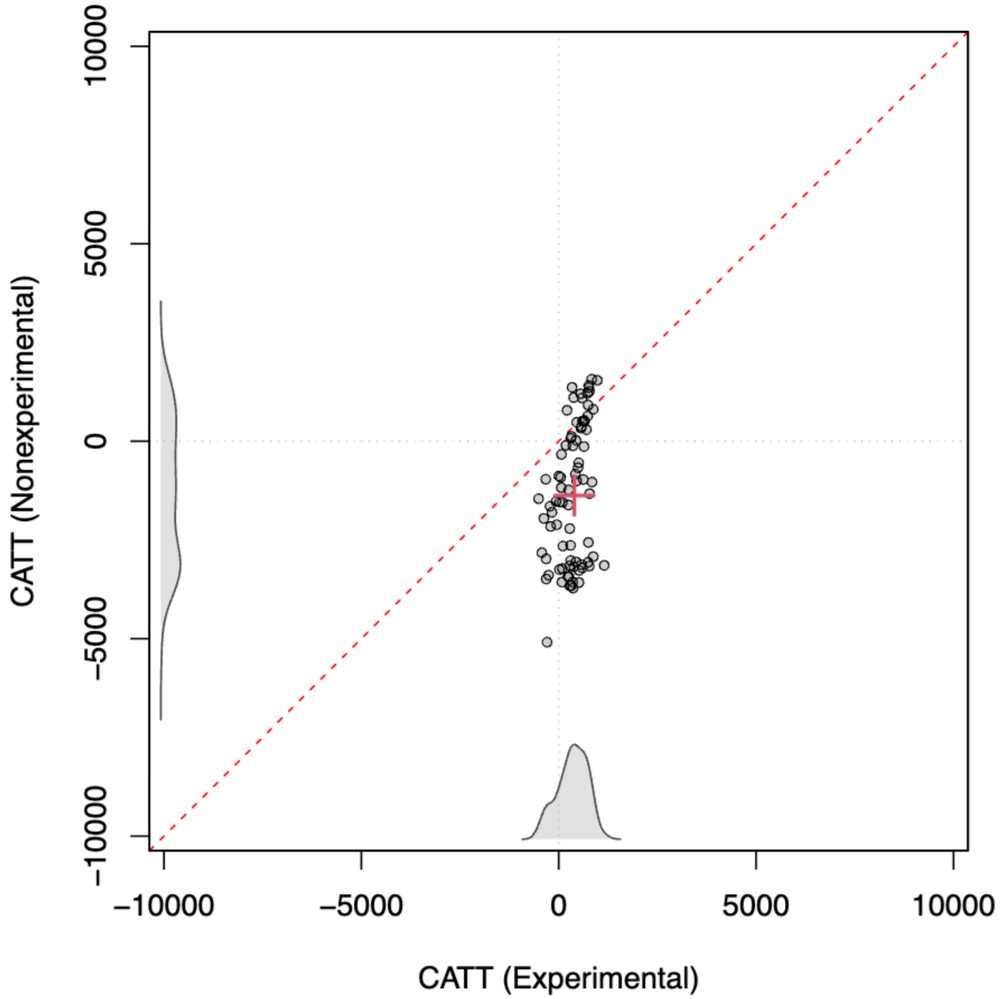


B. LDW-PSID

CATE Estimates (Trimmed)

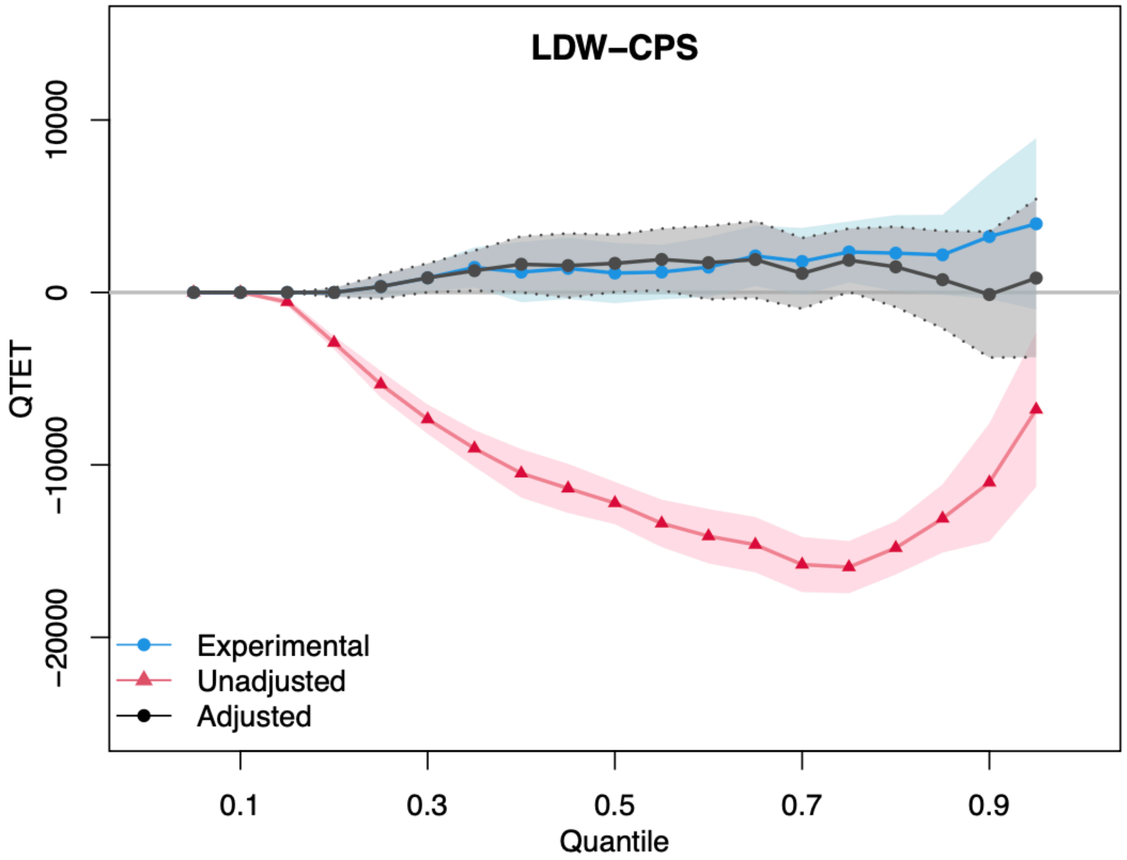


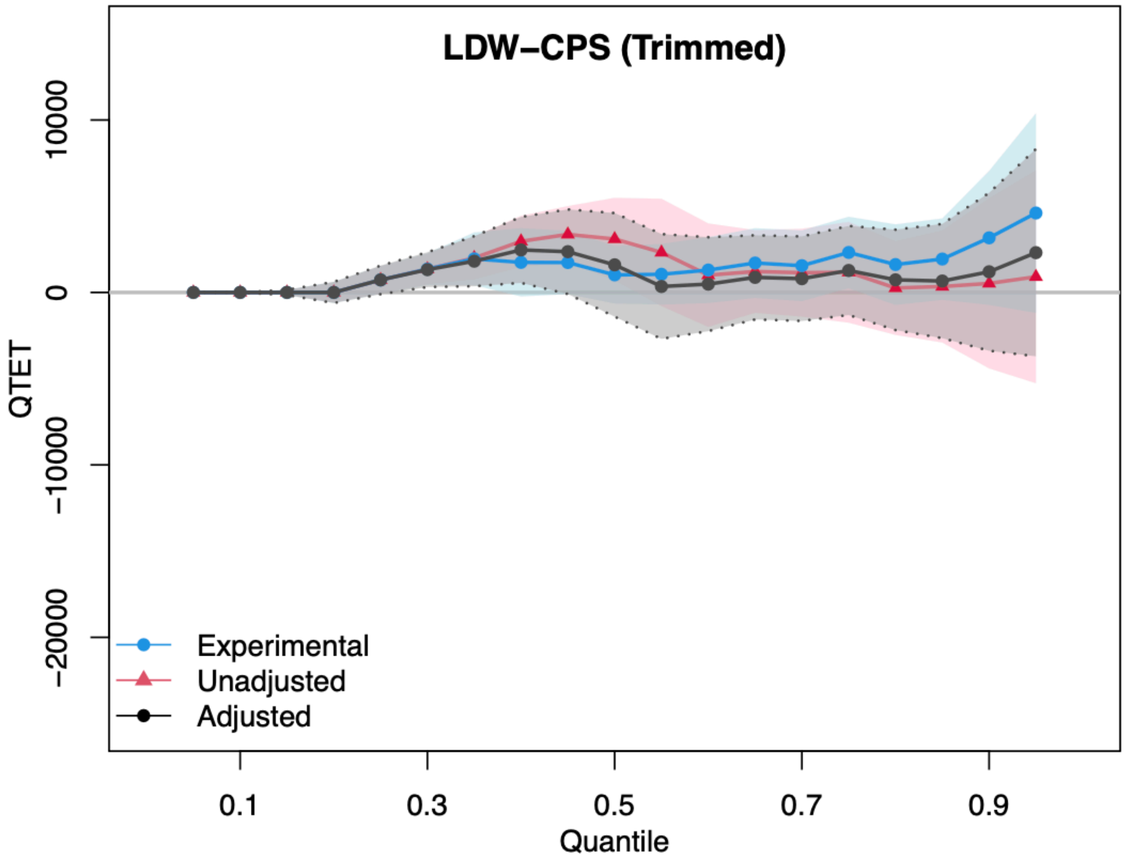
C. LDW-CPS Trimmed



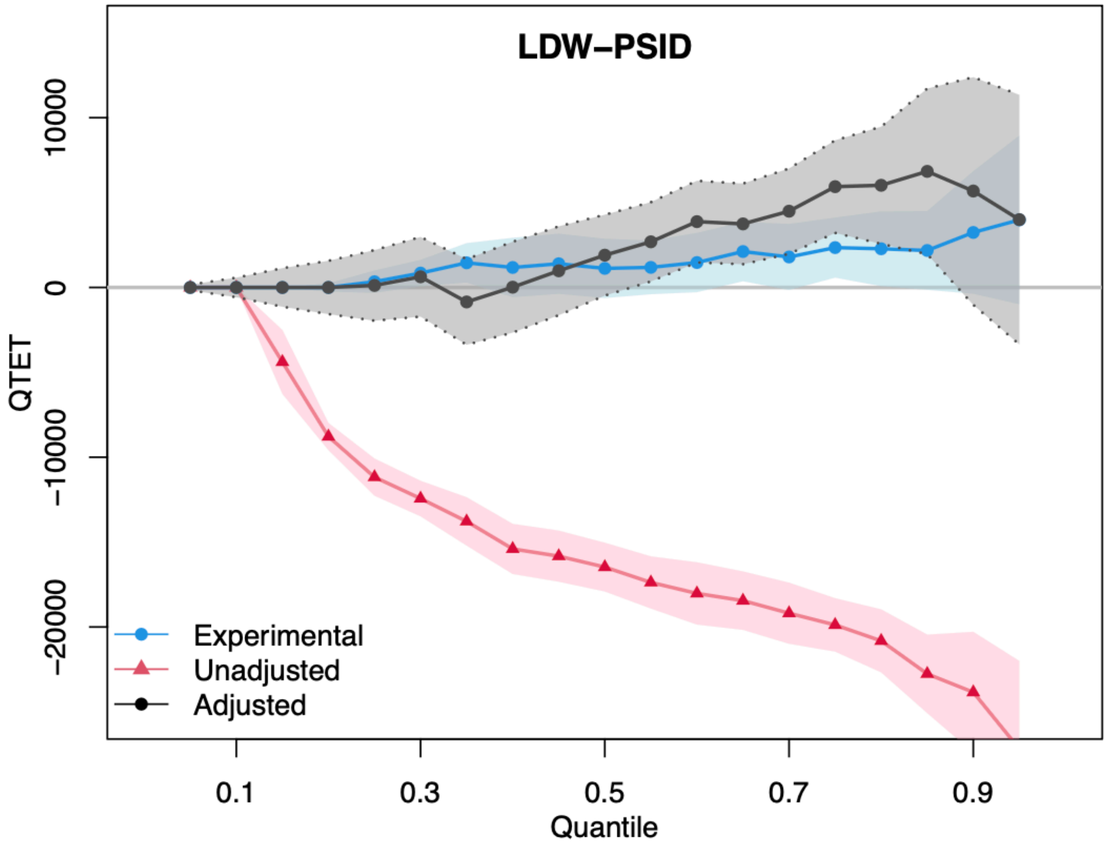
D. LDW-PSID Trimmed

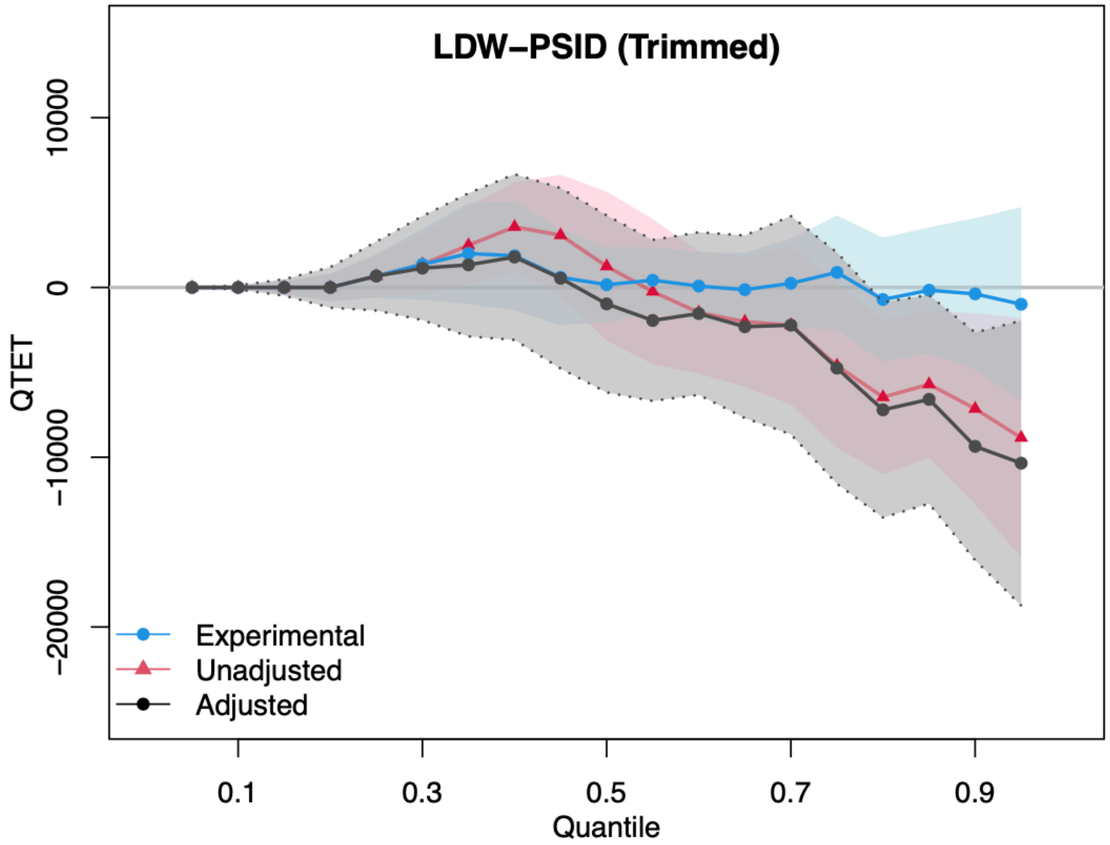
Quantile Treatment Effects — LDW-CPS



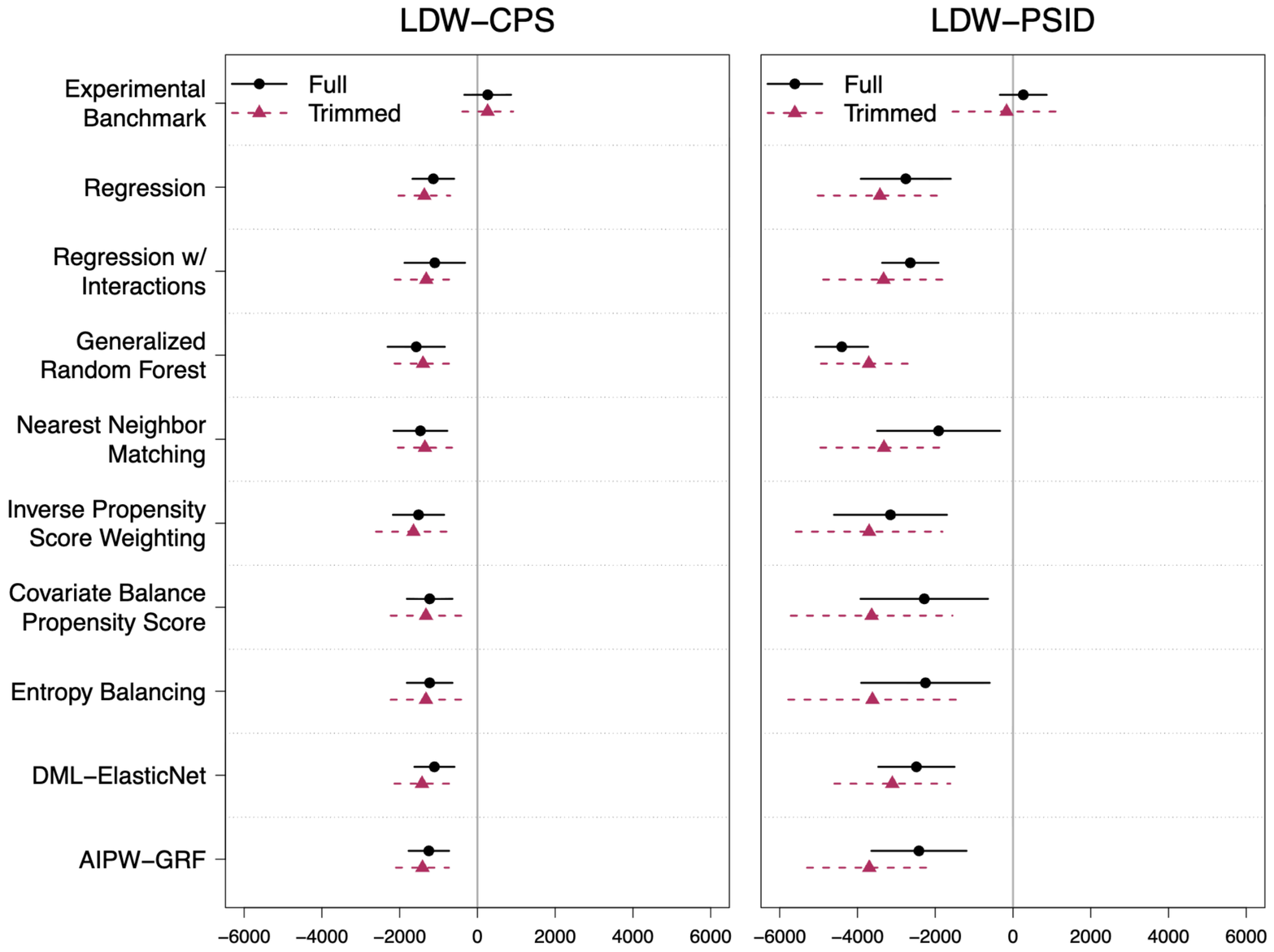


Quantile Treatment Effects — LDW-PSID

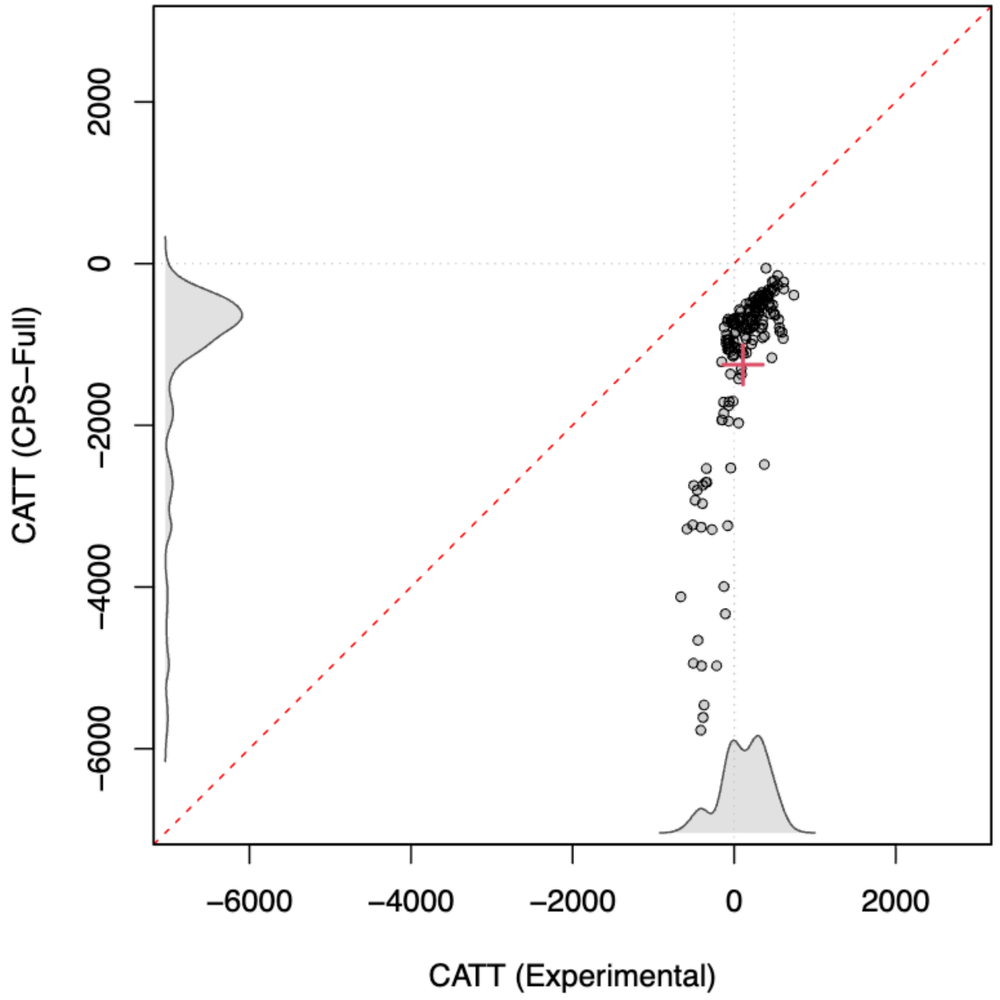




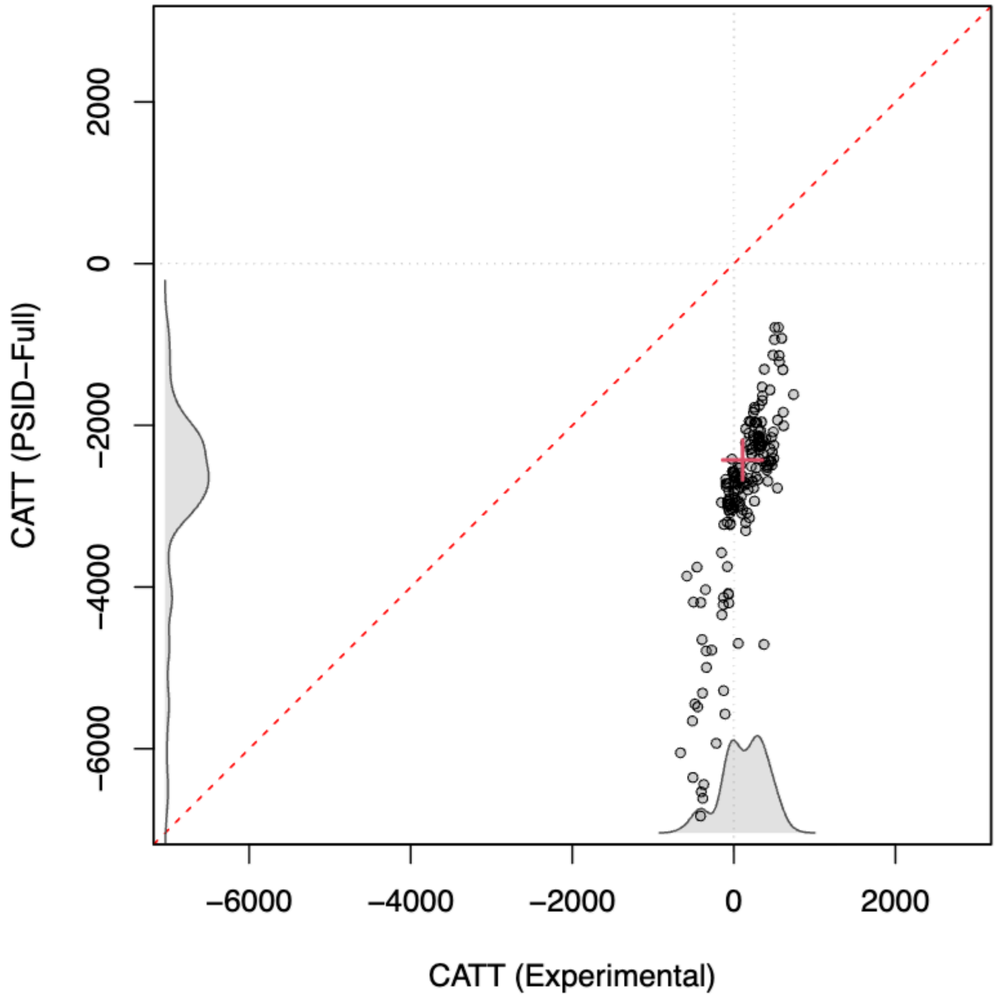
Placebo Estimates ('75 Earnings)



CATE Estimates for the Placebo Outcome ('75 Earnings)



A. LDW-CPS



B. LDW-PSID

Imbens-Rubin-Sacerdote (2001): Lottery Study

- Imbens, Rubin & Sacerdote (IRS 2001) were interested in estimating the effect of unearned income on economic behavior including effects on labor supply, consumption and savings.

- They surveyed individuals who had played and won large sums of money in the lottery — big winners ($\geq$\$100,000) and small winners (\$5,000-\$99,000). As a comparison group they collected data on a second set of individuals who also played the lottery but who had not won big prizes ("losers").

- Data. 43 big winners; 194 small winners, 259 "losers."

- Covariates: the year individuals played the lottery, the number of tickets they typically bought, age, sex, education, and their Social Security earnings for the six years before their winning

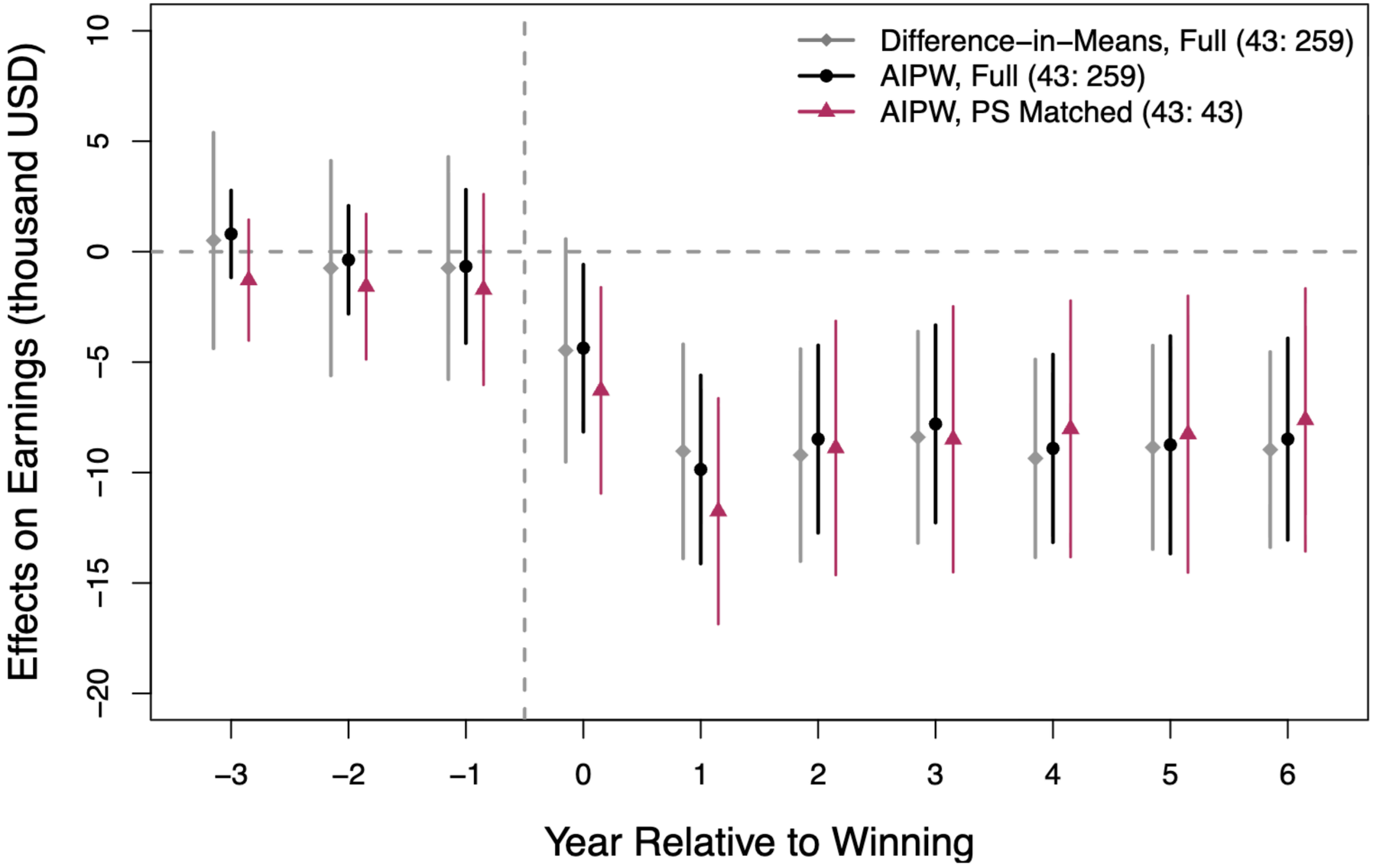
  - We keep earnings in -3 to -1 years as placebo outcomes

Imbens-Rubin-Sacerdote (2001): Lottery Study

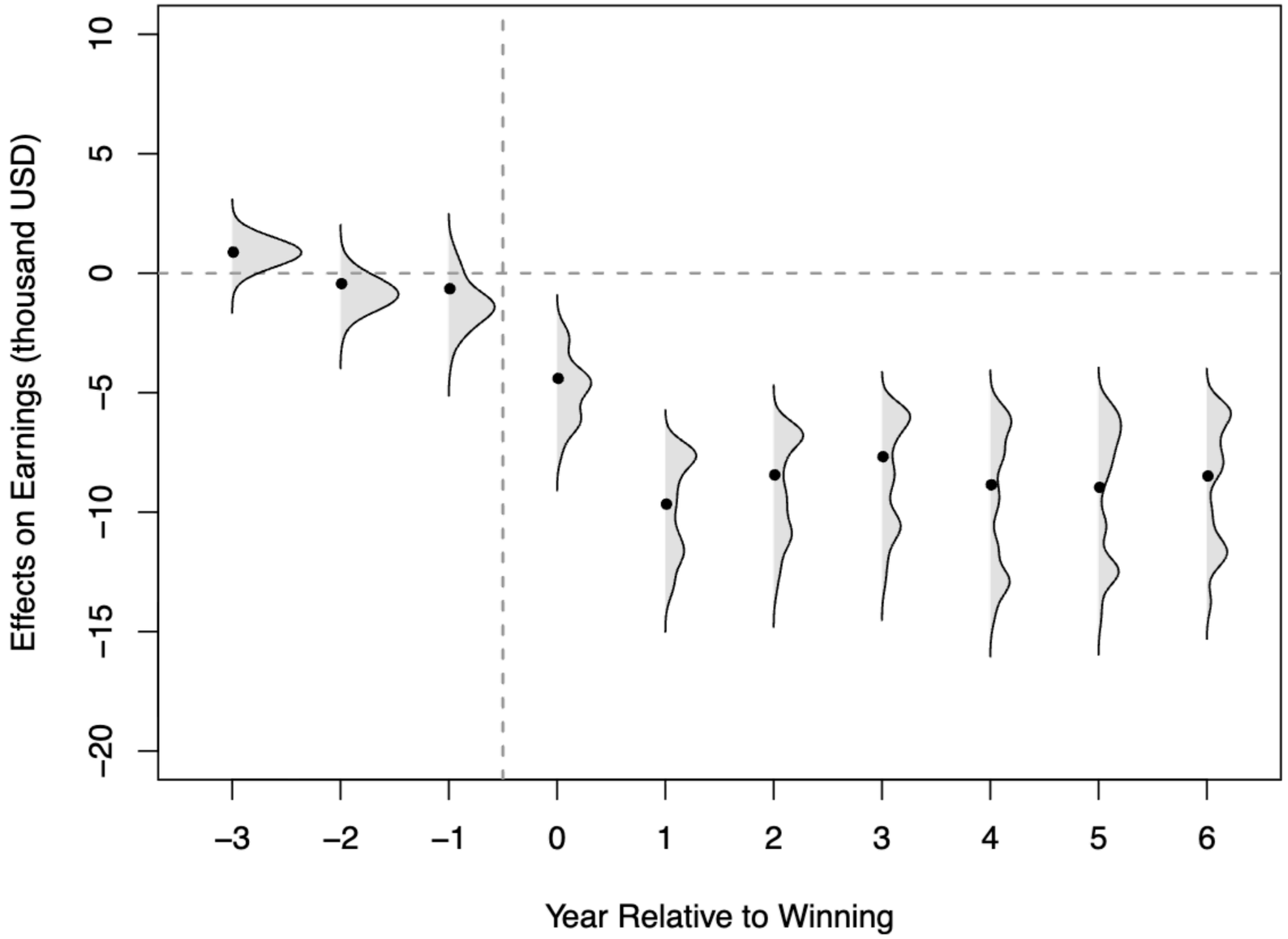
- Substantial differences between winners and "losers," including in pre-winning earnings.
- How come?
  - People buying more tickets are more likely to win.
  - Nonresponse may differ by prize and individual characteristics (including labor income)
- Is unconfoundedness plausible?

Variable	All N=496		Losers N_t=259	Winners N_c=237		Norm.
	mean	(s.d.)	mean	mean	[t-stat]	Dif.
Year Won	6.23	1.18	6.38	6.06	-3.0	-0.27
# Tickets	3.33	2.86	2.19	4.57	9.9	0.90
Age	50.2	13.7	53.2	47.0	-5.2	-0.47
Male	0.63	0.48	0.67	0.58	-2.1	-0.19
Education	13.73	2.20	14.43	12.97	-7.8	-0.70
Working Then	0.78	0.41	0.77	0.80	0.9	0.08
Earn Y -6	13.8	13.4	15.6	12.0	-3.1	-0.27
⋮						
Earn Y -1	16.3	15.7	18.0	14.5	-2.5	-0.23
Pos Earn Y-6	0.69	0.46	0.69	0.70	0.3	0.03
⋮						
Pos Earn Y-1	0.71	0.45	0.69	0.74	1.2	0.10

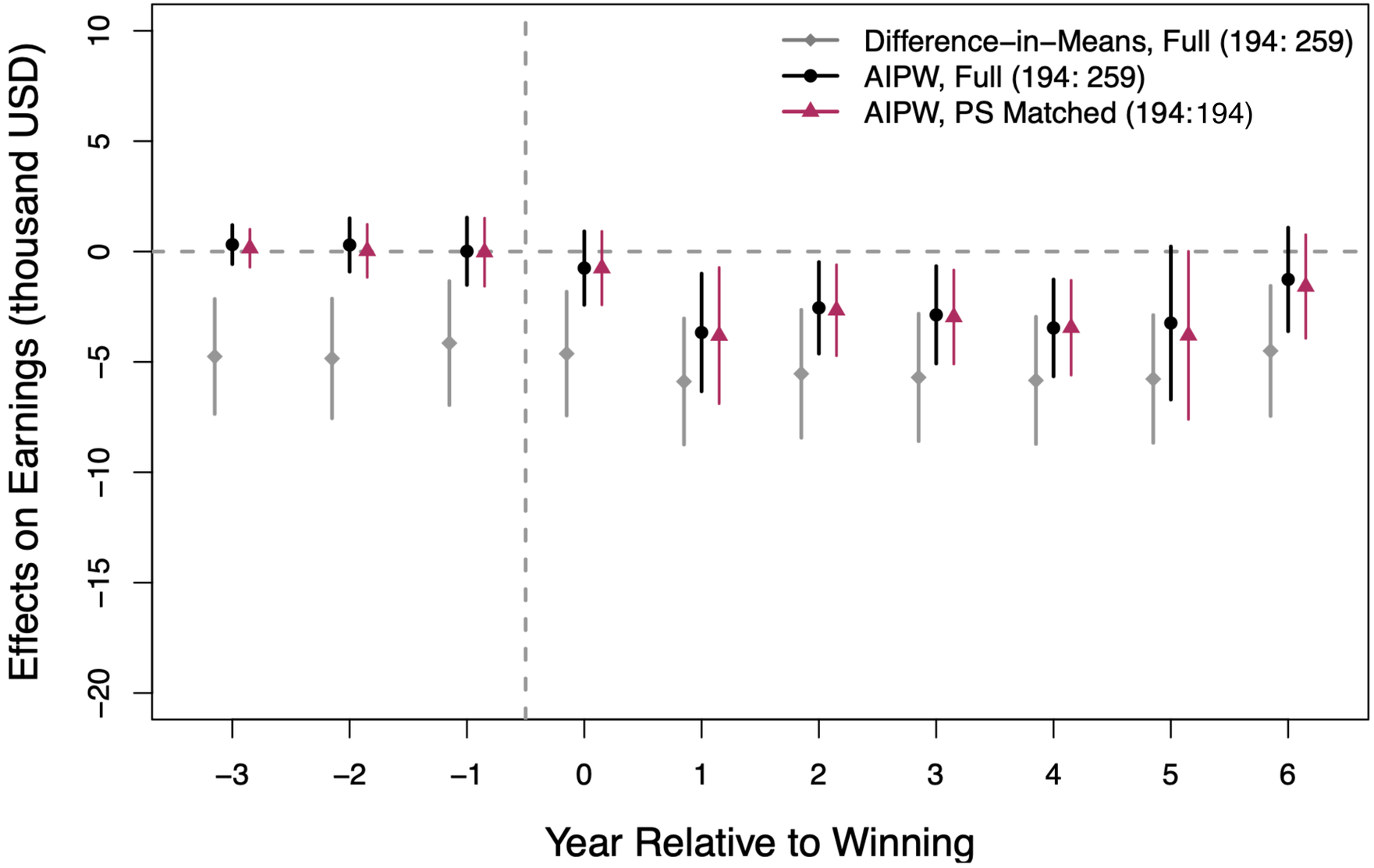
ATT for Big Winners



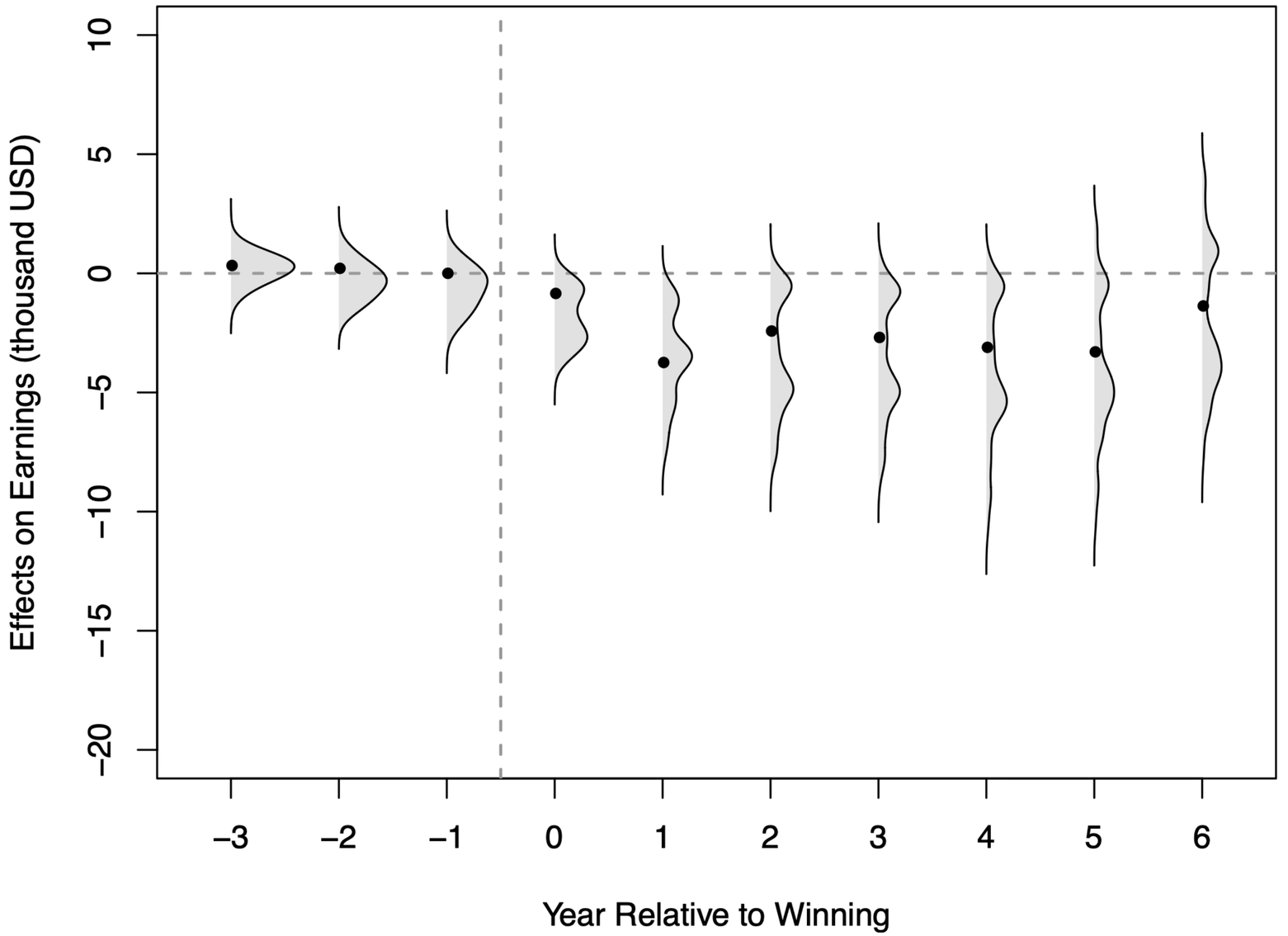
CATT for Big Winners



ATT for Small Winners



CATT for Small Winners



Conclusion (Lessons)

- Observational research, especially in cross-sectional settings, often rests on unconfoundedness and overlap assumptions.
- Unconfoundedness is a strong and inherently untestable assumption; it requires validation using auxiliary information.
- Overlap can be assessed and improved at the design phase.
- With sufficient overlap, a wide range of modern covariate adjustment methods identify a statistical estimand, which can be interpreted as causal when unconfoundedness holds.
- Additional estimands, such as CATE and QTE, can further inform decision making.

- Thank you!