Balancing Methods for Observational Studies

Chen Qiu

Cornell Economics

Data Analytics Colloquium, December 1|2, 2022

References

- Qiu, C. (2022). Approximate minimax estimation of average regression functionals. working paper
- Qiu, C., & Otsu, T. (2022). Information theoretic approach to high-dimensional multiplicative models: Stochastic discount factor and treatment effect. Quantitative Economics, 13(1), 63-94.

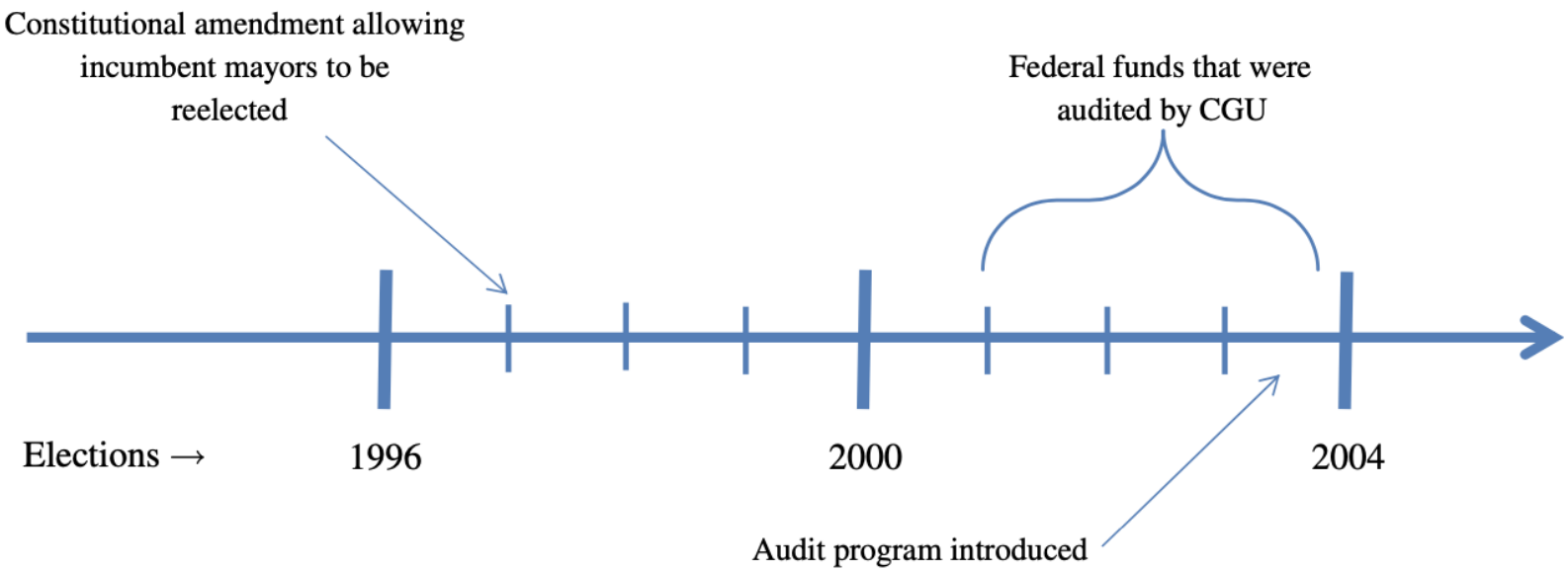
Motivation: empirical analyses with many controls

- Empirical studies in economics often involve observational data with many controls
  - Evaluation of job training programs in labor economics (Heckman et al. 97, 98)
  - Study of long term impact of cash transfer to poor families in economic development (Aizer et al. 16)
  - Explanation of gender gap in earnings for professionals working in finance (Bertrand et al. 10)
- Yet, economic theory is rarely informative regarding which controls or technical terms should be used
- Common practice: estimating parameters using different subsets of controls (Oster, 19) and hoping estimated effects are "stable"

Many controls cause problems: a case in point

- Study of electoral accountability and corruption is important in political economy (Besley and Case, 95; Besley, 06; List and Sturm, 06)
- Ferraz and Finan (11, *AER*) find mayors serving first term are less corrupt
- They exploit a natural experiment where treatment is arguably randomly assigned
- One of the main empirical strategies is OLS with many controls

Basic timeline (Figure 1, Ferraz and Finan, 11)



Basic empirical framework

- Main specification

$$Y = \theta_0 D + X'\varphi + Z'\gamma + \varepsilon$$

where

Y : share of resources related to corrupt activities, collected from audit reports

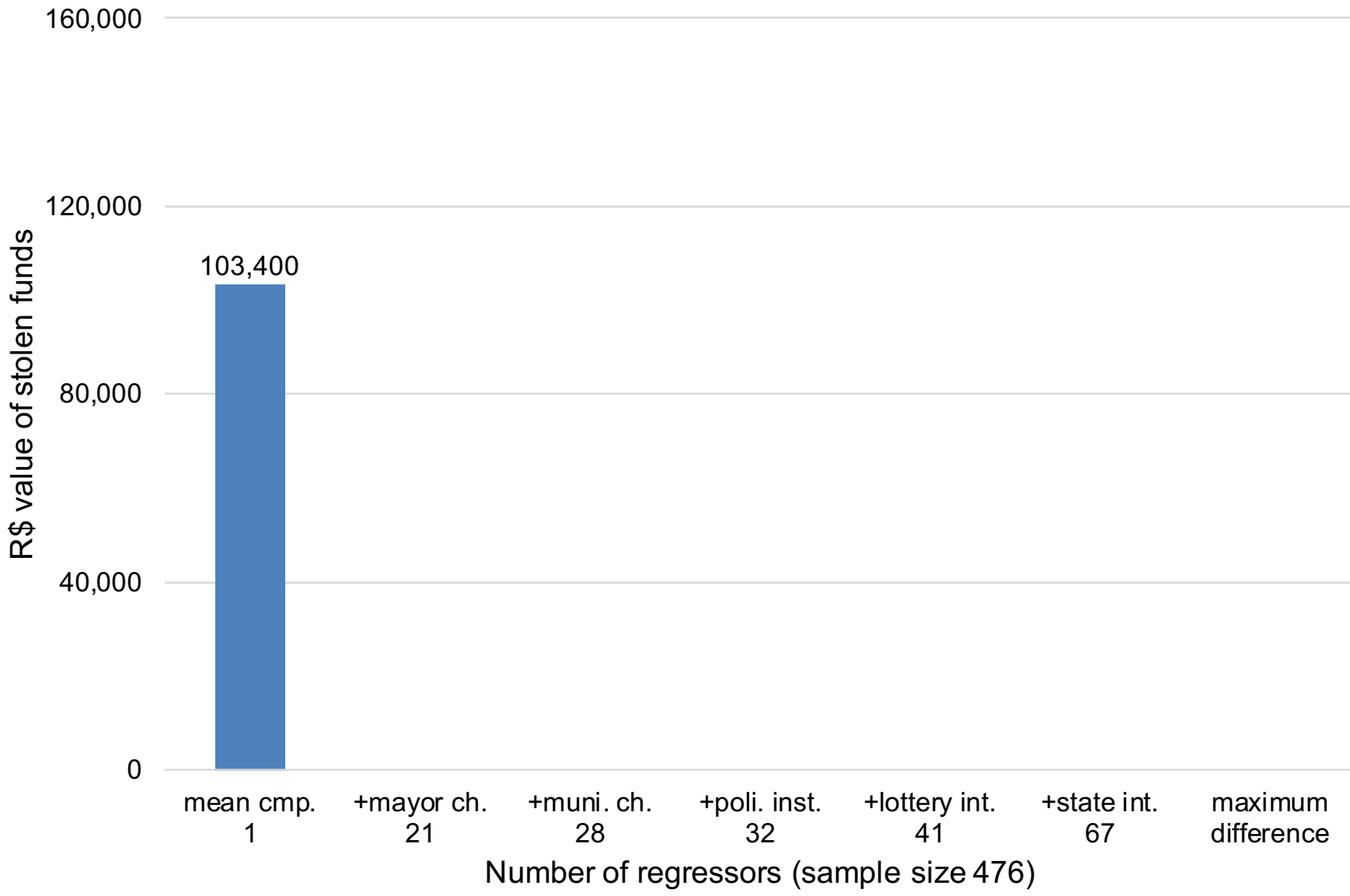
- Treatment

$$D = \begin{cases} 1 & \text{if term not binding(with reelection incentives)} \\ 0 & \text{if term binding(with no reelection incentives)} \end{cases}$$

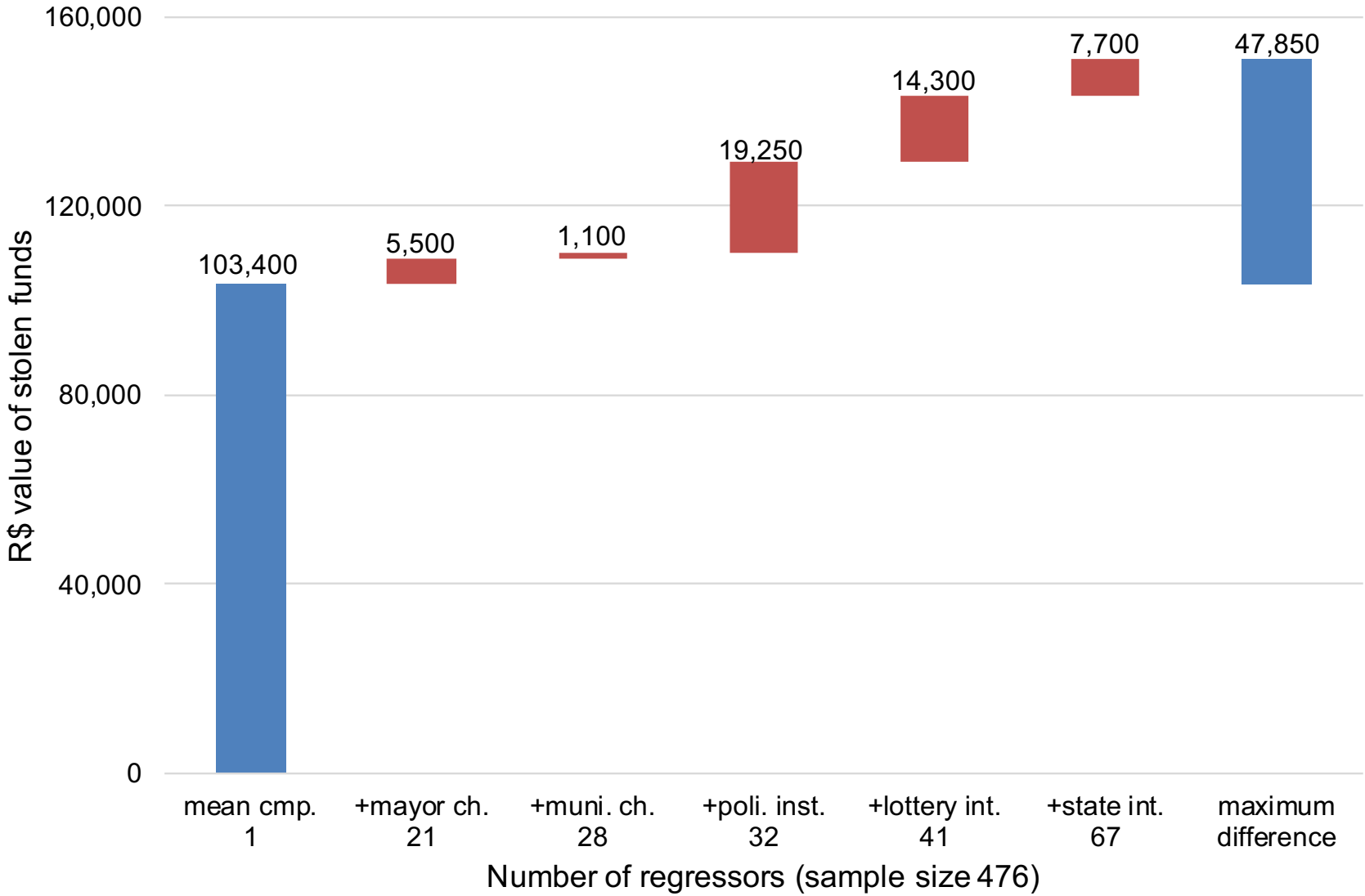
- Sequentially add different sets of controls

X : municipal characteristics
Z : mayoral characteristics

Estimated treatment effect: OLS



Estimated treatment effect: OLS



What happens in this dataset?

1. **Treatment is arguably randomly assigned**...
2. **But point estimates change considerably**
   - estimated treatment effect increases 46.3%
   - robust standard error increases 19.0%

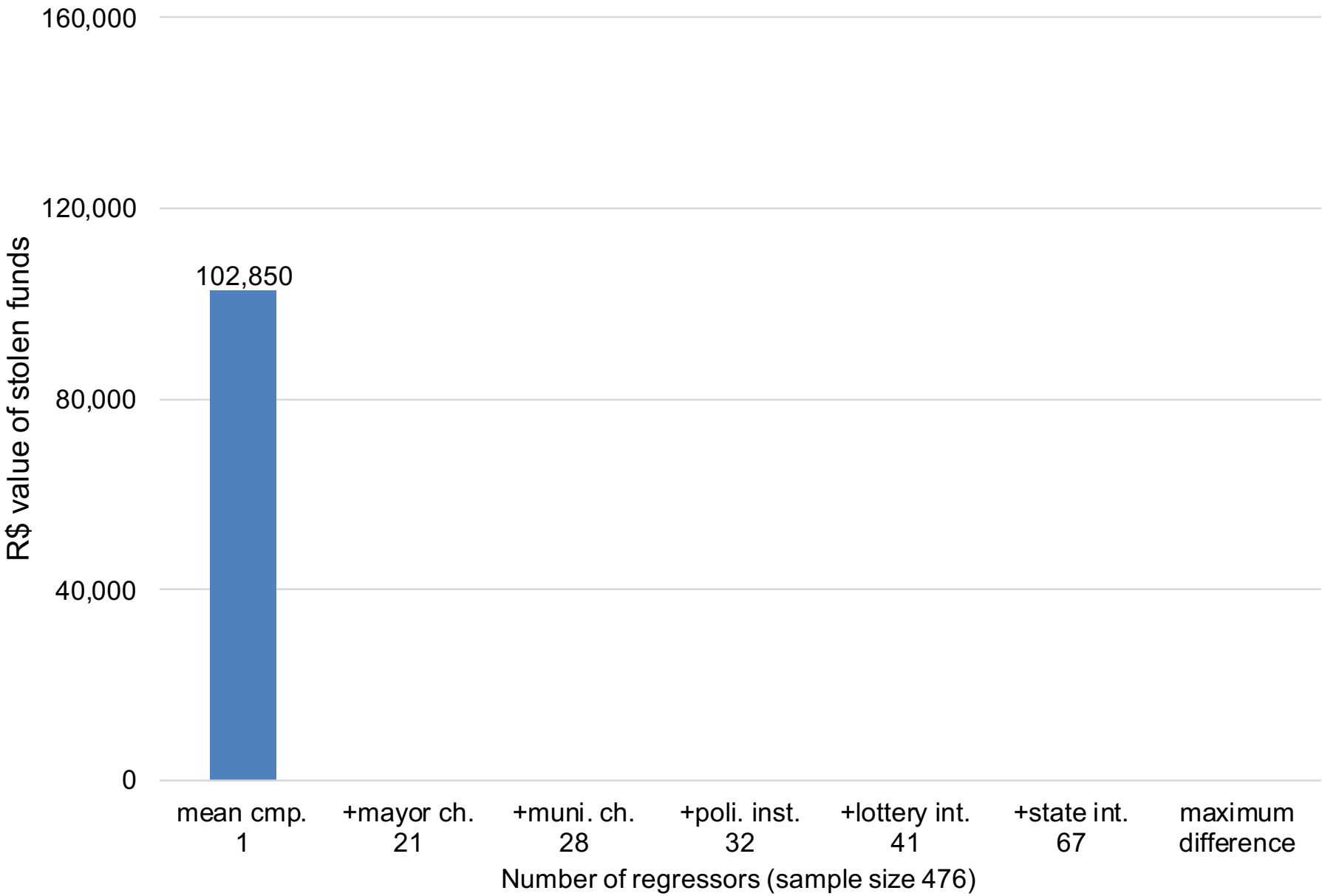
How do we justify these results?

- **Common interpretation: OLS correctly controls omitted variable bias (OVB)**
  1. ignoring added controls underestimates the effect of reelection incentives on corruption
  2. adding more controls partially corrects OVB, thereby providing estimates that more closely approximate the true effect
- **However, such interpretation is unsatisfactory**
  - the sign of the OVB is far from clear
  - a credible natural experiment should contain mild selection bias

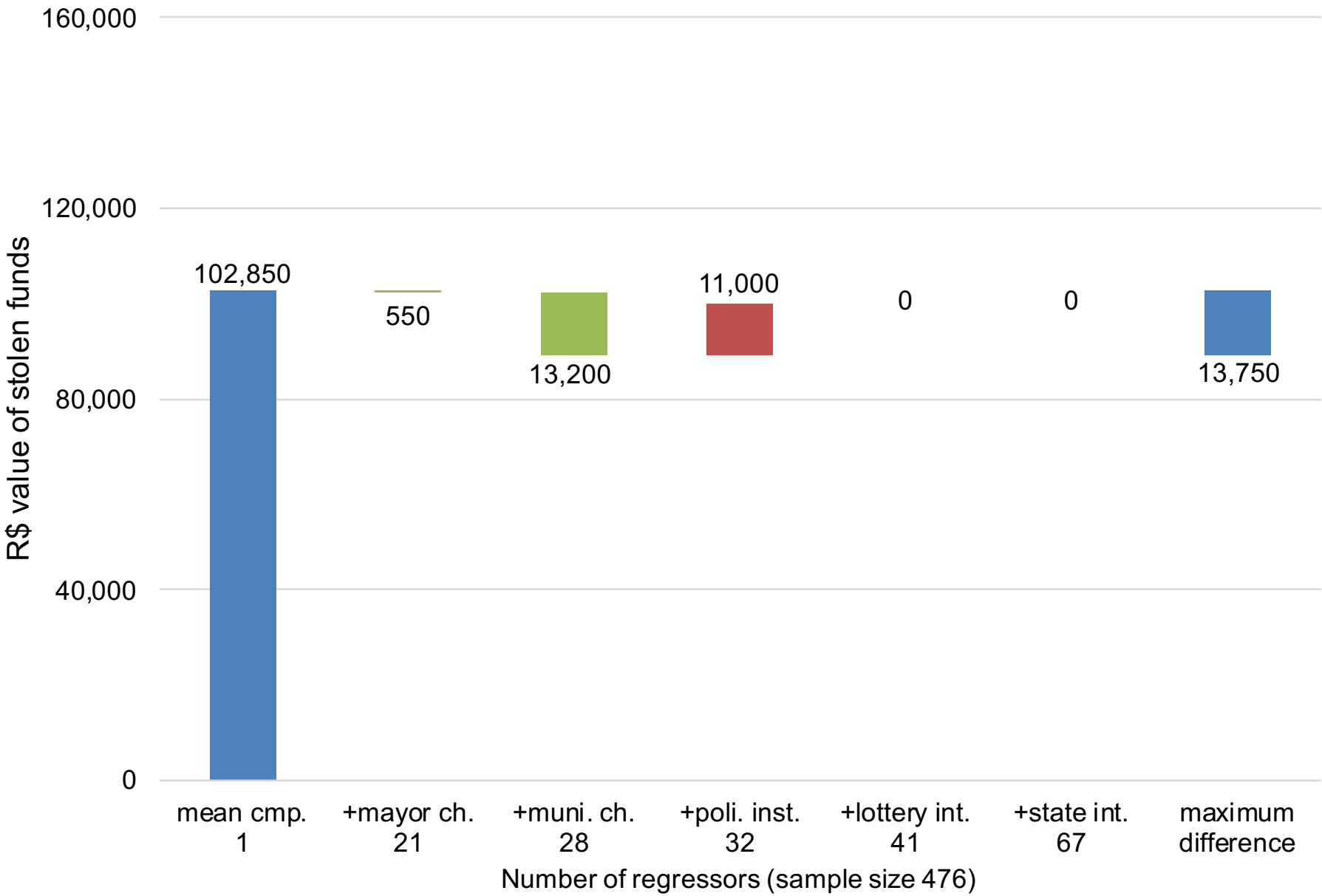
Robust estimation of causal effect with many controls

- Can we estimate a causal effect that is not too sensitive to the use of many different controls?
  - An ideal estimator should produce relatively stable estimates, if the dataset is indeed a good natural experiment with little bias
  - If, in contrast, the dataset features severe bias, adding important controls should correctly reduce the estimation error
- In this talk, we show a new variant of the balancing method that may achieve these goals better than other modern approaches

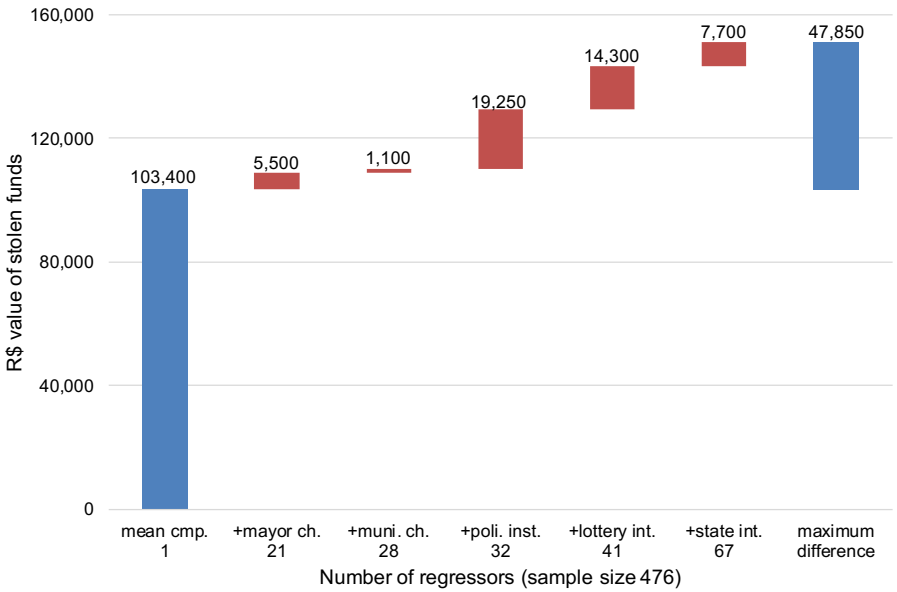
Estimated treatment effect: the new balancing estimator

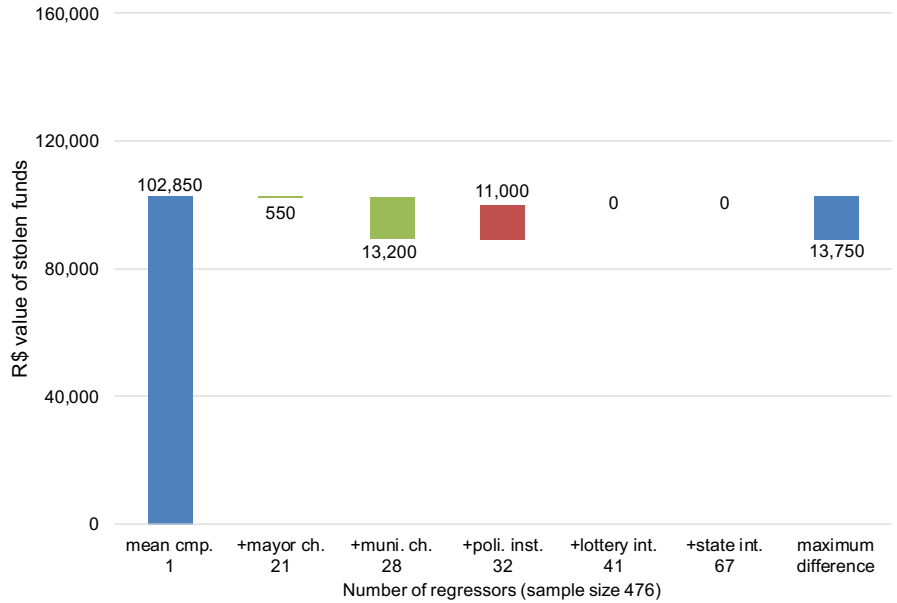


Estimated treatment effect: the new balancing estimator



OLS vs the new balancing estimator





- My interpretation for OLS: with many regressors, OLS:
  - fails to control mean square error (MSE) effectively
  - can cause many empirical results to be misinterpreted

What has the new estimator done differently?

- Imagine a researcher who is contemplating a set of controls, including technical terms, in their analysis
- Given a rich set of conditioning terms, they are willing to assume unconfoundedness
- Our new balancing estimator aims to mitigate the MSE resulting from the presence of many controls or technical terms
    - it acknowledges the need to trade off the finite sample MSE optimally
    - is also asymptotically efficient
- It is very easy to implement

- We call the new estimator as a MSE-optimal balancing estimator
- We find
  - it behaves more robustly for the dataset of Ferraz and Finan (2011)
  - it outperforms other modern approaches in terms of MSE in a variety of simulations
- We believe
  - it is a more robustly estimator
  - is a more suitable estimator for moderately high-dimensional datasets

Rest of the talk

- Review
  - Potential outcome model and average treatment effect
  - Simple difference-in-means estimator
  - Matching
- New method
  - Balancing
  - MSE-optimal balancing
- More empirical and simulation evidence

Potential outcome model and average treatment effect

Rubin's potential outcome model

- A sample of n units from a large population
- For each unit $i = 1 \ldots n$:
  - $D_i \in \{0, 1\}$ is treatment indicator: $D_i = 1$ means treated; $D_i = 0$ means untreated (control)
  - $Y_i(0)$ is potential outcome when i is not treated
  - $Y_i(1)$ is potential outcome when i is treated
- $Y_i(0)$ and $Y_i(1)$ are **counterfactuals**: both potentially observable before treatment but only one of them is observed after treatment

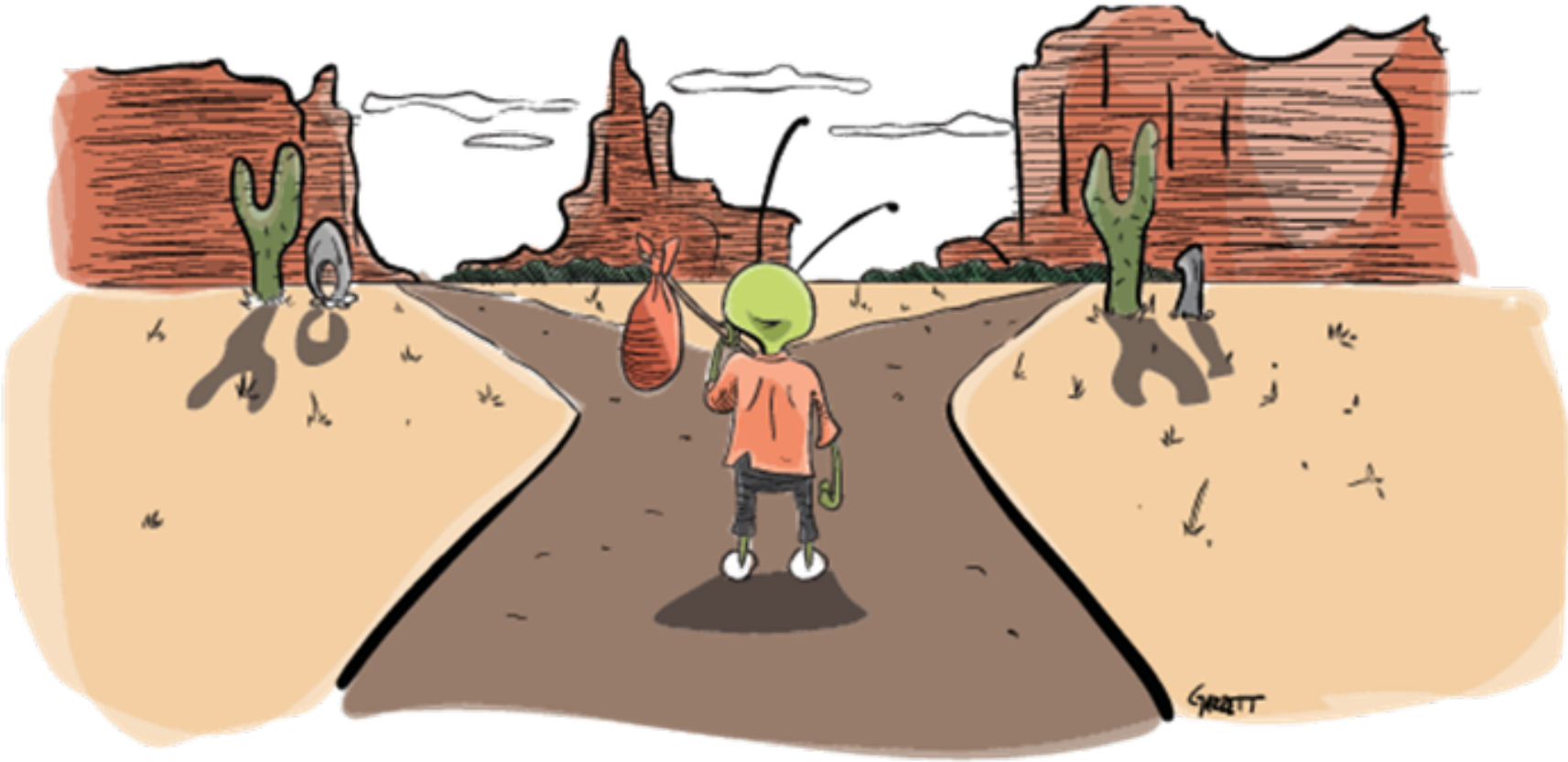


Figure: Two potential outcomes

SUTVA

- Throughout the talk, we maintain the Stable Unit Treatment Value Assumption (SUTVA):
  - The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.
- In short: no interference and no hidden variations of treatments
- Strong assumption

Individual treatment effect

- Individual treatment effect is $\theta_i = Y_i(1) - Y_i(0)$
- Observed outcome is

$$\begin{aligned} Y_i &= \begin{cases} Y_i(0) & \text{if } D_i = 0 \\ Y_i(1) & \text{if } D_i = 1 \end{cases} \\ &= D_i Y_i(1) + (1 - D_i) Y_i(0) \end{aligned}$$

- θ_i can never be learnt. Learning about *some* causal effect requires multiple units

Some treatment effect can be learnt

- Some treatment effects among certain population groups can be learnt. For example
  - Average treatment effect (focus of this talk)

$$\text{ATE} = \mathbb{E}[\theta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

  - Average treatment on the treated

$$\text{ATT} = \mathbb{E}[\theta_i | D_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$$

  - Local average treatment effect
  - Conditional average treatment effect

$$\text{CATE} = \mathbb{E}[\theta_i | X_i = x] = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$$

    where X_i is some observable characteristics of unit i

Inherent difficulty of measuring a causal effect

- Even with multiple units, there are still inherent difficulty of measuring a causal effect
  1. The assignment mechanism of D is unknown in a missing data problem
  2. individual heterogeneity: θ_i could be different from θ_j for $i \neq j$

Example: the importance of the assignment mechanism

- Suppose patients are sent to an emergency room for medical treatment
- Two treatments available: surgery (1), or drugs (0)

Patients	Potential outcomes		Individual treatment effect
	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	1	7	6
2	6	5	-1
3	1	5	4
4	8	7	-1
Average	4	6	2

- Suppose doctors know each individual's optimal treatment and apply the optimal treatment to each patient

Patients	Treatment assignment	Observed outcome
	D_i	Y_i
1	1	7
2	0	6
3	1	5
4	0	8
Average difference between two treatments		$\frac{7+5}{2} - \frac{6+8}{2} = -1$

Example: the importance of outcome heterogeneity

- It is common for practitioners to run a regression

$$Y = \theta_0 D + \beta' X + \varepsilon$$

  and report the estimated coefficient $\hat{\theta}$ for θ_0

- Shown by Angrist (98), $\hat{\theta}$ is **not** estimating the ATE, but a weighted version of the treatment effect

- That is, $\hat{\theta}$ is estimating

$$\int \text{CATE}(x) w(x) dx$$

  where $w(x)$ is proportional to $\text{var}[D|X = x]$

- Exceptions:
  - $Y_i(1) - Y_i(0) =$ constant for all i
  - X are all indicator functions and they partition the whole population (i.e., estimate using a full set of interactions)

Simple difference-in-means estimator

Difference-in-means estimator

- Suppose we are interested in estimating the ATE

$$\text{ATE} = \mathbb{E}[\theta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

- We observe

$$\begin{aligned} D &\in \{0, 1\} && \text{binary treatment} \\ X &\in \mathbb{R}^k && \text{pretreatment covariate vector} \\ Y &= DY_1 + (1 - D)Y_0 && \text{observed outcome} \end{aligned}$$

- The simple difference-in-means estimator for ATE is

$$\hat{\theta}_{\text{diff}} = \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n (1 - D_i) Y_i}{\sum_{i=1}^n (1 - D_i)}$$

Example: last six obs. from Ferraz and Finan (11)

Unit	Potential outcomes		Treatment	Observed outcomes
	$Y_i(1)$	$Y_i(0)$	D_i	$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$
1	?	4.97	0	4.97
2	?	0	0	0
3	0	?	1	0
4	?	1.81	0	1.81
5	?	1.46	0	1.46
6	0	?	1	0
Average	0	2.06		
$\hat{\theta}_{\text{diff}}$		-2.06		

Note: Question marks represent missing potential outcomes

Theoretic property of $\hat{\theta}_{\text{diff}}$

- By law of large numbers

$$\frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} = \frac{\sum_{i=1}^n D_i Y_i(1)}{\sum_{i=1}^n D_i} \xrightarrow{p} \frac{\mathbb{E} D_i Y_i(1)}{\mathbb{E} D_i} = \mathbb{E}[Y_i(1)|D_i = 1]$$

$$\frac{\sum_{i=1}^n (1 - D_i) Y_i}{\sum_{i=1}^n (1 - D_i)} = \frac{\sum_{i=1}^n (1 - D_i) Y_i(0)}{\sum_{i=1}^n (1 - D_i)} \xrightarrow{p} \frac{\mathbb{E}(1 - D_i) Y_i(0)}{\mathbb{E}(1 - D_i)} = \mathbb{E}[Y_i(0)|D_i = 0]$$

- Thus,

$$\hat{\theta}_{\text{diff}} \xrightarrow{p} \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$$

  while what we want is

$$\text{ATE} = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

- If $(Y_i(1), Y_i(0))$ is independent of D_i, i.e., treatment is randomly assigned, then
  - $\mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i(1)|D_i = 1]$, $\mathbb{E}[Y_i(0)] = \mathbb{E}[Y_i(0)|D_i = 0]$
  - $\hat{\theta}_{\text{diff}}$ is a consistent and unbiased estimator of ATE
- However, for observational data, treatment assignment mechanism is rarely known and usually not random
  - $\hat{\theta}_{\text{diff}}$ is usually biased and inconsistent
  - for $\mathbb{E}[Y_i(1)]$, the selection bias is $\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(1)]$
  - for $\mathbb{E}[Y_i(0)]$, the selection bias is $\mathbb{E}[Y_i(0)|D_i = 0] - \mathbb{E}[Y_i(0)]$

Matching

Matching: a motivation

- Selection bias is related to group differences
  - For example, to measure $\mathbb{E}[Y_i(1)]$, we only have data for those who are getting treated
  - An average person from those treated can be very different from an average person from the whole population
  - e.g., in Ferraz and Finan (11), a mayor serving first term might be less educated than an average mayor in the whole population

- If selection is only related to observable variables, we can control selection bias by comparing similar individuals
    - **Unconfoundedness**: Treatment is randomly assigned (chosen) among units that are observationally identical in terms of pretreatment observable covariates X
- To estimate ATE, we can compare the outcomes of the individuals who look identical in terms of X, but differ in their treatment status
- One way of doing this is by matching

Example: last six obs. from Ferraz and Finan (11)

Unit	Potential outcomes		Treatment	Observed outcomes	Schooling level
	$Y_i(1)$	$Y_i(0)$	D_i	Y_i	X_i
1	?	4.97	0	4.97	6
2	?	0	0	0	8
3	0	?	1	0	6
4	?	1.81	0	1.81	2
5	?	1.46	0	1.46	8
6	0	?	1	0	6
Average					
$\hat{\theta}_{\text{matching}}$					

Simple illustration: matching with one covariate

Unit	Potential outcomes		Treatment	Observed outcomes	Schooling level
	$Y_i(1)$	$Y_i(0)$	D_i	Y_i	X_i
1	(0)	4.97	0	4.97	6
2	(0)	0	0	0	8
3	0	(4.97)	1	0	6
4	(0)	1.81	0	1.81	2
5	(0)	1.46	0	1.46	8
6	0	(4.97)	1	0	6
Average	0	3.03			
$\hat{\theta}_{\text{matching}}$		-3.03			

Key insight

- Matching adjusts observed outcomes so that the treatment and control groups are more comparable
- For example, for the estimation of $\mathbb{E}[Y_i(0)]$:
  - If we use simple sample mean (which estimates $\mathbb{E}[Y_i(0)|D_i = 0]$)

$$\hat{\theta} = \frac{4.97 + 0 + 1.81 + 1.46}{4}$$

    each observed outcome gets equal weight $\frac{1}{4}$

  - If we use matching

$$\begin{aligned}\hat{\theta}_{\text{matching}} &= \frac{3 \cdot 4.97 + 0 + 1.81 + 1.46}{6} \\ &= \frac{3 \cdot 4.97 + 0 + 1.81 + 1.46}{6}\end{aligned}$$

    4.97 gets higher weight because there are more units with schooling level 6 in the whole population than in the untreated population

- Distribution of X in the controlled sample ($D_i = 0$)

X_i	2	6	8
Probability	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{4}$

- Distribution of X in the whole sample ($D_i = 0 + D_i = 1$)

X_i	2	6	8
Probability	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$

- Average observed outcomes by a weight that resembles the target population

General matching with more than one covariate

- Take i-th unit. If $D_i = 1$, then we observe $Y_i = Y_i(1)$. We only need to impute $Y_i(0)$
- To impute $Y_i(0)$, we find units similar to i in untreated group based on the value of X_i. Then impute

  $\hat{Y}_i(0) =$ average of $Y_j's$ in untreated group with $\|X_j - X_i\| \leq \varepsilon$

  for some fixed $\varepsilon \geq 0$
- Similarly, if $D_i = 0$, we observe $Y_i(0)$ and impute

  $\hat{Y}_i(1) =$ average of $Y_j's$ in treated group with $\|X_j - X_i\| \leq \varepsilon$
- Estimated ATE is

$$\hat{\theta}_{\text{matching}} = \frac{1}{n}\left\{\sum_{i:D_i=1}(Y_i(1) - \hat{Y}_i(0)) + \sum_{j:D_j=0}(\hat{Y}_j(1) - Y_j(0))\right\}$$

- Usually choose the distance measure $\|\cdot\|$ as the Mahalanobis distance

$$\|X_i - X_j\|_M = \sqrt{(X_i - X_j)' V^{-1} (X_i - X_j)}$$

  where V is the covariance matrix of random vector X

- M-nearest neighbor covariate matching: for each unit, to impute the opposite potential outcome, use a fixed M number of units that are closest in terms of metric $\|\cdot\|$ in the opposite treatment group

- Nearest neighbor matching: 1-nearest neighbor covariate matching

Balancing

Balancing: a motivation

- Causal inference is a missing data problem
- Consider the estimation of average treatment effect

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

- From now on, focus on

$$\theta = \mathbb{E}[Y(1)]$$

- θ is concerned with the mean of $Y(1)$ in the **whole population**
- But we only have information about $Y(1)$ in the **treated population**
- Balancing is a more flexible and more effective way to make the treated population and the whole population more comparable (thus removing selection bias)

From matching to balancing

- Recall the simple sample mean estimator for $\mathbb{E}[Y(1)]$ is

$$\hat{\theta} = \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i} = \sum_{i=1}^{n} \frac{D_i}{\sum_{i=1}^{n} D_i} Y_i$$

$$= \frac{1}{n} \sum_{i=1}^{n} D_i \underbrace{\frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i}}_{\text{weight}} Y_i$$

- Each observed outcome in the treated group gets the same weight
- $\hat{\theta}$ only works when D is randomly assigned (i.e., there is no difference between the treated group and the whole population)
- Matching adjusts the weight $\frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i}$ by a different weight that resembles the distribution of X in the whole population

Key idea of balancing

- Select an ideal set of weight functions $\{w(X_i)\}_{i=1}^n$ to estimate $\mathbb{E}[Y(1)]$ by

$$\frac{1}{n}\sum_{i=1}^{n} D_i \frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i} w(X_i) Y_i$$

  where $\{w(X_i)\}_{i=1}^n$ is chosen so that

  distribution of X in the treated population after adjustment
  =distribution of X in the whole population (1)

- If (1) is met, then we say the distribution of X is balanced
- In plain words: adjust the outcome by a set of weights that make the treated and the whole population indistinguishable

Achieve balancing by equalizing the mean

- We can usually achieve balancing by equalizing the mean of the covariates in the treated and the whole population
- Before balancing

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\text{sample mean of controls in the whole population}} \neq \underbrace{\frac{1}{n}\sum_{i=1}^{n} D_i \frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i} X_i}_{\text{sample mean of cntrols in the treated}}$$

- Choose $w(X_i)$ to force balancing

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\text{sample mean of controls in the whole population}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} D_i \frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i} w(X_i) X_i}_{\text{adjusted sample mean of controls in the treated}} \tag{2}$$

- Then we can estimate $\mathbb{E}[Y(1)]$ by

$$\frac{1}{n}\sum_{i=1}^{n} D_i \underbrace{\frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i} w(X_i)}_{\text{balancing weight}} Y_i \tag{3}$$

- The balancing weight $\frac{w(X_i)}{\frac{1}{n}\sum_{i=1}^{n} D_i}$ shows up in both (2) and (3)
- Let $\alpha(X_i) = \frac{w(X_i)}{\frac{1}{n}\sum_{i=1}^{n} D_i}$ be the normalized weight
- A more streamlined balancing scheme is to estimate $\mathbb{E}[Y(1)]$ by

$$\frac{1}{n}\sum_{i=1}^{n} D_i \alpha(X_i) Y_i \tag{4}$$

where $\alpha(X_i)$ forces balance

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} D_i \alpha(X_i) X_i \tag{5}$$

- Without further restrictions, there are too many choices of $\{\alpha_i\}_{i=1}^n$ that can satisfy (5)
- Two main (related) approaches to pin down a unique set of weights
  - Information theoretic approach (Qiu and Otsu, 22)
  - Linear approximation approach (Qiu, 22 and this talk)

Linear approximation approach

- At the sample level, we aim to choose a unique set of weight so that

$$\frac{1}{n}\sum_{i=1}^{n}[D_i\alpha_i X_i] = \frac{1}{n}\sum_{i=1}^{n}[X_i]$$

- Suppose α_i is linear such that $\alpha_i(X_i) = a'X_i$ for some a, then it should hold

$$\frac{1}{n}\sum_{i=1}^{n}\left[D_i X_i X_i'\right]\hat{a} = \frac{1}{n}\sum_{i=1}^{n}[X_i]$$

- That is, $\hat{a} = \left(\frac{1}{n}\sum_{i=1}^{n}[D_i X_i X_i']\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}[X_i]$. The estimated balancing weight is

$$\hat{\alpha}(X_i) = \hat{a}'X_i \tag{6}$$

- A balancing estimator for $\mathbb{E}[Y(1)]$ is

$$\frac{1}{n}\sum_{i=1}^{n}D_i\hat{\alpha}(X_i)Y_i$$

Remarks

- Balancing is extremely **simple**
  - If we use linear weight function, the estimator has a simple form: as easy as OLS
- Balancing is extremely **flexible**
  - we might want to balance covariates in terms of other features
  - Let $g(X_i)$ be a vector of functions of X that we wish to balance (e.g., first and second moments)
  - Simply choose weight function so that

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = \frac{1}{n}\sum_{i=1}^{n} D_i \alpha(X_i) g(X_i) \tag{7}$$

- Balancing estimator for $\mathbb{E}[Y(0)]$ (and thus for $\mathbb{E}[Y(1) - Y(0)]$) can be constructed analogously

MSE-optimal balancing

What could go wrong for balancing?

- If we have many covariates, achieving balancing can be difficult and in fact, is not desirable
- Intuition: The coefficient for the balancing weight is

$$\hat{a} = \left(\frac{1}{n} \sum_{i=1}^{n} \left[D_i X_i X_i' \right] \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \left[X_i \right]$$

- When X is high dimensional, $\left(\frac{1}{n} \sum_{i=1}^{n} \left[D_i X_i X_i' \right] \right)^{-1}$ might not exist
- Even if $\left(\frac{1}{n} \sum_{i=1}^{n} \left[D_i X_i X_i' \right] \right)^{-1}$ exists, we might not wish to achieve exact balancing (cf., OLS with many covariates)
- This section: how do we construct a good balancing estimator in the presence of many controls?

Key idea: choose weight to minimize MSE

- Still focus on $\theta_0 = \mathbb{E}[Y_i(1)]$. Consider a class of weightes sample average estimator with some weight $\alpha(X_i)$

$$\hat{\theta}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} D_i \alpha(X_i) Y_i$$

- Instead of choosing $\alpha(X_i)$ to achieve balance, select $\alpha(X_i)$ to minimize the finite sample MSE:

$$\begin{aligned}\text{MSE}_n(\hat{\theta}(\alpha)) &= \mathbb{E}\left[\left(\hat{\theta}(\alpha) - \theta_0\right)^2 | X_i, D_i, i = 1 \ldots n\right] \\ &= \text{bias}_n^2(\hat{\theta}(\alpha)) + \text{var}_n(\hat{\theta}(\alpha))\end{aligned}$$

- We can show

$$\text{bias}_n(\hat{\theta}(\alpha)) \approx \frac{1}{n}\sum_{i=1}^{n}\left[D_i\alpha(X_i)\gamma_1(X_i) - \gamma_1(X_i)\right],$$

$$\text{var}_n(\hat{\theta}(\alpha)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}[D_i\alpha^2(X_i)],$$

  and

$$\gamma_1(x) = \mathbb{E}[Y_i|X_i = x, D_i = 1]$$
$$\sigma^2 = \mathbb{E}[e_i^2|X_i, D_i], e_i = Y_i - \gamma_1(X_i)$$

- Assume σ^2 is known
  - $\text{var}_n(\hat{\theta}(\alpha))$ is a known function of α
  - while $\text{bias}_n(\hat{\theta}(\alpha))$ depends on the unknown γ_1

- Choose α to minimize the worst case MSE

$$\min_{\alpha}\left\{\sup_{\gamma_1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(D_i\alpha(X_i)\gamma_1(X_i)-\gamma_1(X_i)\right)\right]^2+\frac{\sigma^2}{n}\sum_{i=1}^{n}[D_i\alpha^2(X_i)]\right\} \tag{8}$$

when we assume

$$\gamma_1(X_i)=\beta' X_i \text{ for some } \|\beta\|=b<\infty$$

and we use a set of linear weights

$$\alpha(X_i)=a' X_i \text{ for some } a\in\mathbb{R}$$

Simple solution

- The solution of the minimax problem (8) has a simple and insightful characterization

$$\min_{\alpha} \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n}[D_i\alpha_i(X)X_i] - \frac{1}{n}\sum_{i=1}^{n}[X_i] \right\|^2}_{\text{degree of imbalance}} + \underbrace{\lambda_1 \frac{1}{n}\sum_{i=1}^{n}\left[\alpha_i^2(X_i)D_i\right]}_{\text{unstableness of the balancing weight}} \tag{9}$$

  where λ_1 is a penalty coefficient

- Key insight: when X is high dimensional, we should balance only approximately

- (9) has an explicit solution

$$\tilde{\alpha}(X_i) = \tilde{a}' X_i, \qquad \tilde{a} = (\hat{G}\hat{G} + \lambda_1 \hat{G})^- \hat{G}\hat{P}$$

$$\hat{G} = \frac{1}{n}\sum_{i=1}^{n}[D_i X_i X_i'], \qquad \hat{P} = \frac{1}{n}\sum_{i=1}^{n}[X_i]$$

- MSE-optimal balancing estimator is

$$\tilde{\theta} = \tilde{\theta}(\tilde{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}[D_i \tilde{\alpha}(X_i) Y_i]$$

Simulation results

Simulation with mild selection bias

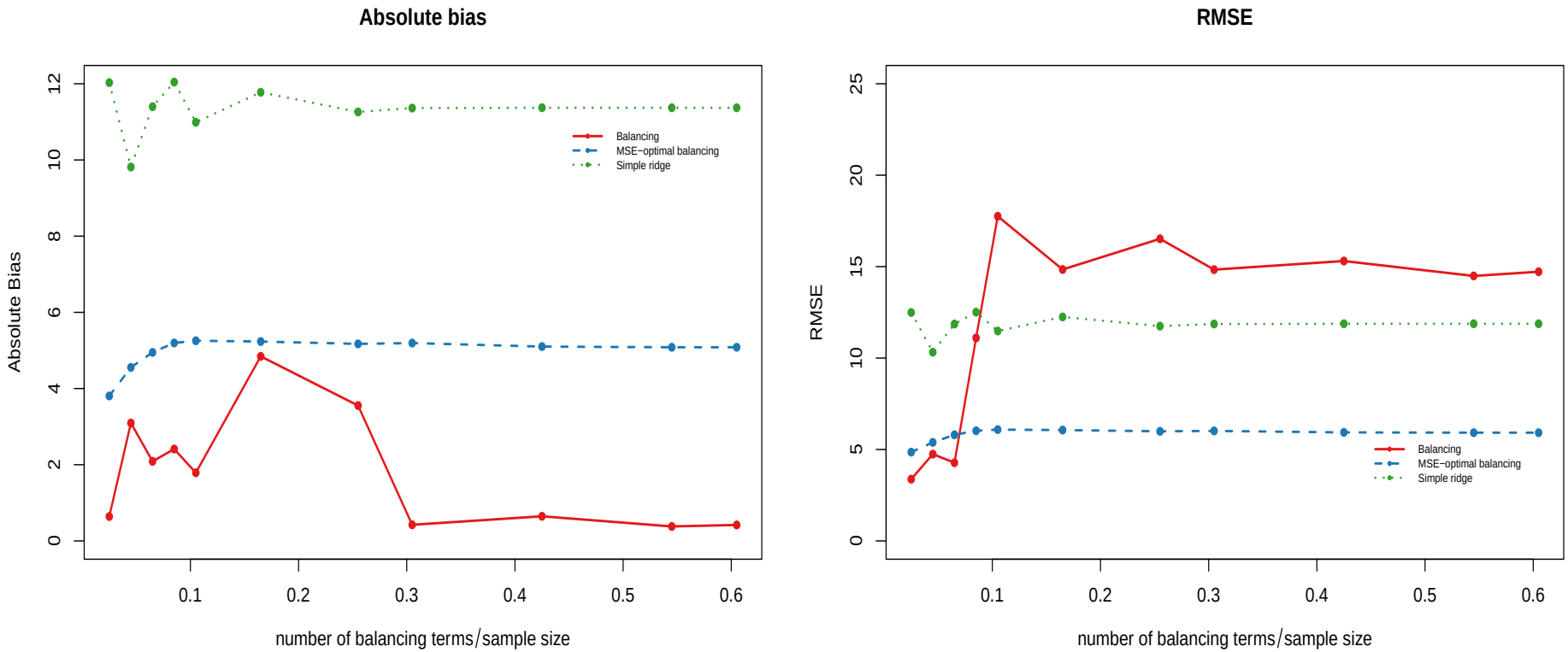
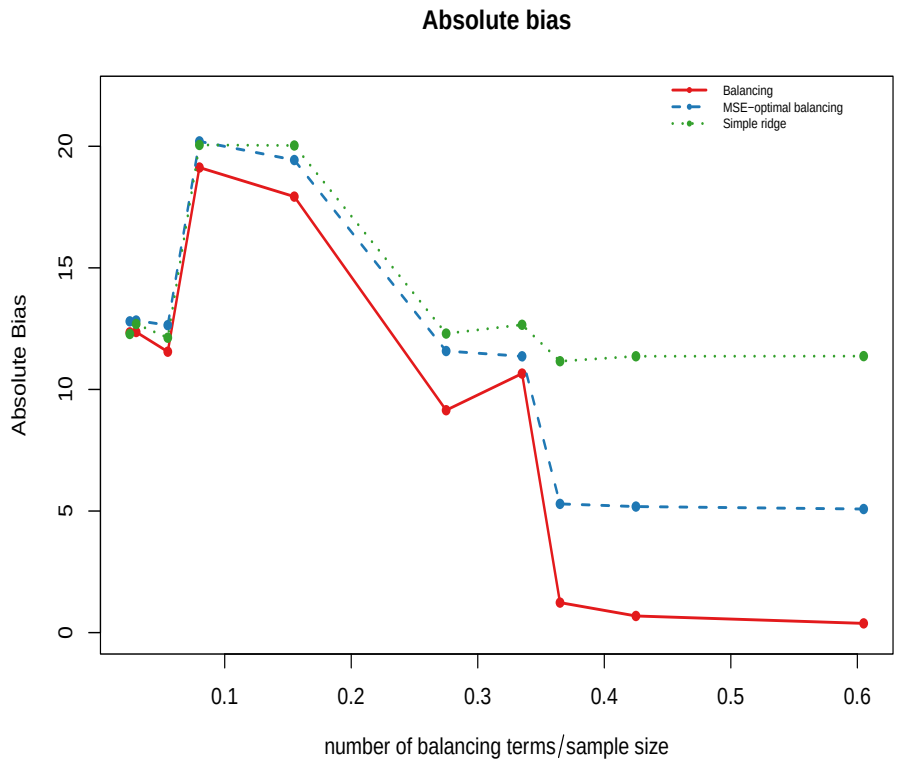


Figure: Bias and Root Mean Square Error (RMSE). Set-up follows Kang and Schafer (07), $n = 200$, 10000 experiments, $\lambda_1 = 0.002$

Simulation with substantial selection bias



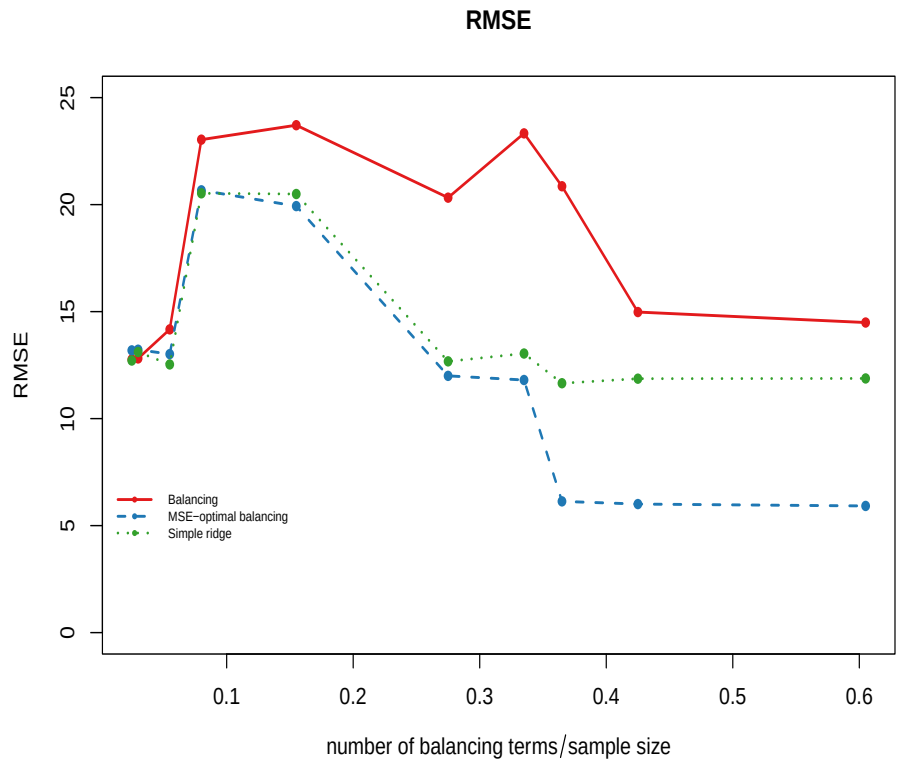


Figure: Bias and RMSE. Set-up follows Kang and Schafer (07), $n = 200$, 10000 experiments, $\lambda_1 = 0.002$

Back

Simulation: debiased machine learning estimator vs balancing estimator

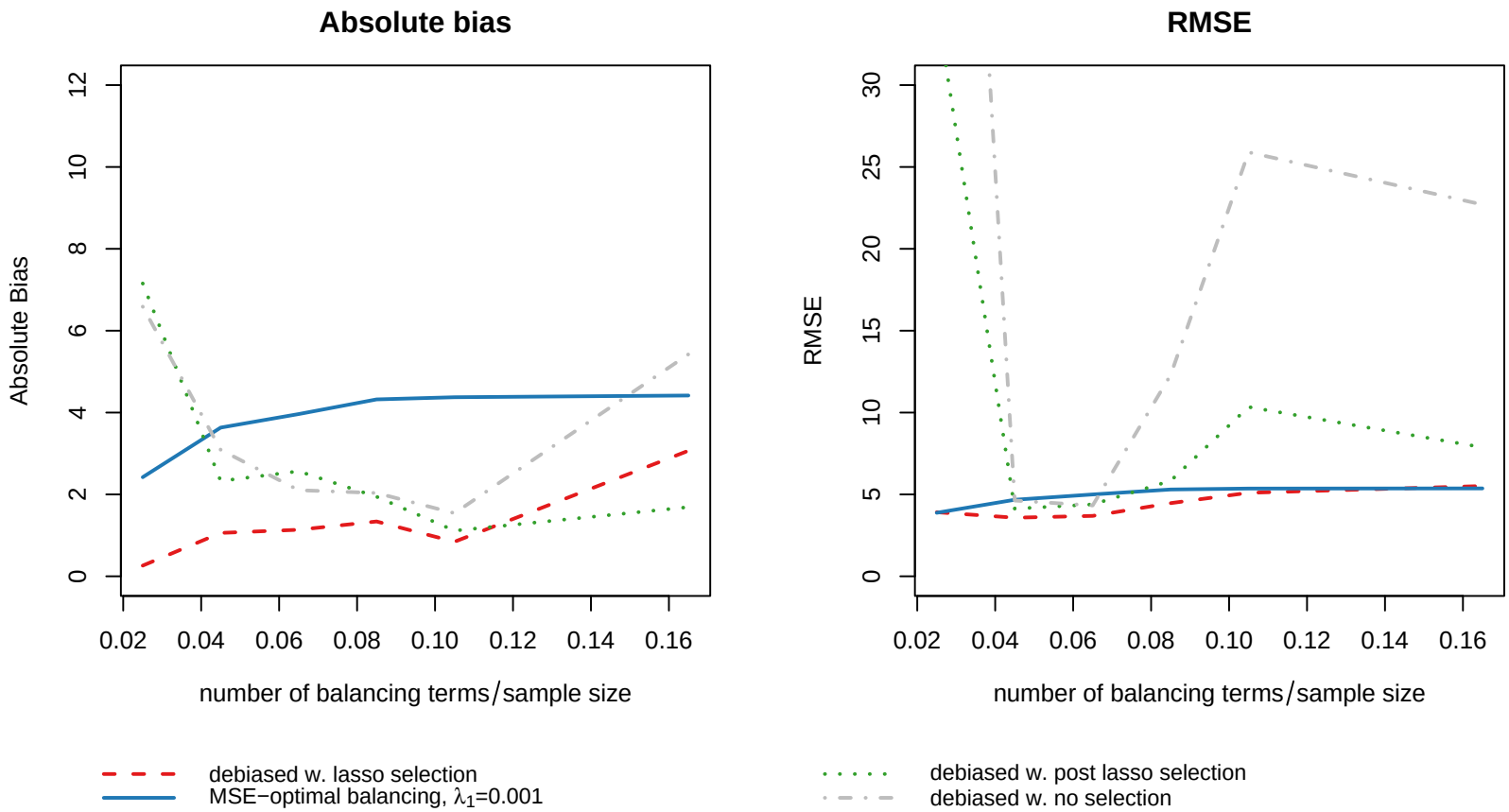
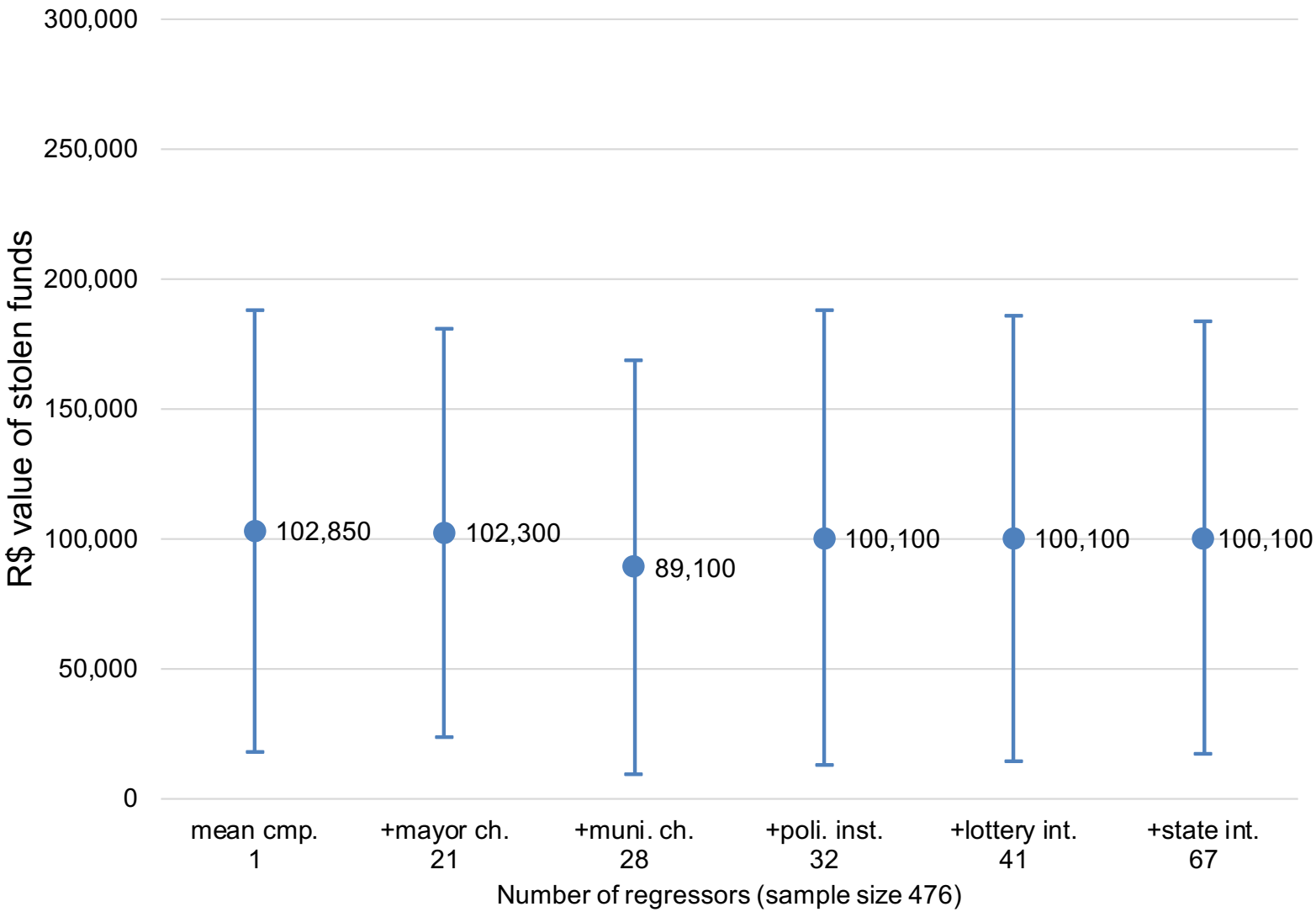


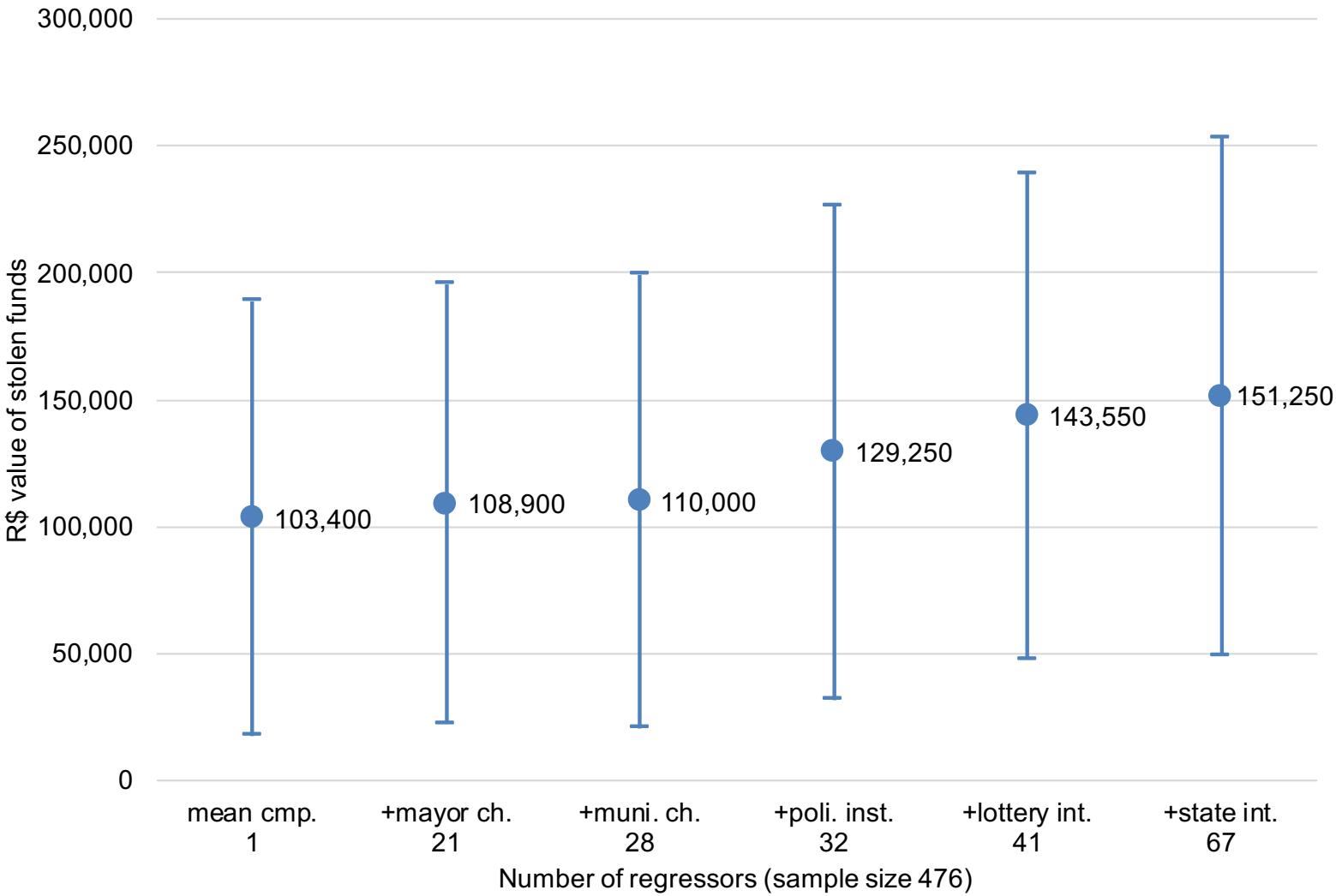
Figure: Bias and RMSE. Set-up follows Kang and Schafer (07), $n = 200$, 10000 experiments

More empirical evidence: Ferraz and Finan (2011)

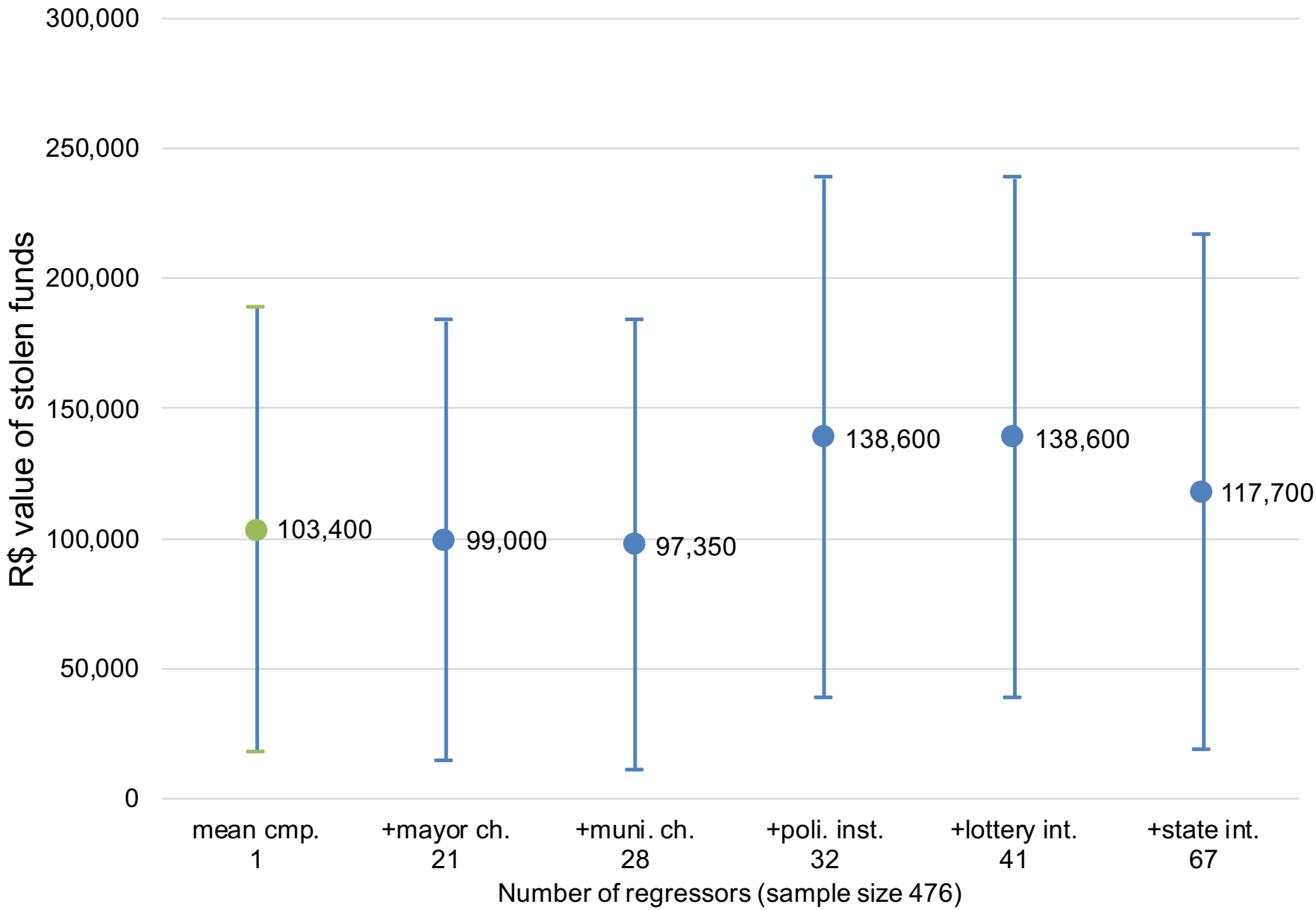
Estimated treatment effect: MSE-optimal balancing estimator



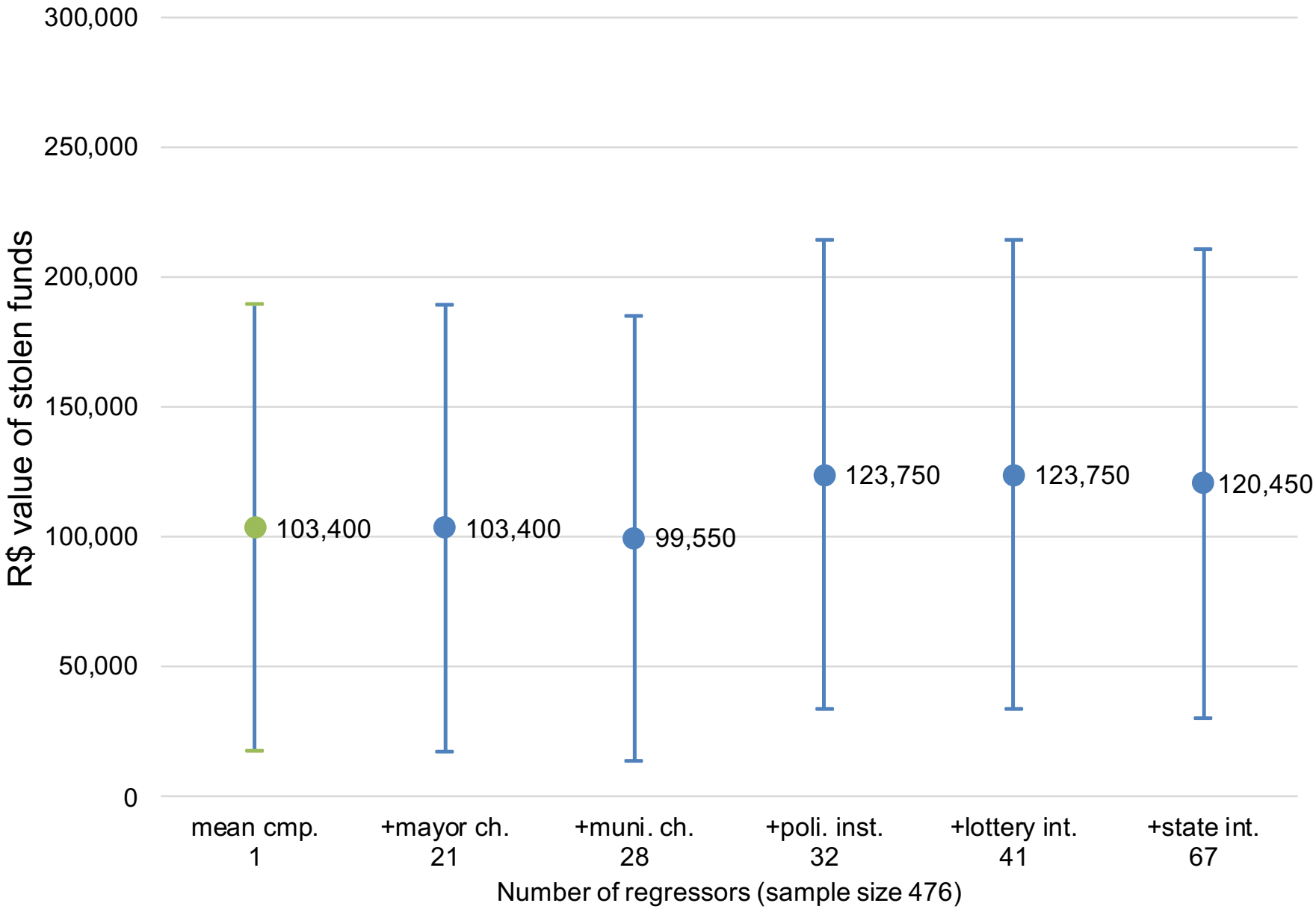
Estimated treatment effect: OLS



Debiased machine learning w. post lasso



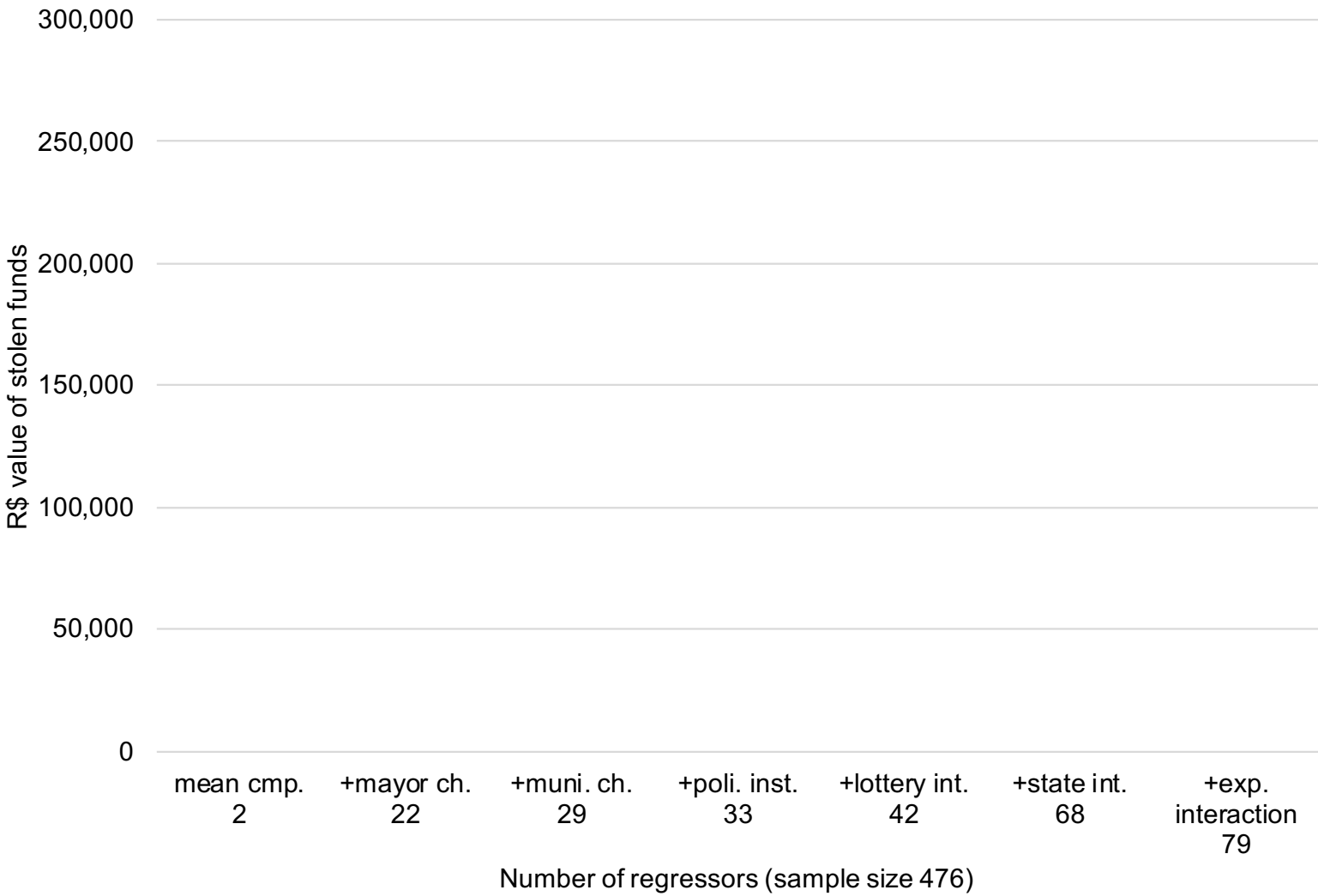
Debiased machine learning w. lasso



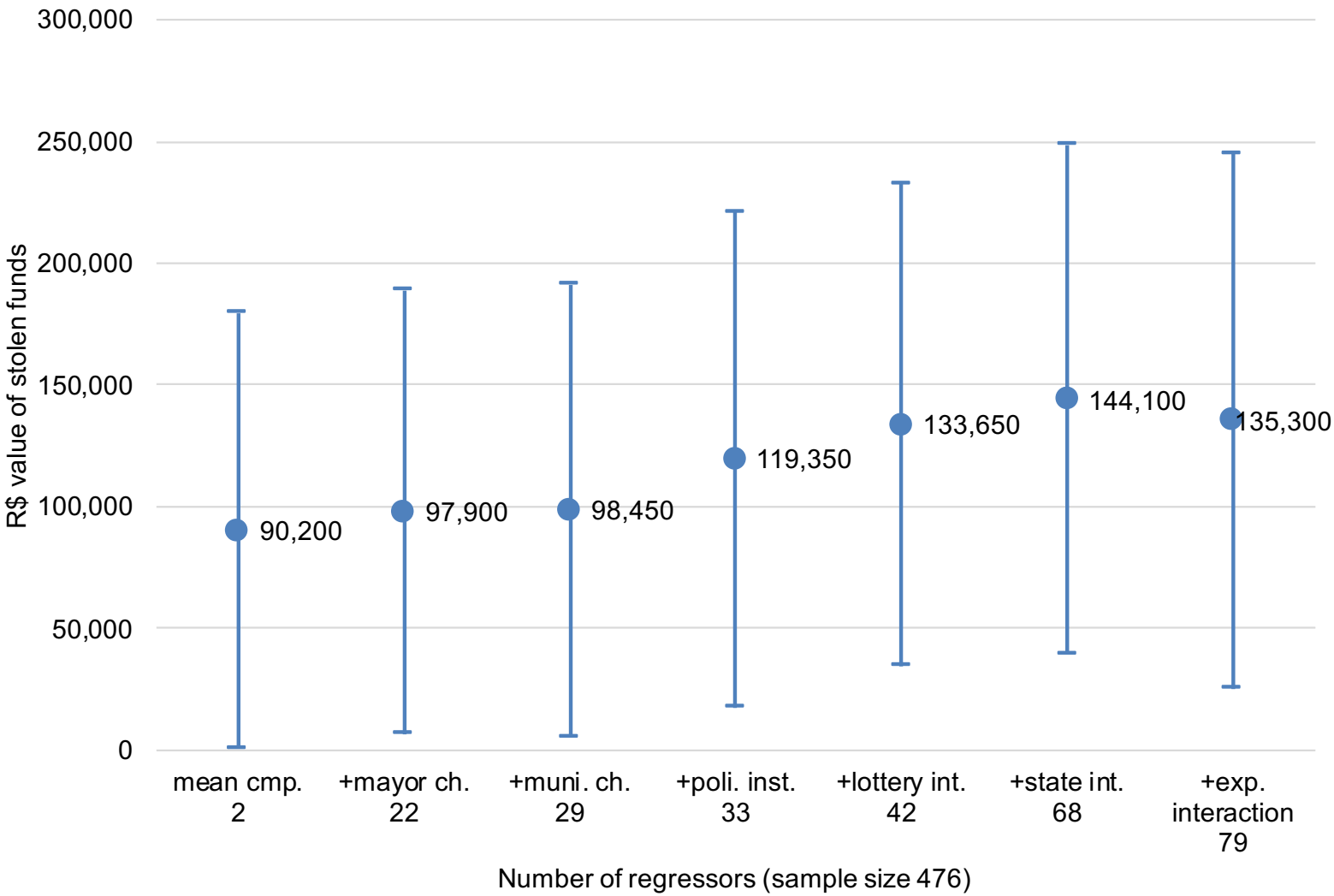
Testing performance by adding a bit more regressors

- Political experience seems a confounding factor
  - If more experienced mayors learn to be more corrupt, we over estimate
  - If more experienced mayors are more honest, we under estimate
- Add an indicator to every original specification showing
  - whether first term mayor was in power in one of three previous terms
- Add interaction of this indicator with other continuous variables in the full specification

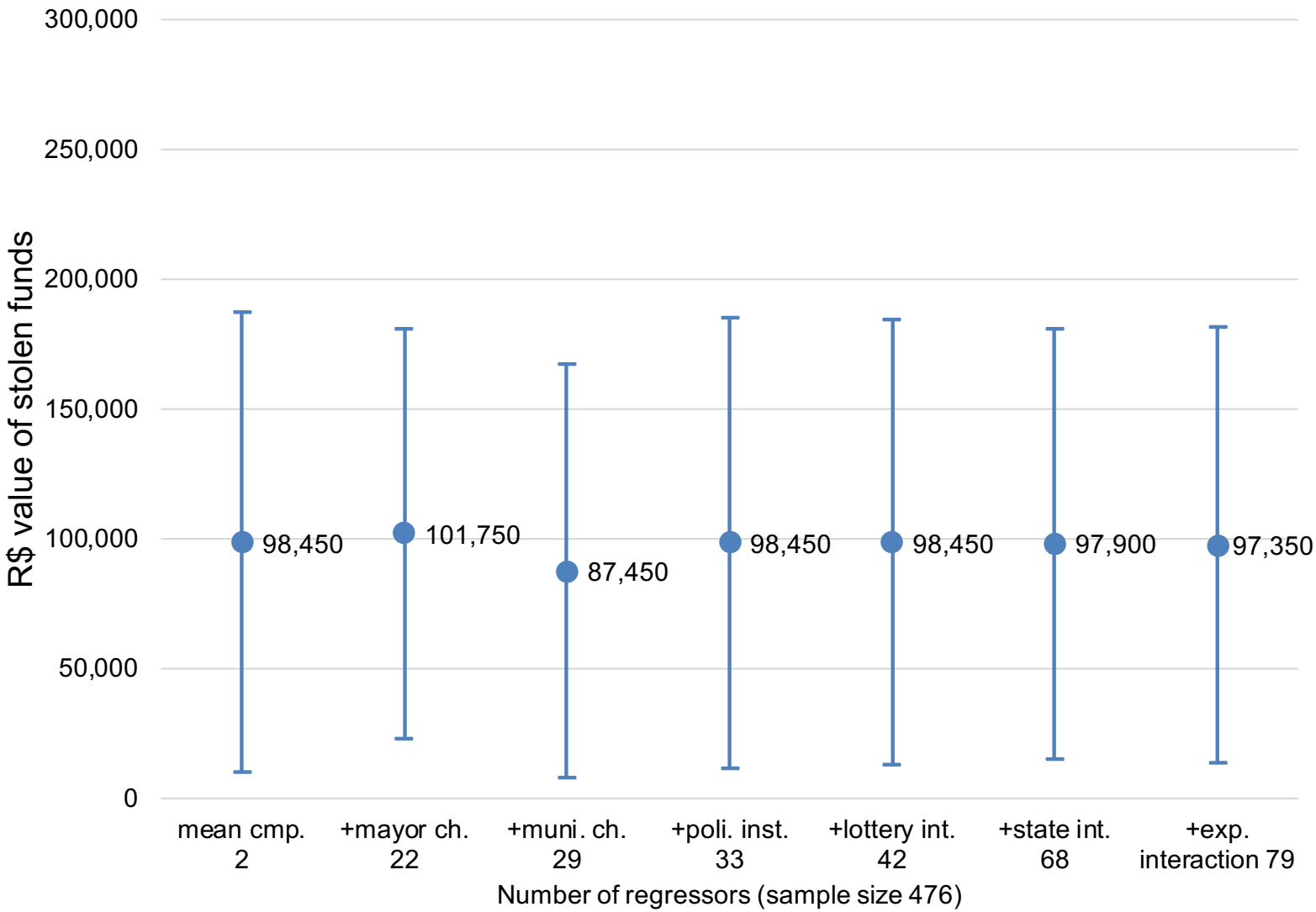
Controlling political experience



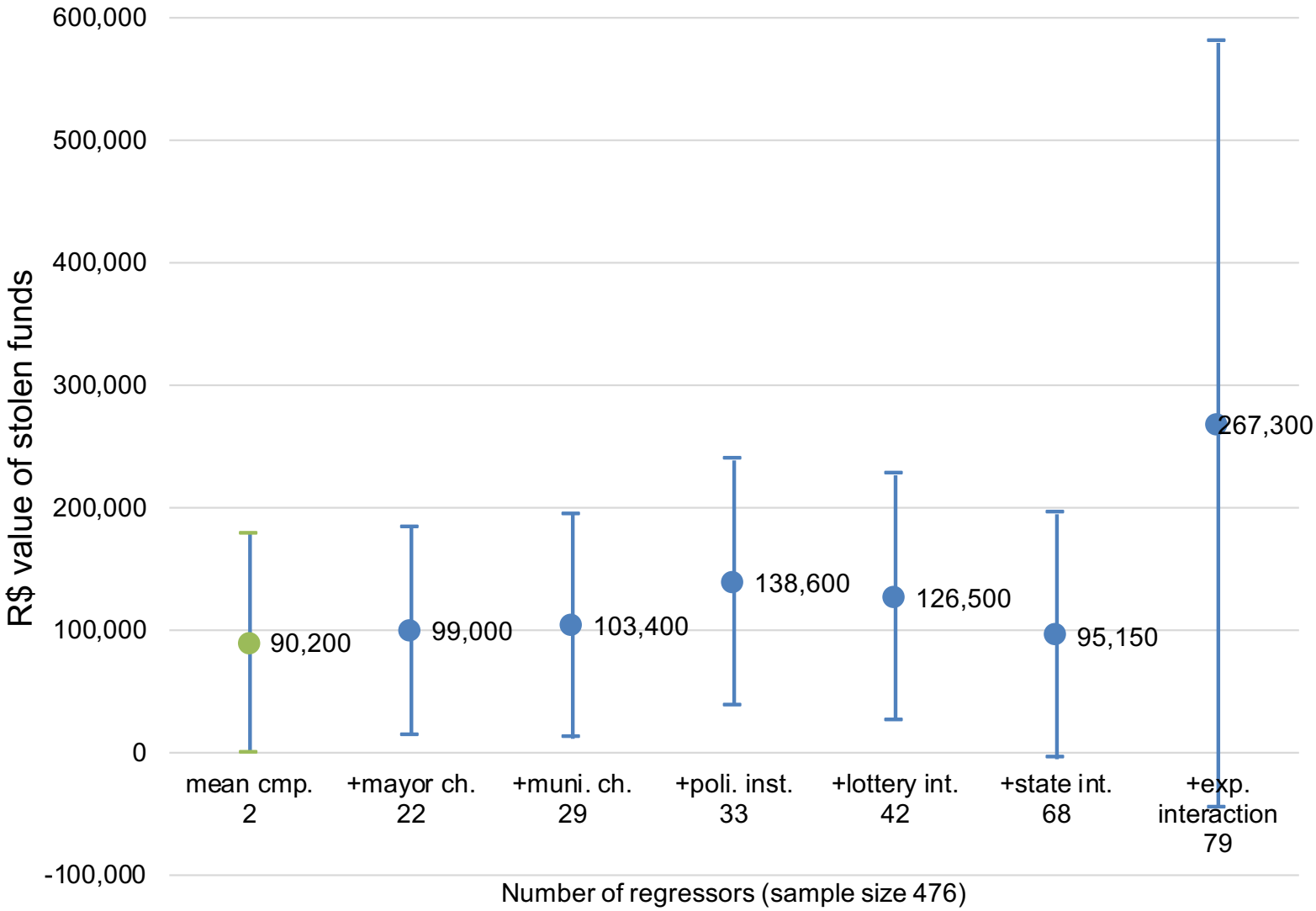
Controlling political experience: OLS



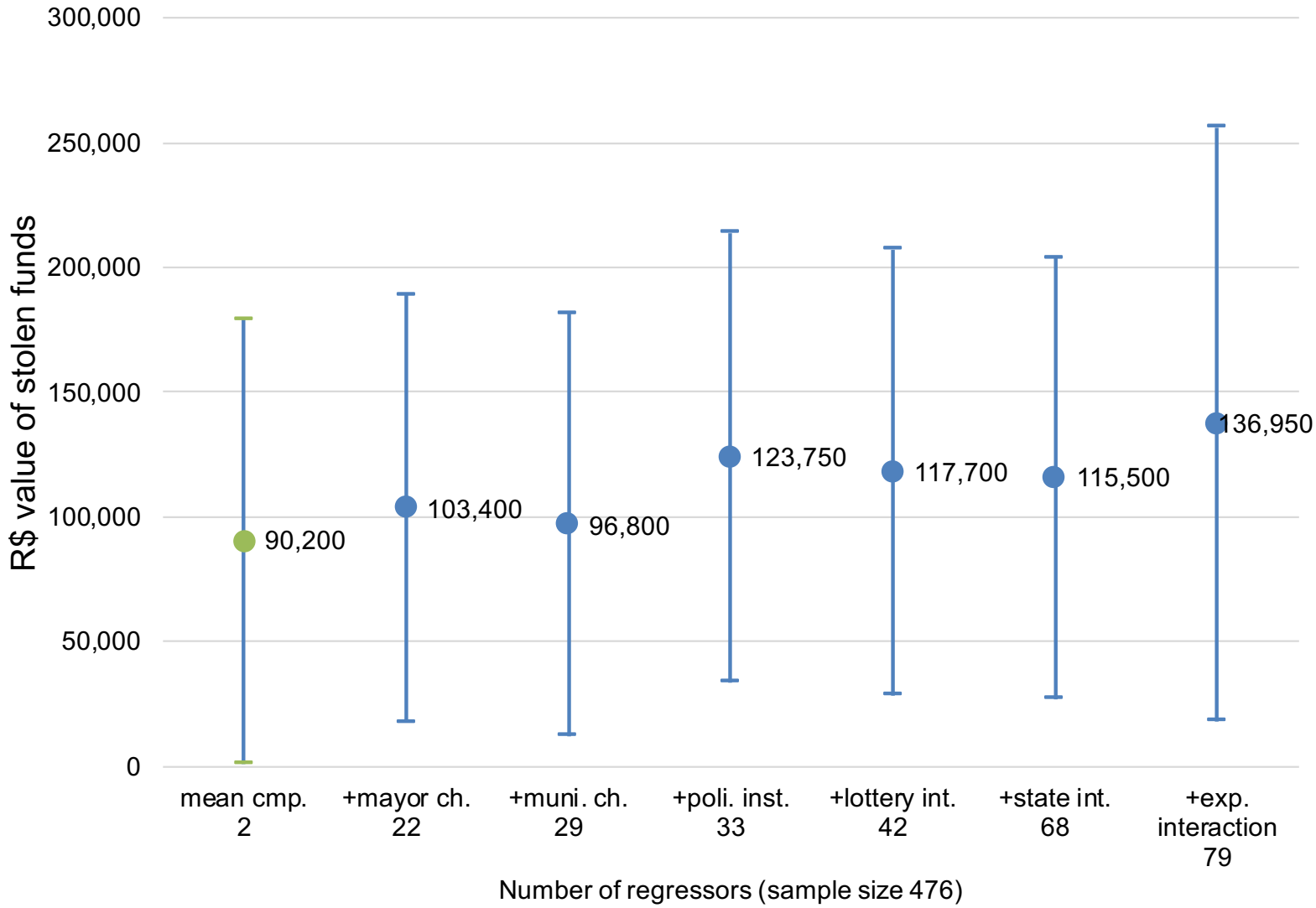
Controlling political experience: MSE-optimal balancing



Debiased machine learning w. post lasso



Debiased machine learning w. lasso



Conclusion

- We review balancing as an effective method to estimate causal effect with observational data
- When the number of covariates is large, the MSE-optimal balancing estimator is even better: it controls finite sample MSE but balances only approximately
- Balancing estimators are flexible, simple and more robust compared to many other modern approaches
- Balancing estimators work particularly well for moderately high dimensional datasets

Additional Materials

Simulation: minimax BP in missing data framework

- Let $U = \{U_1, U_2, U_3, U_4\}'$ be generated from $N(0, I_4)$
- $Y^* = 210 + 27.4U_1 + 13.7U_2 + 13.7U_3 + 13.7U_4 + e$, $e \sim N(0, 1)$
- True propensity score

$$\pi(u) = \Lambda(-u_1 + 0.5u_2 - 0.25u_3 - 0.1u_4), \quad \Lambda(\cdot) = \frac{\exp(\cdot)}{1 + \exp(\cdot)}$$

- Observed outcome variable is then $Y = TY^*$
- The econometrician does not observe U directly but a transformed version, denoted as $X = \{X_1, X_2, X_3, X_4\}'$

$$X_1 = \exp\left(\frac{U_1}{2}\right) \quad X_2 = \frac{U_2}{1 + \exp(U_1)} + 10$$

$$X_3 = \left(\frac{U_1 U_3}{25} + 0.6\right)^3 \quad X_4 = (U_2 + U_4 + 20)^2$$