Journal of Econometrics 239 (2024) 105460

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Inference on the best policies with many covariates

Waverly Wei [a,1], Yuqing Zhou [b,1], Zeyu Zheng [c], Jingshen Wang [a,*]

[a] Division of Biostatistics, UC Berkeley, USA
[b] Department of Statistics and Finance, University of Science and Technology of China, China
[c] Department of Industrial Engineering and Operations Research, UC Berkeley, USA

ARTICLE INFO

Article history:
Received 30 December 2021
Received in revised form 27 April 2022
Accepted 23 June 2022
Available online 17 June 2023

JEL classification:
C12
C13

Keywords:
Winner's curse
High dimensional data
Linear regression
Order statistics

ABSTRACT

Understanding the impact of the most effective policies or treatments on a response variable of interest is desirable in many empirical works in economics, statistics and other disciplines. Due to the widespread winner's curse phenomenon, conventional statistical inference assuming that the top policies are chosen independent of the random sample may lead to overly optimistic evaluations of the best policies. In recent years, given the increased availability of large datasets, such an issue can be further complicated when researchers include many covariates to estimate the policy or treatment effects in an attempt to control for potential confounders. In this manuscript, to simultaneously address the above-mentioned issues, we propose a resampling-based procedure that not only lifts the winner's curse in evaluating the best policies observed in a random sample, but also is robust to the presence of many covariates. The proposed inference procedure yields accurate point estimates and valid frequentist confidence intervals that achieve the exact nominal level as the sample size goes to infinity for multiple best policy effect sizes. We illustrate the finite-sample performance of our approach through Monte Carlo experiments and two empirical studies, evaluating the most effective policies in charitable giving and the most beneficial group of workers in the National Supported Work program.

1. Introduction

1.1. Motivation and our contribution

Many empirical work requires an understanding of the impact of the most effective policies or treatments on a relevant response variable of interest. For instance, in randomized (factorial) experiments with multiple treatments, researchers may be interested in the most effective policies (combinations). In online platforms, decision makers may be interested in the top five advertising strategies. In financial portfolio management, managers might want to learn about the best-performing strategies among many alternatives. In practice, after different policy effect sizes are estimated from a random sample, researchers may naturally look into those policies with the largest effect sizes. Accurately measuring the performance of top policies allows policy makers to deliver better-informed decisions for forecasting the effects of future policy implementations.

* Correspondence to: 5408 Berkeley Way West, Berkeley, 94720, CA, USA.
E-mail address: jingshenwang@berkeley.edu (J. Wang).

[1] Co-first authors. These authors contributed equally to this work and are alphabetically ordered.

https://doi.org/10.1016/j.jeconom.2022.06.013
0304-4076/

Nevertheless, given the well-recognized "winner's curse" phenomenon, there can be considerable uncertainties concerning if the top policies with large estimated effect sizes are indeed effective in the population (see Section 1.2 for a literature review). In fact, due to the winner's curse phenomenon, literature documents that the estimated effect sizes of the best-performing policies without additional adjustments tend to be overly optimistic, rendering under-covered confidence intervals (Lee and Shen, 2018; Andrews et al., 2019). In this manuscript, we refer to the optimistic bias introduced by the winner's curse phenomenon as the winner's curse bias. To mitigate this bias issue, we focus on the problem of constructing accurate point estimates and valid confidence intervals for the true effect sizes of the (observed) best policies. By the best policies, we refer to a user-supplied number of policies that have the largest (estimated) effects among a set of candidate policies (see Section 2.1 for a concrete problem setup), as we would expect that in practice researchers might want to focus on a few top policies of interest.

Other than the winner's curse phenomenon discussed above, an additional consideration gains prominence in the evaluation of the most effective policies. Since policy (or intervention) variables are often not exogenous, researchers may adopt observational methods to estimate their effects. In recent years, given the increased availability of large datasets with rich covariate information, one commonly adopted approach in empirical works is to assume that the policy variables are exogenous after controlling for a sufficiently large set of factors or covariates. Such a consideration demandingly requires empirical researchers to estimate the policy effects in the presence of many covariates.

To simultaneously address the above-mentioned issues, in this article, we propose a procedure that not only is robust to the presence of many covariates, but also provides accurate point estimates and valid frequentist confidence intervals for multiple best policy effect sizes. By many covariates, we allow the number of covariates q_n to diverge with the sample size n as long as $\limsup_{n\to\infty} q_n/n < 1$. Note that this does not rule out the cases where q_n is fixed or $q_n = o(n)$. In other words, our inferential method remains valid when the dimension of the covariates q_n is fixed or $q_n = o(n)$. Our proposed confidence intervals are built upon resampling methods, and we demonstrate that they achieve exact nominal coverage as the sample size goes to infinity under fairly moderate assumptions. Our empirical evidence shows that conventional estimates ignoring the winner's curse issue are substantially upward biased, while our corrections reduce the winner's curse bias and increase coverage. As far as we know, valid statistical inferential tools on multiple best policies that lift the winner's curse while incorporating possibly many covariates have been lacking, and the contribution of our work is to bridge this gap and help policy makers deliver well-informed decisions in practice.

We illustrate our method with two empirical applications. In the first case study, we use the charitable giving data from Karlan and List (2007) to evaluate the best pricing policies that motivate donors to give. Our results suggest that simple methods without adjusting for the winner's curse bias could be potentially overly optimistic in identifying the most effective polices. After accounting for the winner's curse bias, we do not find sufficient evidence to support that the second best pricing policy – asking the donor to give 25% more than his/her highest historical donation – is effective, implying that asking for a more "expensive" donation may not encourage donors to give. We nevertheless note that given our calibration only marginally reduces the effect size of the second best policy, the above conclusion might not warrant a different economic interpretation. In the second case study, we evaluate the effectiveness of the national supported work (NSW) program in different groups of workers. The NSW program is a job training program designed to prepare disadvantaged workers for employment, and it has been investigated in various studies (Dehejia and Wahba, 2002; LaLonde, 1986). We apply the proposed approach to evaluate the performance of the NSW program on the most-affected subgroups of workers observed in the dataset. Our study results potentially suggest that married black workers might benefit from the NSW program with an average increase of $4410 for their annual income.

1.2. Connection to the existing literature

One fundamental trend that drives the motivation of the methodology developed in this manuscript is the increasing availability of massive datasets and the associated increasing dimensionality. Such a trend brings scientists opportunities to deliver better-informed policies but, at the same time, presents challenges in developing econometric and statistical tools; see Fan et al. (2011), Belloni et al. (2014), Fan et al. (2014), Cattaneo et al. (2018b) for example. A recent book (Fan et al., 2020) provides a thorough discussion of analytical methods that aim to address such challenges. Specifically, the increasing data availability brings challenges and also opportunities to better understand various policies whose effects can be inferred from data. Along this line, our manuscript aims at providing understating for policies that are estimated and selected to be the most effective from a pool of policies.

The winner's curse phenomenon and its related issues have been widely recognized in economics, statistics, and data science at large. Seminal works by Fan and Zhou (2016), Fan et al. (2018) point out that spurious discoveries can easily arise when target parameters are selected through data mining and statistical machine learning algorithms. Recent work by Andrews et al. (2019) considers performing conditional and unconditional inference on observed best policy and Andrews et al. (2022) extends the work to more general ranking problems, which is still different from our goal in conducting unconditional inference on multiple top policies. Moreover, while the conditional approaches in Andrews et al. (2019) and Andrews et al. (2022) produce optimal confidence intervals for the observed policy effects, their point estimates and confidence intervals can be conservative when they are applied unconditionally. Efron (2011) considers a method to handle the winner's curse bias with Tweedie's formula concerning the empirical Bayes theory. Lee and Shen (2018) consider a plug-in correction of the winner's curse bias and propose to construct confidence interval based on

bootstrapping in the context of A/B testing, but the proposed method lacks theoretical justifications. In clinical trials for evaluating the largest observed treatment effect in multiple subpopulations, Guo and He (2020) propose a bootstrap-based confidence interval that achieves the exact nominal level as the sample size goes to infinity, though generalizing their method to make inference on several top policies might not be straightforward, especially in the presence of many covariates.

Our manuscript builds upon the literature on linear regression models with many or high dimensional covariates; see Huber (1973), Mammen (1993), Long and Ervin (2000), Anatolyev (2012), El Karoui et al. (2013), Cattaneo et al. (2018a), Cattaneo et al. (2019), Jochmans (2020) and the reference therein. In particular, Mammen (1993) has established the asymptotic normality results for any contrasts of the ordinary least squares (OLS) coefficient vector estimator, when the dimension of the covariates divided by the sample size vanishes asymptotically. More recently, Cattaneo et al. (2018b) have shown that a small subset of the OLS estimators for the regression coefficients are asymptotically normal without restricting the dimension of the covariates to be a vanishing fraction of the sample size. Moreover, Cattaneo et al. (2018b) have proposed a robust covariance matrix estimator for the subset of the OLS estimator under fairly general conditions. Jochmans (2020) has proposed an alternative covariance matrix estimator that can deal with designs with even large number of covariates under additional assumptions (Assumption 4 in the current manuscript).

Making inference on the best-performing policies is related to the literature on constructing confidence intervals for extrema parameters with bootstrap; see Andrews (2000), Fan et al. (2007), Xie et al. (2009), Chernozhukov et al. (2013a), Claggett et al. (2014) and the reference therein. Given the asymptotic distributions of extrema parameter estimators are often not normal, bootstrap-based methods can face serious difficulties when used to replicate the distribution of extrema of parameter estimators (Mammen, 1993, 2012). While subsampling could overcome this issue faced by the classical bootstrap, it can exhibit very poor finite-sample performance because of the noise introduced by the vanishing subsample size. Different from our goal in constructing confidence intervals that achieve the exact nominal level, Hall and Miller (2010) and Chernozhukov et al. (2013a) propose to construct conservative bootstrap confidence intervals for extrema of parameters. In our current problem setup with many covariates, the problem becomes even more acute as El Karoui and Purdom (2018) show through a mix of simulation and theoretical analyses that the bootstrap is fraught with problems in moderate high dimensions. In the context of meta-analyses, Claggett et al. (2014) propose an approach to make inference on ordered fixed study-specific parameters when different parameters are estimated independently from multiple studies.

Our method also contributes to the rapidly growing literature on program evaluations; see Farrell (2015), Belloni et al. (2014), Kueck et al. (2023), Athey and Imbens (2017), Abadie and Cattaneo (2018), Chernozhukov et al. (2018), Fan et al. (2021), Vazquez-Bare (2021) among many others. Under our asymptotic regime where the number of covariates q_n grows with the sample size n, the Neyman orthogonalization based approaches often need to work with models with sparse regression coefficients (Belloni et al., 2014). Rather than imposing such a sparsity assumption, our approach estimates the policy effects with regression adjustments without requiring the regression coefficients to be sparse. Because our approach only requires a consistent covariance matrix estimation for different policy effect estimators, we expect that the proposed framework on evaluating the best policies can be generalized when different policy effects are estimated with other off-shelf methods and we relegate such extensions for future work.

Notation. We work with triangular array data $\{\omega_{i,n} : i = 1, \ldots, n; n = 1, 2, \ldots\}$ where for each n, $\{\omega_{i,n} : i = 1, \ldots, n\}$ is defined on the probability space $(\Omega, \mathcal{S}, P_n)$. All parameters that characterize the distribution of $\{\omega_{i,n} : i = 1, \ldots, n\}$ are implicitly indexed by P_n and thus by n. We write vectors and matrices in bold font, and use regular font for univariate variables and constants.

2. Model setup and methodology

2.1. Problem setup and a revisit to the winner's curse phenomenon

Suppose we have a random sample $\{(y_{i,n}, \boldsymbol{x}_{i,n}', \boldsymbol{w}_{i,n}')'\}_{i=1}^n$, we pose the problem in the framework of a linear regression model under heteroscedasticity

$$y_{i,n} = \boldsymbol{x}_{i,n}'\boldsymbol{\beta} + \boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n + u_{i,n}, \quad i = 1, \ldots, n, \tag{1}$$

where $y_{i,n}$ is the outcome variable, $\boldsymbol{x}_{i,n} \in \mathbb{R}^d$ are the treatment or policy variables of interest, $\boldsymbol{w}_{i,n} \in \mathbb{R}^{q_n}$ contains the confounding factors, $u_{i,n}$ is an unobserved error term, and the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$ contains the treatment effect of $\boldsymbol{x}_{i,n}$ on the outcome $y_{i,n}$. We allow the linear model (1) to hold approximately by allowing $\mathbb{E}[u_{i,n}|\{\boldsymbol{x}_{i,n}\}_{i=1}^n, \{\boldsymbol{w}_{i,n}\}_{i=1}^n] \neq 0$. We are also in a scenario where $\boldsymbol{w}_{i,n}$ is high-dimensional, in the sense that q_n can be a vanishing fraction of the sample size n as long as $\limsup_{n\to\infty} q_n/n < 1$. To simplify notations, we drop subscript n in univariate random variables in the rest of the manuscript. That is, for example, we denote $\gamma_{j,n}$ as γ_j.

We write the ordered values of $\beta_1, \ldots, \beta_d$ as $\beta_{(1)} \geq \cdots \geq \beta_{(d)}$. We adopt the ordinary least-squares (OLS) estimator $\widehat{\boldsymbol{\beta}}$ (see Remark 2 for other possible estimates) to estimate $\boldsymbol{\beta}$ and write the order statistics of $\widehat{\boldsymbol{\beta}}$ as $\widehat{\beta}_{(1)} \geq \cdots \geq \widehat{\beta}_{(d)}$. Because researchers in practice might hope to focus on a few top policies, given that d_0 is a user-supplied positive integer,our goal

is to construct accurate point estimates and valid confidence intervals for two sets of quantities:

(1) the best policy effect sizes in the population: $\beta_{(1)}, \ldots, \beta_{(d_0)}$,
(2) the observed best policy effect sizes: $\beta_{\widehat{j}}$, where $\widehat{j} = \sum_{k=1}^{d} k \cdot \mathbb{1}(\widehat{\beta}_k = \widehat{\beta}_{(j)})$, for $j = 1, \ldots, d_0$.

The first set of quantities characterizes the effects of the top d_0 policies in the population and are thus fixed parameters. The second set of quantities describes the true effect sizes of the best performing policies observed in the random sample, and these quantities are thus "data-dependent parameters". Both sets of quantities can be of interest in different empirical applications (Dawid, 1994; Chernozhukov et al., 2013b; Rai, 2018), and our proposed procedure can be used to deliver valid statistical inference on both quantities (Theorem 1 and Corollary 1).

Remark 1 (*Ties in the Estimated Policy Effects*)**.** The second set of parameters is well defined if the observed policies do not have exact ties in the sense that $\widehat{\beta}_{(1)} > \cdots > \widehat{\beta}_{(d)}$. When the policies effect estimators solve to the interior points of the feasible parameter space, it is likely that no exact ties appear in the random sample. On the other hand, there can exist scenarios where, for example, d_0 is set as 2 but there are multiple policy effect sizes that tie at rank 2. In this case, one may choose instead a data-dependent $\widehat{d}_0 = \max\{k : \widehat{\beta}_{(k)} > \widehat{\beta}_{(2)} - C_1 \cdot n^{-0.25}\}$. This new random d_0 will asymptotically be able to incorporate all the effect sizes that actually are equal to the true effect size associated with $\widehat{\beta}_{(2)}$. In this way, the limiting value of $\widehat{d}_0$ will not necessarily be 2, but can be a larger number than 2 to incorporate "very close" effect sizes with the rank-2 effect size. We also provide some related discussions in Remark 4.

Remark 2 (*Other Possible Estimators of* $\boldsymbol{\beta}$)**.** In the presence of many covariates when q_n is potentially large ($\limsup_{n\to\infty} q_n/n \to 1$ in our asymptotic regime) without assuming the coefficient γ_n to be sparse, we adopt the OLS estimator to estimate $\boldsymbol{\beta}$, because the OLS estimator has been thoroughly studied in the existing literature and enjoys favorable theoretical guarantees. In high dimensions when $q_n \gg n$, other estimators of $\boldsymbol{\beta}$ that incorporate model selection procedures can be adopted as well. Our procedure can produce valid statistical inference as long as the covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be consistently estimated. For example, under the sparsity assumption on $\boldsymbol{\gamma}_n$ documented in the literature (Fan et al., 2020), we may adopt the covariance matrix estimator from the de-sparsified Lasso procedure (Van de Geer et al., 2014; Zhang and Zhang, 2014).

To fully realize the challenges on delivering valid statistical inference on these two sets of parameters in our current problem setup, we revisit the winner's curse phenomenon. When first discussed in common-value auctions, the winner's curse refers to the bidding behavior where bidders systematically overbid, resulting in an expected loss (Charness and Levin, 2009). In our context of policy evaluations, the winner's curse refers to the issue that the observed best policies have the tendency to over-estimate the best policies in the population. We would thus often expect that neither $\mathbb{E}[\widehat{\beta}_{(j)} - \beta_{(j)}]$ nor $\mathbb{E}[\widehat{\beta}_{(j)} - \beta_{\widehat{j}}]$ is close to zero, and the resulting confidence interval may fail to reach the nominal level. Such an issue becomes even more acute as we have many covariates $\boldsymbol{w}_{i,n}$ entering the inferential process.

We next illustrate the winner's curse issue through Example 1 with a simple simulation study, where we observe substantial winner's curse bias and under-covered confidence intervals for the top polices. In particular, Fig. 1(b) demonstrates that coverage probabilities are worsened when a larger number of covariates are incorporated for estimating $\boldsymbol{\beta}$. It is worth pointing out that when $d = 3$, $\widehat{\beta}_{(2)}$ is the median policy effect. Thus, the estimation bias is around 0, and the true standard deviation is much smaller than the estimated standard deviation, resulting in a confidence interval with close to 100% coverage. When d increases, the coverage probability gradually drops due to a larger estimation bias and inaccurately estimated standard deviation.

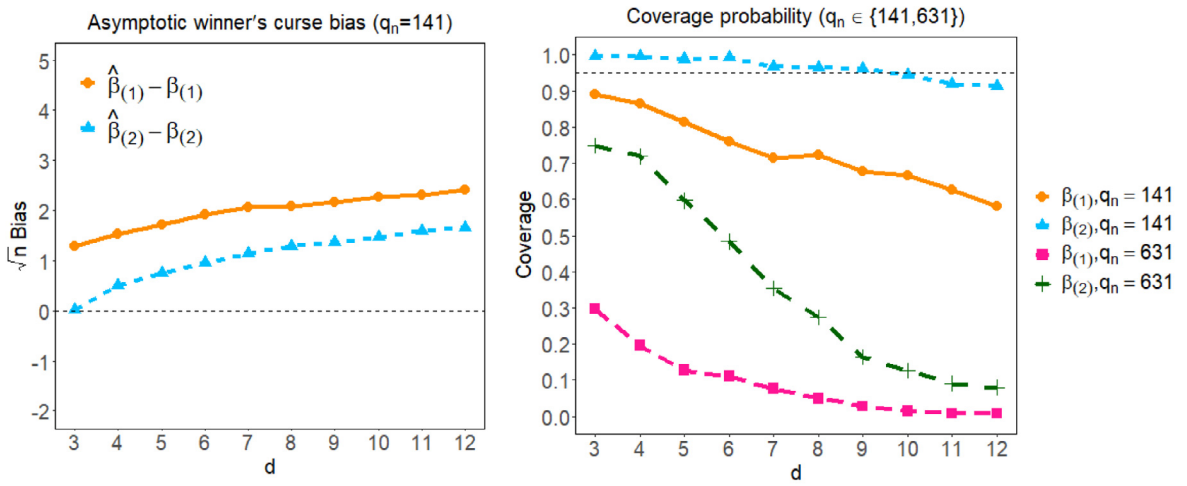
Example 1 (*A Simulation Study Demonstrating the Winner's Curse Phenomenon with Many Covariates*)**.** We generate 1000 Monte Carlo samples following the setup in Model (1). We generate $\boldsymbol{x}_{i,n} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\Sigma_{jk} = 0.5^{|j-k|}$ for $j, k = 1, \ldots, d$, $\boldsymbol{w}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{w}}_{i,n} \geq \Phi^{-1}(0.98))$ with $\widetilde{\boldsymbol{w}}_{i,n} \sim \mathcal{N}(0, \boldsymbol{I}_{q_n})$, where $\boldsymbol{I}_{q_n}$ is a q_n-dimensional identity matrix. We consider the case where no policy is effective (so that $\boldsymbol{\beta} = 0$, $\beta_{\widehat{1}} = \beta_{\widehat{2}} = 0$) and $\gamma_j = 1/j$, for $j = 1, \ldots, q_n$. We report the asymptotic bias of the conventional estimator (i.e., $\sqrt{n} \cdot \mathbb{E}[\widehat{\beta}_{(j)} - \beta_{(j)}]$) as well as the coverage probability of confidence intervals constructed based on normal approximation with the Eicker–White (Eicker, 1963; White, 1980) covariance matrix estimator defined in Eq. (6).

2.2. Methodology

Our method starts with the ordinary least-squares (OLS) estimator of $\boldsymbol{\beta}$, that is

$$\widehat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n} \widehat{\boldsymbol{v}}_{i,n} \widehat{\boldsymbol{v}}_{i,n}'\Big)^{-1} \Big(\sum_{i=1}^{n} \widehat{\boldsymbol{v}}_{i,n} y_{i,n}\Big),$$

where $\widehat{\boldsymbol{v}}_{i,n} = \sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} \boldsymbol{x}_{j,n}$, and $(\boldsymbol{M}_n)_{i,j} \triangleq \mathbb{1}(i=j) - \boldsymbol{w}_{i,n}' \Big(\sum_{k=1}^{n} \boldsymbol{w}_{k,n} \boldsymbol{w}_{k,n}'\Big)^{-1} \boldsymbol{w}_{j,n}$. As we focus on the case when q_n can be a non-vanishing fraction of n, $n \to \infty$, we adopt the robust covariance matrix estimator proposed in Jochmans (2020).



(a) Winner's curse bias ($q_n = 141$) (b) Coverage probability

Fig. 1. Demonstration of the winner's curse phenomenon following the simulation setup in Example 1. The maximum Monte Carlo standard error for the asymptotic bias is 0.88. Panel (a) captures the asymptotic winner's curse bias when $q_n = 141$; Panel (b) captures the coverage probability when $q_n \in \{141, 631\}$ and the nominal level is 0.95.

We try to follow the author's notation as closely as possible:

$$\widehat{\boldsymbol{\Omega}}_n^{\mathrm{KJ}} \triangleq \widehat{\boldsymbol{\Gamma}}_n^{-1} \widehat{\boldsymbol{\Sigma}}_n^{\mathrm{KJ}} \widehat{\boldsymbol{\Gamma}}_n^{-1},$$

where

$$\widehat{\boldsymbol{\Gamma}}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{v}}_{i,n} \widehat{\boldsymbol{v}}_{i,n}', \quad \widehat{\boldsymbol{\Sigma}}_n^{\mathrm{KJ}} \triangleq \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{v}}_{i,n} \widehat{\boldsymbol{v}}_{i,n}' y_{i,n} \acute{u}_{i,n},$$

where $\acute{u}_{i,n} = \frac{\widehat{u}_{i,n}}{(\boldsymbol{M}_n)_{i,i}}$, $\widehat{u}_{i,n} = \sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} (y_{j,n} - \boldsymbol{x}_{j,n}' \widehat{\boldsymbol{\beta}})$, for $i = 1, \ldots, n$. Such an estimator is well-defined as long as $\min_i (\boldsymbol{M}_n)_{i,i} > 0$. If $\min_i (\boldsymbol{M}_n)_{i,i} = 0$, it means that the auxiliary regression produces a perfect prediction. So the observation does not carry information on $\boldsymbol{\beta}$ and can be ignored.

As shown in Example 1, the estimated top policy effect sizes with $\widehat{\beta}_{(1)}, \ldots, \widehat{\beta}_{(d_0)}$ are often biased upward for our target parameters due to the winner's curse phenomenon. Inspired by the procedure proposed by Claggett et al. (2014)[2] for meta-analyses, we generate replicates of $\widehat{\boldsymbol{\beta}}$ from a multivariate normal distribution

$$\widehat{\boldsymbol{\beta}}^* \big| \{(y_{i,n}, \boldsymbol{x}_{i,n}', \boldsymbol{w}_{i,n}')'\}_{i=1}^n \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}}_n^{\mathrm{KJ}}/n), \quad \text{where } \widehat{\boldsymbol{\beta}}^* = (\widehat{\beta}_1^*, \ldots, \widehat{\beta}_d^*)', \tag{2}$$

and we denote the ordered values of the vector $\widehat{\boldsymbol{\beta}}^*$ as $\widehat{\beta}_{(1)}^* \geq \cdots \geq \widehat{\beta}_{(d)}^*$. Note that the above description of $\widehat{\boldsymbol{\beta}}^*$ differs from some previous work on bootstrapping insofar we have suppressed the role of "multiplier variables", and we have defined $\widehat{\boldsymbol{\beta}}^*$ as a sample from $\mathcal{N}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}}_n^{\mathrm{KJ}}/n)$. Different from Claggett et al. (2014) that requires different estimators to be estimated from independent studies with non-overlapping random samples, our approach relaxes such an requirement and allows $\widehat{\beta}_1, \ldots, \widehat{\beta}_d$ to be correlated.

Next, given properly chosen b_{L} and b_{R} so that $b_{\mathrm{L}} - b_{\mathrm{R}} = O(n^{-\delta})$ with $\delta \in (0, \frac{1}{2})$ (see Supplementary Materials Section C.1 for their data-adaptive choices, and robustness to different choices of tuning parameters in Supplementary Materials, Section C.2), we estimate a "near tie" set that captures policies that have similar effect sizes to the jth largest policy:

$$\widehat{\mathcal{H}}_{(j)} = \left\{k : \widehat{\beta}_{(j)}^* - b_{\mathrm{L}} \leq \widehat{\beta}_k^* \leq \widehat{\beta}_{(j)}^* + b_{\mathrm{R}},\ k = 1, \ldots, d\right\}.$$

We then record the averages of $\widehat{\beta}_1^*, \ldots, \widehat{\beta}_d^*$ and of $\widehat{\beta}_1, \ldots, \widehat{\beta}_d$ in the estimated tie set $\widehat{\mathcal{H}}_{(j)}$ as

$$\widetilde{\beta}_{(j)}^* = \frac{\sum_{k \in \widehat{\mathcal{H}}_{(j)}} \widehat{\beta}_k^*}{|\widehat{\mathcal{H}}_{(j)}|}, \text{ and } \widetilde{\beta}_{(j)} = \frac{\sum_{k \in \widehat{\mathcal{H}}_{(j)}} \widehat{\beta}_k}{|\widehat{\mathcal{H}}_{(j)}|}, \tag{3}$$

where $|\widehat{\mathcal{H}}_{(j)}|$ denotes the cardinality of the set $\widehat{\mathcal{H}}_{(j)}$.

[2] Note that there is a typo in Claggett et al. (2014) for the definition of the near tie set. Although the near tie $\mathcal{H}_{(j)}$ in their manuscript was originally defined as $\mathcal{H}_{(j)} = \left\{k : |\beta_k - \beta_{(j)}| = O(n^{-\frac{1}{2}}),\ k = 1, \ldots, d\right\}$, their proof goes through when the near tie set is defined with $\mathcal{H}_{(j)} = \left\{k : |\beta_k - \beta_{(j)}| = o(n^{-\frac{1}{2}}),\ k = 1, \ldots, d\right\}$.

Finally, we apply the above resampling procedure to construct point estimates and confidence intervals for $\beta_{(j)}$ as well as $\beta_{\widehat{j}}$ (as defined in Section 2.1 and in Eq. (5)), $j = 1, \ldots, d_0$. Specifically, for confidence interval construction, we generate B independent samples of $\widetilde{\beta}^*_{(j)}$ as in Eq. (3), and then define $\widehat{q}_{(j)}(\alpha/2)$ to be the empirical $\alpha/2$-quantile of the $B \geq 1$ samples (and similarly for $\widehat{q}_{(j)}(1-\alpha/2)$), leading to a level-$\alpha$ confidence interval for $\beta_{(j)}$ with

$$\big[\widehat{q}_{(j)}(\alpha/2),\ \widehat{q}_{(j)}(1-\alpha/2)\big], \quad j = 1, \ldots, d_0.$$

Corollary 1 demonstrates that the above confidence interval also serves as an asymptotically exact level-α prediction interval for $\beta_{\widehat{j}}$. For point estimates, we may either use $\widetilde{\beta}_{(j)}$ or the averaged resampled statistics $\widetilde{\beta}^*_{(j)}$ to estimate $\beta_{(j)}$ and $\beta_{\widehat{j}}$.

3. Theoretical investigation

3.1. Notations and assumptions

Before discussing the theoretical results in detail, we revisit and introduce some notations and assumptions adopted in the manuscript. We denote the sample $\{(y_{i,n}, \boldsymbol{x}'_{i,n}, \boldsymbol{w}'_{i,n})'\}_{i=1}^n$ as $\{\boldsymbol{z}_{i,n}\}_{i=1}^n$. Recall $u_{i,n}$ is the random error in the considered linear model (1), we define

$$\varepsilon_{i,n} = u_{i,n} - \mathbb{E}[u_{i,n}|\{\boldsymbol{w}_{i,n}\}_{i=1}^n, \{\boldsymbol{x}_{i,n}\}_{i=1}^n], \quad \boldsymbol{v}_{i,n} = \boldsymbol{x}_{i,n} - \mathbb{E}[\boldsymbol{x}_{i,n}|\{\boldsymbol{w}_{i,n}\}_{i=1}^n], \quad i = 1, \ldots, n. \tag{4}$$

Let $e_{i,n} = \mathbb{E}[u_{i,n}|\{\boldsymbol{w}_{i,n}\}_{i=1}^n, \{\boldsymbol{x}_{i,n}\}_{i=1}^n]$, we further denote

$$\sigma^2_{i,n} = \mathbb{E}[\varepsilon^2_{i,n}|\{\boldsymbol{w}_{i,n}\}_{i=1}^n, \{\boldsymbol{x}_{i,n}\}_{i=1}^n], \quad \widetilde{\boldsymbol{v}}_{i,n} = \sum_{j=1}^n (\boldsymbol{M}_n)_{i,j}\boldsymbol{v}_{j,n},$$

$$\rho^1_n = \frac{\sum_{i=1}^n \mathbb{E}\big[e^2_{i,n}\big]}{n}, \quad \rho^2_n = \frac{\sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}\big(e_{i,n}|\{\boldsymbol{w}_{i,n}\}_{i=1}^n\big)^2\Big]}{n},$$

$$\boldsymbol{Q}_{i,n} = \mathbb{E}\Big[\boldsymbol{x}_{i,n} - \big(\sum_{j=1}^n \mathbb{E}[\boldsymbol{w}_{j,n}\boldsymbol{w}'_{j,n}]\big)^{-1} \sum_{j=1}^n \mathbb{E}[\boldsymbol{w}_{j,n}\boldsymbol{x}'_{j,n}]\Big|\{\boldsymbol{w}_{i,n}\}_{i=1}^n\Big],$$

$$\widetilde{\boldsymbol{Q}}_{i,n} = \sum_{j=1}^n (\boldsymbol{M}_n)_{i,j}\boldsymbol{Q}_{i,n}.$$

For a policy j, we define the near tie set in the population as:

$$\mathcal{H}_{(j)} = \big\{k : |\beta_k - \beta_{(j)}| = o(n^{-\frac{1}{2}}),\ k = 1, \ldots, d\big\}.$$

Next, let $\widehat{\boldsymbol{e}}_j$ denote a d-dimensional (sparse) vector with

$$\widehat{\boldsymbol{e}}_j = (\widehat{e}_{j,1}, \ldots, \widehat{e}_{j,d}), \quad \widehat{e}_{jk} = \frac{\mathbb{1}(k \in \widehat{\mathcal{H}}_{(j)})}{|\widehat{\mathcal{H}}_{(j)}|}, \quad k = 1, \ldots, d.$$

We will use the notation $\mathbb{P}(\cdot|\{\boldsymbol{z}_{i,n}\}_{i=1}^n)$ to refer to the probability that is conditional on the random variables $\{\boldsymbol{z}_{i,n}\}_{i=1}^n$.

We make following assumptions throughout this section. Note that Assumptions 1–4 listed below largely follow the assumptions in Cattaneo et al. (2018b) and Jochmans (2020), we list these assumptions along with their interpretations to present a full picture for our readers.

Assumption 1 (*Sampling*)**.** The errors $\varepsilon_{i,n}$ are uncorrelated across i conditional on $\{\boldsymbol{x}_{i,n}\}_{i=1}^n$ and $\{\boldsymbol{w}_{i,n}\}_{i=1}^n$. Let $\{N_1, \ldots, N_{G_n}\}$ represents a partition of $\{1, \ldots, n\}$ with $\max_{1\leq g\leq G_n} |N_g| = O(1)$ such that $\{(\varepsilon_{i,n}, \boldsymbol{v}_{i,n}), i \in N_g\}$ (defined in (4)) are independent across g conditional on $\{\boldsymbol{w}_{i,n}\}_{i=1}^n$.

Assumption 1 generalizes the classical independent and identically distributed (i.i.d.) setting to allow for repeated measurements or group structures in the observed data. For example, Assumption 1 allows the observed data to form clusters of finite sample sizes, and within-cluster dependency is allowed as long as the observations between clusters are independent.

Assumption 2 (*Design*)**.** The dimension of the covariates $\boldsymbol{w}_{i,n}$ satisfies that $\limsup_{n\to\infty} q_n/n < 1$. The minimum eigenvalue of the matrix $\sum_{i=1}^n \boldsymbol{w}_{i,n}\boldsymbol{w}'_{i,n}$ is bounded away from 0 with probability approaching one, that is

$$\lim_{n\to\infty} \mathbb{P}\Big(\lambda_{\min}\big(\sum_{i=1}^n \boldsymbol{w}_{i,n}\boldsymbol{w}'_{i,n}\big) > 0\Big) = 1.$$

Lastly,

$$\max_{1\le i\le n}\Big\{\mathbb{E}[\varepsilon_{i,n}^4|\{\boldsymbol{w}_{i,n}\}_{i=1}^n,\{\boldsymbol{x}_{i,n}\}_{i=1}^n],\ \frac{1}{\sigma_{i,n}^2},\ \mathbb{E}[\boldsymbol{v}_{i,n}^4|\{\boldsymbol{w}_{i,n}\}_{i=1}^n],\ 1/\lambda_{\min}\Big(\frac{\sum_{i=1}^n\mathbb{E}\big[\widetilde{\boldsymbol{v}}_{i,n}\widetilde{\boldsymbol{v}}_{i,n}'|\{\boldsymbol{w}_{i,n}\}_{i=1}^n\big]}{n}\Big)\Big\}=O_p(1).$$

Assumption 2 contains three conditions. The first condition allows the dimension of the covariates $\boldsymbol{w}_{i,n}$ to grow at the sample rate as the sample size n. The second condition requires the matrix $\sum_{i=1}^n\boldsymbol{w}_{i,n}\boldsymbol{w}_{i,n}'$ to be full rank, which is necessary otherwise the OLS estimator would not be able to calculate the matrix $\boldsymbol{M}_n$. Furthermore, as noted in Cattaneo et al. (2018b), such an assumption can be imposed by dropping any covariates in $\boldsymbol{w}_{i,n}$ that are collinear. The third condition contains conventional moment conditions for the covariates and heteroscedasticity.

Assumption 3 (*Linear Model Approximation*). $\sum_{i=1}^n\mathbb{E}[\|\mathbf{Q}_{i,n}\|^2]/n=O(1)$, $\rho_n^1+n(\rho_n^1-\rho_n^2)+\rho_n^1\cdot\sum_{i=1}^n\mathbb{E}[\|\mathbf{Q}_{i,n}\|^2]=o(1)$, and $\max_{1\le i\le n}\|\widehat{\boldsymbol{v}}_{i,n}\|/\sqrt{n}=o_p(1)$, $n\rho_n^1=O(1)$.

Assumption 3 mainly characterizes the difference between the mean squares of the conditional errors ρ_n^1 and the projection ρ_n^2 into the covariate space $\{\boldsymbol{w}_{i,n}\}$'s. The characterization of this difference involves $\sum_{i=1}^n\mathbb{E}[\|\mathbf{Q}_{i,n}\|^2]$ where $\mathbf{Q}_{i,n}$ describes the deviation of $\boldsymbol{x}_{i,n}$ from its population linear projection. Residuals of this linear projection, represented by $\widehat{\boldsymbol{v}}_{i,n}$'s, are assumed to satisfy a negligibility condition after a maximization over all i's. This negligibility condition regularizes the distributional connection between $\boldsymbol{x}_{i,n}$'s and $\boldsymbol{w}_{i,n}$'s. We note that if the mean squares of $\boldsymbol{x}_{i,n}$'s are bounded and that an exogeneity condition $e_{i,n}=0$ holds for all i and n, then the linear model approximation assumption naturally holds. Otherwise, if the exogeneity condition does not hold, Assumption 3 requires a small-bias condition $n\rho_n^1=O(1)$.

Assumption 4 (*Variance Estimation*). $\lim_{n\to\infty}\mathbb{P}(\min_i\big(\boldsymbol{M}_n\big)_{i,i}>0)=1$,

$$\mathbb{P}\Big(\min_i\big(\boldsymbol{M}_n\big)_{i,i}>0\Big)=O_p(1),\quad\frac{\sum_{i=1}^n\|\widetilde{\mathbf{Q}}_{i,n}\|^4}{n}=O_p(1),$$

and $\max_i\|\mu_{i,n}\|/\sqrt{n}=o_p(1)$ with $\mu_{i,n}=\mathbb{E}\big[y_{i,n}|\{\boldsymbol{x}_{i,n}\}_{i=1}^n,\{\boldsymbol{w}_{i,n}\}_{i=1}^n\big]$.

Assumption 4 has two major parts. The first part regularizes the diagonal elements $\big(\boldsymbol{M}_n\big)_{i,i}$'s, essentially requiring the smallest diagonal element to be consistently bounded away from zero when n tends to infinity. Even though it is difficult to provide broadly general primitives to validate this assumption, Assumption 2 of Cattaneo et al. (2018b), Assumption 4 of Jochmans (2020), and the discussions therein provide sufficient conditions for this assumption to hold. The second part regularizes $\mu_{i,n}$'s and $\widetilde{\mathbf{Q}}_{i,n}$'s in order to control the variance of $y_{i,n}$'s and the variance of $\mathbb{E}(\boldsymbol{v}_{i,n}|\{\boldsymbol{w}_{i,n}\}_{i=1}^n)$'s.

Assumption 5 (*Policy Effect Sizes*). For $\delta\in(0,\frac{1}{2})$, the asymptotic distance between the effects of policy $k\notin\mathcal{H}_{(j)}$ and $j\in\mathcal{H}_{(j)}$ diverges as $n\to\infty$:

$$n^\delta\cdot\min_{k\notin\mathcal{H}_{(j)}}\big|\beta_{(j)}-\beta_k\big|\to\infty,\ \text{as } n\to\infty,\quad j=1,\dots,d.$$

Assumption 5 requires that any policies outside the near tie set $\mathcal{H}_{(j)}$ have effect sizes sufficiently different from the ones in $\mathcal{H}_{(j)}$. In fixed dimensions when q_n does not grow with n, the underlying policy effect sizes $\beta_1,\dots,\beta_d$ are constant with respect to the sample size n. The near tie set reduces to a "precise" tie set $\mathcal{H}_{(j)}=\big\{k:\beta_k=\beta_{(j)},\ k=1,\dots,d\big\}$, suggesting that $\min_{k\notin\mathcal{H}_{(j)}}\big|\beta_{(j)}-\beta_k\big|$ is a positive constant bounded away from zero. In such a case, Assumption 5 is automatically satisfied.

3.2. Properties of the proposed estimator

For the proposed estimator, we show that the following theorem holds:

Theorem 1. *Under Assumptions 1–5, for any $t\in\mathbb{R}$, for the resampled statistics, the following holds*

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{\sqrt{n}\big(\widetilde{\beta}_{(j)}^*-\widetilde{\beta}_{(j)}\big)}{(\widehat{\boldsymbol{e}}_j'\widehat{\boldsymbol{\Omega}}_n^{KJ}\widehat{\boldsymbol{e}}_j)^{\frac{1}{2}}}\le t\Big|\{(y_{i,n},\boldsymbol{x}_{i,n}',\boldsymbol{w}_{i,n}')'\}_{i=1}^n\right)=\Phi(t).$$

For the original statistics, it holds that

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{\sqrt{n}\big(\widetilde{\beta}_{(j)}-\beta_{(j)}\big)}{(\widehat{\boldsymbol{e}}_j'\widehat{\boldsymbol{\Omega}}_n^{KJ}\widehat{\boldsymbol{e}}_j)^{\frac{1}{2}}}\le t\right)=\Phi(t).$$

Furthermore, we have that $\lim_{n\to\infty}\mathbb{P}\Big(\mathbb{P}\big(\widetilde{\beta}_{(j)}^*\le\beta_{(j)}|\{\boldsymbol{z}_{i,n}\}_{i=1}^n\big)\le s\Big)=s.$

Theorem 1 confirms that our proposed confidence interval for $\beta_{(j)}$ achieves exact $1-\alpha$ coverage probability as the sample size goes to infinity when B is sufficiently large, which distinguishes the proposed inference procedure from simultaneous methods. Furthermore, Theorem 1 says that $\widetilde{\beta}_{(j)}$ is a root-n consistent estimator of $\beta_{(j)}$, in the sense that $\forall \varepsilon > 0$, there exists $M > 0$ such that $\mathbb{P}\big(|\sqrt{n}(\widetilde{\beta}_{(j)} - \beta_{(j)})| > M\big) \leq \varepsilon$, for $n \geq 1$.

As for the observed best policies, recall that we denote the observed jth largest policy as

$$\widehat{j} = \sum_{k=1}^{d} k \cdot \mathbb{1}(\widehat{\beta}_k = \widehat{\beta}_{(j)}). \tag{5}$$

The following corollary suggests that the proposed confidence interval for $\beta_{(j)}$ can also serve as an exact prediction interval for $\beta_{\widehat{j}}$. Therefore, the proposed procedure in Section 2.2 can also be used to make inference on the observed top policies in a random sample:

Corollary 1. *Under Assumptions 1–5, we have that* $\lim_{n\to\infty} \mathbb{P}\Big(\mathbb{P}\big(\widetilde{\beta}^*_{(j)} \leq \beta_{\widehat{j}} | \{\boldsymbol{z}_{i,n}\}_{i=1}^n\big) \leq s\Big) = s$. *Furthermore,* $\widetilde{\beta}_{(j)}$ *is a "root-n consistent" estimator of the data-dependent parameter* $\beta_{\widehat{j}}$ *in the sense that* $\forall \varepsilon > 0$, *there exists* $M > 0$ *such that* $\mathbb{P}\big(|\sqrt{n}(\widetilde{\beta}_{(j)} - \beta_{\widehat{j}})| > M\big) \leq \varepsilon$, *for* $n \geq 1$.

Remark 3 (*Regression Models with Fixed Effects*)**.** The proposed resampling-based approach can be used to calibrate multiple best policies when fixed effects are introduced in linear regression models (see Verdier (2020) for comprehensive discussion). This suggests that our approach not only applies to independently sampled data, but also remains valid when there are repeated-measurements present in the data. These may include short panel data, and datasets in which, for example, two individuals have sampled from each household. To conserve space in the main manuscript, we have leave the detailed discussion in the Supplementary Materials (Section D).

Remark 4 (*Data Dependent Choice of* d_0)**.** In addition to a deterministic choice of d_0, another practically relevant scenario is a data dependent choice of d_0. An example of such a data dependent choice is $\widehat{d}_0 = \max\{k : \widehat{\beta}_{(k)} > C\}$, where C is a user-specified threshold for the effect size. A relatively complicated situation is that C coincides with some of the policy sizes in $\beta_1, \beta_2, \ldots, \beta_d$. In this situation, it is possible that no matter how large n is, $\widehat{d}_0$ does not converge to a deterministic value but instead to a non-degenerate random variable. For the purpose of separation, we may adjust $\widehat{d}_0 = \max\{k : \widehat{\beta}_{(k)} > C\}$ to be $\widehat{d}'_0 = \max\{k : \widehat{\beta}_{(k)} > C + C_1 \cdot n^{-0.25}\}$, where C_1 is a constant that does not depend on n. The choice of -0.25 is tunable and may be of independent interest. By this new choice of $\widehat{d}'_0$, the policy effects that exactly equal C will be eliminated almost surely when n tends to infinity. This elimination exactly matches the target to select all the policy sizes that are larger than C. In the limit of n tending to infinity, $\max\{k : \widehat{\beta}_{(k)} > C + C_1 \cdot n^{-0.25}\}$ will converge almost surely to a set that contains all effect sizes larger than C. Therefore, the large-sample theory results for a pre-specified deterministic integer would still hold by plugging in $\widehat{d}'_0$.

4. Simulation studies

4.1. Simulation design

We generate i.i.d. Monte Carlo samples of $\{(y_{i,n}, \boldsymbol{x}'_{i,n}, \boldsymbol{w}'_{i,n})\}_{i=1}^n$ from the model

$$y_{i,n} = \boldsymbol{x}'_{i,n}\boldsymbol{\beta} + \boldsymbol{w}'_{i,n}\gamma_n + \varepsilon_{i,n}, \quad i = 1, \ldots, n.$$

We consider various data generating processes (DGP) for different choices of the policy variable $\boldsymbol{x}_{i,n}$, the covariates $\boldsymbol{w}_{i,n}$ and the random noise $\varepsilon_{i,n}$. The first DGP follows a similar setup taking from Jochmans (2020) and Cattaneo et al. (2018b), where we generate many (sparse) dummy variables entering the estimation of $\boldsymbol{\beta}$. We generate $\boldsymbol{x}_{i,n} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{jk} = 0.5^{|j-k|}$ for $j, k = 1, \ldots, d$, $\boldsymbol{w}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{w}}_{i,n} \geq \Phi^{-1}(0.98))$ with $\widetilde{\boldsymbol{w}}_{i,n} \sim \mathcal{N}(0, \boldsymbol{I}_{q_n})$ and $\boldsymbol{I}_{q_n}$ is a q_n-dimensional identity matrix, and $\varepsilon_{i,n} \sim \mathcal{N}(0, 1)$. The second DGP considers a case with dummy policy random variables and normal covariates, where we generate $\boldsymbol{x}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{x}}_{i,n} > 0)$ with $\widetilde{\boldsymbol{x}}_{i,n} \sim \mathcal{N}(0, \Sigma)$, $\boldsymbol{w}_{i,n} \sim \mathcal{N}(0, \boldsymbol{I}_{q_n})$ and $\varepsilon_{i,n} \sim \mathcal{N}(0, 1)$. In the Supplementary Materials, we have further included DGPs with more realistic error terms beyond normal distribution, including error terms with asymmetric and bimodal distributions. For most of the DGPs, we investigate both homoscedastic as well as heteroscedastic models. See Supplementary Materials Section C for detailed description and simulation results.

As for the coefficients, we consider three DGPs that vary in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_n$. The first DGP considers the case in which no policy is effective (meaning that $\boldsymbol{\beta} = 0$), and the coefficient $\gamma_j = 1/j$, for $j = 1, \ldots, q_n$. We refer to this case as the "homogeneity" case since β_j's take the same value zero. The second and the third DGPs consider cases where policy effects are generated from $\beta_j = \Phi^{-1}\big(\frac{j}{d+1}\big)$ for $j = 1, \ldots, d$, and the coefficients are either $\boldsymbol{\gamma}_n = 0$ or $\gamma_j = 1/j$, for $j = 1, \ldots, q_n$. We refer to this case as the "heterogeneity(1)" case and "heterogeneity(2)" case, respectively, since different policies have heterogeneous effects.

We set the sample size $n \in \{700, 2000\}$ to mimic the sample size in our case studies, the number of policies $d \in \{5, 10\}$, and the dimension of the covariates q_n from $q_n \in \{1, 141, 281, 421, 561, 631\}$. All statistics reported below are computed

based on over 1000 Monte Carlo replications. To avoid redundancy, we present the results for $n = 700$ and $d = 5$ in the main manuscript, and rests are provided in the Supplementary Materials (Section C).

To demonstrate the robustness of the adopted covariance matrix estimator, we compare our proposal with three alternative covariance matrix estimators. The first one we compare with is the covariance matrix estimator proposed by Cattaneo et al. (2018b): $\widehat{\boldsymbol{\Omega}}_n^{\mathtt{HCK}} = \widehat{\boldsymbol{\Gamma}}_n^{-1}\widehat{\boldsymbol{\Sigma}}_n^{\mathtt{HCK}}\widehat{\boldsymbol{\Gamma}}_n^{-1}$, where $\widehat{\boldsymbol{\Sigma}}_n^{\mathtt{HCK}} \triangleq \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \kappa_{ij,n}^{\mathtt{HCK}}\widehat{\boldsymbol{v}}_{i,n}\widehat{\boldsymbol{v}}_{i,n}'\widehat{u}_{j,n}^2$, $\widehat{u}_{j,n} = \sum_{k=1}^n(\boldsymbol{M}_n)_{j,k}(y_{k,n} - \boldsymbol{x}_{k,n}'\widehat{\boldsymbol{\beta}})$, and

$$\kappa_n^{\mathtt{HCK}} = \begin{pmatrix} M_{11,n}^2 & \cdots & M_{1n,n}^2 \\ \vdots & \ddots & \vdots \\ M_{n1,n}^2 & \cdots & M_{nn,n}^2 \end{pmatrix}^{-1} = (\boldsymbol{M}_n \odot \boldsymbol{M}_n)^{-1},$$

with $\odot$ denoting the Hadamard product. The estimator $\widehat{\boldsymbol{\Sigma}}_n^{\mathtt{HCK}}$ is well-defined whenever $(\boldsymbol{M}_n \odot \boldsymbol{M}_n)$ is invertible. We use the acronym "HCK" to denote this estimator in the following parts. The second one we compare with is the classical Eicker–White covariance matrix estimator (Eicker, 1963; White, 1980) of the form:

$$\widehat{\boldsymbol{\Omega}}_n^{\mathtt{EW}} = \widehat{\boldsymbol{\Gamma}}_n^{-1}\widehat{\boldsymbol{\Sigma}}_n^{\mathtt{EW}}\widehat{\boldsymbol{\Gamma}}_n^{-1}, \tag{6}$$

where $\widehat{\boldsymbol{\Sigma}}_n^{\mathtt{EW}} \triangleq \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{v}}_{i,n}\widehat{\boldsymbol{v}}_{i,n}'\widehat{u}_{i,n}^2$ and $\widehat{u}_{i,n} = \sum_{j=1}^n(\boldsymbol{M}_n)_{i,j}(y_{j,n} - \boldsymbol{x}_{j,n}'\widehat{\boldsymbol{\beta}})$. We use the acronym "EW" to denote this estimator in our simulation results section. Huber–Eicker–White standard error is also known as the HC0 standard error, where HC stands for "heteroskedasticity robust". The last covariance matrix estimator we adopted is a variant of the HC0 estimator:

$$\widehat{\boldsymbol{\Omega}}_n^{\mathtt{HC3}} = \widehat{\boldsymbol{\Gamma}}_n^{-1}\widehat{\boldsymbol{\Sigma}}_n^{\mathtt{HC3}}\widehat{\boldsymbol{\Gamma}}_n^{-1}, \quad \text{where } \widehat{\boldsymbol{\Sigma}}_n^{\mathtt{HC3}} \triangleq \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{v}}_{i,n}\widehat{\boldsymbol{v}}_{i,n}'\frac{\widehat{u}_{i,n}^2}{(\boldsymbol{M}_n)_{i,j}^2}. \tag{7}$$

The above estimator upward reweights regression residuals, and we use the acronym "HC3" to denote this estimator in our simulation results section.

4.2. Simulation results

We summarize our main takeaways from the simulation results presented in Tables 1–3, where we have compared our proposed approach ("Proposed + KJ") in Section 2.2 with four other methods. "Proposed + EW", "Proposed + HC3", and "Proposed + HCK" refer to methods adjusting for the winner's curse bias but use $\widehat{\boldsymbol{\Omega}}_n^{\mathtt{EW}}$, $\widehat{\boldsymbol{\Omega}}_n^{\mathtt{HC3}}$, and $\widehat{\boldsymbol{\Omega}}_n^{\mathtt{HCK}}$, respectively, to estimate the covariance matrix of $\boldsymbol{\beta}$. "No adjustment+KJ" refers to the approach with no adjustment for the winner's curse bias and adopts the robust covariance matrix estimator proposed by Jochmans (2020) to make inference on the best policies. We present the coverage probabilities and $\sqrt{n}$-scaled biases for the top two policies in the population, i.e., $\beta_{(1)}$ and $\beta_{(2)}$. As the simulation results are rather similar for the observed top two policies in the random sample, i.e., $\beta_{\widehat{1}}$, $\beta_{\widehat{2}}$, we present these results in the Supplementary Materials (Section C.4).

Our simulation results confirm our theoretical results presented in Theorem 1. When no policy is effective, our proposed method not only successfully suppresses the winner's curse bias for the top two policies but also attains near nominal coverage (Table 2). Similar pattern can also be observed when top policies are effective (i.e., β_j's are heterogeneous, and Table 1 in particular). In nearly all designs and for a range of considered values of q_n, our proposal yields close to nominal confidence interval, though some under coverage is observed for large values of q_n. The method with no adjustment is obviously biased upward due to the winner's curse phenomenon, thus it provides under-covered confidence intervals and point estimates with rather large biases. In all considered cases, both the EW-based method and the HC3-based method tend to lose coverage when $q_n \geq 141$, and the HCK-based method tends to produce under-covered confidence interval whenever $q_n \geq 561$. In moderately high dimensions so that q_n/n is approximately one half, the proposed method with the HCK variance estimator has comparable performances with our approach.

5. Case studies

5.1. Case study I: Charitable giving

In the past half century, charitable giving by individuals in the United States has grown and it has contributed to more than two percent of the annual GDP since 1998 (List, 2011). Charitable giving is often driven by altruism, while as suggested by many field experiments, improper policies adopted by the demand side–fundraisers–may impair the supply side's (individual donors) motivation of giving (Andreoni and Miller, 2002). Therefore, to effectively attract resources from individual donors, fundraisers need to properly design donation incentives. One of the donation incentives is matching grant which means that a matching donor pledges to match any donation from other donors with certain ratio and up to some threshold. As the price elasticity of matching donation may differ from other donation incentives, we hope to carefully investigate different pricing policies in a matching donation and study if the observed top two performing policies are indeed effective.

Table 1
Simulation results ($d = 5$, heterogeneity, $\beta_{(1)}$).

		$\beta_j = \Phi^{-1}(\frac{j}{d+1})$, $\gamma_n = 0$, j = 1,..., d				
		$\boldsymbol{x}_{i,n} \sim \mathcal{N}(0, \Sigma)$, $\boldsymbol{w}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{w}}_{i,n} \geq \Phi^{-1}(0.98))$				
		Proposed+KJ	Proposed+HCK	Proposed+HC3	Proposed+EW	No adjustment+KJ
$q_n = 1$	Cover	0.97(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.97(0.01)
	$\sqrt{n}$Bias	−0.04(0.06)	−0.03(0.04)	−0.03(0.04)	−0.04(0.05)	0.05(0.06)
$q_n = 141$	Cover	0.96(0.01)	0.96(0.01)	0.94(0.01)	0.95(0.01)	0.95(0.01)
	$\sqrt{n}$Bias	−0.04(0.05)	−0.04(0.04)	−0.04(0.04)	0.06(0.06)	0.06(0.07)
$q_n = 281$	Cover	0.96(0.01)	0.95(0.01)	0.82(0.02)	0.80(0.01)	0.94(0.01)
	$\sqrt{n}$Bias	−0.05(0.06)	−0.06(0.05)	−0.06(0.03)	−0.10(0.07)	−0.08(0.08)
$q_n = 421$	Cover	0.95(0.02)	0.94(0.01)	0.79(0.01)	0.76(0.01)	0.78(0.01)
	$\sqrt{n}$Bias	−0.05(0.05)	−0.06(0.05)	−0.07(0.05)	−0.12(0.09)	0.11(0.09)
$q_n = 561$	Cover	0.95(0.01)	0.92(0.01)	0.65(0.02)	0.63(0.01)	0.68(0.01)
	$\sqrt{n}$Bias	−0.07(0.07)	−0.09(0.07)	−0.17(0.10)	−0.20(0.12)	0.15(0.13)
$q_n = 631$[a]	Cover	0.93(0.01)	0.91(0.01)	0.51(0.02)	0.48(0.01)	0.55(0.01)
	$\sqrt{n}$Bias	−0.17(0.08)	−0.19(0.10)	−0.28(0.11)	−0.35(0.22)	−0.26(0.13)
		$\boldsymbol{x}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{x}}_{i,n} > 0)$, $\boldsymbol{w}_{i,n} \sim \mathcal{N}(0, I)$				
		Proposed+KJ	Proposed+HCK	Proposed+HC3	Proposed + EW	No adjustment+KJ
$q_n = 1$	Cover	0.97(0.01)	0.97(0.01)	0.95(0.01)	0.96(0.01)	0.97(0.01)
	$\sqrt{n}$Bias	−0.02(0.07)	−0.02(0.04)	−0.01(0.03)	−0.07(0.11)	−0.05(0.09)
$q_n = 141$	Cover	0.96(0.01)	0.95(0.01)	0.94(0.01)	0.94(0.01)	0.96(0.01)
	$\sqrt{n}$Bias	−0.02(0.03)	−0.02(0.02)	−0.03(0.02)	0.11(0.12)	−0.06(0.12)
$q_n = 281$	Cover	0.95(0.01)	0.94(0.01)	0.87(0.01)	0.85(0.01)	0.95(0.01)
	$\sqrt{n}$Bias	−0.03(0.04)	−0.03(0.03)	−0.04(0.02)	0.14(0.12)	−0.08(0.13)
$q_n = 421$	Cover	0.95(0.01)	0.94(0.01)	0.78(0.01)	0.76(0.01)	0.75(0.01)
	$\sqrt{n}$Bias	−0.03(0.03)	−0.05(0.04)	−0.08(0.05)	−0.19(0.17)	0.19(0.14)
$q_n = 561$	Cover	0.95(0.01)	0.92(0.01)	0.63(0.02)	0.61(0.01)	0.63(0.01)
	$\sqrt{n}$Bias	−0.04(0.04)	−0.08(0.06)	−0.19(0.10)	−0.30(0.26)	−0.24(0.22)
$q_n = 631$	Cover	0.94(0.01)	0.91(0.01)	0.49(0.02)	0.45(0.01)	0.68(0.01)
	$\sqrt{n}$Bias	−0.06(0.07)	−0.11(0.09)	−0.23(0.13)	−0.42(0.20)	0.39(0.29)

Note: "Cover" is the empirical coverage of the 95% confidence interval for $\beta_{(1)}$ and "$\sqrt{n}$Bias" captures the root-n scaled Monte Carlo bias for estimating $\beta_{(1)}$.
[a]Indicates that $\widehat{\Omega}_n^{KJ}$ is not positive semi-definite in some Monte Carlo samples.

We work with the charitable giving data in Karlan and List (2007). Karlan and List (2007) conduct a field experiment that explores the price elasticity in a matching donation. The field experiment involves 50,083 previous donors to a political charity. Individuals are randomly assigned to two groups: treatment ($n = 33{,}396$) and control ($n = 16{,}687$). In the control group, individuals receive a standard letter with no matching details. In the treatment group, each potential donor receives a letter with three strategies: (1) match ratio, (2) match size, and (3) ask amount. Within each strategy, individuals are randomly assigned to a sub-policy detailed below.

For the match ratio strategy, there are three sub-policies: (1) 1:1 (the matching donor contributes the same amount as the individual donor), (2) 2:1 (the matching donor contributes twice as many as the individual donor), (3) 3:1 (the matching donor contributes three times as many as the individual donor). For the match size strategy, there are four sub-policies with different pledge amounts: (1) $25,000, (2) $50,000, (3) $100,000, and (4) unstated amount. For the ask amount strategy, individual donors are asked to give same amount, 25% more or 50% more than their largest past donation.

In our study, we focus on the treatment "ask amount" with three pricing policies, and we study the subpopulation ($n = 7938$) of unmarried males living in red counties or red states. Red county (state) refers to a county (state) in which residents predominantly vote for the Republican Party. The outcome of interest is the donation amount. We have adjusted $q_n = 1049$ covariates including the donors' demographic information (26 variables), census information (27 variables), and their two-way interaction terms. Our results are summarized in Table 4.

Results in Table 4 suggest that, without any calibration, asking the donor either to give the same amount or to give 25% more than their highest past donation seems to be the best policies that significantly increase the donation amount. Specifically, our results from running a simple linear regression model suggest that asking the individual donor to give the same amount of their largest past donation appears to be the most effective pricing policy, and it on average raises $0.67 (95% CI = (0.09, 1.25), p-value = 0.023) per donor. Asking the individual donor to give 25% more than their largest past donation is the second most effective policy, with an increased donation by $0.66 (95% CI = (0.01, 1.31), p-value = 0.046) per donor.

Because we pick the most effective policies from a random sample, these estimates are potentially subject to the winner's curse bias. We thus apply the proposed method to carefully examine these seemly effective policies. After

Table 2
Simulation results ($d = 5$, homogeneity, $\beta_{(1)}$).

		$\boldsymbol{\beta} = 0, \quad \boldsymbol{\gamma}_j = 1/j$				
		$\boldsymbol{x}_{i,n} \sim \mathcal{N}(0, \Sigma), \quad \boldsymbol{w}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{w}}_{i,n} \geq \Phi^{-1}(0.98))$				
		Proposed+KJ	Proposed+HCK	Proposed+HC3	Proposed+EW	No adjustment+KJ
$q_n = 1$	Cover	0.97(0.01)	0.96(0.01)	0.96(0.01)	0.93(0.02)	0.90(0.01)
	$\sqrt{n}$Bias	0.02(0.03)	0.02(0.03)	0.03(0.03)	0.03(0.03)	1.64(0.04)
$q_n = 141$	Cover	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.89(0.02)	0.88(0.01)
	$\sqrt{n}$Bias	0.03(0.04)	0.04(0.04)	0.03(0.04)	0.12(0.04)	1.78(0.04)
$q_n = 281$	Cover	0.96(0.01)	0.94(0.01)	0.90(0.02)	0.85(0.01)	0.83(0.01)
	$\sqrt{n}$Bias	0.03(0.04)	0.04(0.04)	0.05(0.04)	0.22(0.03)	2.03(0.05)
$q_n = 421$	Cover	0.95(0.01)	0.93(0.01)	0.82(0.02)	0.79(0.01)	0.74(0.02)
	$\sqrt{n}$Bias	0.05(0.05)	0.18(0.05)	0.24(0.06)	0.36(0.03)	2.63(0.06)
$q_n = 561$	Cover	0.95(0.01)	0.93(0.01)	0.67(0.02)	0.73(0.01)	0.63(0.02)
	$\sqrt{n}$Bias	0.08(0.09)	0.51(0.05)	0.74(0.06)	0.44(0.04)	3.74(0.09)
$q_n = 631$[a]	Cover	0.93(0.01)	0.89(0.01)	0.53(0.02)	0.50(0.01)	0.45(0.02)
	$\sqrt{n}$Bias	0.18(0.09)	1.21(0.09)	1.84(0.11)	2.42(0.06)	5.10(0.12)
		$\boldsymbol{x}_{i,n} = \mathbb{1}(\widetilde{\boldsymbol{x}}_{i,n} > 0), \quad \boldsymbol{w}_{i,n} \sim \mathcal{N}(0, I)$				
		Proposed+KJ	Proposed+HCK	Proposed+HC3	Proposed+EW	No adjustment+KJ
$q_n = 1$	Cover	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.94(0.01)	0.90(0.01)
	$\sqrt{n}$Bias	0.03(0.04)	0.04(0.05)	0.05(0.05)	0.07(0.07)	2.75(0.06)
$q_n = 141$	Cover	0.96(0.01)	0.96(0.01)	0.93(0.01)	0.90(0.01)	0.83(0.01)
	$\sqrt{n}$Bias	0.05(0.05)	0.05(0.06)	0.17(0.07)	0.31(0.07)	3.29(0.08)
$q_n = 281$	Cover	0.95(0.01)	0.95(0.01)	0.90(0.01)	0.88(0.01)	0.75(0.02)
	$\sqrt{n}$Bias	0.07(0.08)	0.07(0.07)	0.31(0.08)	0.54(0.05)	3.59(0.08)
$q_n = 421$	Cover	0.95(0.01)	0.95(0.01)	0.85(0.01)	0.86(0.01)	0.65(0.02)
	$\sqrt{n}$Bias	0.04(0.04)	0.10(0.11)	0.73(0.13)	0.64(0.06)	4.58(0.11)
$q_n = 561$	Cover	0.93(0.01)	0.90(0.02)	0.59(0.02)	0.61(0.01)	0.53(0.02)
	$\sqrt{n}$Bias	0.13(0.07)	0.19(0.13)	2.00(0.13)	1.73(0.08)	5.90(0.13)
$q_n = 631$	Cover	0.90(0.01)	0.78(0.02)	0.38(0.02)	0.30(0.01)	0.33(0.02)
	$\sqrt{n}$Bias	0.50(0.12)	2.47(0.16)	5.16(0.19)	7.68(0.18)	6.51(0.19)

Note: "Cover" is the empirical coverage of the 95% confidence interval for $\beta_{(1)}$ and " $\sqrt{n}$Bias " captures the root-n scaled Monte Carlo bias for estimating $\beta_{(1)}$.
[a]Indicates that $\widehat{\Omega}_n^{\mathrm{KJ}}$ is not positive semi-definite in some Monte Carlo samples.

calibrating for the winner's curse bias, we confirm that the asking for the same amount policy remains as the most effective policy, though with a slightly smaller estimated effect size (Est = \$0.63, 95% CI = (0.10, 1.20), p-value = 0.025). This result is moderately aligned with the analysis in Karlan and List (2007), whose results suggest that donors from red states or red counties are more willing to contribute, partially because the collaborating charity is politically oriented. However, for the effect of the second best policy–asking to donate 25% more than past donation–is shifted downward, and it no longer has significant impact in promoting the donation amount (Est = \$0.56, 95% CI = (−0.01, 1.07), p-value = 0.052). This result might be partially explained by the observation that donors are more motivated by a lower "price" of donation (Warwick, 2003). In sum, our analyses suggest that the best pricing policy of charitable giving for unmarried males living in the Republican Party dominated voting regions could be asking for the same amount as their highest previous donation, and asking for more donations may not incentivize the donors to give. Though given the obtained p-value before and after calibration for the second best policy is rather close to the 5 percent threshold, we note that such a conclusion should also be viewed with caution.

5.2. Case study II: National supported work (NSW) program

In this case study, we revisit a dataset from the National Supported Work (NSW) program. The NSW program is a labor training program implemented in 1970's that provides work experience to disadvantaged workers. Our proposed method can also be used to evaluate if the job training program is indeed beneficial for certain groups of workers. To do so, the structural component $\boldsymbol{x}_{i,n}$ in the model (1) would include variables representing the interactions between the treatment variable (the job training program) and different subgroup indicator variables of interest.

We use the field experiment dataset adopted in Dehejia and Wahba (2002) ($n = 455$), in which 185 workers are in the treatment group and 260 workers are in the control group. This dataset consists of a treatment indicator variable, an outcome variable measured by the participant post-treatment earnings in 1978, and eight baseline variables (including age, years of education, an indicator for high school degree, indicators for Black and Hispanic, marital status, and pre-treatment earnings in 1974 and 1975). We further add three sets of additional covariates following the setup in Farrell

Table 3
Simulation results ($d = 5$, heterogeneity, $\beta_{(2)}$).

		$\beta_j = \Phi^{-1}(\frac{j}{d+1})$, $\gamma_n = 0$				
		$x_{i,n} \sim \mathcal{N}(0, \Sigma)$, $w_{i,n} = \mathbb{1}(\widetilde{w}_{i,n} \geq \Phi^{-1}(0.98))$				
		Proposed+KJ	Proposed+HCK	Proposed+HC3	Proposed+EW	No adjustment+KJ
$q_n = 1$	Cover	0.97(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)	0.96(0.01)
	$\sqrt{n}$Bias	0.02(0.06)	0.02(0.06)	0.01(0.06)	−0.05(0.07)	−0.04(0.07)
$q_n = 141$	Cover	0.97(0.01)	0.96(0.01)	0.93(0.01)	0.94(0.01)	0.94(0.01)
	$\sqrt{n}$Bias	−0.02(0.04)	−0.02(0.04)	−0.04(0.03)	−0.06(0.06)	−0.07(0.08)
$q_n = 281$	Cover	0.95(0.01)	0.94(0.01)	0.88(0.02)	0.89(0.01)	0.85(0.01)
	$\sqrt{n}$Bias	−0.03(0.04)	−0.03(0.03)	−0.07(0.03)	−0.11(0.10)	0.19(0.11)
$q_n = 421$	Cover	0.95(0.01)	0.94(0.02)	0.81(0.02)	0.80(0.01)	0.77(0.01)
	$\sqrt{n}$Bias	−0.03(0.03)	−0.03(0.04)	−0.10(0.06)	−0.15(0.12)	−0.23(0.15)
$q_n = 561$	Cover	0.95(0.01)	0.94(0.02)	0.67(0.02)	0.65(0.01)	0.63(0.01)
	$\sqrt{n}$Bias	0.03(0.03)	0.05(0.06)	0.12(0.08)	−0.17(0.13)	−0.26(0.18)
$q_n = 631$[a]	Cover	0.94(0.01)	0.93(0.01)	0.56(0.02)	0.53(0.01)	0.50(0.01)
	$\sqrt{n}$Bias	−0.07(0.07)	−0.18(0.06)	−0.23(0.08)	−0.26(0.17)	0.47(0.22)
		$x_{i,n} = \mathbb{1}(\widetilde{x}_{i,n} > 0)$, $w_{i,n} \sim \mathcal{N}(0, I)$				
		Proposed+KJ	Proposed+HCK	Proposed+HC3	Proposed+EW	No adjustment+KJ
$q_n = 1$	Cover	0.98(0.01)	0.98(0.01)	0.98(0.01)	0.96(0.01)	0.98(0.01)
	$\sqrt{n}$Bias	−0.08(0.12)	−0.09(0.12)	−0.10(0.13)	0.09(0.12)	−0.07(0.10)
$q_n = 141$	Cover	0.97(0.01)	0.97(0.01)	0.95(0.01)	0.95(0.01)	0.97(0.01)
	$\sqrt{n}$Bias	−0.09(0.12)	−0.10(0.13)	−0.12(0.13)	0.09(0.13)	−0.08(0.10)
$q_n = 281$	Cover	0.97(0.01)	0.97(0.01)	0.90(0.01)	0.87(0.01)	0.96(0.01)
	$\sqrt{n}$Bias	−0.11(0.14)	−0.10(0.14)	−0.15(0.14)	−0.18(0.14)	−0.10(0.11)
$q_n = 421$	Cover	0.96(0.01)	0.95(0.02)	0.80(0.02)	0.75(0.01)	0.94(0.01)
	$\sqrt{n}$Bias	0.14(0.16)	−0.16(0.18)	−0.20(0.18)	−0.22(0.17)	−0.15(0.15)
$q_n = 561$	Cover	0.96(0.01)	0.94(0.02)	0.63(0.02)	0.60(0.01)	0.92(0.01)
	$\sqrt{n}$Bias	0.14(0.18)	0.20(0.23)	−0.24(0.22)	−0.30(0.23)	0.19(0.20)
$q_n = 631$	Cover	0.94(0.01)	0.93(0.02)	0.58(0.02)	0.55(0.01)	0.52(0.01)
	$\sqrt{n}$Bias	0.15(0.20)	0.24(0.26)	0.28(0.25)	−0.35(0.13)	0.56(0.24)

Note: "Cover" is the empirical coverage of the 95% confidence interval for $\beta_{(2)}$ and "$\sqrt{n}$Bias" captures the root-n scaled Monte Carlo bias for estimating $\beta_{(2)}$.
[a]Indicates that $\widehat{\Omega}_n^{KJ}$ is not positive semi-definite in some Monte Carlo samples.

Table 4
Estimated treatment effects (Est), 95% confidence intervals (95% CI), and two-sided p-values for the three "ask amount" policies.

Method	Policies(Ask amount)	Est (95% CI)	p-value
Uncalibrated	Same	0.67 (0.09, 1.25)	0.023*
	25% more	0.66 (0.01, 1.31)	0.046*
	50% more	0.33 (−0.21, 0.86)	0.235
Calibrated	Same	0.63 (0.10, 1.20)	0.025*
	25% more	0.56(−0.01, 1.07)	0.052

"Uncalibrated" refers to the study results obtained without any adjustment, and the confidence intervals are constructed based on normal approximation with the estimated covariance matrix $\widehat{\Omega}_n^{KJ}$. "Calibrated" refers to our proposed methodology. The computational time is 741 seconds on a Lenovo NeXtScale nx360m5 node (24 cores per node) equipped with Intel Xeon Haswell processor.

(2015): (1) $\mathbb{1}$(1974 earnings =0) and $\mathbb{1}$(1975 earnings =0); (2) all first-order interactions; (3) all polynomials up to the 2nd-order. The final dataset includes 51 covariates. We aim to investigate the effectiveness of the NSW program in four groups of workers: (1) married Black workers, (2) unmarried Black workers, (3) married Non-Black workers, and (4) unmarried Non-Black workers. The summarized results are shown in Table 5.

Table 5 demonstrates that without adjusting for the winner's curse bias, married Black workers (estimated treatment effect = 4.35, 95% CI = (0.89, 7.81), p-value = 0.014, in units $\$10^3$) seem to benefit from the program the most. After accounting for the winner's curse bias issue, our approach potentially confirms that the treatment effect of the NSW program for the married Black workers is still significant, and the calibrated treatment effect remains roughly the same (Est= 4.41, 95% CI = (1.74, 8.50), p-value = 0.009, in units $\$10^3$/year).

The dataset collected from the NSW program has been frequently analyzed in the past decade, and our results are largely in-line with current understandings gathered in past studies. For example, although not focusing on the same

Table 5
Estimated treatment effects (Est), 95% confidence intervals (CI), in units $\$10^3$/year, and two-sided p-values for the four subgroups in the NSW study ($n = 445$, $q_n = 51$).

Method	Subgroups	Est (95% CI) ($\$10^3$)	p-value
Uncalibrated	Black, married	4.35 (0.89, 7.81)	0.014*
	Black, unmarried	1.10(−0.55, 2.75)	0.190
	Non-Black, married	1.33(−6.63, 9.29)	0.743
	Non-Black, unmarried	1.40(−2.61, 5.40)	0.494
Calibrated	Black, married	4.41(1.74, 8.50)	0.009*

"Uncalibrated" refers to the study results obtained without any adjustment, and the confidence intervals are constructed based on normal approximation with the estimated covariance matrix $\widehat{\Omega}_n^{\mathrm{KJ}}$. "Calibrated" refers to our proposed methodology. The computational time is 122 seconds on a Lenovo NeXtScale nx360m5 node (24 cores per node) equipped with Intel Xeon Haswell processor.

groups of workers, Imai and Ratkovic (2013) suggest that married and unemployed Black workers with some college education have increased their post-treatment earnings for about 38%. Dehejia and Wahba (2002) show that the job training program yields positive treatment effect on the overall Black participants. In this case study, our approach may help to confirm the seemly effective subgroup observed in a random sample while providing a statistically justified estimate accounting for the winner's curse bias.

6. Concluding remarks

In this article, we have introduced an approach to evaluate multiple best policies based on resampling in the context of a linear model with many covariates. While our approach is numerically reliable and theoretically grounded, it is worthwhile to generalize our framework so that the policy effects can be estimated with other off-shelf methods that are, for example, robust to the high-dimensional confounders or to the presence of interference and noncompliance. Our current theoretical analysis suggests that our proposed approach can be readily extended as long as the covariance matrix between different policies can be consistently estimated. It is thus desirable for us to provide a general framework to broaden future applications for other disciplines in general.

Acknowledgments

The research of Jingshen Wang is supported in part by the National Science Foundation, United States (DMS 2015325) and the National Institute of Health (R01MH125746).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.06.013.

References

Abadie, Alberto, Cattaneo, Matias D., 2018. Econometric methods for program evaluation. Annu. Rev. Econ. 10, 465–503.
Anatolyev, Stanislav, 2012. Inference in regression models with many regressors. J. Econometrics 170 (2), 368–382.
Andreoni, James, Miller, John, 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica 70 (2), 737–753.
Andrews, Donald W.K., 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. Econometrica 399–405.
Andrews, Isaiah, Bowen, Dillon, Kitagawa, Toru, McCloskey, Adam, 2022. Inference for losers. Am. Econ. Assoc. Pap. Proc..
Andrews, Isaiah, Kitagawa, Toru, McCloskey, Adam, 2019. Inference on winners. Technical report, National Bureau of Economic Research.
Athey, Susan, Imbens, Guido W., 2017. The state of applied econometrics: Causality and policy evaluation. J. Econ. Perspect. 31 (2), 3–32.
Belloni, Alexandre, Chernozhukov, Victor, Hansen, Christian, 2014. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81 (2), 608–650.
Cattaneo, Matias D., Jansson, Michael, Ma, Xinwei, 2019. Two-step estimation and inference with possibly many included covariates. Rev. Econom. Stud. 86 (3), 1095–1122.
Cattaneo, Matias D., Jansson, Michael, Newey, Whitney K., 2018a. Alternative asymptotics and the partially linear model with many regressors. Econom. Theory 34 (2), 277–301.
Cattaneo, Matias D., Jansson, Michael, Newey, Whitney K., 2018b. Inference in linear regression models with many covariates and heteroscedasticity. J. Amer. Statist. Assoc. 113 (523), 1350–1361.
Charness, Gary, Levin, Dan, 2009. The origin of the winner's curse: a laboratory study. Am. Econ. J. Microecon. 1 (1), 207–236.
Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, Newey, Whitney, Robins, James, 2018. Double/debiased machine learning for treatment and structural parameters.
Chernozhukov, Victor, Chetverikov, Denis, Kato, Kengo, 2013a. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Statist. 41 (6), 2786–2819.
Chernozhukov, Victor, Lee, Sokbae, Rosen, Adam M., 2013b. Intersection bounds: estimation and inference. Econometrica 81 (2), 667–737.
Claggett, Brian, Xie, Minge, Tian, Lu, 2014. Meta-analysis with fixed, unknown, study-specific parameters. J. Amer. Statist. Assoc. 109 (508), 1660–1671.

Dawid, A.P., 1994. Selection paradoxes of Bayesian inference. In: Multivariate Analysis and its Applications. Institute of Mathematical Statistics, pp. 211–220.
Dehejia, Rajeev H., Wahba, Sadek, 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. 84 (1), 151–161.
Efron, Bradley, 2011. Tweedie's formula and selection bias. J. Amer. Statist. Assoc. 106 (496), 1602–1614.
Eicker, Friedhelm, 1963. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. Ann. Math. Stat. 447–456.
El Karoui, Noureddine, Bean, Derek, Bickel, Peter J, Lim, Chinghway, Yu, Bin, 2013. On robust regression with high-dimensional predictors. Proc. Natl. Acad. Sci. 110 (36), 14557–14562.
El Karoui, Noureddine, Purdom, Elizabeth, 2018. Can we trust the bootstrap in high-dimensions? The case of linear models. J. Mach. Learn. Res. 19 (1), 170–235.
Fan, Jianqing, Hall, Peter, Yao, Qiwei, 2007. To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? J. Amer. Statist. Assoc. 102 (480), 1282–1288.
Fan, Jianqing, Han, Fang, Liu, Han, 2014. Challenges of big data analysis. Natl. Sci. Rev. 1 (2), 293–314.
Fan, Jianqing, Imai, Kosuke, Lee, Inbeom, Liu, Han, Ning, Yang, Yang, Xiaolin, 2021. Optimal covariate balancing conditions in propensity score estimation. J. Bus. Econom. Statist. 1–14.
Fan, Jianqing, Li, Runze, Zhang, Cun-Hui, Zou, Hui, 2020. Statistical Foundations of Data Science. Chapman and Hall/CRC.
Fan, Jianqing, Lv, Jinchi, Qi, Lei, 2011. Sparse high-dimensional models in economics. Annu. Rev. Econ. 3 (1), 291–317.
Fan, Jianqing, Shao, Qi-Man, Zhou, Wen-Xin, 2018. Are discoveries spurious? Distributions of maximum spurious correlations and their applications. Ann. Statist. 46 (3), 989.
Fan, Jianqing, Zhou, Wen-Xin, 2016. Guarding against spurious discoveries in high dimensions. J. Mach. Learn. Res. 17 (1), 7123–7156.
Farrell, Max H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econometrics 189 (1), 1–23.
Van de Geer, Sara, Bühlmann, Peter, Ritov, Ya'acov, Dezeure, Ruben, 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.
Guo, Xinzhou, He, Xuming, 2020. Inference on selected subgroups in clinical trials. J. Amer. Statist. Assoc. 1–19.
Hall, Peter, Miller, Hugh, 2010. Bootstrap confidence intervals and hypothesis tests for extrema of parameters. Biometrika 97 (4), 881–892.
Huber, Peter J., 1973. Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist. 799–821.
Imai, Kosuke, Ratkovic, Marc, 2013. Estimating treatment effect heterogeneity in randomized program evaluation. Ann. Appl. Stat. 7 (1), 443–470.
Jochmans, Koen, 2020. Heteroscedasticity-robust inference in linear regression models with many covariates. J. Amer. Statist. Assoc. 1–10.
Karlan, Dean, List, John A., 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. Amer. Econ. Rev. 97 (5), 1774–1793.
Kueck, Jannis, Luo, Ye, Spindler, Martin, Wang, Zigan, 2023. Estimation and inference of treatment effects with l2-boosting in high-dimensional settings. J. Econometrics 234 (2), 714–731.
LaLonde, Robert J., 1986. Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. 604–620.
Lee, Minyong R., Shen, Milan, 2018. Winner's curse: Bias estimation for total effects of features in online controlled experiments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 491–499.
List, John A., 2011. The market for charitable giving. J. Econ. Perspect. 25 (2), 157–180.
Long, J. Scott, Ervin, Laurie H., 2000. Using heteroscedasticity consistent standard errors in the linear regression model. Amer. Statist. 54 (3), 217–224.
Mammen, Enno, 1993. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist. 255–285.
Mammen, Enno, 2012. When Does Bootstrap Work?: Asymptotic Results and Simulations, vol. 77. Springer Science & Business Media.
Rai, Yoshiyasu, 2018. Statistical inference for treatment assignment policies. Unpublished Manuscript.
Vazquez-Bare, Gonzalo, 2021. Identification and estimation of spillover effects in randomized experiments. J. Econometrics Forthcoming.
Verdier, Valentin, 2020. Estimation and inference for linear models with two-way fixed effects and sparsely matched data. Rev. Econ. Stat. 102 (1), 1–16.
Warwick, Mal, 2003. Testing, Testing 1, 2, 3: Raise more Money with Direct Mail Tests. John Wiley & Sons.
White, Halbert, 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 817–838.
Xie, Minge, Singh, Kesar, Zhang, Cun-Hui, 2009. Confidence intervals for population ranks in the presence of ties and near ties. J. Amer. Statist. Assoc. 104 (486), 775–788.
Zhang, Cun-Hui, Zhang, Stephanie S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 217–242.