Endogenous Stratification in Randomized Experiments

Alberto Abadie
Harvard University

Matthew M. Chingos
Brookings Institution

Martin R. West
Harvard University

April 2014

Abstract

Researchers and policy makers are often interested in estimating how treatments or policy interventions affect the outcomes of those most in need of help. This concern has motivated the increasingly common practice of disaggregating experimental data by groups constructed on the basis of an index of baseline characteristics that predicts the values that individual outcomes would take on in the absence of the treatment. This article shows that substantial biases may arise in practice if the index is estimated, as is often the case, by regressing the outcome variable on baseline characteristics for the full sample of experimental controls. We analyze the behavior of leave-one-out and repeated split sample estimators and show they behave well in realistic scenarios, correcting the large bias problem of the full sample estimator. We use data from the National JTPA Study and the Tennessee STAR experiment to demonstrate the performance of alternative estimators and the magnitude of their biases.

Alberto Abadie, John F. Kennedy School of Government, Harvard University, alberto_abadie@harvard.edu. Matthew Chingos, Brown Center on Education Policy, Brookings Institution, mchingos@brookings.edu. Martin West, Harvard Graduate School of Education, Harvard University, martin_west@gse.harvard.edu.
We thank Beth Akers, Josh Angrist, Matias Cattaneo, Gary Chamberlain, David Deming, Sara Goldrick-Rab, Josh Goodman, Jerry Hausman, Guido Imbens, Max Kasy, Larry Katz, Amanda Pallais, Paul Peterson, Russ Whitehurst, and seminar participants at Harvard/MIT for helpful comments and discussions, and Jeremy Ferwerda for developing `estrat` (available at SSC), a Stata package that calculates the leave-one-out and repeated split sample endogenous stratification estimators considered in this study.

1. Introduction

Recent years have seen rapid growth in the use of randomized experiments in social science research. In part, this has been motivated by the "credibility revolution" in which researchers have devoted much attention to the study of the conditions that allow estimation of treatment effects (Angrist and Pischke 2010; Murnane and Willett 2011). The main advantage of a large and well-executed randomized experiment is that the researcher can confidently rule out the possibility that unobserved differences between the treatment and control groups could explain the study's results.

In addition to allowing estimation of average treatment effects, experiments also make it possible to obtain unbiased estimates of treatment effects for subgroups. Subgroup treatment effects are of particular interest to policymakers seeking to target policies on those most likely to benefit. As a general rule, subgroups must be created based on characteristics that are either immutable (e.g., race) or observed before randomization (e.g., on a baseline survey) so that they could not possibly have been affected by the treatment.

However, many researchers and policy makers are interested in estimating how treatments affect those most in need of help, that is, those who would attain unfavorable outcomes in the absence of the treatment. Treatment parameters of this nature depend on the joint distribution of potential outcomes with and without treatment, which is not identified by randomization (see, e.g., Heckman et al., 1997). A solution to this problem is to combine baseline characteristics into a single index that reflects each participant's predicted outcome without treatment, and conduct separate analysis for subgroups of participants defined in terms of intervals of the predicted outcome without treatment.

A well-known implementation of this idea is the use of data on out-of-sample untreated units to estimate a prediction model for the outcome variable, which can then be applied to predict outcomes without treatment for the experimental units. This approach is common in medical research, where validated risk models are available to stratify experimental subjects based on their predicted probability of certain health outcomes (Kent and Hayward 2007).

However, experimental studies in the social sciences often lack externally validated models that can be employed to predict the outcomes that experimental units would attain in the absence of the treatment. A potential approach to this problem that is gaining popularity among empirical researchers is to use in-sample information on the relationship between the outcome of interest and covariates for the experimental controls to estimate potential outcomes without treatment for all experimental units. We call this practice endogenous stratification, because it uses in-sample data on the outcome variable to stratify the sample.

Endogenous stratification is typically implemented in practice by first regressing the outcome variable on baseline characteristics using the full sample of experimental controls, and then using the coefficients from this regression to generate predicted potential outcomes without treatment for all sample units.

Unfortunately, as we show below, this procedure generates estimators of treatment effects that are substantially biased, and the bias follows a predictable pattern: results are biased upward for individuals with low predicted outcomes and biased downward for individuals with high predicted outcomes.

This bias pattern matches the results of several recent experimental studies that use this procedure and estimate strong positive effects for individuals with low predicted outcomes and, in some cases, negative effects for individuals with high predicted outcomes. For example, a 2011 working paper by Goldrick-Rab et al. reports that a Wisconsin need-based financial aid program for post-secondary education had no overall impacts on college enrollment or college persistence among eligible students as a whole. Looking separately at subgroups based on predicted persistence, however, the study finds large positive effects on enrollment after three years for students in the bottom third of predicted persistence and almost equally large negative effects for students in the top third of predicted persistence.[1]

[1] Goldrick-Rab et al. (2011) report that, for students in the bottom third group of predicted persistence, grant receipt was associated with an increase of 17 percentage points in enrollment three years after they started college. Conversely, for students in the top third group of predicted persistence, grant receipt was associated with a decrease of 15 percentage points in enrollment three years after the start of college. These findings were characterized by the authors as "exploratory" but received widespread media coverage, including articles in the *Chronicle of Higher Education*, *Inside Higher Education*, and *Education Week*. In

A 2011 working paper by Dynarski et al. analyzing long-term impacts of the Project STAR experiment similarly finds that assignment to a small class in grades K-3 increased college enrollment rates among the quintile of students with the lowest ex-ante probability to enroll by 11 percentage points, but had no impact on students in the top four quintiles. Pane et al. (2013) report experimental estimates of the effects of a technology-based algebra curriculum on the test scores of middle and high school students disaggregated by quintiles of predicted test scores. For middle school students exposed to the program in the first year of its implementation they find "potentially moderately large positive treatment effects in the lowest quintile and small negative effects of treatment in the highest two quintiles". Hemelt et al. (2012) find no significant average impacts in a experimental evaluation of the effects of two elementary school interventions on college enrolment or degree receipt. They report, however, significant positive impacts on two-year college enrollment for both interventions and on associate's degree completion for one of the interventions when they restrict the sample to students in the bottom quartile of in-sample predicted probability of college-attendance. Rodriguez-Planas (2012) reports that a mentoring program for adolescents reduced risky behavior and improved educational attainment for students in the top half of the risk distribution but increased risky behavior in the bottom half.[2]

Endogenous stratification also plays a supporting role in Angrist and Lavy's (2009) experimental evaluation of a cash incentive program aimed at increasing matriculation certification rates for Israeli high school students. In order to test whether the program was most effective for girls on the certification margin, the researchers first group students by baseline test scores. They also, however, report results for students grouped by ex-ante certification probability based on a broader set of background characteristics as "a check on the notion that high lagged scores identify students who have a shot at classification"

a related paper on the design of randomized experiments, Harris and Goldrick-Rab (2012) discuss potential explanations for the unexpected heterogeneity in their impact estimates based on full-sample endogenous stratification.

[2]We should note that endogenous stratification estimates do not appear in the published versions of two of the studies described here, see Dynarski et al. (2013) and Hemelt et al. (2013), or in a subsequent working paper on the grant program evaluated in Goldrick-Rab et al. 2011 by the same authors, see Goldrick-Rab et al. 2012.

(p. 1396).

The possibility of bias arising from endogenous stratification has been previously acknowledged in the evaluation literature (see, e.g., Peck, 2003), in statistics (Hansen, 2008), and in economics (Sanbonmatsu et al., 2006, and Giné et al, 2012), but the size and significance of the bias in realistic evaluation settings is not well understood.[3] A deceivingly comforting property of the bias is that it vanishes as sample size increases, under weak regularity conditions. However, as we demonstrate below using data from the National JTPA Study and the Tennessee STAR experiment, biases resulting from endogenous stratification can completely alter the quantitative and qualitative conclusions of empirical studies.

In the remainder of this article, we first describe in more detail the increasingly popular practice of stratifying experimental data by groups constructed on the basis of the predicted values from a regression of the outcome on baseline covariates for the full sample of experimental controls. We next explain why this method generates biases and describe the direction of those biases. We then describe leave-one-out and repeated split sample procedures that generate consistent estimators and show that the biases of these estimators are substantially lower than the bias of the full sample estimator in two realistic scenarios. We use data from the National JTPA Study and the Tennessee STAR experiment to demonstrate the performance of endogenous stratification estimators and the magnitude of their biases. We restrict our attention to randomized experiments, because this is the setting where endogenous stratification is typically used. However, similarly large biases may arise from endogenous stratification in observational studies.

2. Using Control Group Data to Create Predicted Outcomes

We begin by describing in detail the endogenous stratification method outlined above, which aims to classify study participants into groups based on their predicted value of the outcome variable in the absence of the treatment. Suppose that the sample consists of N observations

[3] Hausman and Wise (1977) and Hausman and Wise (1981), from which we borrow the term "endogenous stratification", study the related problem of biased sampling in randomized experiments. Altonji and Segal (1996) study biases that arise in the context of efficient generalized methods of moments estimation for reasons that are related to those that explain the bias of the full sample endogenous stratification estimator.

of the triple $(y, w, \boldsymbol{x})$, where y is an outcome variable, w is the treatment, and $\boldsymbol{x}$ is a vector of baseline characteristics. When the object of interest is the average treatment effect, which in a randomized experiment is equal to $\tau = E[y|w=1] - E[y|w=0]$, researchers typically compare sample average outcomes for the treated and the control groups:

$$\widehat{\tau} = \frac{\sum_{i=1}^{N} y_i w_i}{\sum_{i=1}^{N} w_i} - \frac{\sum_{i=1}^{N} y_i (1-w_i)}{\sum_{i=1}^{N} (1-w_i)}.$$

As discussed above, researchers sometimes aim to compare treated and non-treated after stratifying on a predictor of the outcome in the absence of the treatment. To our knowledge, most studies that use endogenous stratification implement it roughly as follows:

(1) Regress the outcome variable on a set of baseline characteristics using the control group only. The regression coefficients are:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i (1-w_i) \boldsymbol{x}_i' \right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i (1-w_i) y_i.$$

(2) Use the estimated coefficients to generate predicted outcome values for all participants (both treatment and control groups), $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$.

(3) Divide participants into groups based on their predicted outcomes. Typically, unit i is assigned to group k if $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ falls in some interval delimited by c_{k-1} and c_k. The interval limits may be fixed or could be quantiles of the empirical distribution of $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Many authors use a three-bin classification scheme of low, medium, and high predicted outcomes.

(4) Estimate treatment effects for each of the subgroups,

$$\widehat{\tau}_k = \frac{\sum_{i=1}^{N} y_i I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c_k]}}{\sum_{i=1}^{N} I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_i=0, c_{k-1} < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c_k]}}{\sum_{i=1}^{N} I_{[w_i=0, c_{k-1} < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c_k]}},$$

where I_A is the indicator function that takes values one if event A is realized, and value zero otherwise. Alternatively, treatment effect estimates could be computed after controlling for a set of covariates using regression.

For example, Goldrick-Rab et al. (2011) in their study of the impact of a need-based grant regress college persistence on baseline characteristics using only observations from the control group, generate predicted probabilities of college persistence for all students, classify students into three equal-sized groups based on their ex-ante predicted probability, and then estimate treatment effects for each of the three groups.

This is a simple and direct approach to stratification, which has great intuitive appeal. Moreover, it is easy to show that under usual regularity conditions, $\widehat{\tau}_k$ converges to

$$\tau_k = E[y|w = 1, c_{k-1} < \boldsymbol{x}'\boldsymbol{\beta} \leq c_k] - E[y|w = 0, c_{k-1} < \boldsymbol{x}'\boldsymbol{\beta} \leq c_k].$$

As we will see next, however, $\widehat{\tau}_k$ is biased in finite samples, and the bias follows a predictable pattern.

To simplify the exposition, suppose that predicted outcomes are divided into three groups (low, medium, high). Let $\boldsymbol{\beta} = (E[\boldsymbol{x}\boldsymbol{x}'|w = 0])^{-1}E[\boldsymbol{x}y|w = 0]$ be the population counterpart of $\widehat{\boldsymbol{\beta}}$, and let $e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ be the regression error. In a finite sample, untreated observations with large negative values for e_i tend to be over-fitted, so we expect $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} < \boldsymbol{x}_i'\boldsymbol{\beta}$, which pushes these observations towards the lower interval of predicted outcomes. This creates a negative bias in the average outcome among control observations that fall into the lower interval for $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ and, therefore, a positive bias in the average treatment effect estimated for that group. Analogously, average treatment effect estimators for the upper intervals of predicted outcomes are biased downward. Endogenous stratification results in a predictable pattern: average treatment effect estimators are biased upward for individuals with low predicted outcomes and biased downward for individuals with high predicted outcomes. As we will demonstrate below, because the finite sample bias of the endogenous stratification estimator is created by over-fitting, this bias tends to be more pronounced when the number of observations is small and the dimensionality of $\boldsymbol{x}_i$ is large.

A natural solution to the over-fitting issue is provided by leave-one-out estimators.[4] Let

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \left(\sum_{j \neq i} \boldsymbol{x}_i (1 - w_i) \boldsymbol{x}_i' \right)^{-1} \sum_{j \neq i} \boldsymbol{x}_i (1 - w_i) y_i,$$

be the regression coefficients estimators that discard observation i. Over-fitting is precluded by not allowing the outcome, y_i, of each observation to contribute to the estimation of its own predicted value, $\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_{(-i)}$. Because only untreated observations are employed in the estimation of $\widehat{\boldsymbol{\beta}}_{(-i)}$ and $\widehat{\boldsymbol{\beta}}$, if i is a treated observation then $\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}}$. We consider the following leave-one-out estimator of τ_k:

$$\widehat{\tau}_k^{LOO} = \frac{\sum_{i=1}^N y_i I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} \leq c_k]}}{\sum_{i=1}^N I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} \leq c_k]}} - \frac{\sum_{i=1}^N y_i I_{[w_i=0, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_{(-i)} \leq c_k]}}{\sum_{i=1}^N I_{[w_i=0, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_{(-i)} \leq c_k]}}.$$

Under weak assumptions, it can be seen that both $\widehat{\tau}_k$ and $\widehat{\tau}_k^{LOO}$ are consistent estimators of τ_k. Moreover, $\widehat{\tau}_k$ and $\widehat{\tau}_k^{LOO}$ have the same large sample distribution.[5] However, we show in section 4 that $\widehat{\tau}_k$ is substantially biased in two realistic scenarios, while $\widehat{\tau}_k^{LOO}$ is not.[6]

Another way to avoid over-fitting is sample splitting. We consider a repeated split sample estimator. In each repetition, m, the untreated sample is randomly divided into two groups, which we will call the prediction and the estimation groups. Let $v_{im} = 0$ if untreated observation i is assigned the prediction group in repetition m, and $v_{im} = 1$ if it is assigned to the estimation group. In each repetition, m, we estimate $\boldsymbol{\beta}$ using only the observations in the prediction group:

$$\widehat{\boldsymbol{\beta}}_m = \left(\sum_{i=1}^N \boldsymbol{x}_i (1 - w_i)(1 - v_{im}) \boldsymbol{x}_i' \right)^{-1} \sum_{i=1}^N \boldsymbol{x}_i (1 - w_i)(1 - v_{im}) y_i.$$

[4]This is the approach followed in Sanbonmatsu et al. (2006). Harvill et al. (2013) propose a variant of this approach based on 10-fold cross-validation.

[5]Proofs of these and other formal statements made in this paper are provided in Appendix 1.

[6]A separate issue in the estimation of τ_k is that first step estimation of $\boldsymbol{\beta}$ affects the large sample distribution of the estimator (see Appendix 1 for a derivation of the large sample distribution of $\widehat{\tau}_k$ and $\widehat{\tau}_k^{LOO}$). The contribution of the estimation of $\boldsymbol{\beta}$ to the variance of $\widehat{\tau}_k$ has been ignored in empirical practice.

For each repetition, m, the split sample estimator of τ_k is

$$\widehat{\tau}_{km}^{SS} = \frac{\sum_{i=1}^{N} y_i I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_m \leq c_k]}}{\sum_{i=1}^{N} I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_m \leq c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_i=0, v_{im}=1, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_m \leq c_k]}}{\sum_{i=1}^{N} I_{[w_i=0, v_{im}=1, c_{k-1} < \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_m \leq c_k]}}.$$

We then average $\widehat{\tau}_{km}^{SS}$ over M repetitions to obtain the repeated split sample estimator:

$$\widehat{\tau}_{k}^{RSS} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\tau}_{km}^{SS}.$$

The repeated split sample estimator is asymptotically unbiased and Normal but, unlike the leave-one-out estimator, its large sample distribution does not coincide with the large sample distribution of the full-sample endogenous stratification estimator. For large M, however, the difference between the large sample distribution of the repeated split sample estimator and the large sample distribution of the full-sample and leave-one-out estimators is small.[7,8]

In the next section, we apply the estimators described above to the analysis of data from two well-know experimental studies: the National JTPA Study and the Tennessee Project STAR experiment.

3. Evidence of Large Biases in Two Actual Applications

To demonstrate the performance of the estimators described in the previous section and the magnitude of their biases in realistic scenarios we use data from two randomized evaluations: the National JTPA Study, an evaluation of a vocational training program in the U.S., and the kindergarten cohort of the Tennessee Project STAR class-size experiment.

3.1. The National JTPA Experiment

We first examine data from the National JTPA Study. The National JTPA Study was a large experimental evaluation of a job training program commissioned by the U.S. Depart-

[7]This is proven in Appendix 1.

[8]The methods described in this section do not exhaust the possible approaches to the bias of the full-sample endogenous stratification estimator. Bootstrap/jackknife bias corrections on $\widehat{\tau}_k$ and shrinkage estimation of $\boldsymbol{\beta}$ are potentially fruitful approaches that we are starting to explore.

ment of Labor in the late 1980's. The National JTPA Study data have been extensively analyzed by Orr et al. (1996), Bloom, et al. (1997), and many others. The National JTPA Study randomized access to vocational training to applicants in 16 service delivery areas, or SDAs, across the U.S. Randomized assignment was done after applicants were deemed eligible for the program and recommended to one of three possible JTPA service strategies: on the job training/job search assistance, classroom training, and other services. Individuals in the treatment group were provided with access to JTPA services, and individuals in the control group were excluded from JTPA services for an 18-month period after randomization. We use data for the sample of male applicants recommended to the job training/job search assistance service strategy, and discard three SDAs with few observations. Our sample consists of 1681 treated observations and 849 untreated observations, for a total of 2530 observations in 13 SDAs.[9] In this example, w_i is an indicator of a randomized offer of JTPA services, y_i is nominal 30-month earnings in U.S. dollars after randomization, and $\boldsymbol{x}_i$ includes age, age squared, marital status, previous earnings, indicators for having worked less than 13 weeks during the year previous to randomization, having a high-school degree, being African-American, and being Hispanic, as well as SDA indicators.

Table 1 reports estimates for the JTPA sample. The first row reports two treatment effect estimates. The "unadjusted" estimate is the difference in outcome means between treated and controls. The "adjusted" estimate is the coefficient on the treatment indicator in a linear regression of the outcome variable, y_i, on the treatment indicator, w_i, and the covariates, $\boldsymbol{x}_i$, listed above. The unadjusted estimate suggests a \$1516 effect on 30-month earnings. This estimate is significant at the 10 percent level. Regression adjustment reduces the point estimate to \$1207, which becomes marginally non-significant at the 10 percent level. The rest of Table 1 reports average treatment effects by predicted outcome group. The first set of estimates correspond to $\widehat{\tau}_k$, the full-sample endogenous stratification estimator. This estimator produces a large and significant effect for the low predicted outcome group. The unadjusted estimate is \$2380 and significant at the 5 percent level. This rep-

[9]See Appendix 2 for detailed information on sample selection and estimation methods.

resents a 12.6 percent effect on 30-month earnings, once we divide it by the average value of 30-month earnings among the experimental controls. It also represents an effect that is 57 percent higher than the corresponding unadjusted estimated for the average treatment effect in the first row of the table. The adjusted estimate is \$2012, similarly large, and significant at the 10 percent level. For the high predicted outcome group, the estimates are also large, but not statistically significant at conventional test levels. For the middle predicted outcome group, the estimates are negative, but of moderate magnitude and not statistically significant. All in all, the full-sample endogenous stratification estimates provide a much more favorable picture of JTPA effectiveness relative to the average treatment effects reported on the first row. Now, the bulk of the effect seems to be concentrated on the low predicted outcome group, precisely the one in most need of help, with more diffuse effects estimated for the middle and high predicted outcome groups.

The next two sets of estimates reported in Table 1 correspond to the leave-one-out estimator, $\widehat{\tau}_k^{LOO}$, and the repeated split sample estimator $\widehat{\tau}_k^{RSS}$, with number of repetitions, M, equal to 100. These two estimators, which avoid over-fitting bias arising from the estimation of $\boldsymbol{\beta}$, produce results that are substantially different than those obtained with the full-sample endogenous stratification estimator, $\widehat{\tau}_k$. Relative to the $\widehat{\tau}_k$ estimates, the $\widehat{\tau}_k^{LOO}$ and $\widehat{\tau}_k^{RSS}$ estimates are substantially smaller for the low predicted outcome group, and substantially larger for the high predicted outcome group. For the high predicted outcome group we obtain unadjusted estimates of \$3647 (leave-one-out) and \$3569 (split sample) both significant at the 5 percent level, and adjusted estimates of \$3118 (leave-one-out) and \$2943 (split sample) significant at the 10 percent and 5 percent levels respectively. The estimates for the low and middle predicted outcome groups are small in magnitude and not statistically significant. These results place the bulk of the treatment effect on the high predicted outcome group and do not provide substantial statistical evidence of beneficial effects for the low and middle predicted outcome groups.[10] The comparison of

[10] This is loosely consistent with the findings in Abadie, Angrist, and Imbens (2002) who report large JTPA effects at the upper tail of the distribution of earnings for male trainees, and no discernible effects at the middle or lower parts of the distribution.

estimates produced with the full sample endogenous stratification estimator and the leave-one-out and split sample estimators suggests that the over-fitting bias of the full sample endogenous stratification estimator is of substantial magnitude and dramatically changes the qualitative and quantitative interpretations of the results.

As a further check on the magnitude of endogenous stratification biases in the analysis of the National JTPA Study data, Table 1 reports a last set of treatment effects estimates, which are stratified using data on earnings before randomization. The National JTPA Study data include individual earnings during the 12 months before randomization. We use the sorting of the experimental subjects in terms of pre-randomization earnings to approximate how the experimental subjects would have been sorted in terms of earnings in the absence of the treatment. We construct the estimator $\widehat{\tau}_k^{PREV}$ in the same way as $\widehat{\tau}_k$ but using previous earnings, instead of predicted earnings, to divide the individuals into three groups of approximately equal size. Notice that, because previous earnings is a baseline characteristic, $\widehat{\tau}_k^{PREV}$ is not affected by over-fitting bias. As shown on the bottom of Table 1, stratification on previous earning produces results similar to those obtained with $\widehat{\tau}_k^{LOO}$ and $\widehat{\tau}_k^{RSS}$: large and significant effects for the high predicted outcome group and smaller and non-significant effects for the middle and low predicted outcome groups.

3.2. The Tennessee Project STAR Experiment

Our second example uses data from the Tennessee Project STAR class-size study. In the Project STAR experiment, students in 79 schools were randomly assigned to small, regular-size, and regular-size classes with a teacher's aide. Krueger (1999) analyzes the STAR data set and provides detailed explanations of the STAR experiment. For our analysis, we use the 3764 students who entered the study in kindergarten and were assigned to small classes or to regular-size classes (without a teacher's aide). Our outcome variable is standardized end-of-the-year kindergarten math test scores.[11] The covariates are indicators for African-American, female, eligibility for the free lunch program, and school attended. We discard

[11]Standardized test scores are computed dividing raw test scores by the standard deviation of the distribution of the scores in regular-size classes.

observations with missing values in any of these variables.

Results for the STAR experiment data are reported in Table 2. The adjusted and unadjusted estimators of the average treatment effect on the first row of Table 2 show positive and significant effects. Using a simple difference in means, the effect of small classes is estimated as 0.1659 of the regular class standard deviation in math test scores, and 0.1892 of the same standard deviation when we use a regression-adjusted estimator.[12] In both cases, the estimates are significant at the 5 percent level. For the low and middle predicted outcomes groups, the full-sample endogenous stratification estimator, $\widehat{\tau}_k$, produces estimates that are positive and roughly double the average treatment effects estimates on the first row of the table. Counter-intuitively, however, the full sample endogenous stratification estimates for the high predicted outcome group are negative and significant. They seem to suggest that being assigned to small classes was detrimental for those students predicted to obtain high math scores if all students had remained in regular-size classes. We deem this result counter-intuitive because it implies that reductions in the student/teacher ratio have detrimental effects on average for a large group of students. Notice that the magnitudes of the negative effects estimated for high predicted outcome group are substantial: smaller, but not far from the positive average treatment effects reported in the first row of the table. We will see that the large and significant negative effect for the high predicted outcome group disappears when the leave-one-out or the repeated split sample procedures are used for estimation. Indeed, the leave-one-out and repeated split sample estimates on the two bottom rows of Table 2 suggest positive, significant, and large effects on the low and middle predicted outcome groups and effects of small magnitude and not reaching statistical significance at conventional test levels for the high predicted outcome group. Like for the JTPA, the qualitative and quantitative interpretations of the STAR experiment results change dramatically when the leave-one-out or the repeated split sample estimators

[12] To be consistent with much of the previous literature on the STAR experiment, we report both regression-adjusted and unadjusted estimates. Because the probability of assignment to a small class varied by school in the STAR experiment, the regression-adjusted estimator is most relevant in this setting. Like in Krueger (1999), however, covariate regression adjustment does not substantially change our estimates.

are used instead of the full-sample endogenous stratification estimator.

In this section, we have used data from two well-known and influential experimental studies to investigate the magnitude of the distortion that over-fitting may induce on endogenous stratification estimators. In the next section, we use Monte Carlo simulations to assess the magnitude of the biases of the different estimators considered Section 2. To keep the exercise as realistic as possible, in two of our simulations we choose data generating processes that mimic the features of the JTPA and STAR data sets.

4. Simulation Evidence on the Behavior of Endogenous Stratification Estimators

This section reports simulation evidence on the finite sample behavior of endogenous stratification estimators. We run Monte Carlo simulations in three different settings. In the first two Monte Carlo simulations, we make use of the JTPA and STAR data sets to assess the magnitudes of biases and other distortions to inference in realistic scenarios. The third and fourth Monte Carlo simulations use computer-generated data to investigate how the bias of endogenous stratification estimators changes when the sample size or the number of covariates changes.

In the JTPA-based simulation, we first use the JTPA control units to estimate a two-part model for the distribution of earnings conditional on the covariates of the adjusted estimates in Table 1. The two-part model consists of a Logit specification for the probability of zero earnings and a Box-Cox model for positive earnings.[13] In each Monte Carlo iteration we draw 2530 observations, that is, the same number of observations as in the JTPA sample, from the empirical distribution of the covariates in the JTPA sample. Next, we use the estimated two-part model to generate earnings data for each observation in the Monte Carlo sample. Then, we randomly assign 1681 observations to the treatment group and 849 observations to the control group, to match the numbers of treated and control units in the original JTPA sample. Finally, in each Monte Carlo iteration, we compute the full-sample, leave-one-out, and repeated split sample endogenous stratification estimates. We

[13]Additional details about the simulation models can be found in Appendix 2.

also compute the value of the unfeasible estimator, $\widehat{\tau}_k^{UNF}$, obtained by stratification on the population regression function (which can be calculated from the estimated parameters of the two-part model by simulation). We conduct a total of 10000 Monte Carlo iterations.

Figure 1 reports the Monte Carlo distributions of the endogenous stratification estimators that divide the experimental sample into three categories of predicted earnings of roughly equal size (bottom third, middle third, and top third). To economize space this figure shows only the distribution of the unadjusted estimators.[14] Because assignment to the treatment and control groups is randomized in our simulation and because the process that generates earnings data is the same for treated and controls, it follows that the average effect of the treatment in the simulations is equal to zero unconditionally as well as conditional on the covariates. As a result, unbiased estimators should generate Monte Carlo distributions centered around zero. The first plot of Figure 1 shows the Monte Carlo distribution of the full-sample endogenous stratification estimator of average treatment effects conditional on predicted earnings group. The pattern of the distribution of the average treatment effect estimator for the bottom, middle, and top third predicted earnings groups matches the directions of the biases discussed in Section 2. That is, $\widehat{\tau}_k$ is biased upwards for the low predicted earnings group and downwards for the high predicted earnings group. The remaining three plots of Figure 1 do not provide evidence of substantial biases for the leave-one-out, repeated split sample, or unfeasible estimators. These three estimators produce Monte Carlo distributions that are centered around zero for each predicted earnings category.

Table 3 reports biases, coverage rates for nominal 0.05 confidence intervals based on the Normal approximation, and root mean square error (root-MSE) values for endogenous stratification estimators in the JTPA-based Monte Carlo simulation. In addition to the estimators considered in Figure 1, we compute a single split sample estimator, $\widehat{\tau}_k^{SSS}$, which is defined like the repeated split sample estimator but with $M = 1$. The full-sample endogenous stratification estimator is subject to substantial distortions for the low and

[14]Simulation results for unadjusted and adjusted estimators are very similar, as reflected in Tables 3 to 6 below.

high predicted earnings group. The magnitude of the bias in each these two groups is more than \$1000, which is substantial compared to the \$1516 and \$1207 unadjusted and adjusted average effect estimates in the JTPA data. As reflected in Figure 1, the bias is positive for the low predicted earnings group and negative for the high predicted earnings group. Biases are uniformly small for the leave-one-out, repeated split sample, and unfeasible estimators, but the leave-one-out estimator has higher biases than the repeated split sample and the unfeasible estimator. Similar results emerge for coverage rates and mean square errors. The full-sample endogenous stratification estimator produces substantially higher than nominal coverage rates and substantially higher root-MSE than the leave-one-out and repeated split sample estimators for the low and high predicted income categories. The repeated split sample estimator dominates in terms of root-MSE. The single split sample estimators produce small biases and close to nominal coverage rates, but have root-MSE values consistently higher than the full-sample endogenous stratification estimator.

Figure 2 and Table 4 report simulation results for the STAR-based Monte Carlo simulation. For this simulation, the data generating process is based on a linear model with Normal errors. The model is estimated using data for STAR students in regular size classes. The results are qualitatively identical to those obtained in the JTPA-based simulation. The biases of $\widehat{\tau}_k$ are around 0.05 and -0.05 for the low and medium predicted test score groups, respectively. These are sizable magnitudes, compared to the STAR effect estimates in Table 2. Also, like in the JTPA-based simulation, for the low and high predicted outcome groups coverage rates of the full-sample endogenous stratification estimator are heavily distorted and root-MSE values are larger than for the leave-one-out and the repeated split sample estimators. The repeated split sample estimator has the lowest root-MSE, and single sample splits produce root-MSE values larger than any other estimator with the exception of the full-sample endogenous stratification estimator.

The analysis of how average treatment effects co-vary with predicted outcomes without the treatment can also be based on a regression equation with interaction terms, such as:

$$y_i = \alpha_0 + (\boldsymbol{x}_i'\boldsymbol{\beta})\alpha_1 + w_i\alpha_2 + w_i(\boldsymbol{x}_i'\boldsymbol{\beta})\alpha_3 + u_i,$$

where u_i is a regression error orthogonal to the included regressors. A negative sign of α_3 would indicate average treatment effects inversely related to $\boldsymbol{x}'\boldsymbol{\beta}$. Under the data generating processes employed in our simulations α_3 is equal to zero. Figure 3 reports Monte Carlo distributions of estimators of α_3 for the JTPA-based and STAR-based simulations. The full sample and leave-one-out estimators use $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(-i)}$, respectively, instead of $\boldsymbol{\beta}$, and estimate the regression equation by ordinary least squares. The unfeasible estimator uses the true value, $\boldsymbol{\beta}$. For $m = 1, \ldots, M$, the repeated split sample estimator uses $\widehat{\boldsymbol{\beta}}_m$ and average the resulting estimates of α_3 over the M repetitions. Finally, we also report the distribution of the estimator of α_3 given by one-step nonlinear least squares estimation of the regression equation above. The one-step nature of the nonlinear least squares estimator implies that predicted outcomes are fitted to all experimental units, and not only to the units in the control group.[15] The results in Figure 3 are consistent with our previous evidence on the performance of estimators that stratify on subgroups of predicted values. The Monte Carlo distributions of the leave-one-out, repeated split sample, nonlinear least squares and unfeasible estimators are all centered around zero. In contrast, the full-sample endogenous stratification estimator of α_3 is negatively biased.

The third and fourth Monte Carlo simulations use computer generated data only. The purpose of these simulations is to demonstrate how the bias of endogenous stratification estimators changes with changes in the sample size and in the number of covariates. The data generating model for the third simulation is

$$y_i = 1 + \sum_{l=1}^{40} z_{li} + v_i$$

for $i = 1, \ldots, N$, where the variables z_{li} have independent Standard Normal distributions, and the variable v_i has a independent Normal distribution with variance equal to 60. As a result, the unconditional variance of y_i is equal to 100. In each Monte Carlo simulation the sample is divided at random into two equally-sized treated and control groups. Predicted

[15]We thank Gary Chamberlain for suggesting this estimator. Nonlinear least squares estimation of the regression equation above uses the normalization $\alpha_0 = 0$ and $\alpha_1 = 1$ to ensure that the regression parameters are properly defined.

outcomes are computed using data for the control group to estimate

$$y_i = \alpha + \boldsymbol{x}'_{Ki}\boldsymbol{\beta}_K + u_{Ki}$$

by least squares, where $\boldsymbol{x}_{Ki}$ is the $(K \times 1)$-vector $(z_{1i}, \dots, z_{Ki})'$, for $K \leq 40$. That is, $\boldsymbol{x}_{Ki}$ contains the values of the first K regressors in $z_{1i}, \dots, z_{40i}$. The data generating process implies that α is equal to one, $\boldsymbol{\beta}_K$ is a $(K \times 1)$-vector of ones, $u_{Ki} = z_{K+1i} + \cdots + z_{40i} + v_i$ if $K < 40$ and $u_{40i} = v_i$. We run Monte Carlo simulations for samples sizes $N = 200$, $N = 1000$, and $N = 5000$, and numbers of included regressors $K = 10$, $K = 20$, and $K = 40$.

The results are reported in Table 5. To economize space, we omit results on the single split sample estimator and report bias results only. Coverage rate and root-MSE results are available upon request. The magnitude of the biases in the Table 5 are easily understood when compared to the standard deviation of the outcome, which is equal to 10. As expected, the bias of the full-sample endogenous stratification estimator is particularly severe when the sample size is small or when the number of included regressors is large, because in both cases significant over-fitting may occur. The increase in bias resulting from increasing the number of regressors is particularly severe when the sample size is small, $N = 200$. The biases of the leave-one-out, repeated split sample, and unfeasible estimators are negligible in most cases and consistently smaller than the bias of the full-sample endogenous stratification estimators, although the leave-one-out estimator tends to produce larger biases than the repeated split sample and unfeasible estimators.

The bias of the full-sample endogenous stratification estimator increases with K in spite of the fact that, as K increases, each additional included regressor has the same explanatory power as each of the regressors included in simulations with smaller K. Our final simulation studies a setting where each additional included regressor has lower explanatory power than the previously included ones. Consider:

$$y_i = 1 + \sum_{l=1}^{40} \rho^{l-1} z_{li} + \tilde{v}_i,$$

where the variables z_{li} have independent Standard Normal distributions as before, and the variable $\tilde{v}_i$ has a independent Normal distribution with a variance such that the variance of y_i is equal to 100. Table 6 reports the biases of the endogenous stratification estimators across Monte Carlo simulations under the new data generating process (with $\rho = 0.80$). The biases of the endogenous stratification estimator are larger than in the previous simulation. Their magnitudes increase faster than in the previous simulation when the number of included covariates increases, and decrease slower than in the previous simulation when the number of observations increases. The biases of the leave-one-out, repeated split sample, and unfeasible estimators are smaller and less sensitive to changes in the number of included covariates and sample size.

Overall, among the estimators that address the over-fitting problem of full-sample endogenous stratification, the repeated split sample estimator out-performs leave-one-out in the simulations. Moreover, the leave-one-out estimator can behave erratically in settings where the regressors take on only a few values and the variance of e_i is large.[16] The single split sample estimator has low bias and produces close-to-nominal coverage rates, but also large dispersion induced by the reduction in sample size. The increased variance of the single split sample estimator can make root-MSE of this estimator larger than the root-MSE of the full-sample endogenous stratification estimator (Table 3). All in all, the repeated split sample estimator displays the best performance in our simulations. It has low bias, accurate coverage rates, and out-performs alternative estimators in terms of root-MSE.

[16] This is the case, for example, in the Tennessee STAR experiment if school indicators are excluded from the vector $\boldsymbol{x}_i$. In that case, $\boldsymbol{x}_i$ only includes three indicator variables for race, gender, and eligibility for a free lunch program. As a result, $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ takes on only eight different values. In this setting, over-fitting is not an issue and the full-sample endogenous stratification estimator produces small biases in simulations. However, the leave-one-out estimator generates large biases. The reason is that, in this setting, choosing c_1 and c_2 to be the quantiles 1/3 and 2/3 of the distribution of the predicted outcomes results in a large number of observations being located exactly at the boundaries of the values of $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ that define the predicted outcome groups. To be concrete, consider the untreated observations with $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} = c_1$. These observations are classified by the full-sample endogenous stratification estimator as members of the low predicted outcome group. However, it is easy to see that if $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} = c_1$, then $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} > c_1$ if $y_i < c_1$ and $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \leq c_1$ if $y_i \geq c_1$, which induces biases in the leave-one-out estimator.

5. Conclusions

In this paper, we have argued that the increasingly popular practice of stratifying experimental units on the basis of a prediction of the outcome without treatment estimated using full sample data from the control group leads to substantially biased estimates of treatment effects. We illustrate the magnitude of this bias using data from two well-known social experiments: the National JTPA Study and the Tennessee STAR Class Size Experiment. The full-sample endogenous stratification approach is most problematic in studies with small sample sizes and many regressors, where the predictor of the outcome without treatment may be severely over-fitted in the control sample. We demonstrate that alternative endogenous stratification estimators based on leave-one-out and repeated split sample techniques display substantially improved small sample behavior in our simulations relative to the full-sample endogenous stratification estimator.

Some questions remain open to future research. First, the methods described in this article do not exhaust the possible approaches to the bias of the full-sample endogenous stratification estimator. Bootstrap/jackknife bias corrections on $\widehat{\tau}_k$ and shrinkage estimation of $\boldsymbol{\beta}$ are potentially fruitful approaches that we are starting to explore. In addition, a question of interest is whether the good small-sample behavior of the repeated split sample estimator generalizes to other settings, like the two-step generalized method of moments setting analyzed by Altonji and Segal (1996).

Appendix 1: Proofs

Suppose that we have data from an experiment where a fraction p of experimental units are assigned to a treatment group and a fraction $1-p$ to a control group, with $0 < p < 1$. Let N_1 be the number of units assigned to the treatment group and N_0 the number of units assigned to the control group, with $N = N_0 + N_1$. We will derive the limit distributions of endogenous stratification estimators as $N \to \infty$. For each experimental unit, i, we observe the triple $\boldsymbol{u}_{Ni} = (y_{Ni}, w_{Ni}, \boldsymbol{x}_{Ni})$, where w_{Ni} is a binary indicator that takes value one if observation i is in the treatment group, and value zero otherwise, y_{Ni} is the outcome of interest for observation i, and $\boldsymbol{x}_{Ni}$ is a vector of baseline characteristics for observation i. We conduct our analysis assuming that the experimental units are sampled at random from some large population of interest, so the values of $(y_{Ni}, \boldsymbol{x}_{Ni})$ for the treated and the non-treated can be regarded as independent i.i.d. samples of sizes N_1 and N_0 from some distributions P_1 and P_0, respectively. Probability statements about $\boldsymbol{u} = (y, w, \boldsymbol{x})$ are understood to refer to the distribution induced by first sampling w at random from a Bernoulli with parameter p and then sampling $(y, \boldsymbol{x})$ from P_1 with probability p and from P_0 with probability $1-p$. Because w is randomized, the marginal distribution of $\boldsymbol{x}$ is the same under P_1 and P_0. Let

$$\boldsymbol{\beta} = (E[\boldsymbol{x}\boldsymbol{x}'|w=0])^{-1}E[\boldsymbol{x}y|w=0].$$

That is, $\boldsymbol{x}'\boldsymbol{\beta}$ is the linear least-squares predictor of $E[y|\boldsymbol{x}, w=0]$. Let c be a known constant such that $\Pr(\boldsymbol{x}'\boldsymbol{\beta} \leq c) > 0$. We aim to estimate:

$$\tau = E[y|w=1, \boldsymbol{x}'\boldsymbol{\beta} \leq c] - E[y|w=0, \boldsymbol{x}'\boldsymbol{\beta} \leq c].$$

This is the average effect of the treatment for individuals with $\boldsymbol{x}'\boldsymbol{\beta} \leq c$. Consider the full-sample endogenous stratification estimator:

$$\widehat{\tau}(\widehat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^{N} y_{Ni} I_{[w_{Ni}=1, \boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \leq c]}}{\sum_{i=1}^{N} I_{[w_{Ni}=1, \boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \leq c]}} - \frac{\sum_{i=1}^{N} y_{Ni} I_{[w_{Ni}=0, \boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \leq c]}}{\sum_{i=1}^{N} I_{[w_{Ni}=0, \boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \leq c]}},$$

where $\widehat{\boldsymbol{\beta}}$ is a first-step estimator of the linear regression parameters that uses the untreated sample only:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{x}_{Ni}(1-w_{Ni})\boldsymbol{x}'_{Ni}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_{Ni}(1-w_{Ni})y_{Ni}.$$

Notice that, because w_{Ni} is randomized, the entire sample could be used to estimate $E[\boldsymbol{x}\boldsymbol{x}'|w=0]$. To our knowledge, this is not done in empirical research, so we do not follow that route in the derivations. However, all our large sample results would remain unchanged if we used both treated and untreated observations to estimate $E[\boldsymbol{x}\boldsymbol{x}'|w=0]$. We will assume that $\boldsymbol{x}$ has bounded support and that $E[y^2] < 0$. Under these assumptions, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, and

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (E[\boldsymbol{x}(1-w)\boldsymbol{x}'])^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{x}_{Ni}(1-w_{Ni})(y_{Ni} - \boldsymbol{x}'_{Ni}\boldsymbol{\beta}) + o_p(1)$$

$$= O_p(1).$$

For $j = 0, 1$ and any l, let

$$\mu_{jl} = E[I_{[\boldsymbol{x}'\boldsymbol{\beta} \le c]} y^l | w = j].$$

Notice that $\mu_{10} = \mu_{00} = \Pr(\boldsymbol{x}'\boldsymbol{\beta} \le c) > 0$. Let $\widehat{\mu}_{jl}(\widehat{\boldsymbol{\beta}})$ be

$$\widehat{\mu}_{0l}(\widehat{\boldsymbol{\beta}}) = \frac{1}{N_0} \sum_{i=1}^{N} (1 - w_{Ni}) I_{[\boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \le c]} y^l_{Ni}$$
$$= \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - w_{Ni})}{(1 - p)} I_{[\boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \le c]} y^l_{Ni},$$

for $j = 0$, and

$$\widehat{\mu}_{1l}(\widehat{\boldsymbol{\beta}}) = \frac{1}{N_1} \sum_{i=1}^{N} w_{Ni} I_{[\boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \le c]} y^l_{Ni}$$
$$= \frac{1}{N} \sum_{i=1}^{N} \frac{w_{Ni}}{p} I_{[\boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \le c]} y^l_{Ni},$$

for $j = 1$. Then,

$$\widehat{\tau}(\widehat{\boldsymbol{\beta}}) = \frac{\widehat{\mu}_{11}(\widehat{\boldsymbol{\beta}})}{\widehat{\mu}_{10}(\widehat{\boldsymbol{\beta}})} - \frac{\widehat{\mu}_{01}(\widehat{\boldsymbol{\beta}})}{\widehat{\mu}_{00}(\widehat{\boldsymbol{\beta}})}.$$

Notice that

$$\sqrt{N}\big(\widehat{\mu}_{0l} - \mu_{0l}(\boldsymbol{\beta})\big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{(1 - w_{Ni})}{(1 - p)} \Big(I_{[\boldsymbol{x}'_{Ni}\boldsymbol{\beta} \le c]} y^l_{Ni} - \mu_{0l}(\boldsymbol{\beta}) \Big)$$
$$+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{(1 - w_{Ni})}{(1 - p)} \Big[\Big(I_{[\boldsymbol{x}'_{Ni}\widehat{\boldsymbol{\beta}} \le c]} y^l_{Ni} - \mu_{0l}(\widehat{\boldsymbol{\beta}}) \Big) - \Big(I_{[\boldsymbol{x}'_{Ni}\boldsymbol{\beta} \le c]} y^l_{Ni} - \mu_{0l}(\boldsymbol{\beta}) \Big) \Big]$$
$$+ \sqrt{N} \Big(\mu_{0l}(\widehat{\boldsymbol{\beta}}) - \mu_{0l}(\boldsymbol{\beta}) \Big).$$

Consider now $\mathcal{M}_1 = \{ y I_{[\boldsymbol{x}'\boldsymbol{b} \le c]} : \boldsymbol{b} \in \Theta \}$ and $\mathcal{M}_0 = \{ I_{[\boldsymbol{x}'\boldsymbol{b} \le c]} : \boldsymbol{b} \in \Theta \}$. If follows from Andrews (1994, Theorems 2 and 3) that $\mathcal{M}_1$ satisfies Pollard's entropy condition with envelope $\max\{1, y\}$, while $\mathcal{M}_0$ satisfies Pollard's entropy condition with envelope 1. By Andrews (1994, Theorem 1), if $E[|y|^{2+\delta}] < \infty$ for some $\delta > 0$, we obtain that the second term on the right hand side of last equation converges in probability to zero. As a result:

$$\sqrt{N}\big(\widehat{\mu}_{0l} - \mu_{0l}(\boldsymbol{\beta})\big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{(1 - w_{Ni})}{(1 - p)} \Big(I_{[\boldsymbol{x}'_{Ni}\boldsymbol{\beta} \le c]} y^l_{Ni} - \mu_{0l}(\boldsymbol{\beta}) \Big)$$
$$+ \sqrt{N} \Big(\mu_{0l}(\widehat{\boldsymbol{\beta}}) - \mu_{0l}(\boldsymbol{\beta}) \Big) + o_p(1).$$

For $j, l = 0, 1$, we will assume that $\mu_{jl}(\boldsymbol{b})$ is differentiable at $\boldsymbol{\beta}$ (Kim and Pollard 1990 section 5 provides high-level sufficient conditions). Then,

$$\mu_{jl}(\boldsymbol{\beta} + \boldsymbol{h}) - \mu_{jl}(\boldsymbol{\beta}) - \boldsymbol{r}_{jl}(\boldsymbol{\beta})\boldsymbol{h} = o(\|\boldsymbol{h}\|)$$

where

$$\boldsymbol{r}_{jl}(\boldsymbol{\beta}) = \frac{\partial \mu_{jl}(\boldsymbol{b})}{\partial \boldsymbol{b}'}(\boldsymbol{\beta}).$$

As a result (see, e.g., Lemma 2.12 in van der Vaart, 1998),

$$\mu_{jl}(\widehat{\boldsymbol{\beta}}) - \mu_{jl}(\boldsymbol{\beta}) - \boldsymbol{r}_{jl}(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = o_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|).$$

Therefore, because $\sqrt{N}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(1)$, we obtain

$$\begin{aligned}\sqrt{N}\Big(\mu_{0l}(\widehat{\boldsymbol{\beta}}) - \mu_{0l}(\boldsymbol{\beta})\Big) &= \sqrt{N}\boldsymbol{r}_{0l}(\boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1) \\ &= \boldsymbol{r}_{0l}(\boldsymbol{\beta})'\boldsymbol{Q}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{x}_{Ni}(1 - w_{Ni})e_{Ni} + o_p(1),\end{aligned}$$

where $\boldsymbol{Q} = E[\boldsymbol{x}(1-w)\boldsymbol{x}']$. Then,

$$\begin{aligned}\sqrt{N}\big(\widehat{\mu}_{0l} - \mu_{0l}(\boldsymbol{\beta})\big) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{(1 - w_{Ni})}{(1-p)}\Big(I_{[\boldsymbol{x}'_{Ni}\boldsymbol{\beta} \le c]}y_{Ni}^{l} - \mu_{0l}(\boldsymbol{\beta})\Big) \\ &\quad + \boldsymbol{r}_{0l}(\boldsymbol{\beta})'\boldsymbol{Q}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{x}_{Ni}(1 - w_{Ni})e_{Ni} + o_p(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\sqrt{N}\big(\widehat{\mu}_{1l} - \mu_{1l}(\boldsymbol{\beta})\big) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{w_{Ni}}{p}\Big(I_{[\boldsymbol{x}'_{Ni}\boldsymbol{\beta} \le c]}y_{i}^{l} - \mu_{1l}(\boldsymbol{\beta})\Big) \\ &\quad + \boldsymbol{r}_{1l}(\boldsymbol{\beta})'\boldsymbol{Q}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{x}_{Ni}(1 - w_{Ni})e_{Ni} + o_p(1).\end{aligned}$$

Consider now,

$$\xi_{jl,Ni} = \frac{1}{\sqrt{N}}\left[\left(\frac{1 - w_{Ni}}{1-p}\right)^{1-j}\left(\frac{w_{Ni}}{p}\right)^{j}\Big(I_{[\boldsymbol{x}'_{Ni}\boldsymbol{\beta} \le c]}y_{Ni}^{l} - \mu_{jl}(\boldsymbol{\beta})\Big) + \boldsymbol{r}_{jl}(\boldsymbol{\beta})'\boldsymbol{Q}^{-1}\boldsymbol{x}_{Ni}(1 - w_{Ni})e_{Ni}\right],$$

and let

$$\boldsymbol{\xi}_{Ni} = \begin{pmatrix}\xi_{11,Ni} \\ \xi_{10,Ni} \\ \xi_{01,Ni} \\ \xi_{00,Ni}\end{pmatrix}.$$

For any $\boldsymbol{a} \in \mathbb{R}^4$, the terms $\boldsymbol{a}'\boldsymbol{\xi}_{Ni}$ are martingale differences with respect to the filtration $\mathcal{F}_{Ni}$ spanned by $(y_{N1}, w_{N1}, \boldsymbol{x}_{N1}), \ldots, (y_{Ni}, w_{Ni}, \boldsymbol{x}_{Ni})$. For $j, l = 0, 1$ and $j', l' = 0, 1$, let,

$$s_{jlj'l'} = E\left[\left(\frac{1 - w_{Ni}}{1-p}\right)^{(2-j-j')}\left(\frac{w_{Ni}}{p}\right)^{j+j'}(I_{[\boldsymbol{x}'_{i}\boldsymbol{\beta} \le c, w_i = j]}y_{i}^{l} - \mu_{jl})(I_{[\boldsymbol{x}'_{i}\boldsymbol{\beta} \le c, w_i = j']}y_{i}^{l'} - \mu_{j'l'})\right]$$

and let $\boldsymbol{S}$ be a (4×4) matrix with element $(2(1-j)+(2-l), 2(1-j')+(2-l'))$ equal to $s_{jlj'l'}$. Then,

$$\boldsymbol{S} = \begin{pmatrix} (\mu_{12}-\mu_{11}^2)/p & \mu_{11}(1-\mu_{10})/p & 0 & 0 \\ \mu_{11}(1-\mu_{10})/p & \mu_{10}(1-\mu_{10})/p & 0 & 0 \\ 0 & 0 & (\mu_{02}-\mu_{01}^2)/(1-p) & \mu_{01}(1-\mu_{00})/(1-p) \\ 0 & 0 & \mu_{01}(1-\mu_{00})/(1-p) & \mu_{00}(1-\mu_{00})/(1-p) \end{pmatrix}.$$

Let

$$\boldsymbol{D} = \left(\frac{1}{\mu_{10}}, \frac{-\mu_{11}}{\mu_{10}^2}, -\frac{1}{\mu_{00}}, \frac{\mu_{01}}{\mu_{00}^2} \right)'.$$

The asymptotic distribution of the unfeasible estimator that employs the treated and control units with $\boldsymbol{x}_i'\boldsymbol{\beta} \le c$ is

$$\sqrt{N}\left(\widehat{\tau}_k^{UNF} - \tau_k\right) \xrightarrow{d} N(0, \boldsymbol{D}'\boldsymbol{S}\boldsymbol{D}),$$

where

$$\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D} = \frac{1}{\mu_{10}p}\left[\frac{\mu_{12}}{\mu_{10}} - \left(\frac{\mu_{11}}{\mu_{10}}\right)^2\right] + \frac{1}{\mu_{00}(1-p)}\left[\frac{\mu_{02}}{\mu_{00}} - \left(\frac{\mu_{01}}{\mu_{00}}\right)^2\right].$$

Let

$$\boldsymbol{C} = E\left[\begin{pmatrix} 0 \\ 0 \\ I_{[\boldsymbol{x}'\boldsymbol{\beta}\le c]}y(y-\boldsymbol{x}'\boldsymbol{\beta})(1-w)/(1-p) \\ I_{[\boldsymbol{x}'\boldsymbol{\beta}\le c]}(y-\boldsymbol{x}'\boldsymbol{\beta})(1-w)/(1-p) \end{pmatrix}\boldsymbol{x}'\right]\boldsymbol{Q}^{-1}\boldsymbol{r}',$$

$$\boldsymbol{\Sigma} = E[(1-w_i)\boldsymbol{x}_i e_i^2 \boldsymbol{x}_i'],$$

and

$$\boldsymbol{V} = \boldsymbol{S} + \boldsymbol{r}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}\boldsymbol{r}' + \boldsymbol{C} + \boldsymbol{C}'.$$

We obtain that the asymptotic distribution of the full-sample endogenous stratification estimator is

$$\sqrt{N}\left(\widehat{\tau}_k - \tau_k\right) \xrightarrow{d} N(0, \boldsymbol{D}'\boldsymbol{V}\boldsymbol{D}).$$

Notice that estimation of the derivative vector $\boldsymbol{r}$ can be accomplished using numerical methods (see, e.g., Newey and McFadden 1994, Theorem 7.4).

Similar derivations can be used to find the large sample distribution of $\widehat{\tau}_k^{RSS}$. Let p_s be the fraction of untreated observations that are assigned to the estimation group. We present the result next, omitting details. Consider the matrix partition

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix},$$

where each of the sub-matrices is (2×2). Let

$$\boldsymbol{S}^* = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22}^* \end{pmatrix},$$

where

$$\boldsymbol{S}_{22}^* = \left[1 + \frac{1}{M}\left(\frac{1-p_s}{p_s}\right)\right]\boldsymbol{S}_{22}.$$

Let

$$\boldsymbol{V}^* = \boldsymbol{S}^* + \left[1 + \frac{1}{M}\left(\frac{p_s}{1-p_s}\right)\right]\boldsymbol{r}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}\boldsymbol{r}' + \left[1 - \frac{1}{M}\right](\boldsymbol{C}+\boldsymbol{C}').$$

The asymptotic distribution of the repeated split sample estimator is:

$$\sqrt{N}\left(\widehat{\tau}_k^{RSS} - \tau_k\right) \xrightarrow{d} N(0, \boldsymbol{D}'\boldsymbol{V}^*\boldsymbol{D}).$$

The following intermediate lemma will be useful to derive the properties of the leave-one-out estimator.

Lemma A.1: *Let x_N be a sequence of random variables, a_N a sequence of real numbers, and I_{A_N} be the indicator function for the event A_N. Suppose that $a_N \Pr(I_{A_N}=0) \to 0$ and $a_N \Pr\left(I_{A_N}x_N > \varepsilon\right) \to 0$ for some $\varepsilon > 0$. Then, $a_N \Pr\left(x_N > \varepsilon\right) \to 0$.*

Proof:

$$\begin{aligned} a_N \Pr(x_N > \varepsilon) &= a_N \Pr(x_N > \varepsilon, I_{A_N}=1) + a_N \Pr(x_N > \varepsilon, I_{A_N}=0) \\ &= a_N \Pr(I_{A_N}x_N > \varepsilon, I_{A_N}=1) + a_N \Pr(x_N > \varepsilon, I_{A_N}=0) \\ &\le a_N \Pr(I_{A_N}x_N > \varepsilon) + a_N \Pr(I_{A_N}=0) \to 0. \end{aligned}$$

□

Next, we prove that the leave-one-out estimator has the same large sample distribution as the full-sample estimator. For simplicity and because it does not play any role in the calculations below, we omit the subscript N from the notation for sample units. Consider the leave-one-out estimator:

$$\begin{aligned} \widehat{\tau}^{LOO} &= \frac{\frac{1}{N}\sum_{i=1}^N y_i I_{[w_i=1,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}{\frac{1}{N}\sum_{i=1}^N I_{[w_i=1,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}} - \frac{\frac{1}{N}\sum_{i=1}^N y_i I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}{\frac{1}{N}\sum_{i=1}^N I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}} \\ &= \frac{\frac{1}{N}\sum_{i=1}^N y_i I_{[w_i=1,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}}{\frac{1}{N}\sum_{i=1}^N I_{[w_i=1,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}} - \frac{\frac{1}{N}\sum_{i=1}^N y_i I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}{\frac{1}{N}\sum_{i=1}^N I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}. \end{aligned}$$

Therefore,

$$\sqrt{N}\big(\widehat{\tau}(\widehat{\boldsymbol{\beta}}) - \widehat{\tau}^{LOO}\big) = \sqrt{N}\left(\frac{\frac{1}{N}\sum_{i=1}^N y_i I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}{\frac{1}{N}\sum_{i=1}^N I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}} - \frac{\frac{1}{N}\sum_{i=1}^N y_i I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}}{\frac{1}{N}\sum_{i=1}^N I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}}\right)$$

$$= \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} y_i\left(I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]} - I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}\right)}{\frac{1}{N}\sum_{i=1}^{N} I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}$$

$$- \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]} - I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}\right)}{\frac{1}{N}\sum_{i=1}^{N} I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}\frac{1}{N}\sum_{i=1}^{N} I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}}\frac{1}{N}\sum_{i=1}^{N} y_i I_{[w_i=0,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]}.$$

We will show that $\sqrt{N}(\widehat{\tau}(\widehat{\boldsymbol{\beta}}) - \widehat{\tau}^{LOO}) \xrightarrow{p} 0$. Suppose that the r-th moment of $|y|$ exists, where $r > 1$ (later we will strengthen this requirement to $r > 2$ and eventually to $r > 3$). Then, by Holder's Inequality:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}|y_i|\left|I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]} - I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}\right| \le N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}|y_i|^r\right)^{1/r}\left(\frac{1}{N}\sum_{i=1}^{N}\left|I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]} - I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}\right|\right)^{(r-1)/r}.$$

The first sample average on the right hand side of last equation is bounded in probability. Now we need to show that the second sample average on the right hand side of last equation goes to zero fast enough to beat $N^{1/2}$ after taking the $(r-1)/r$ power. Because the distribution of $(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}, \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)})$ does not depend on i, we obtain:

$$\begin{aligned} E\left[\frac{1}{N}\sum_{i=1}^{N}\left|I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c]} - I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c]}\right|\right] &= E\left[\frac{1}{N}\sum_{i=1}^{N} I_{[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\le c<\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\,\cup\,\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\le c<\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}]}\right] \\ &= \Pr\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \cup \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right). \end{aligned}$$

Therefore, by Markov's inequality, it is enough to show that:

$$N^{\frac{r}{2(r-1)}}\Pr\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \cup \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) \xrightarrow{p} 0.$$

Let $\zeta_N = N^{\alpha}$, where $\alpha > r/(2(r-1))$. Notice that,

$$\begin{aligned} &\Pr\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \cup \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) \\ &\qquad \le \Pr\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \cup \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \le c < \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}, |\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - c| > 1/\zeta_N\right) \\ &\qquad + \Pr\left(|\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - c| \le 1/\zeta_N\right) \\ &\qquad \le \Pr\left(|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})| > 1/\zeta_N\right) + \Pr\left(|\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - c| \le 1/\zeta_N\right). \end{aligned}$$

Suppose that there exists $\epsilon > 0$ such that for $\boldsymbol{b} \in B(\boldsymbol{\beta}, \epsilon)$, the distribution of $\boldsymbol{x}'\boldsymbol{b}$ is absolutely continuous with density bounded (uniformly) by a constant C. Assume that $r > 2$. Consider a sequence $\epsilon_N = N^{-\gamma}$, where $0 < \gamma < (r-2)/4(r-1)$. Then, for large enough N (so $\epsilon_N < \epsilon$)

$$\Pr\left(|\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - c| \le 1/\zeta_N\right) \le \Pr\left(|\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - c| \le 1/\zeta_N, \widehat{\boldsymbol{\beta}} \in B(\boldsymbol{\beta}, \epsilon_N)\right) + \Pr\left(\widehat{\boldsymbol{\beta}} \notin B(\boldsymbol{\beta}, \epsilon_N)\right)$$

$$\leq \sup_{\boldsymbol{b}\in B(\boldsymbol{\beta},\epsilon_N)} \Pr\left(|\boldsymbol{x}_i'\boldsymbol{b} - c| \leq 1/\zeta_N\right) + \Pr\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > \epsilon_N\right)$$
$$\leq \frac{2C}{\zeta_N} + \Pr\left(N^\gamma \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > 1\right).$$

The first term on the right hand side of last equation multiplied by $N^{r/(2(r-1))}$ converges to zero because $N^{r/(2(r-1))}/\zeta_N \to 0$. For any arbitrary real square matrix $\boldsymbol{A}$, let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ be the minimum and maximum eigenvalue of $\boldsymbol{A}$, respectively. Assume $\lambda_{\min}(\boldsymbol{Q}) > 0$. Because $\|\boldsymbol{x}_i\|$ is bounded by some constant, C, it follows that

$$\lambda_{\max}\left(\frac{1}{N}\boldsymbol{x}_i(1-w_i)\boldsymbol{x}_i'\right) = \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}'\boldsymbol{x}_i(1-w_i)\boldsymbol{x}_i'\boldsymbol{v}/N \leq \frac{C^2}{N}.$$

Let

$$\boldsymbol{Q}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i(1-w_i)\boldsymbol{x}_i'.$$

Now, Corollary 5.2 in Tropp (2012) implies

$$\Pr\left(\lambda_{\min}\left(\boldsymbol{Q}_N\right) \leq t\lambda_{\min}(\boldsymbol{Q})\right) \leq K\, e^{-\frac{N(1-t)^2\lambda_{\min}(\boldsymbol{Q})}{2C^2}}, \tag{A.1}$$

where K is the length of $\boldsymbol{x}_i$ and $t \in [0,1]$. Define the event

$$A_N = \left\{\lambda_{\min}(\boldsymbol{Q}_N) \geq C_\lambda\right\},$$

for some $0 < C_\lambda < \lambda_{\min}(\boldsymbol{Q})$, and let I_{A_N} be the indicator function for the event A_N. The concentration inequality in (A.1) implies

$$N^{\frac{r}{2(r-1)}} \Pr(I_{A_N} = 0) \to 0. \tag{A.2}$$

Let $e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$. Notice that,

$$\begin{aligned}
I_{A_N} N\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 &= I_{A_N} N(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= I_{A_N}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N e_i(1-w_i)\boldsymbol{x}_i'\right)\left(\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i(1-w_i)\boldsymbol{x}_i'\right)^{-2}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{x}_i(1-w_i)e_i\right) \\
&\leq I_{A_N}\lambda_{\min}^{-2}(\boldsymbol{Q}_N)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N e_i(1-w_i)\boldsymbol{x}_i'\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{x}_i(1-w_i)e_i\right) \\
&\leq C_\lambda^{-2}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N e_i(1-w_i)\boldsymbol{x}_i'\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{x}_i(1-w_i)e_i\right).
\end{aligned}$$

Given that $\boldsymbol{x}$ is bounded, $E[e_i^2] < \infty$ (which follows from $r > 2$), and given that $E[e_i(1 - w_i)\boldsymbol{x}_i'\boldsymbol{x}_j(1-w_j)e_j] = E[e_i(1-w_i)\boldsymbol{x}_i']E[\boldsymbol{x}_j(1-w_j)e_j] = 0$ for any $1 \leq i < j \leq N$, we obtain

$$\limsup_{N\to\infty} E[I_{A_N} N\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] < \infty. \tag{A.3}$$

By Markov's inequality, equation (A.3), and because $\gamma < (r-2)/4(r-1)$

$$N^{\frac{r}{2(r-1)}} \Pr\left(I_{A_N} N^{\gamma} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > 1\right) \leq N^{2\gamma - 1 + \frac{r}{2(r-1)}} E[I_{A_N} N \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] \rightarrow 0. \tag{A.4}$$

From equations (A.2), (A.4), and Lemma A.1, it follows that

$$N^{\frac{r}{2(r-1)}} \Pr\left(N^{\gamma} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > 1\right) \rightarrow 0.$$

Therefore,

$$N^{\frac{r}{2(r-1)}} \Pr\left(|\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} - c| \leq 1/\zeta_N\right) \rightarrow 0.$$

Next, we will prove $N^{\frac{r}{2(r-1)}} \Pr(|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})| > 1/\zeta_N) \rightarrow 0$. Notice that (see Hansen, 2012 sections 4.12 and 4.13)

$$\begin{aligned} |\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})| &= \frac{h_{Ni}}{1 - h_{Ni}} |y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}| \\ &\leq \left(\frac{\max_{i=1,\ldots,N} h_{Ni}}{1 - \max_{i=1,\ldots,N} h_{Ni}} \right) |y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}|, \end{aligned}$$

where the leverage values h_{Ni} are

$$h_{Ni} = (1 - w_i) \boldsymbol{x}_i' \left(\sum_{i=1}^{N} \boldsymbol{x}_i (1 - w_i) \boldsymbol{x}_i' \right)^{-1} \boldsymbol{x}_i. \tag{A.5}$$

Therefore,

$$\max_{1 \leq i \leq N} h_{Ni} \leq \lambda_{\min}^{-1} (\boldsymbol{Q}_N) \frac{C^2}{N}.$$

Notice also that,

$$\begin{aligned} & I_{A_N} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^r \\ & \leq I_{A_N} \left(\left(\frac{1}{N} \sum_{i=1}^{N} (1 - w_i) e_i \boldsymbol{x}_i' \right) \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i (1 - w_i) \boldsymbol{x}_i' \right)^{-2} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i e_i (1 - w_i) \right) \right)^{r/2} \\ & \leq I_{A_N} \lambda_{\min}^{-r}(\boldsymbol{Q}_N) \left(\left(\frac{1}{N} \sum_{i=1}^{N} (1 - w_i) e_i \boldsymbol{x}_i' \right) \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i e_i (1 - w_i) \right) \right)^{r/2} \\ & \leq C_{\lambda}^{-r} \left(\left(\frac{1}{N} \sum_{i=1}^{N} (1 - w_i) e_i \boldsymbol{x}_i' \right) \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i e_i (1 - w_i) \right) \right)^{r/2} \\ & = C_{\lambda}^{-r} C^r \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |e_i e_j| \right)^{r/2}. \end{aligned}$$

If $r \geq 2$, by Minkowski's and Cauchy-Schwartz's inequalities:

$$\left(E\left[\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|e_ie_j|\right)^{r/2}\right]\right)^{2/r} \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(E\left[|e_ie_j|^{r/2}\right]\right)^{2/r}$$
$$\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(E\left[|e_1|^r\right]\right)^{2/r}.$$

Because $E[|e_i|^r] = E[|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|^r]$, $E[|y_i|^r] < \infty$ and $\|\boldsymbol{x}_i\|$ is bounded, we obtain

$$\limsup_{N\to\infty} E[I_{A_N}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^r] < \infty,$$

which, by Minkowski's inequality, implies

$$\limsup_{N\to\infty} E[I_{A_N}|y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}|^r] < \infty. \tag{A.6}$$

By Markov's inequality:

$$N^{\frac{r}{2(r-1)}}\Pr\left(I_{A_N}|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})| > 1/\zeta_N\right) \leq N^{\frac{r}{2(r-1)}+r\alpha}E\left[I_{A_N}|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})|^r\right].$$

The condition $\alpha < (2r-3)/(2(r-1))$ implies $r/(2(r-1)) + r\alpha < r$. So, under that condition, it is left to be proven that for any $\vartheta > 0$

$$N^{r-\vartheta}E\left[I_{A_N}|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})|^r\right] \to 0. \tag{A.7}$$

Consider N large enough so that there is a positive constant C_d, such that $1/(1-C_\lambda^{-1}C^2/N) < C_d$. Then,

$$I_{A_N}|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})| \leq C_d\, C_\lambda^{-1}\frac{C^2}{N}|y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}|.$$

This result, along with equation (A.6) implies:

$$\limsup_{N\to\infty} E[I_{A_N}N^r|\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})|^r] < \infty,$$

so equation (A.7) holds. Notice that for the condition

$$\frac{r}{2(r-1)} < \alpha < \frac{2r-3}{2(r-1)}$$

to hold we need $r > 3$.

Appendix 2: Estimation and Simulation Details

Leave-one-out predictions can be efficiently calculated using:

$$\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - \frac{h_{Ni}}{1-h_{Ni}}(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}),$$

where h_{Ni} is the leverage value defined in equation (A.5). Let $\widehat{y}_i$ be generic notation for a predicted outcome without treatment. The prediction, $\widehat{y}_i$, may come from full-sample, leave-one-out, or split sample endogenous stratification, stratification on previous earnings in the JTPA example of Section 3, or stratification on the true regression value in the simulations of Section 4. We group observations on the basis of predicted outcomes, $\hat{y}_i$, in the following way. First we sort the observations based on predicted outcomes: $\widehat{y}_{(1)} \leq \widehat{y}_{(2)} \leq \cdots \leq \widehat{y}_{(N)}$. We then define t_1 and t_2 as $N/3$ and $2N/3$ rounded to the nearest integer, respectively. We classify unit i in the low, medium, and high predicted outcome groups if $\widehat{y}_i \leq \widehat{y}_{(t_1)}$, $\widehat{y}_{(t_1)} < \widehat{y}_i \leq \widehat{y}_{(t_2)}$, and $\widehat{y}_{(t_2)} < \widehat{y}_i$, respectively.

For the repeated split sample estimator estimation is as follows. For the JTPA data we randomly select 425 control observations and use them to estimate $\boldsymbol{\beta}$. We use the remaining 424 observations and all the treated JTPA units for the second step estimation of τ_k. For the STAR data, we use 1009 untreated observations to estimate $\boldsymbol{\beta}$ and the remaining 1008 and all the treated observations in the second step. We average the split sample estimators over 100 repetitions to obtain $\widehat{\tau}_k^{RSS}$.

As explained in Section 3, the JTPA sample consists of male applicants assigned to on the job training/job search assistance. We discard three of the sixteen SDAs, Jersey City (21 observations), Butte (15 observations), and Oakland (5 observations) because of small sample sizes. The STAR sample consists of use 3764 students who entered the study in kindergarten, were assigned to small classes or to regular-size classes without a teacher's aide, and for whom there is complete information on all the variables used in our analysis.

Standard errors in Tables 1 and 2 are calculated using the nonparametric bootstrap (conditioning on the number of treated and untreated observations in the original samples).

In the JTPA-based simulation we first estimate a Logit model, $p(\boldsymbol{x}, \boldsymbol{\gamma}) = e^{\boldsymbol{x}'\boldsymbol{\gamma}}/(1 + e^{\boldsymbol{x}'\boldsymbol{\gamma}})$, for the probability of employment, measured as positive labor market earnings, using the sample of experimental controls. Next, using only the experimental control with positive earnings, we estimate a Box-Cox regression model

$$\frac{y^\lambda - 1}{\lambda} = \boldsymbol{x}'\boldsymbol{\theta} + \sigma u,$$

where u has a Standard Normal distribution. We will use $\boldsymbol{\gamma}^*$ to refer to the estimate of $\boldsymbol{\gamma}$, and analogous notation for the estimated parameters of the Box-Cox model. We create each simulated data set in the following manner. We first resample 2530 observations from the empirical distribution of $\boldsymbol{x}$ among all the JTPA sample units. We assign zero earnings with probability $1 - p(\boldsymbol{x}, \boldsymbol{\gamma}^*)$. With probability $p(\boldsymbol{x}, \boldsymbol{\gamma}^*)$ we assign earnings using

$$y = \left(\max\{1 + \lambda^*(\boldsymbol{x}'\boldsymbol{\theta}^* + \sigma^* u), 0\}\right)^{1/\lambda^*},$$

where u has a Standard Normal distribution. We randomly label 1681 observations as treated and 849 as untreated. As a result, all treatment effects are equal to zero by construction. The coefficients of the regression function of y on $\boldsymbol{x}$ under this data generating process, which are needed to compute $\widehat{\tau}_k^{UNF}$, are calculated by simulation.

For the STAR-based simulations, we estimate the linear model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \sigma u,$$

where u has a Standard Normal distribution, using the sample of experimental controls. We use least squares to estimate $\boldsymbol{\beta}$ and the variance of the regression residuals corrected for degrees of freedom to estimate σ. To construct each simulated sample, we first randomly resample 3764 observations from the empirical distribution of $\boldsymbol{x}$ in the STAR sample. We simulate math scores using

$$y = \boldsymbol{x}'\boldsymbol{\beta}^* + \sigma^* u,$$

where u has a Standard Normal distribution, and $\boldsymbol{\beta}^*$ and σ^* are the estimates of $\boldsymbol{\beta}$ and σ.

Section 4 contains detailed information on data generating processes for the simulations of Tables 5 and 6.

Bias, coverage rates, and root-MSE are calculated as follows. Because the simulations impose zero treatments effects, the bias and the MSE are calculated as the mean of the estimates and the mean of the square of the estimates, respectively, across all simulation repetitions. We calculate t-ratios dividing the estimates from each simulation repetition by the standard deviation of the estimates across repetitions. Coverage rates are the frequencies of the t-ratios falling outside the $[-1.96, 1.96]$ interval across repetitions. Root-MSE is the square-root of the MSE.

References

Abadie, Alberto, Joshua D. Angrist and Guido W. Imbens. 2002. "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings." *Econometrica* 70(1): 91-117.

Altonji, Joseph G. and Lewis M. Segal. 1996. "Small-Sample Bias in GMM Estimation of Covariance Structures." *Journal of Business and Economic Statistics* 14(3): 353-366.

Andrews, Donald W. K. 1994. "Empirical Process Methods in Econometrics." In R.F. Engle and D.L. McFadden (Eds.), *Handbook of Econometrics, Vol. IV*. Amsterdam: Elsevier Science.

Angrist, Joshua D. and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99(4): 1384-1414.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3-30.

Billingsley, Patrick. 1995. *Probability and Measure*, third edition. New York: Wiley.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin and Johannes M. Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs." *The Journal of Human Resources*, 32(3): 549-576.

Dynarski, Susan, Joshua M. Hyman and Diane Schanzenbach. 2011. "Experimental Evidence on the Effect of Childhood Investments on Post-secondary Attainment and Degree Completion." National Bureau of Economic Research Working Paper No. 17533.

Dynarski, Susan, Joshua M. Hyman and Diane Schanzenbach. 2013. "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion." *Journal of Policy Analysis and Management* 32(4): 692-717.

Giné, Xavier, Jessica Goldberg and Dean Yang. 2012. "Credit Market Consequences of Improved Personal Identification." *American Economic Review* 102(6): 2923-54.

Goldrick-Rab, Sara, Douglas N. Harris, James Benson and Robert Kelchen. 2011. "Conditional cash transfers and college persistence: Evidence from a randomized need-based grant program." Institute for Research on Poverty Discussion Paper No. 1393-11.

Goldrick-Rab, Sara, Douglas N. Harris, James Benson and Robert Kelchen. 2012. "Need-based financial aid and college persistence: Experimental evidence from Wisconsin." Available at SSRN 1887826.

Hansen, Ben B. 2008. "The Prognostic Analogue of the Propensity Score." *Biometrika* 95(2): 481-488.

Hansen, Bruce E. 2012. *Econometrics.* Available at: `http://www.ssc.wisc.edu/~bhansen/`

Harris, Douglas N. and Sara Goldrick-Rab. 2012. "Improving the Productivity of Education Experiments: Lessons from a Randomized Study of Need-Based Financial Aid." *Education Finance and Policy* 7(2): 143-169.

Harvill, Eleanor L., Laura R. Peck and Stephen H. Bell. 2013. "On Overfitting in Analysis of Symmetrically Predicted Endogenous Subgroups From Randomized Experimental Samples: Part Three of a Method Note in Three Parts." *American Journal of Evaluation* 34(4): 545-566.

Hausman, Jerry A., and David A. Wise. 1977. "Social Experimentation, Truncated Distributions, and Efficient Estimation." *Econometrica* 45(4): 919-938.

Hausman, Jerry A., and David A. Wise. 1981. "Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment." In: Manski, C.F., McFadden, D.L. (Eds.), *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge: MIT Press.

Heckman, James J., Jeffrey Smith and Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487-535.

Hemelt, Steven W., Kimberly B. Roth and William W. Eaton. 2012. "Childhood Educational Interventions: Experimental Evidence on Post-secondary Outcomes." Paper presented at the Association for Education Finance and Policy Conference, Boston, MA.

Kent, David M. and Rodney A. Hayward. 2007. "Limitations of Applying Summary Results of Clinical Trials to Individual Patients: The Need for Risk Stratification." *Journal of the American Medical Association* 298(10): 1209-12.

Kim, Jeankyung and David Pollard. 1990. "Cube Root Asymptotics." *The Annals of Statistics* 18(1): 191-219.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

Murnane, Richard J. and John B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* Oxford: Oxford University Press.

Newey, Whitney K. and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In: Engle, R.F., McFadden, D.L. (Eds.), *Handbook of Econometrics, Vol. IV.* Amsterdam: Elsevier Science.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin and George Cave. 1996. *Does Training for the Disadvantaged Work?* Washington, DC: The Urban Institute.

Pane, John F., Beth Ann Griffin, Daniel F. McCaffrey and Rita Karam. 2013. "Effectiveness of Cognitive Tutor Algebra I at Scale." *Educational Evaluation and Policy Analysis*, Online-First: `http://epa.sagepub.com/content/early/2013/11/08/0162373713507480`.

Peck, Laura R. 2003. "Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post-Treatment Choice." *American Journal of Evaluation* 24(2): 157-187.

Rodriguez-Planas, Nuria. 2012. "Schools and Drugs: Closing the Gap. Evidence from a Randomized Trial in the US." IZA Discussion Paper No. 6770.

Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan and Jeanne Brooks-Gunn. 2006. "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment." *Journal of Human Resources* 41(4): 649-691.

Tropp, Joel A. 2012. "User-Friendly Tail Bounds for Sums of Random Matrices." *Foundations of Computational Mathematics* 12: 389-434.

van der Vaart, Aad W. 1998. *Asymptotic Statistics.* New York: Cambridge University Press.

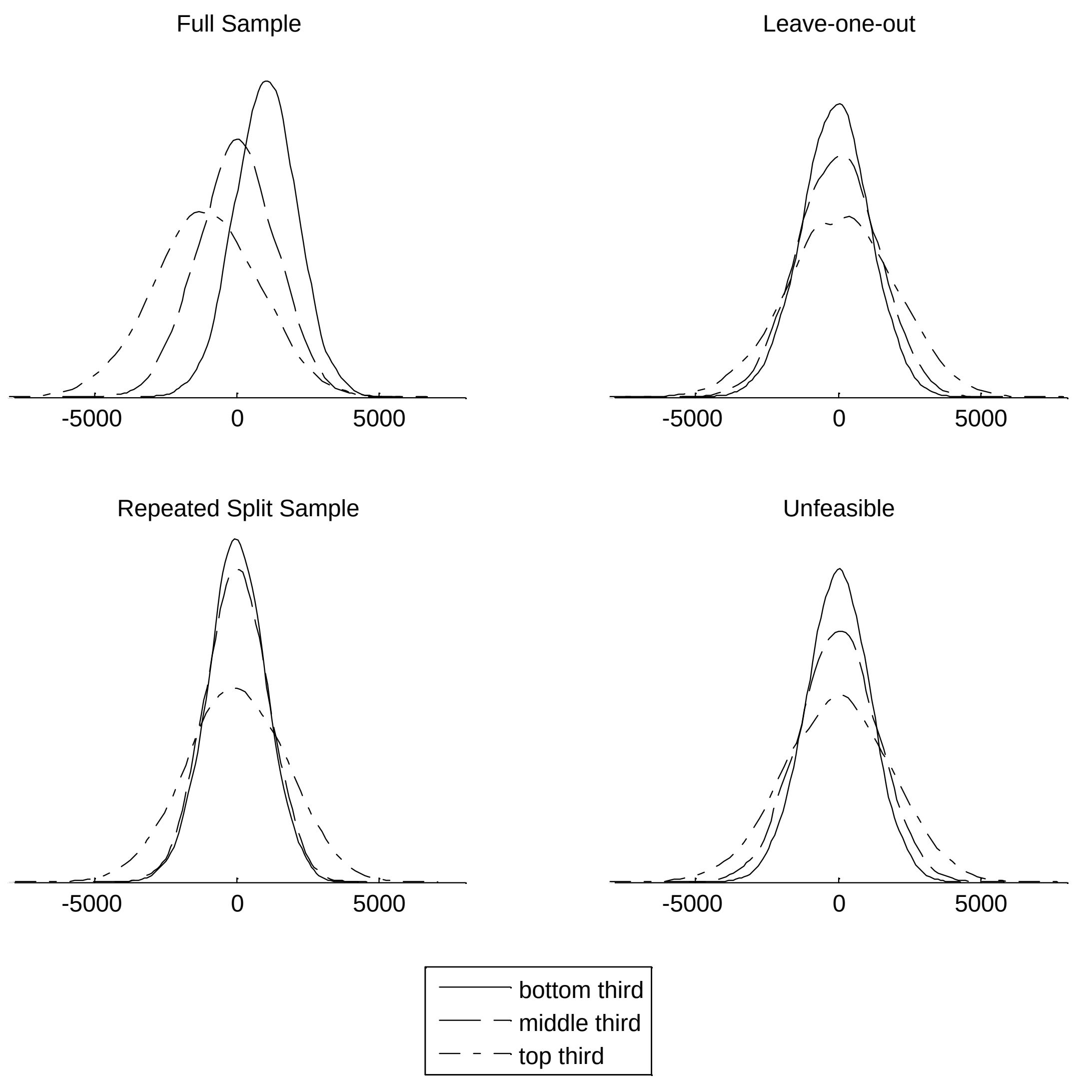

Figure 1
Distributions of the Estimators in the JTPA Simulation

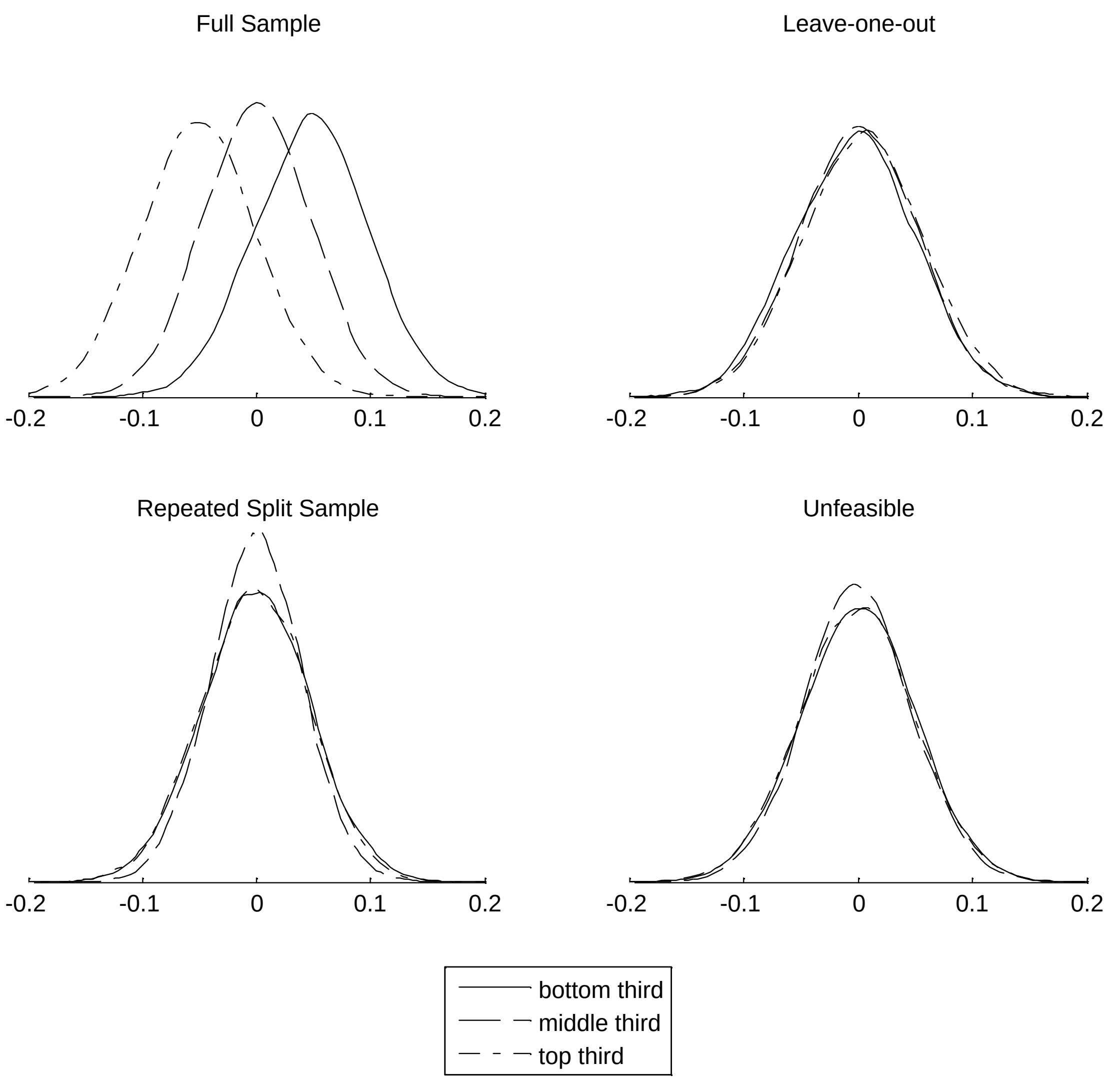

Figure 2
Distributions of the Estimators in the STAR Simulation

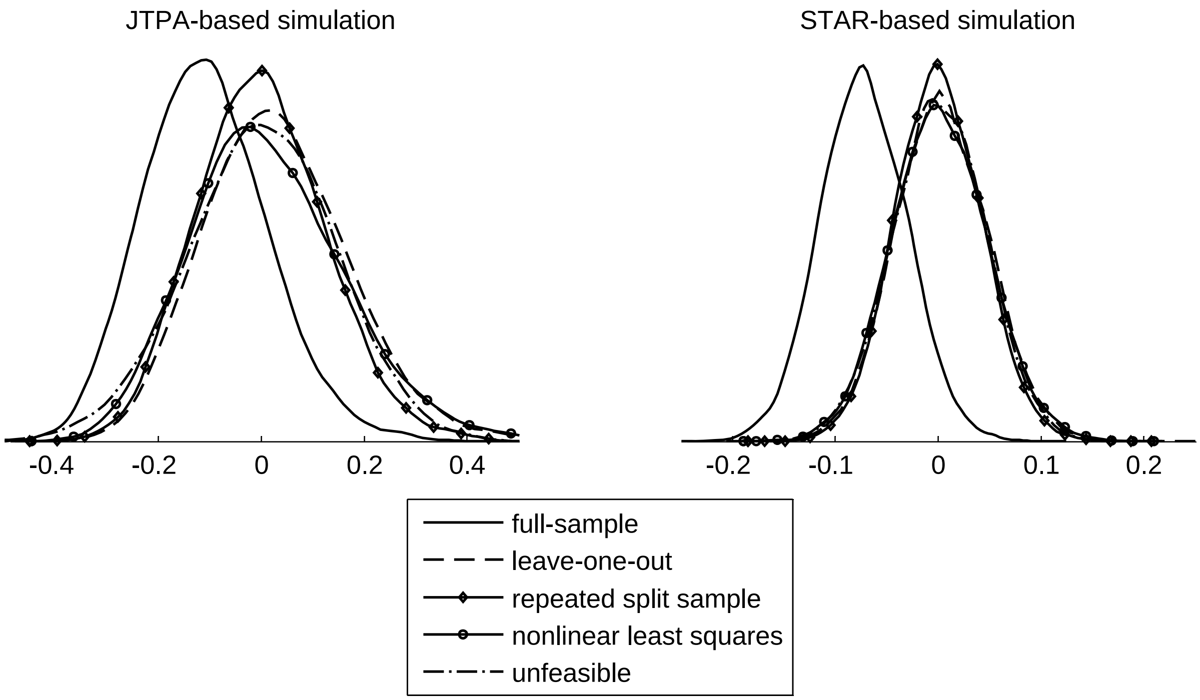

Figure 3
Distribution of the Regression Interaction Estimator

Table 1
JTPA Estimation Results

Panel A: Average treatment effect

	unadjusted	adjusted
$\widehat{\tau}$	1516.49*	1207.22
	(807.27)	(763.54)

Panel B: Average treatment effect by predicted outcome group

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	2379.65**	−719.38	2397.26	2011.70*	−554.65	1769.03
	(1151.07)	(1474.81)	(1672.62)	(1150.68)	(1482.32)	(1639.06)
$\widehat{\tau}_k^{LOO}$	573.74	35.31	3646.53**	173.45	172.28	3118.17*
	(1201.33)	(1509.30)	(1727.08)	(1213.25)	(1513.70)	(1679.62)
$\widehat{\tau}_k^{RSS}$	788.75	254.25	3569.41**	412.01	181.81	2942.69**
	(1027.47)	(1092.85)	(1496.73)	(1042.17)	(1087.51)	(1454.16)
$\widehat{\tau}_k^{PREV}$	1278.88	−67.95	3972.21**	822.05	−150.89	3146.85**
	(1221.96)	(1284.77)	(1497.47)	(1235.13)	(1274.45)	(1430.37)

Note: The JTPA sample includes 1681 treated observations and 849 untreated observations, for a total of 2530 observations. Bootstrap standard errors, based on 1000 bootstrap repetitions, are reported in parentheses. The repeated split sample estimator, $\widehat{\tau}_k^{RSS}$, uses 100 repetitions. Each repetition randomly permutes the order of the untreated observations. Then, the first 425 untreated observations after re-ordering are used to estimate $\boldsymbol{\beta}$. The remaining 424 untreated observations and the 1681 treated observations are used in the second step estimation of τ_k. The "unadjusted" estimates are differences in mean outcomes between treated and non-treated. The "adjusted" estimates are regression coefficients on the treatment variable in a linear regression that includes the list of covariates detailed in Section 3.
* indicates statistical significance at the 0.10 level.
** indicates statistical significance at the 0.05 level.

Table 2

STAR Estimation Results

Panel A: Average treatment effect

	unadjusted	adjusted
$\widehat{\tau}$	0.1659^{**}	0.1892^{**}
	(0.0329)	(0.0294)

Panel B: Average treatment effect by predicted outcome group

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	0.3705^{**}	0.2688^{**}	-0.1330^{**}	0.3908^{**}	0.3023^{**}	-0.1242^{**}
	(0.0521)	(0.0655)	(0.0636)	(0.0509)	(0.0678)	(0.0614)
$\widehat{\tau}_k^{LOO}$	0.3277^{**}	0.2499^{**}	-0.0486	0.3440^{**}	0.2730^{**}	-0.0660
	(0.0547)	(0.0670)	(0.0654)	(0.0519)	(0.0696)	(0.0634)
$\widehat{\tau}_k^{RSS}$	0.3152^{**}	0.2617^{**}	-0.0520	0.3130^{**}	0.3005^{**}	-0.0374
	(0.0467)	(0.0505)	(0.0567)	(0.0459)	(0.0526)	(0.0552)

Note: The STAR sample includes 1747 treated observations and 2017 untreated observations, for a total of 3764 observations. Bootstrap standard errors, based on 1000 bootstrap repetitions, are reported in parentheses. The repeated split sample estimator, $\widehat{\tau}_k^{RSS}$, uses 100 repetitions. Each repetition randomly permutes the order of the untreated observations. Then, the first 1009 untreated observations after re-ordering are used to estimate $\boldsymbol{\beta}$. The remaining 1008 untreated observations and the 1747 treated observations are used in the second step estimation of τ_k. The "unadjusted" estimates are differences in mean outcomes between treated and non-treated. The "adjusted" estimates are regression coefficients on the treatment variable in a linear regression that includes the list of covariates detailed in Section 3.

* indicates statistical significance at the 0.10 level.

** indicates statistical significance at the 0.05 level.

Table 3
JTPA Simulation Results

Panel A: Bias

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	1017.51	-4.81	-1082.42	1017.60	-0.98	-1062.39
$\widehat{\tau}_k^{LOO}$	-88.23	-23.28	96.57	-59.08	-54.01	42.86
$\widehat{\tau}_k^{RSS}$	-2.74	-2.34	-20.75	-3.30	-5.96	-17.56
$\widehat{\tau}_k^{SSS}$	5.62	-9.88	6.62	1.34	-11.24	1.39
$\widehat{\tau}_k^{UNF}$	-1.50	-8.56	-16.85	-2.50	-9.04	-11.67

Panel B: Coverage rates for nominal 0.05 C.I.

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	0.152	0.049	0.089	0.154	0.050	0.089
$\widehat{\tau}_k^{LOO}$	0.051	0.048	0.050	0.051	0.049	0.051
$\widehat{\tau}_k^{RSS}$	0.051	0.048	0.049	0.052	0.048	0.050
$\widehat{\tau}_k^{SSS}$	0.050	0.048	0.051	0.052	0.049	0.050
$\widehat{\tau}_k^{UNF}$	0.053	0.050	0.050	0.053	0.051	0.051

Panel C: Root-MSE

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	1492.26	1364.27	2145.78	1489.89	1375.04	2065.87
$\widehat{\tau}_k^{LOO}$	1192.35	1399.74	1895.93	1180.13	1398.76	1800.86
$\widehat{\tau}_k^{RSS}$	1031.61	1101.14	1751.43	1022.53	1103.53	1660.50
$\widehat{\tau}_k^{SSS}$	1500.97	1797.74	2383.17	1493.40	1792.14	2271.90
$\widehat{\tau}_k^{UNF}$	1119.34	1372.67	1867.51	1118.52	1383.76	1792.25

Note: Averages over 10000 simulations. See Section 4 and Appendix 2 for details.

Table 4

STAR Simulation Results

Panel A: Bias

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	0.0483	0.0006	-0.0511	0.0487	0.0010	-0.0506
$\widehat{\tau}_k^{LOO}$	-0.0025	0.0005	0.0046	0.0028	-0.0025	-0.0075
$\widehat{\tau}_k^{RSS}$	0.0001	-0.0000	-0.0012	0.0002	0.0001	-0.0010
$\widehat{\tau}_k^{SSS}$	-0.0005	0.0004	-0.0017	-0.0002	0.0003	-0.0015
$\widehat{\tau}_k^{UNF}$	0.0004	-0.0003	-0.0009	0.0002	-0.0002	-0.0006

Panel B: Coverage rates for nominal 0.05 C.I.

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	0.161	0.051	0.178	0.178	0.049	0.191
$\widehat{\tau}_k^{LOO}$	0.048	0.050	0.051	0.050	0.049	0.056
$\widehat{\tau}_k^{RSS}$	0.053	0.051	0.048	0.052	0.051	0.049
$\widehat{\tau}_k^{SSS}$	0.051	0.050	0.052	0.052	0.052	0.050
$\widehat{\tau}_k^{UNF}$	0.051	0.052	0.049	0.051	0.050	0.050

Panel C: Root-MSE

	unadjusted			adjusted		
	low	medium	high	low	medium	high
$\widehat{\tau}_k$	0.0695	0.0472	0.0716	0.0677	0.0471	0.0691
$\widehat{\tau}_k^{LOO}$	0.0526	0.0509	0.0530	0.0492	0.0507	0.0494
$\widehat{\tau}_k^{RSS}$	0.0473	0.0402	0.0470	0.0444	0.0399	0.0439
$\widehat{\tau}_k^{SSS}$	0.0617	0.0589	0.0615	0.0577	0.0583	0.0571
$\widehat{\tau}_k^{UNF}$	0.0501	0.0469	0.0503	0.0480	0.0475	0.0476

Note: Averages over 10000 simulations. See Section 4 and Appendix 2 for details.

Table 5
Bias in Simulations Using Artificial Data
(constant regression coefficients)

	$K = 10$						$K = 20$						$K = 40$					
	unadjusted			adjusted			unadjusted			adjusted			unadjusted			adjusted		
	low	med.	high	low	med.	high	low	med.	high	low	med.	high	low	med.	high	low	med.	high
$N = 200$																		
$\widehat{\tau}_k$	2.24	-0.05	-2.28	2.23	-0.06	-2.27	3.15	-0.00	-3.13	3.12	-0.01	-3.08	4.09	0.02	-4.06	3.89	0.04	-3.88
$\widehat{\tau}_k^{LOO}$	-0.32	-0.03	0.25	-0.07	-0.04	0.01	-0.18	0.00	0.17	0.30	0.01	-0.26	-0.09	0.00	0.13	0.89	0.02	-0.88
$\widehat{\tau}_k^{RSS}$	-0.03	-0.04	-0.04	-0.02	-0.04	-0.05	0.01	-0.00	-0.01	0.03	0.01	0.01	0.03	0.01	-0.00	0.04	0.02	-0.01
$\widehat{\tau}_k^{UNF}$	-0.04	-0.04	-0.03	-0.02	-0.03	-0.04	0.03	0.00	0.00	0.02	-0.01	0.03	0.02	0.00	0.01	-0.01	0.05	0.04
$N = 1000$																		
$\widehat{\tau}_k$	0.55	0.01	-0.54	0.55	0.01	-0.54	0.71	-0.01	-0.71	0.70	-0.00	-0.70	0.82	0.01	-0.82	0.81	0.01	-0.82
$\widehat{\tau}_k^{LOO}$	-0.05	0.01	0.07	-0.05	0.01	0.06	-0.05	-0.01	0.05	-0.05	-0.00	0.05	-0.01	0.01	0.01	-0.00	0.01	-0.01
$\widehat{\tau}_k^{RSS}$	0.02	0.00	-0.00	0.02	0.01	-0.00	-0.01	-0.01	0.01	-0.01	-0.00	0.01	0.01	0.00	-0.01	0.01	0.00	-0.01
$\widehat{\tau}_k^{UNF}$	0.01	0.01	0.00	0.00	0.01	0.00	-0.00	-0.01	-0.01	-0.00	-0.00	0.00	0.01	0.00	-0.01	0.01	0.01	-0.01
$N = 5000$																		
$\widehat{\tau}_k$	0.12	-0.00	-0.12	0.12	-0.00	-0.12	0.15	0.00	-0.16	0.15	0.00	-0.15	0.16	-0.00	-0.16	0.16	-0.00	-0.16
$\widehat{\tau}_k^{LOO}$	-0.01	-0.00	0.01	-0.01	-0.00	0.01	-0.01	0.00	0.00	-0.00	0.00	0.00	-0.01	-0.00	0.01	-0.01	-0.00	0.01
$\widehat{\tau}_k^{RSS}$	0.01	-0.00	-0.01	0.01	-0.00	-0.01	0.00	-0.00	-0.01	0.01	-0.00	-0.01	-0.00	-0.00	0.00	0.00	-0.00	0.00
$\widehat{\tau}_k^{UNF}$	0.01	-0.00	-0.01	0.01	-0.00	-0.01	0.01	-0.01	-0.01	0.01	-0.01	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.01

Note: Averages over 10000 simulations. K is the number of included regressors. See Section 4 and Appendix 2 for details.

Table 6
Bias in Simulations Using Artificial Data
(decaying regression coefficients)

	$K=10$						$K=20$						$K=40$					
	unadjusted			adjusted			unadjusted			adjusted			unadjusted			adjusted		
	low	med.	high	low	med.	high	low	med.	high	low	med.	high	low	med.	high	low	med.	high
$N=200$																		
$\widehat{\tau}_k$	2.94	-0.00	-2.97	2.97	-0.00	-2.99	4.57	-0.01	-4.60	4.63	-0.02	-4.69	6.86	0.04	-6.87	7.20	-0.01	-7.24
$\widehat{\tau}_k^{LOO}$	-0.41	-0.01	0.40	0.01	-0.01	-0.02	-0.26	0.01	0.23	0.76	-0.04	-0.79	-0.14	-0.00	0.17	2.48	-0.00	-2.48
$\widehat{\tau}_k^{RSS}$	-0.03	-0.01	0.01	-0.02	-0.01	0.01	-0.01	-0.00	-0.01	-0.02	-0.02	-0.02	0.00	0.01	0.01	-0.02	0.01	0.02
$\widehat{\tau}_k^{UNF}$	-0.01	-0.01	-0.01	-0.02	-0.01	0.00	-0.01	-0.03	0.02	0.00	-0.05	-0.01	-0.03	0.02	0.04	-0.03	0.02	0.01
$N=1000$																		
$\widehat{\tau}_k$	0.95	0.01	-0.94	0.95	0.01	-0.94	1.64	-0.01	-1.64	1.64	-0.00	-1.64	2.64	-0.01	-2.64	2.66	-0.01	-2.65
$\widehat{\tau}_k^{LOO}$	-0.11	0.01	0.12	-0.10	0.01	0.11	-0.10	-0.01	0.11	-0.03	-0.01	0.04	-0.07	-0.00	0.08	0.14	0.00	-0.13
$\widehat{\tau}_k^{RSS}$	0.01	0.00	0.01	0.01	0.00	0.01	-0.00	-0.00	0.00	-0.00	-0.00	0.01	0.00	-0.00	0.00	-0.00	-0.00	0.01
$\widehat{\tau}_k^{UNF}$	0.01	0.00	0.01	0.01	0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.01	0.00	-0.01	-0.01	0.02	-0.01	-0.01	0.03
$N=5000$																		
$\widehat{\tau}_k$	0.23	-0.01	-0.22	0.23	-0.01	-0.22	0.43	-0.00	-0.43	0.43	-0.01	-0.43	0.81	-0.00	-0.81	0.81	-0.00	-0.81
$\widehat{\tau}_k^{LOO}$	-0.02	-0.01	0.04	-0.02	-0.01	0.03	-0.03	-0.01	0.03	-0.03	-0.01	0.03	-0.03	-0.00	0.02	-0.02	-0.01	0.02
$\widehat{\tau}_k^{RSS}$	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00
$\widehat{\tau}_k^{UNF}$	0.00	-0.00	0.01	0.00	-0.00	0.01	0.00	-0.01	0.00	0.00	-0.01	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00

Note: Averages over 10000 simulations. K is the number of included regressors. See Section 4 and Appendix 2 for details.