Online supplement for "Optimal Data Collection for Randomized Control Trials"

By Pedro Carneiro, Sokbae Lee, and Daniel Wilhelm*

May 1, 2019

Abstract

In Appendix S1, we justify our data collection rule in a decision-theoretic framework. In Appendix S2, we describe an orthogonal greedy algorithm (OGA) and in Appendix S3, we show that this algorithm possesses desirable theoretical properties. Appendix S4 provides a detailed description of all components of the calibrated cost functions for both empirical examples. In Appendix S5, we consider a simplified setup in which all covariates are orthogonal to each other, and the budget constraint has a very simple form. Appendix S6 presents the results of Monte Carlo simulations, and shows that all three methods considered in the main text select more covariates and smaller sample sizes as we increase the predictive power of some covariates. Appendix S7 shows the full list and definitions of selected covariates for the baseline outcome in the school grants example. In Appendix S8, we perform an out-of-sample evaluation by splitting the dataset into training samples for the covariate selection step and evaluation samples for the computation of the performance measures in both empirical examples. Appendix S9 provides details

*Carneiro: University College London, Institute for Fiscal Studies (IFS), and Centre for Microdata Methods and Practice (CeMMAP); Lee: Columbia University, IFS and CeMMAP; Wilhelm: University College London and CeMMAP. We thank Frank Diebold, Kirill Evdokimov, Michal Kolesar, David McKenzie, Ulrich Müller, Andriy Norets, Imran Rasul, and participants at various seminars for helpful discussions. An early version of this paper was presented at Columbia University and Princeton University in September 2014, and at New York University and University of Pennsylvania in December 2014. This work was supported in part by the European Research Council (ERC-2014-CoG-646917-ROMIA and ERC-2015-CoG-682349) and by the UK Economic and Social Research Council (ESRC) through a grant (ES/P008909/1) to the ESRC CeMMAP.

regarding how to deal with a vector of outcomes when we Appendix S10 gives a detailed description of how we increase the correlation of the baseline score with the follow-up score in the school grants example. select the common set of regressors for all outcomes.

S1 Decision Theory Applied to Data Collection

To describe our decision theoretic problem, we first introduce some notation. Let $\Theta := \Theta_1 \times \Gamma$, where $\Theta_1 \subseteq \mathbb{R}$, $\Gamma \subseteq \mathbb{R}^M$, be a parameter space and decompose a typical element θ as $\theta = (\beta, \gamma')'$ where $\beta \in \Theta_1$ corresponds to possible average treatment effect parameters and $\gamma \in \Gamma$ to possible coefficients for the additional covariates. Let $\mathcal{S}_{pre} := \{Y_i, X_i\}_{i=1}^N$ and $\mathcal{S}_{exp} := \{Y_i, D_i, X_i\}_{i=1}^n$ denote the pre-experimental and experimental random samples that take values in the sample spaces $\mathbb{R}^{N(M+1)}$ and $\mathbb{R}^{n(M+2)}$, respectively. Let $\mathbb{A} \subseteq \mathbb{R}$ be the action space and $\mathbb{S}_{exp}$ the space of possible $\mathcal{S}_{exp}$. The goal is to make a decision, i.e. choose an estimator, $\hat{\beta} : \mathbb{S}_{exp} \to \mathbb{A}$, which maps the experimental sample into an action, i.e. into an estimate of the average treatment effect β_0. We evaluate the performance of the decision rule $\hat{\beta}$ by average risk $R_n : \mathbb{D}_n \to \mathbb{R}$ defined as

$$R_n(\hat{\beta}) := \int_\Theta L_n(\theta, \hat{\beta}(\mathcal{S}_{exp}))d\mu(\theta),$$

where $\mu : \Theta \to \mathbb{R}$ is a weight function that integrates to one, $L_n : \Theta \times \mathbb{A} \to \mathbb{R}$ is a loss function, and $\mathbb{D}_n$ a set of feasible decision rules. To make a decision we minimize average risk:

$$\min_{\hat{\beta} \in \mathbb{D}_n} R_n(\hat{\beta}). \tag{S.1}$$

In the context of our data collection problem, we can implement our proposed data collection procedure only after making some specific choices about the loss function, the set of feasible decision rules and the weighting function μ. First, we pick average squared loss (or mean-squared error):

$$L_n(\theta, \hat{\beta}(\mathcal{S}_{exp})) := E_{exp,\theta}\left[\left(\beta - \hat{\beta}(\mathcal{S}_{exp})\right)^2\right],$$

where $E_{exp,\theta}$ denotes the expectation with respect to the distribution of the experimental random sample $\mathcal{S}_{exp}$, which might depend on the unknown state θ. This is a common loss function for the evaluation of estimators and leads to a particularly simple formulation of the resulting data collection problem that can easily be implemented in practice.

For any random sample $\mathcal{S} \in \mathbb{S}_{exp}$ and any $\gamma \in \Gamma$, let $b(\gamma, \mathcal{S})$ be the OLS estimator from a regression of $Y - \gamma' X$ on a constant and D. We choose the set of feasible decision rules to contain all $\hat{\beta}$ that can be written as the OLS estimator $b(\gamma, \cdot)$ for some value of

$\gamma \in \Gamma$ and also satisfy the budget constraint:

$$\mathbb{D}_n := \Big\{ b(\gamma, \cdot) : \mathbb{S}_{exp} \to \mathbb{A} \ \text{ s.t. } \ \gamma \in \Gamma \ \text{ and } \ c(\mathcal{I}(\gamma), n) \leq B \Big\}.$$

This definition means that the decision rules in $\mathbb{D}_n$ can be indexed by $\gamma \in \Gamma$ that satisfy the budget constraint. Therefore, choosing a decision rule in $\mathbb{D}_n$ is equivalent to choosing a feasible value of γ.

The weighting function μ is chosen as a product of two densities, $\mu := \mu_\beta \times \mu_\gamma$, where μ_β is a density on Θ_1 and μ_γ a density on Γ.

Since prior to the experiment we do not observe the experimental sample and the distribution of the experimental data is unknown, in practice it is not possible to minimize $R_n(\hat{\beta})$ over $\hat{\beta} \in \mathbb{D}_n$. However, we now show that, under a homoskedasticity assumption, this minimization problem is equivalent to a minimization problem that only involves the pre-experimental sample and thus can be solved prior to the experiment. To this end define the variance $Var_{pre,\gamma}$ based on $E_{pre,\gamma}$, the expectation with respect to the distribution of $\mathcal{S}_{pre}$, which may depend on γ, but not on β. Similarly, let $Var_{exp,\theta}$ be the variance based on $E_{exp,\theta}$. Define $\Delta(n) := (E[(\bar{D}_n(1-\bar{D}_n))^{-1}])^{-1}$ with $\bar{D}_n := n^{-1}\sum_{i=1}^n D_i$. We will assume that the treatment assignment D_i is completely randomized, so that in particular the distribution of D_i does not depend on θ and thus the expectation in the definition of $\Delta(n)$ is independent of θ. Given a distribution for treatment assignment (e.g. Bernoulli with success probability $p = 1/2$), the quantity $\Delta(n)$ is known.

Assumption 1.1. **(i)** $\mathcal{S}_{exp}$ is an i.i.d. sample and D_i is completely randomized, i.e. $P(D_i = 1|X_i) = p$ a.s. for some $p \in (0,1)$. **(ii)** $Var_{exp,\theta}(Y - \hat{\gamma}'X|D = 1) = Var_{exp,\theta}(Y - \hat{\gamma}'X|D = 0)$, where $\theta = (\beta, \gamma')'$, for all $\beta \in \Theta_1$ and $\gamma, \hat{\gamma} \in \Gamma$. **(iii)** $E_{exp,\theta}[X_i X_i'] = E_{pre,\gamma}[X_i X_i']$, where $\theta = (\beta, \gamma')'$, for all $\beta \in \Theta_1$ and $\gamma, \hat{\gamma} \in \Gamma$.

Theorem 1.1. Under Assumption 1.1, the minimizer of (S.1) is equal to $b(\hat{\gamma}^*, \cdot)$, where $\hat{\gamma}^*$ is the minimizer of

$$\min_{\hat{\gamma} \in \Gamma :\, c(\mathcal{I}(\hat{\gamma}), n) \leq B} \ \frac{1}{n\Delta(n)} \int_\Gamma (\gamma - \hat{\gamma})' E_{pre,\gamma}[X_i X_i'](\gamma - \hat{\gamma}) d\mu_\gamma(\gamma).$$

Proof. Consider:

$$\min_{\hat{\beta} \in \mathbb{D}_n} R_n(\hat{\beta}) = \min_{\hat{\beta} \in \mathbb{D}_n} \int_\Theta L_n(\theta, \hat{\beta}(\mathcal{S}_{exp})) d\mu(\theta)$$

$$
\begin{aligned}
&= \min_{\hat{\beta} \in \mathbb{D}_n} \int_\Theta E_{exp,\theta}\left[\left(\beta - \hat{\beta}(\mathcal{S}_{exp})\right)^2\right] d\mu(\theta) \\
&= \min_{\hat{\gamma} \in \Gamma:\ c(\mathcal{I}(\hat{\gamma}),n) \le B} \int_\Theta E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2\right] d\mu(\theta)
\end{aligned}
$$

The homoskedastic error assumption, Assumption 1.1(ii), implies that conditional on $D_1, \ldots, D_n$, the estimator $b(\hat{\gamma}, \mathcal{S}_{exp})$ is unbiased and thus a straight-forward calculation shows that its finite-sample MSE (conditional on $D_1, \ldots, D_n$) is

$$
\begin{aligned}
E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2 \mid D_1, \ldots, D_n\right] &= Var_{exp,\theta}\left(b(\hat{\gamma}, \mathcal{S}_{exp}) \mid D_1, \ldots, D_n\right) \\
&= \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i \mid D_i = 0\right)}{n\bar{D}_n(1-\bar{D}_n)} \\
&= \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\bar{D}_n(1-\bar{D}_n)}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\int_\Theta E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2\right] d\mu(\theta) \\
&\quad = \int_\Theta E_{exp,\theta}\left\{E_{exp,\theta}\left[(\beta - b(\hat{\gamma}, \mathcal{S}_{exp}))^2 \mid D_1, \ldots, D_n\right]\right\} d\mu(\theta) \\
&\quad = \int_\Theta E_{exp,\theta}\left\{\frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\bar{D}_n(1-\bar{D}_n)}\right\} d\mu(\theta) \\
&\quad = \int_\Theta \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\Delta(n)} d\mu(\theta)
\end{aligned}
$$

The usual least-squares calculation then shows that minimizing the residual variance is the same as minimizing the quadratic distance of $\hat{\gamma}$ from γ:

$$
\begin{aligned}
&\text{argmin}_{\hat{\gamma} \in \Gamma:\ c(\mathcal{I}(\hat{\gamma}),n) \le B} \int_\Theta \frac{Var_{exp,\theta}\left(Y_i - \hat{\gamma}' X_i\right)}{n\Delta(n)} d\mu(\theta) \\
&\quad = \text{argmin}_{\hat{\gamma} \in \Gamma:\ c(\mathcal{I}(\hat{\gamma}),n) \le B} \int_\Theta \frac{E_{exp,\theta}[((\gamma - \hat{\gamma})' X_i)^2]}{n\Delta(n)} d\mu(\theta) \\
&\quad = \text{argmin}_{\hat{\gamma} \in \Gamma:\ c(\mathcal{I}(\hat{\gamma}),n) \le B} \int_\Theta \frac{(\gamma - \hat{\gamma})' E_{exp,\theta}[X_i X_i'](\gamma - \hat{\gamma})}{n\Delta(n)} d\mu(\theta) \\
&\quad = \text{argmin}_{\hat{\gamma} \in \Gamma:\ c(\mathcal{I}(\hat{\gamma}),n) \le B} \int_\Gamma \frac{(\gamma - \hat{\gamma})' E_{pre,\gamma}[X_i X_i'](\gamma - \hat{\gamma})}{n\Delta(n)} d\mu_\gamma(\gamma)
\end{aligned}
$$

which is the desired expression. Q.E.D.

This theorem implies that we can minimize risk of the treatment effect estimator by finding the optimal combination of covariates in the pre-experimental sample subject to the budget constraint. This optimization problem can be solved in practice for any user-chosen weight function μ_γ.

It is reasonable to consider the case where $E_{exp,\theta}[Y_i X_i'] = E_{pre,\gamma}[Y_i X_i']$. With this assumption it is reasonable to choose a weight function μ_γ that puts all weight on the available information from the pre-experimental sample. Hence, we make our procedure operational by choosing such a weight function. This leads to the formulation given in (3.8), if we restrict $\bar{D}_n = 1/2$.

S2 A Simple Greedy Algorithm

In practice, the vector X of potential covariates is typically high-dimensional, which makes it challenging to solve the optimization problem (3.8). In this section, we propose a computationally feasible algorithm that is both conceptually simple and performs well in our simulations. In particular, it requires only running many univariate, linear regressions and can therefore easily be implemented in popular statistical packages.

We split the joint optimization problem in (3.8) over n and γ into two nested problems. The outer problem searches over the optimal sample size n, while the inner problem determines the optimal selection of covariates for each sample size n:

$$\min_{n\in\mathcal{N}} \frac{1}{n\Delta(n)} \min_{\gamma\in\mathbb{R}^M} \frac{1}{N}\sum_{i=1}^{N}(Y_i - \gamma' X_i)^2 \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B. \tag{S.2}$$

To convey our ideas in a simple form, suppose for the moment that the budget constraint has the following linear form,

$$c(\mathcal{I}(\gamma), n) = n \cdot |\mathcal{I}(\gamma)| \leq B,$$

where $|\mathcal{I}(\gamma)|$ denotes the number of non-zero elements of γ. Note that the budget constraint puts the restriction on the number of selected covariates, that is, $|\mathcal{I}(\gamma)| \leq B/n$.

It is known to be NP-hard (non-deterministic polynomial time hard) to find a solution

to the inner optimization problem in (S.2) subject to the constraint that γ has m non-zero components, also called an m-term approximation, where m is the integer part of B/n in our problem. In other words, solving (S.2) directly is not feasible unless the dimension of covariates, M, is sufficiently small (Natarajan, 1995; Davis, Mallat, and Avellaneda, 1997).

There exists a class of computationally attractive procedures called greedy algorithms that are able to approximate the infeasible solution. See Temlyakov (2011) for a detailed discussion of greedy algorithms in the context of approximation theory. Tropp (2004), Tropp and Gilbert (2007), Barron, Cohen, Dahmen, and DeVore (2008), Zhang (2009), Huang, Zhang, and Metaxas (2011), Ing and Lai (2011), and Sancetta (2016), among many others, demonstrate the usefulness of greedy algorithms for signal recovery in information theory, and for the regression problem in statistical learning. We use a variant of OGA that can allow for selection of groups of variables (see, for example, Huang, Zhang, and Metaxas (2011)).

To formally define our proposed algorithm, we introduce some notation. For a vector v of N observations $v_1, \ldots, v_N$, let $\|v\|_N := (1/N \sum_{i=1}^N v_i^2)^{1/2}$ denote the empirical L^2-norm and let $\mathbf{Y} := (Y_1, \ldots, Y_N)'$.

Suppose that the covariates $X^{(j)}$, $j = 1, \ldots, M$, are organized into p pre-determined groups $X_{G_1}, \ldots, X_{G_p}$, where $G_k \subseteq \{1, \ldots, p\}$ indicates the covariates of group k. We denote the corresponding matrices of observations by bold letters (i.e., $\mathbf{X}_{G_k}$ is the $N \times |G_k|$ matrix of observations on X_{G_k}, where $|G_k|$ denotes the number of elements of the index set G_k). By a slight abuse of notation, we let $\mathbf{X}_k := \mathbf{X}_{\{k\}}$ be the column vector of observations on X_k when k is a scalar. One important special case is that in which each group consists of a single regressor. Furthermore, we allow for overlapping groups; in other words, some elements can be included in multiple or even all groups. The group structure occurs naturally in experiments where data collection is carried out through surveys whose questions can be grouped in those concerning income, those concerning education, and so on. This can also occur naturally when we consider multivariate outcomes. See Appendix S9 for details.

Suppose that the largest group size $J_{\max} := \max_{k=1,\ldots,p} |G_k|$ is small, so that we can implement orthogonal transformations *within each group* such that $(\mathbf{X}_{G_j}'\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$, where $\mathbf{I}_d$ is the d-dimensional identity matrix. In what follows, assume that $(\mathbf{X}_{G_j}'\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$ without loss of generality. Let $|\cdot|_2$ denote the ℓ_2 norm. The following procedure

describes our algorithm for a general cost function c.

STEP 1. Set the initial sample size $n = n_0$.

STEP 2. Group OGA for a given sample size n:

(a) initialize the inner loop at $k = 0$ and set the initial residual $\hat{\mathbf{r}}_{n,0} = \mathbf{Y}$, the initial covariate indices $\hat{\mathcal{I}}_{n,0} = \emptyset$ and the initial group indices $\hat{\mathcal{G}}_{n,0} = \emptyset$;

(b) separately regress $\hat{\mathbf{r}}_{n,k}$ on each group of regressors in $\{1, \ldots, p\} \backslash \hat{\mathcal{G}}_{n,k}$; call $\hat{j}_{n,k}$ the group of regressors with the largest ℓ_2 regression coefficients,

$$\hat{j}_{n,k} := \arg \max_{j \in \{1,\ldots,p\} \backslash \hat{\mathcal{G}}_{n,k}} \left| \mathbf{X}'_{G_j} \hat{\mathbf{r}}_{n,k} \right|_2 ;$$

add $\hat{j}_{n,k}$ to the set of selected groups, $\hat{\mathcal{G}}_{n,k+1} = \hat{\mathcal{G}}_{n,k} \cup \{\hat{j}_{n,k}\}$;

(c) regress $\mathbf{Y}$ on the covariates $\mathbf{X}_{\hat{\mathcal{I}}_{n,k+1}}$ where $\hat{\mathcal{I}}_{n,k+1} := \hat{\mathcal{I}}_{n,k} \cup G_{\hat{j}_{n,k}}$; call the regression coefficient $\hat{\gamma}_{n,k+1} := (\mathbf{X}'_{\hat{\mathcal{I}}_{n,k+1}} \mathbf{X}_{\hat{\mathcal{I}}_{n,k+1}})^{-1} \mathbf{X}'_{\hat{\mathcal{I}}_{n,k+1}} \mathbf{Y}$ and the residual $\hat{\mathbf{r}}_{n,k+1} := \mathbf{Y} - \mathbf{X}_{\hat{\mathcal{I}}_{n,k+1}} \hat{\gamma}_{n,k+1}$;

(d) increase k by one and continue with (b) as long as $c(\hat{\mathcal{I}}_{n,k}, n) \leq B$ is satisfied;

(e) let k_n be the number of selected groups; call the resulting submatrix of selected regressors $\mathbf{Z} := \mathbf{X}_{\hat{\mathcal{I}}_{n,k_n}}$ and $\hat{\gamma}_n := \hat{\gamma}_{n,k_n}$, respectively.

STEP 3. Set n to the next sample size in $\mathcal{N}$, and go to Step 2 until (and including) $n = n_K$.

STEP 4. Set $\hat{n}$ as the sample size that minimizes the residual variance:

$$\hat{n} := \arg \min_{n \in \mathcal{N}} \frac{1}{nN} \sum_{i=1}^{N} \left(Y_i - \mathbf{Z}_i \hat{\gamma}_n \right)^2 .$$

The algorithm above produces the selected sample size $\hat{n}$, the selection of covariates $\hat{\mathcal{I}} := \hat{\mathcal{I}}_{\hat{n},k_{\hat{n}}}$ with $k_{\hat{n}}$ selected groups and $\hat{m} := m(\hat{n}) := |\hat{\mathcal{I}}_{\hat{n},k_{\hat{n}}}|$ selected regressors. Here, $\hat{\gamma} := \hat{\gamma}_{\hat{n}}$ is the corresponding coefficient vector on the selected regressors Z.

Remark 2.1. The minimal sample size n_0 in $\mathcal{N}$ could, for example, be determined by power calculations (see, e.g. Duflo, Glennerster, and Kremer, 2007; McConnell and Vera-Hernandez, 2015) that guarantee a certain power level for an hypothesis test of $\beta = 0$.

Remark 2.2. In Appendix S3, we provide theoretical properties of the OGA approximation. Theorem 3.2 in Appendix S3 gives a finite-sample bound on the difference between the best possible MSE for the estimator of the average treatment effect and the MSE for the OGA approximation. If we assume that the pre-experimental and experimental samples are from the same population, the pre-experimental sample size N is large, and the budget B is relatively small, then the difference between two MSEs decreases at a rate of $1/k$ as k increases, where k is the number of the steps in the OGA. It is known in a similar setting that this rate $1/k$ cannot generally be improved (see, e.g., Barron, Cohen, Dahmen, and DeVore, 2008). In this sense, we show that our proposed method has a desirable property. See Appendix S3 for further details.

Remark 2.3. There are many important reasons for collecting covariates, such as checking whether randomization was carried out properly and identifying heterogeneous treatment effects, among others. If a few covariates are essential for the analysis, we can guarantee their selection by including them in every group G_k, $k = 1, \ldots, p$.

Remark 2.4. In a simple model such as the one in Appendix S5, the optimal combination of covariates equalizes the percent marginal contribution of an additional variable to the residual variance with the percent marginal contribution of the additional variable to the costs per interview. Step 2 of the OGA selects the next covariate as the one that has the highest predictive power independent of its cost. Outside a class of very simple models as in Appendix S5, it is difficult to determine an OGA approximation to the optimum that jointly takes into account both predictive power as it requires comparison of all possible covariate combinations. In our empirical application of Section V.B, we study a case with heterogeneous costs and propose a sensitivity analysis that assesses whether the OGA solution significantly changes with perturbations of the set of potential covariates.

S3 Theoretical Properties of the OGA

In this appendix, we provide theoretical properties of the OGA approximation $\hat{\mathbf{f}} := \mathbf{Z}\hat{\gamma}$. Following Barron, Cohen, Dahmen, and DeVore (2008), we define

$$\|f\|_{\mathcal{L}_1^N} := \inf \left\{ \sum_{k=1}^{p} |\beta_k|_2 : \beta_k \in \mathbb{R}^{|G_k|} \text{ and } f = \sum_{k=1}^{p} X'_{G_k} \beta_k, \right.$$

where the elements of X_{G_k} are normalized in the empirical norm $\|\cdot\|_N$}.

When the expression $f = \sum_{k=1}^{p} X'_{G_k}\beta_k$ is not unique, we take the true f_0 to be one with the minimum value of $\|f\|_{\mathcal{L}_1^N}$. This gives $f_0 := \gamma_0' X$ and $\mathbf{f_0} := \mathbf{X}\gamma_0$ for some γ_0. Note that $\mathbf{f_0}$ is defined by $\mathbf{X}$ with the true parameter value γ_0, while $\hat{\mathbf{f}}$ is an OGA estimator of $\mathbf{f_0}$ using only $\mathbf{Z}$.

Define

$$\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}) := \|\mathbf{Y} - \hat{\mathbf{f}}\|_N^2/\hat{n},$$

which is equal to the objective function in (3.8). Note that $\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})$ can also be called the "empirical risk". In addition, define

$$MSE_{\hat{n},\infty}(\mathbf{f_0}) := E_{\exp}\left[(Y - f_0)^2\right]/\hat{n}.$$

Note that $E_{\exp}[(Y - f_0)^2]$ is the counterpart of $\|\mathbf{Y} - \mathbf{f_0}\|_N^2$ using the population in the experiment. We assume that f_0 minimizes $f \mapsto E_{\exp}[(Y - f)^2]$, where $f = \gamma' X$. This implies that

$$\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0}) := MSE_{\hat{n},\infty}(\hat{\mathbf{f}}) - MSE_{\hat{n},\infty}(\mathbf{f_0})$$

is always non-negative. For each OGA step $k \geq 1$, let $\mathcal{G}_k$ denote the following class of functions

$$\mathcal{G}_k := \left\{X_{\mathcal{I}_k} \mapsto \gamma'_{\mathcal{I}_k} X_{\mathcal{I}_k} : \gamma_{\mathcal{I}_k} \in \mathbb{R}^{|\mathcal{I}_k|}\right\}.$$

Let $R(g) := E_{\exp}[(Y - g)^2]$ for any function g. The following theorem gives a bound for $\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0})$.

Theorem 3.2. Assume that $(\mathbf{X}'_{G_j}\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$ for each $j = 1, \ldots, p$. Suppose $\mathcal{N}$ is a finite subset of $\mathbb{N}_+$, $c : \{0,1\}^M \times \mathbb{N}_+ \to \mathbb{R}$ some function, and $B > 0$ some constant. Further, assume that f_0 minimizes $f \mapsto E_{\exp}[(Y - f)^2]$, where $f = \gamma' X$. Then the following bound holds:

$$0 \leq \mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0}) \leq T_1 + T_2,$$

where

$$T_1 := \frac{4\|f\|^2_{\mathcal{L}_1^N}}{\hat{n}} \left(\frac{1}{\min\{p, k_{\hat{n}}\}} \right),$$
$$T_2 := \frac{2}{\hat{n}} \sup_{g \in \mathcal{G}_{k_{\hat{n}}}} \left| \left\{ \|\mathbf{Y} - \mathbf{g}\|_N^2 - R(g) \right\} \right|.$$

The theorem provides a bound on $\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0})$ using the OGA approximation. On one hand, the square root of $MSE_{\hat{n},\infty}(\mathbf{f_0})$ is the best possible standard error for the estimator of the average treatment effect when the sample size is $\hat{n}$. This corresponds to the oracle case that $f_0 = \gamma_0' X$ is known. On the other hand, the square root of $MSE_{\hat{n},\infty}(\hat{\mathbf{f}})$ is the standard error for the estimator of the average treatment effect using the OGA approximation. Theorem 3.2 shows that the difference between these two standard errors is small for a given $\hat{n}$ if both terms T_1 and T_2 are small. For a given $\hat{n}$, T_1 gets smaller as $k_{\hat{n}}$ gets larger; however, T_2 does not decrease if $k_{\hat{n}}$ gets larger. In fact, it may get larger since the complexity of $\mathcal{G}_{k_{\hat{n}}}$ increases as $k_{\hat{n}}$ gets larger.

To understand the former term T_1 intuitively, suppose that $\|f\|_{\mathcal{L}_1^N} < \infty$ (Remark 3.6 discusses this condition) and that $p < k_{\hat{n}}$ (that is, some groups of covariates are not selected by the OGA algorithm). Consider, for example, the simple case in which, for a given sample size n, data collection on every covariate incurs the same costs (i.e., $\widetilde{c}(n)$) and each group consists of a single covariate. Then the total data collection costs are equal to the number of covariates selected multiplied by $\widetilde{c}(n)$ (i.e., $c(S, n) = \widetilde{c}(n) \sum_{j=1}^{M} S_j$). Assuming that $\widetilde{c}(n)$ is non-decreasing in n, we then have

$$\frac{1}{k_{\hat{n}}} = \frac{1}{\lfloor B/\widetilde{c}(\hat{n}) \rfloor},$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than x. This shows that for a given $\hat{n}$, as the budget B increases, a larger $k_{\hat{n}}$ will be chosen and the efficiency of the estimator of the average treatment effect will improve.

We now consider the latter term T_2, which is due to the difference between the pre-experimental estimation sample and the population in the experiment. This term will be small only if the following conditions are met: (i) the population for the pre-experimental sample and the population for the experiment need to have the same expectations for $(Y - g)^2$, (ii) the sample size N in the pre-experimental sample has to be large enough,

and (iii) the complexity of $\mathcal{G}_{k_{\hat{n}}}$ is not too large. In the main text of this paper, we have assumed that the pre-experimental and experimental samples are from the same population, N is large, and the budget B is small. In this scenario, the latter term is negligible compared to the former term.

Remark 3.5. Suppose that either (i) the pre-experimental sample size N is relatively small, or (ii) the budget B is large enough such that the possibility of overfitting is present. In this scenario, it might be desirable to solve a penalized version of (S.2) instead of solving (S.2). For example, in the k-th OGA step, one may solve

$$\min_{n\in\mathcal{N}} \frac{1}{n} \min_{\gamma\in\mathbb{R}^M} \left[\frac{1}{N}\sum_{i=1}^{N}(Y_i - \gamma' X_i)^2 + \kappa \frac{k\log N}{N} \right] \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B, \tag{S.3}$$

where $\kappa \geq 0$ is an extra tuning parameter that needs to be determined by the researcher. Theoretical properties of this penalized OGA estimator can be obtained using the arguments similar to those used in Barron, Cohen, Dahmen, and DeVore (2008) in conjunction with the use of the truncation operator.

Remark 3.6. The condition $\|f\|_{\mathcal{L}_1^N} < \infty$ is trivially satisfied when p is finite. In the case $p \to \infty$, the condition $\|f\|_{\mathcal{L}_1^N} < \infty$ requires that not all groups of covariates are equally important in the sense that the coefficients β_k, when their ℓ_2 norms are sorted in decreasing order, need to converge to zero fast enough to guarantee that $\sum_{k=1}^{\infty} |\beta_k|_2 < \infty$. If suitable laws of large numbers apply, we can also replace the condition $\|f\|_{\mathcal{L}_1^N} < \infty$ by its population counterpart.

S3.1 Proofs

Proof of Theorem 3.2: Write

$$\mathcal{R}_{\hat{n},\infty}(\hat{\mathbf{f}}, \mathbf{f_0}) = T_1 + T_2 + T_3,$$

where

$$T_1 := MSE_{\hat{n},\infty}(\hat{\mathbf{f}}) - \widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}),$$
$$T_2 := \widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}) - \widehat{MSE}_{\hat{n},N}(\mathbf{f_0}),$$

$$T_3 := \widehat{MSE}_{\hat{n},N}(\mathbf{f_0}) - MSE_{\hat{n},\infty}(\mathbf{f_0}).$$

Note that for each $k \geq 1$,

$$|T_1 + T_3| \leq \frac{2}{\hat{n}} \sup_{g \in \mathcal{G}_k} \left| \left\{ \|\mathbf{Y} - \mathbf{g}\|_N^2 - R(g) \right\} \right|.$$

Then the desired result follows immediately from Lemma 3.1, which is given below.

Lemma 3.1. Assume that $(\mathbf{X}'_{G_j}\mathbf{X}_{G_j})/N = \mathbf{I}_{|G_j|}$ for each $j = 1, \ldots, p$. Suppose $\mathcal{N}$ is a finite subset of $\mathbb{N}_+$, $c : \{0,1\}^M \times \mathbb{N}_+ \to \mathbb{R}$ some function, and $B > 0$ some constant. Then the following bound holds:

$$\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}}) - \widehat{MSE}_{\hat{n},N}(\mathbf{f_0}) \leq \frac{4\|f_0\|_{\mathcal{L}_1^N}^2}{\hat{n}} \left(\frac{1}{\min\{p, k_{\hat{n}}\}} \right). \tag{S.4}$$

Proof. This lemma can be proved by arguments similar to those used in the proof of Theorem 2.3 in Barron, Cohen, Dahmen, and DeVore (2008). The main difference between our Lemma 3.1 and Theorem 2.3 of Barron, Cohen, Dahmen, and DeVore (2008) is that we pay explicit attention to the group structure. In the subsequent arguments, we fix n and leave indexing by n implicit.

First, letting $\hat{\mathbf{r}}_{k-1,i}$ denote the ith component of $\hat{\mathbf{r}}_{k-1}$, we have

$$\begin{aligned}
\|\hat{\mathbf{r}}_{k-1}\|_N^2 &= N^{-1} \sum_{i=1}^N \hat{\mathbf{r}}_{k-1,i} Y_i \\
&= N^{-1} \sum_{i=1}^N \hat{\mathbf{r}}_{k-1,i} U_i + N^{-1} \sum_{i=1}^N \hat{\mathbf{r}}_{k-1,i} \sum_{j=1}^\infty X'_{G_j,i} \beta_j \\
&\leq \|\hat{\mathbf{r}}_{k-1}\|_N \Big\| \mathbf{Y} - \sum_{k=1}^\infty \mathbf{X}'_{G_k} \beta_k \Big\|_N + \Big[\sum_{j=1}^\infty |\beta_j|_2 \Big] N^{-1} |\hat{\mathbf{r}}'_{k-1} \mathbf{X}_{G_k}|_2 \\
&\leq \frac{1}{2} \Big(\|\hat{\mathbf{r}}_{k-1}\|_N^2 + \Big\| \mathbf{Y} - \sum_{k=1}^\infty \mathbf{X}'_{G_k} \beta_k \Big\|_N^2 \Big) + \Big[\sum_{j=1}^\infty |\beta_j|_2 \Big] N^{-1} |\hat{\mathbf{r}}'_{k-1} \mathbf{X}_{G_k}|_2,
\end{aligned}$$

which implies that

$$\|\hat{\mathbf{r}}_{k-1}\|_N^2 - \Big\|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2 \leq 2\Big[\sum_{j=1}^{\infty}|\beta_j|_2\Big]N^{-1}|\hat{\mathbf{r}}'_{k-1}\mathbf{X}_{G_k}|_2. \tag{S.5}$$

Note that if the left-hand side of (S.5) is negative for some $k = k_0$, then the conclusion of the theorem follows immediately for all $m \geq k_0 - 1$. Hence, we assume that the left-hand side of (S.5) is positive, implying that

$$\Big(\|\hat{\mathbf{r}}_{k-1}\|_N^2 - \Big\|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2\Big)^2 \leq 4\Big[\sum_{j=1}^{\infty}|\beta_j|_2\Big]^2 N^{-2}|\hat{\mathbf{r}}'_{k-1}\mathbf{X}_{G_k}|_2^2. \tag{S.6}$$

Let P_k denote the projection matrix $P_k := \mathbf{X}_{G_k}(\mathbf{X}'_{G_k}\mathbf{X}_{G_k})^{-1}\mathbf{X}'_{G_k} = N^{-1}\mathbf{X}_{G_k}\mathbf{X}'_{G_k}$, where the second equality comes from the assumption that $(\mathbf{X}'_{G_k}\mathbf{X}_{G_k})/N = \mathbf{I}_{|G_k|}$. Hence, it follows from the fact that P_k is the projection matrix that

$$\|\hat{\mathbf{r}}_{k-1} - P_k\hat{\mathbf{r}}_{k-1}\|_N^2 = \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \|P_k\hat{\mathbf{r}}_{k-1}\|_N^2. \tag{S.7}$$

Because $\hat{\mathbf{r}}_k$ is the best approximation to $\mathbf{Y}$ from $\hat{\mathcal{I}}_{n,k}$, we have

$$\|\hat{\mathbf{r}}_k\|_N^2 \leq \|\hat{\mathbf{r}}_{k-1} - P_k\hat{\mathbf{r}}_{k-1}\|_N^2. \tag{S.8}$$

Combining (S.8) with (S.7) and using the fact that $P_k^2 = P_k$, we have

$$\begin{aligned}
\|\hat{\mathbf{r}}_k\|_N^2 &\leq \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \|P_k\hat{\mathbf{r}}_{k-1}\|_N^2 \\
&= \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \|N^{-1}\mathbf{X}_{G_k}\mathbf{X}'_{G_k}\hat{\mathbf{r}}_{k-1}\|_N^2 \\
&= \|\hat{\mathbf{r}}_{k-1}\|_N^2 - N^{-2}|\hat{\mathbf{r}}'_{k-1}\mathbf{X}_{G_k}|_2^2,
\end{aligned} \tag{S.9}$$

Now, combining (S.9) and (S.6) together yields

$$\|\hat{\mathbf{r}}_k\|_N^2 \leq \|\hat{\mathbf{r}}_{k-1}\|_N^2 - \frac{1}{4}\Big(\|\hat{\mathbf{r}}_{k-1}\|_N^2 - \Big\|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2\Big)^2\Big[\sum_{j=1}^{\infty}|\beta_j|_2\Big]^{-2}. \tag{S.10}$$

As in the proof of Theorem 2.3 in Barron, Cohen, Dahmen, and DeVore (2008), let

$a_k := \|\hat{\mathbf{r}}_k\|_N^2 - \|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\|_N^2$. Then (S.10) can be rewritten as

$$a_k \leq a_{k-1}\Big(1 - \frac{a_{k-1}}{4}\Big[\sum_{j=1}^{\infty} |\beta_j|_2\Big]^{-2}\Big). \tag{S.11}$$

Then the induction method used in the proof of Theorem 2.1 in Barron, Cohen, Dahmen, and DeVore (2008) gives the desired result, provided that $a_1 \leq 4[\sum_{j=1}^{\infty} |\beta_j|_2]^2$. As discussed at the end of the proof of Theorem 2.3 in Barron, Cohen, Dahmen, and DeVore (2008), the initial condition is satisfied if $a_0 \leq 4[\sum_{j=1}^{\infty} |\beta_j|_2]^2$. If not, we have that $a_0 > 4[\sum_{j=1}^{\infty} |\beta_j|_2]^2$, which implies that $a_1 < 0$ by (S.11). Hence, in this case, we have that $\|\hat{\mathbf{r}}_1\|_N^2 \leq \|\mathbf{Y} - \sum_{k=1}^{\infty} \mathbf{X}'_{G_k}\beta_k\|_N^2$ for which there is nothing else to prove.

Then, we have proved that the error of the group OGA satisfies

$$\|\hat{\mathbf{r}}_m\|_N^2 \leq \Big\|\mathbf{Y} - \sum_{k=1}^{p} \mathbf{X}'_{G_k}\beta_k\Big\|_N^2 + \frac{4}{m}\Big[\sum_{j=1}^{p} |\beta_j|_2\Big]^2, \ m = 1, 2, \ldots.$$

Equivalently, we have, for any $n \in \mathcal{N}$ and any $k \geq 1$,

$$\|\mathbf{Y} - \hat{\mathbf{f}}_{n,k}\|_N^2 - \|\mathbf{Y} - \mathbf{f_0}\|_N^2 \leq \frac{4\|f_0\|_{\mathcal{L}_1^N}^2}{k}.$$

Because $\mathcal{N}$ is a finite set, the desired result immediately follows by substituting in the definition of $\hat{\mathbf{f}}$ and $k_{\hat{n}}$. Q.E.D.

S4 Cost Functions

In this appendix, we provide detailed descriptions of the cost functions used in Section V.

S4.1 Calibration of the Cost Function in Section V.A

Here, we give a detailed description of components of the cost function used in Section V.A.

- **Administration costs.** The administration costs in the survey were R$10,000 and the average survey took two hours per household to conduct (i.e., $T(S) = 120$

measured in minutes). Therefore,

$$c_{\text{admin}}(S, n) = \phi(120)^{\alpha} = 10{,}000.$$

If we assume that, say, $\alpha = 0.4$ (which means that the costs of 60 minutes are about 75.8 percent of the costs of 120 minutes), we obtain $\phi \approx 1{,}473$.

- **Training costs.** The training costs in the survey were R$25,000, that is,

$$c_{\text{train}}(S, n) = \kappa(1, 466) \cdot 120 = 25{,}000,$$

so that $\kappa(1, 466) \approx 208$. It is reasonable to assume that there exists some lumpiness in the training costs. For example, there could be some indivisibility in hotel rooms that are rented, and in the number of trainers required for each training session. To reflect this lumpiness, we assume that

$$\kappa(n) = \begin{cases} 150 & \text{if} \quad 0 < n \leq 1{,}400 \\ 208 & \text{if} \quad 1{,}400 < n \leq 3{,}000 \\ 250 & \text{if} \quad 3{,}000 < n \leq 4{,}500 \\ 300 & \text{if} \quad 4{,}500 < n \leq 6{,}000 \\ 350 & \text{if} \quad 6{,}000 < n \end{cases}.$$

Note that, in this specification, $\kappa(1{,}466) \approx 208$, as calculated above. We take this as a point of departure to calibrate $\kappa(n)$. Increases in sample size n are likely to translated into increases in the required number of field workers for the survey, which in turn lead to higher training costs. Our experience in the field (based on running surveys in different settings, and on looking at different budgets for different versions of this same survey) suggests that, in our example, there is some concavity in this cost function, because an increase in the sample size, in principle, will not require a proportional increase in the number of interviewers, and an increase in the number of interviewers will probably require a less than proportion increase in training costs. For example, we assume that a large increase in the size of the sample, from 1,500 to 6,000, leads to an increase in $\kappa(n)$ from 208 to 300 (i.e., an increase in overall training costs of about 50 percent).

- **Interview costs.** Interview costs were R$630,000, accounting for the majority of the total survey costs, that is,

$$c_{\text{interv}}(S, n) = 1{,}466 \cdot \eta + 1{,}466 \cdot p \cdot 120 = 630{,}000,$$

so that $\eta + 120p \approx 429.74$. The costs of traveling to each household in this survey were approximately half of the total costs of each interview. If we choose $\eta = 200$, then the fixed costs η amount to about 47 percent of the total interview costs, which is consistent with the actual costs of the survey. Then we obtain the price per unit of survey time as $p \approx 1.91$. It is also reasonable to assume that half of the variable costs per individual are due to the collection of the three outcomes in the survey, because their administration was quite lengthy. The costs of collecting the outcomes could also be seen as fixed costs (equal to $0.955 \times 120 = 114.6$), which means that the price per unit of survey time for each of the remaining covariates is about 0.955. In sum, we can rewrite interview costs as

$$c_{\text{interv}}(S, n) = 1{,}466 \times (200 + 114.6) + 1,466 \times 0.955 \times 120 = 630{,}000.$$

- **Price per covariate.** We treat the sample obtained from the original experiment as $\mathcal{S}_{\text{pre}}$, a pilot study or the first wave of a data collection process, based on which we want to decide which covariates and what sample size to collect in the next wave. We perform the selection procedure for each outcome variable separately, and thus adjust $T(S) = \tau(1 + \sum_{j=1}^{M} S_j)$. For simplicity, we assume that to ask each question on the questionnaire takes the same time, so that $\tau_0 = \tau_j = \tau$ for every question; therefore, $T(S) = \tau(1 + \sum_{j=1}^{M} S_j) = 120$. Note that we set $\tau_0 = \tau$ here, but the high costs of collecting the outcome variables are reflected in the specification of η above. This results in $\tau = 120/(1 + \sum_{j=1}^{M} S_j)$. The actual number of covariates collected in the experiment was 40; so $\sum_{j=1}^{M} S_j = 40$, and thus $\tau \approx 3$.

- **Rescaled budget.** Because we use only a subsample of the original experimental sample, we also scale down the original budget of R$665,000 down to R$569,074, which corresponds to the costs of selecting all 36 covariates in the subsample; that is, $c(\mathbf{1}, 1{,}330)$ where $\mathbf{1}$ is a 36-dimensional vector of ones and $c(S, n)$ is the calibrated cost function.

S4.2 Calibration of the Cost Function in Section V.B

Here, we present a detailed description of components of the cost function used in Section V.B.

- **Administration costs.** The administration costs for the low- and high-cost covariates were estimated to be about \$5,000 and \$24,000, respectively. The high-cost covariates were four tests that took about 15 minutes each (i.e., $T_{\text{high}}(S) = 60$). For the low-cost covariates (teacher and principal survey), the total survey time was around 60 minutes, so $T_{\text{low}}(S) = 60$. High- and low-cost variables were collected by two different sets of enumerators, with different levels of training and skills. Therefore,

$$\phi_{\text{low}}(60)^{\alpha_{\text{low}}} = 5{,}000 \qquad \text{and} \qquad \phi_{\text{high}}(60)^{\alpha_{\text{high}}} = 24{,}000.$$

  If we assume that, say, $\alpha_{\text{low}} = \alpha_{\text{high}} = 0.7$, we obtain $\phi_{\text{low}} \approx 285$ and $\phi_{\text{high}} \approx 1{,}366$.

- **Training costs.** μ_{high} and μ_{low} are the numbers of enumerators collecting high- and low-cost variables, respectively. The training costs for enumerators in the high and low groups increase by 20 for each set of additional 20 low-cost enumerators, and by 12 for each set of 4 high-cost enumerators:

$$\kappa_{\text{low}}(c, n_c) := 20 \sum_{k=1}^{19} k \cdot \mathbb{1}\{20(k-1) < \mu_{\text{low}}(c, n_c) \leq 20k\}$$

  and

$$\kappa_{\text{high}}(c, n_c) := 12 \sum_{k=1}^{17} k \cdot \mathbb{1}\{4(k-1) < \mu_{\text{high}}(c, n_c) \leq 4k\}.$$

  This is reasonable because enumerators for low-cost variables can be trained in large groups (i.e., groups of 20), while enumerators for high-cost variables need to be trained in small groups (i.e., groups of 4). However, training a larger group demands a larger room, and, in our experience, more time in the room. The lumpiness comes from the costs of hotel rooms and the time of the trainers. The numbers 20 and 12 as the average costs of each cluster of enumerators were chosen based on our experience with this survey (even if the design of the training and the organization

of the survey was not exactly the same as the stylized version presented here), and reflect both the time of the trainer and the costs of hotel rooms for each type of enumerators. Because the low-cost variables are questionnaires administered to principals and teachers, in principle the number of required enumerators only depends on c (i.e., $\mu_{\text{low}}(c, n_c) = \lfloor \lambda_{\text{low}} c \rfloor$). High-cost variables are collected from students, and therefore the number of required enumerators should depend on c and n_c, so $\mu_{\text{high}}(c, n_c) = \lfloor \lambda_{\text{high}} c \mu_{n,\text{high}}(n_c) \rfloor$. We assume that the latter increases again in steps, in this case of 10 individuals per cluster, that is,

$$\mu_{n,\text{high}}(n_c) := \sum_{k=1}^{7} k \cdot \mathbb{1}\{10(k-1) < n_c \leq 10k\}.$$

We let $\lambda_{\text{low}} = 0.14$ (capturing the idea that one interviewer could do about seven schools) and $\lambda_{\text{high}} = 0.019$ (capturing the idea that one enumerator could perhaps work with about 50 children). The training costs in the survey were \$1,600 for the low-cost group of covariates and \$1,600 for the high-cost group of covariates.

- **Interview costs.** We estimate that interview costs in the survey were \$150,000 and \$10,000 for the high- and low-cost variables, respectively, i.e.

$$\psi_{\text{low}}(350)\eta_{\text{low}} + 350 \cdot p_{\text{low}} \cdot 60 = 10{,}000$$

and

$$\psi_{\text{high}}(350, 24)\eta_{\text{high}} + 350 \cdot 24 p_{\text{high}} \cdot 60 = 150{,}000.$$

We set $\psi_{\text{low}}(c) = \mu_{\text{low}}(c)$ and $\psi_{\text{high}}(c, n_c) = \mu_{\text{high}}(c, n_c)$, the number of required enumerators for the two groups, so that η_{low} and η_{high} can be interpreted as fixed costs per enumerator. From the specification of $\mu_{\text{low}}(c)$ and $\mu_{\text{high}}(c, n_c)$ above, we obtain $\mu_{\text{low}}(350) = 50$ and $\mu_{\text{high}}(350, 24) = 20$. The fixed costs in the survey were about $\psi_{\text{low}}(350)\eta_{\text{low}} = 500$ and $\psi_{\text{high}}(350, 24)\eta_{\text{high}} = 1{,}000$ for low- and high-cost covariates. Therefore, $\eta_{\text{low}} = 500/50 = 10$ and $\eta_{\text{high}} = 1{,}000/20 = 50$. Finally, we can solve for the prices $p_{\text{low}} = (10{,}000 - 500)/(350 \times 60) \approx 0.45$ and $p_{\text{high}} = (150{,}000 - 1{,}000)/(350 \times 24 \times 60) \approx 0.3$.

- **Price per covariate.** For simplicity, we assume that to ask each low-cost ques-

tion takes the same time, so that $\tau_j = \tau_{\text{low}}$ for every low-cost question (i.e., $j = 1, \ldots, M_{\text{low}}$), and that each high-cost question takes the same time (i.e., $\tau_j = \tau_{\text{high}}$) for all $j = M_{\text{low}} + 1, \ldots, M$. The experimental budget contains funding for the collection of one outcome variable, the high-cost test results at follow-up, and three high-cost covariates at baseline. We modify $T_{\text{high}}(S)$ accordingly: $T_{\text{high}}(S) = \tau_{\text{high}}(1+\sum_{j=M_{\text{low}}+1}^{M} S_j) = 4\tau_{\text{high}}$ so that $\tau_{\text{high}} = 60/4 = 15$. Similarly, originally there were 255 low-cost covariates, which leads to $\tau_{\text{low}} = 120/255 \approx 0.47$.

- **Rescaled budget.** As in the previous subsection, we use only a subsample of the original experimental sample. Therefore, we scale down the original budget to the amount that corresponds to the costs of collecting all covariates used in the subsample. As a consequence, the rescaled budget is \$25,338 in the case of baseline outcomes and \$33,281 in the case of the follow-up outcomes.

S5 A Simple Formulation of the Problem

S5.1 Uniform Cost per Covariate

Take the following simple example where: (1) all covariates are orthogonal to each other; (2) all covariates have the same price, and the budget constraint is just $B = nk$, where n is sample size and k is the number of covariates. Order the covariates by the contribution to the MSE, so that the problem is to choose the first k covariates (and the corresponding n).

Define $\sigma^2(k) = (1/N)\sum_{i=1}^{N}(Y_i - \gamma_{0,k}' X_i)^2$, where $\gamma_{0,k}$ is the same as the vector of true coefficients γ_0 except that all coefficients after the $(k+1)$th coefficient are set to be zeros, and let $MSE(k, n) = (1/n)\sigma^2(k)$. For the convenience of using simple calculus, suppose that k is continuous, ignoring that k is a positive integer, and that $\sigma^2(k)$ is twice continuously differentiable. This would be a reasonable first-order approximation when there are a large number of covariates, which is our set-up in the paper. Because we ordered the covariates by the magnitude of their contribution to a reduction in the MSE, we have $\partial\sigma^2(k)/\partial k < 0$, and $\partial^2\sigma^2(k)/\partial k^2 > 0$.

The problem we solve in this case is just

$$\min_{n,k} \frac{1}{n}\sigma^2(k) \quad \text{s.t.} \quad nk \leq B.$$

Assume we have an interior solution and that n is also continuous. Replace the budget constraint in the objective function and we obtain

$$\min_{n,k} \frac{k}{B}\sigma^2(k).$$

This means that k is determined by

$$\sigma^2(k) + k\frac{\partial\sigma^2(k)}{\partial k} = 0,$$

or

$$\frac{\sigma^2(k)}{k} + \frac{\partial\sigma^2(k)}{\partial k} = 0, \tag{S.12}$$

which in this particular case does not depend on B. Then, n is given by the budget constraint (i.e., $n = B/k$).

Another way to see where this condition comes from is just to start from the budget constraint. If we want to always satisfy it then, starting from a particular choice of n and k yields

$$n \cdot dk + k \cdot dn = 0,$$

or

$$\frac{dn}{dk} = -\frac{n}{k}.$$

Now, suppose we want to see what happens when k increases by a small amount. In that case, keeping n fixed, the objective function falls by

$$\frac{1}{n}\frac{\partial\sigma^2(k)}{\partial k}dk.$$

This is the marginal benefit of increasing k. However, n cannot stay fixed, and needs to decrease by $(n/k)dk$ to keep the budget constraint satisfied. This means that the objective function will increase by

$$\left(-\frac{1}{n^2}\right)\sigma^2(k)\left(-\frac{n}{k}\right)dk.$$

This is the marginal cost of increasing k.

At the optimum, in an interior solution, marginal costs and marginal benefits need to

balance out, so

$$\frac{1}{nk}\sigma^2(k)\,dk = -\frac{1}{n}\frac{\partial\sigma^2(k)}{\partial k}dk$$

or

$$\frac{\sigma^2(k)}{k} + \frac{\partial\sigma^2(k)}{\partial k} = 0,$$

which reproduces (S.12).

There are a few things to notice in this simple example.

(1) The marginal costs of an increase in k are increasing in $\sigma^2(k)$. This is because increases in n are more important role for the MSE when $\sigma^2(k)$ is large than when it is small.

(2) The marginal costs of an increase in k are decreasing in k. This is because when k is large, adding an additional covariate does not cost much in terms of reductions in n.

(3) A large n affects the costs and benefits of increasing k in similar way. Having a large n reduces benefits of additional covariates because it dilutes the decrease in $\sigma^2(k)$. Then, on one hand, it increases costs through the budget constraint, as a larger reduction in n is needed to compensate for the same change in k. However, on the other hand, it reduces costs, because when n is large, a particular reduction in n makes much less difference for the MSE than in the case where n is small.

(4) We can rewrite this condition as

$$\frac{1}{k} + \frac{\partial\sigma^2(k)/\partial k}{\sigma^2(k)} = 0,$$

where the term $(\partial\sigma^2(k)/\partial k)/\sigma^2(k)$ is the percentage change in the unexplained variance from an increase in k.

If we combine

$$\frac{dn}{n} = \frac{dk}{k},$$

which comes from the budget constraint, and

$$\frac{1}{MSE(n,k)}\frac{\partial MSE(n,k)}{\partial n} = -\frac{1}{n},$$

we notice that the percentage decrease in MSE from an increase in n is just $(dn)/n$, the percentage change in n, which in turn is just equal to $(dk)/k$. So what the condition above says is that we want to equate the percentage change in the unexplained variance from a change in k to the percentage change in the MSE from the corresponding change in n.

Perhaps even more interesting is to notice that k is the survey cost per individual in this very simple example. Then this condition says that we want to choose k to equate the percentage change in the survey costs per individual $((dk)/k)$ to the percentage change in the residual variance

$$\frac{\partial \sigma^2(k)/\partial k}{\sigma^2(k)} dk.$$

This condition explicitly links the impacts of k on the survey costs and on the reduction in the MSE.

Adding fixed costs F of visiting each individual is both useful and easy in this very simple framework. Suppose there are a fixed costs F of going to each individual, so the budget constraint is $n(F+k) = B$. Proceeding as above, we can rewrite our problem as

$$\min_{n,k} \frac{F+k}{B} \sigma^2(k).$$

This means that k is determined by

$$\sigma^2(k) + (F+k) \frac{\partial \sigma^2(k)}{\partial k} = 0,$$

or

$$\frac{1}{F+k} + \frac{\partial \sigma^2(k)/\partial k}{\sigma^2(k)} = 0.$$

Note that, when there are large fixed costs of visiting each individual, increasing k is not going to be that costly at the margin. It makes it much easier to pick a positive k. However, other than that, the main lessons (1)–(4) of this simple model remain unchanged.

S5.2 Variable Cost per Covariate

If covariates do not have uniform costs, then the problem is much more complicated. Consider again a simple set-up where all the regressors are orthogonal, and we order them by their contribution to the MSE. However, suppose that the magnitude of each covariate's contribution the MSE takes a discrete finite number of values. Let $\mathcal{R}$ denote the set of these discrete values. Let r denote an element of $\mathcal{R}$ and $R = |\mathcal{R}|$ (the total number of all elements in $\mathcal{R}$). There are many potential covariates within each r group, each with a different price p. The support of p could be different for each r. So, within each r, we will then order variables by p. The problem will be to determine the optimal k for each r group. Let $\mathbf{k} \equiv \{k_r : r \in \mathcal{R}\}$.

The problem is

$$\min_{n,\mathbf{k}} \frac{1}{n}\sigma^2(\mathbf{k}) \quad \text{s.t.} \sum_{r\in\mathcal{R}} c_r(k_r) \leq B,$$

where $c_r(k_r) = \sum_{l=1}^{k_r} p_l$ are the costs of variables of type r used in the survey. We can also write it as $c_r(k_r) = p_r(k_r)\, k_r$, where $p_r(k_r) = (\sum_{l=1}^{k_r} p_l)/k_r$. Because we order the variables by price (from low to high), $\partial p_r(k_r)/\partial k_r > 0$. Let $\sigma_r^2 = \partial\sigma^2(\mathbf{k})/\partial k_r$, which is a constant (this is what defines a group of variables).

Then, assume we can approximate $p_l(k_r)$ by a continuous function and that we have an interior solution. Then, substituting the budget constraint in the objective function:

$$\min_{n,\mathbf{k}} \frac{1}{B}\left[\sum_{r\in\mathcal{R}} c_r(k_r)\right]\sigma^2(\mathbf{k}).$$

From the first-order condition for k_r,

$$\frac{\partial c_r(k_r)}{\partial k_r}\sigma^2(\mathbf{k}) + \left[\sum_{r\in\mathcal{R}} c_r(k_r)\right]\frac{\partial\sigma^2(\mathbf{k})}{\partial k_r} = 0,$$

or

$$\frac{\partial c_r(k_r)/\partial k_r}{\sum_{r\in\mathcal{R}} c_r(k_r)} = -\frac{\partial\sigma^2(\mathbf{k})/\partial k_r}{\sigma^2(\mathbf{k})}.$$

What this says is that, for each r, we choose variables up to the point where the percent marginal contribution of the additional variable to the residual variance equals the percent marginal contribution of the additional variable to the costs per interview, just as in the

previous subsection.

S6 Simulations

In this appendix, we study the finite sample behavior of our proposed data collection procedure, and compare its performance to other variable selection methods. We consider the linear model from above, $Y = \gamma' X + \varepsilon$, and mimic the data-generating process in the day-care application of Section V.A with the cognitive test outcome variable.

First, we use the dataset to regress Y on X. Call the regression coefficients $\hat{\gamma}_{\text{emp}}$ and the residual variance $\hat{\sigma}^2_{\text{emp}}$. Then, we regress Y on the treatment indicator to estimate the treatment effect $\hat{\beta}_{\text{emp}} = 0.18656$. We use these three estimates to generate Monte Carlo samples as follows. For the pre-experimental data $\mathcal{S}_{\text{pre}}$, we resample X from the empirical distribution of the $M = 36$ covariates in the dataset and generate outcome variables by $Y = \gamma' X + \varepsilon$, where $\varepsilon \sim N(0, \hat{\sigma}^2_{\text{emp}})$ and

$$\gamma = \hat{\gamma}_{\text{emp}} + \frac{1}{2}\text{sign}(\hat{\gamma}_{\text{emp}})\kappa\bar{\gamma}.$$

We vary the scaling parameter $\kappa \in \{0, 0.3, 0.7, 1\}$ and $\bar{\gamma} := (\bar{\gamma}_1, \ldots, \bar{\gamma}_{36})'$ is specified in three different fashions, as follows:

- "lin-sparse", where the first five coefficients linearly decrease from 3 to 1, and all others are zero, that is,

$$\bar{\gamma}_k := \begin{cases} 3 - 2(k-1)/5, & 1 \leq k \leq 5 \\ 0, & \text{otherwise} \end{cases};$$

- "lin-exp", where the first five coefficients linearly decrease from 3 to 1, and the remaining decay exponentially, that is,

$$\bar{\gamma}_k := \begin{cases} 3 - 2(k-1)/5, & 1 \leq k \leq 5 \\ e^{-k}, & k > 5 \end{cases};$$

- "exp", where exponential decay $\bar{\gamma}_k := 10e^{-k}$.

When $\kappa = 0$, the regression coefficients γ are equal to those in the empirical application. When $\kappa > 0$, we add one of the three specifications of $\bar{\gamma}$ to the coefficients found in the dataset, thereby increasing (in absolute value) the first few coefficients[1] more than the others, and thus increasing the importance of the corresponding regressors for prediction of the outcome. Figure 1 displays the regression coefficients in the dataset (i.e., when $\kappa = 0$, denoted by the blue line labeled "data"), and γ for the three different specifications when $\kappa = 0.3$.

[1]Because all estimated coefficients in the dataset ($\hat{\gamma}_{\text{emp}}$) are close to zero and roughly of the same magnitude, we simply pick the first five covariates that have the highest correlation with the outcome variable.

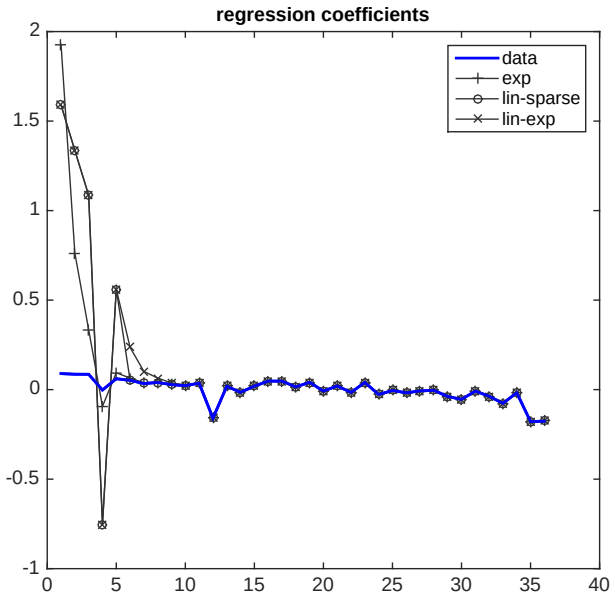

Figure 1: Regression coefficients in the simulation when $\kappa = 0.3$

Table 1: Simulation results: lin-sparse

Scale	Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	$\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$	$bias(\hat{\beta})$	$sd(\hat{\beta})$	RMSE($\hat{\beta}$)	EQB
0	Experiment	1,330	36	1	0.02498	−0.0034284	0.049981	0.050048	\$56,9074
	OGA	2,508	1.4	0.99543	0.019249	0.00034598	0.038838	0.038801	\$34,8586
	LASSO	2,587	0.1	0.99278	0.019418	0.0020874	0.039372	0.039388	\$35,6781
	POST-LASSO	2,529	1.0	0.99457	0.019222	0.00069394	0.037758	0.037727	\$35,1659
0.3	Experiment	1,330	36	1	0.02494	0.00036992	0.049501	0.049453	\$56,9074
	OGA	2,350	3.9	0.99443	0.019751	0.0013905	0.038275	0.038262	\$37,7076
	LASSO	2,228	5.9	0.988	0.020346	−0.00093089	0.041713	0.041682	\$39,3017
	POST-LASSO	2,320	4.4	0.99321	0.019696	−0.0020751	0.038746	0.038763	\$37,1730
0.7	Experiment	1,330	36	1	0.024953	−0.00086563	0.050992	0.050948	\$56,9074
	OGA	2,346	4.0	0.99433	0.020552	−0.0020151	0.041475	0.041483	\$39,7722
	LASSO	2,218	6.1	0.98747	0.021145	0.00057516	0.043957	0.043917	\$42,3971
	POST-LASSO	2,246	5.7	0.98929	0.020177	0.0019693	0.042683	0.042686	\$38,7095
1	Experiment	1,330	36	1	0.024938	−0.0021535	0.051146	0.05114	\$56,9074
	OGA	2,346	4.0	0.99433	0.021566	0.00044162	0.043389	0.043348	\$43,8536
	LASSO	2,172	6.9	0.98513	0.021383	−0.00058106	0.045378	0.045336	\$43,1589
	POST-LASSO	2,172	6.9	0.98513	0.019956	−0.0048726	0.040967	0.041215	\$38,0053

Table 2: Simulation results: lin-exp

Scale	Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	$\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$	$bias(\hat{\beta})$	$sd(\hat{\beta})$	RMSE($\hat{\beta}$)	EQB
0	Experiment	1,330	36	1	0.024965	0.0027033	0.051564	0.051583	\$569,074
	OGA	2,509	1.3	0.99541	0.019249	−0.00042961	0.03723	0.037195	\$348,682
	LASSO	2,588	0.1	0.99275	0.01941	−0.003374	0.03845	0.03856	\$357,261
	POST-LASSO	2,530	1.0	0.9946	0.019215	0.00076956	0.037924	0.037894	\$351,755
0.3	Experiment	1,330	36	1	0.02492	−0.0015645	0.049457	0.049432	\$569,074
	OGA	2,343	4.0	0.99421	0.019868	−0.0014349	0.040197	0.040182	\$379,540
	LASSO	2,186	6.7	0.98569	0.020652	0.0019377	0.04084	0.040845	\$403,004
	POST-LASSO	2,313	4.5	0.99288	0.019816	−0.0025812	0.039587	0.039631	\$377,876
0.7	Experiment	1,330	36	1	0.024936	0.0041527	0.050436	0.050556	\$569,074
	OGA	2,301	4.7	0.99247	0.020805	−0.0017267	0.041303	0.041297	\$408,990
	LASSO	2,134	7.7	0.98551	0.02162	−0.00071182	0.042716	0.042679	\$440,232
	POST-LASSO	2,206	6.5	0.98955	0.020522	0.0013055	0.043358	0.043334	\$400,219
1	Experiment	1,330	36	1	0.024964	−0.0034064	0.049484	0.049551	\$569,074
	OGA	2,286	5.0	0.99187	0.021874	−0.0025106	0.042304	0.042336	\$451,756
	LASSO	2,080	9.0	0.98793	0.021987	−0.0015746	0.044218	0.044201	\$454,765
	POST-LASSO	2,078	9.0	0.98787	0.020374	0.00077488	0.041977	0.041942	\$396,218

Table 3: Simulation results: exp

Scale	Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	$\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$	$bias(\hat{\beta})$	$sd(\hat{\beta})$	RMSE($\hat{\beta}$)	EQB
0	Experiment	1,330	36	1	0.024953	0.00083077	0.054043	0.053996	\$569,074
	OGA	2,511	1.3	0.99538	0.019234	0.0016616	0.037237	0.037236	\$348,426
	LASSO	2,588	0.1	0.99278	0.019394	−0.00049328	0.038849	0.038813	\$356,941
	POST-LASSO	2,529	1.0	0.99452	0.019203	−0.00044404	0.039549	0.039512	\$351,403
0.3	Experiment	1,330	36	1	0.024947	−0.00089522	0.051246	0.051202	\$569,074
	OGA	2,411	2.9	0.99605	0.019426	0.0016951	0.038729	0.038727	\$359,950
	LASSO	2,291	4.9	0.9911	0.020184	−0.0022094	0.040243	0.040263	\$389,560
	POST-LASSO	2,380	3.5	0.99514	0.019377	0.0014552	0.039996	0.039982	\$359,662
0.7	Experiment	1,330	36	1	0.024946	−0.0012694	0.050947	0.050912	\$569,074
	OGA	2,408	3.0	0.99605	0.019457	0.0015399	0.040789	0.040778	\$362,287
	LASSO	2,279	5.1	0.99039	0.020233	0.0011166	0.042491	0.042463	\$391,128
	POST-LASSO	2,376	3.5	0.99515	0.019405	−0.0023208	0.037252	0.037287	\$361,903
1	Experiment	1,330	36	1	0.024948	−0.0034014	0.051898	0.051957	\$569,074
	OGA	2,407	3.0	0.99603	0.019494	0.0022031	0.038846	0.038869	\$364,015
	LASSO	2,271	5.2	0.99008	0.020298	0.0016393	0.039024	0.039019	\$392,857
	POST-LASSO	2,377	3.5	0.99516	0.019448	−0.00085645	0.039135	0.039106	\$363,023

For each Monte Carlo sample from $\mathcal{S}_{\text{pre}}$, we apply the OGA, LASSO, and POST-LASSO methods, as explained in Section V.A. The cost function and budget are specified exactly as in the empirical application. We store the sample size and covariate selection produced by each of the three procedures, and then mimic the randomized experiment $\mathcal{S}_{\text{exp}}$ by first drawing a new sample of X from the same data-generating process as in $\mathcal{S}_{\text{pre}}$. Then we generate random treatment indicators D, so that outcomes are determined by

$$Y = \hat{\beta}_{\text{emp}} D + \gamma' X + \varepsilon,$$

where ε is randomly drawn from $N(0, \hat{\sigma}^2_{\text{emp}})$. We then compute the treatment effect estimator $\hat{\beta}$ of β by regressing Y_i on $(1, D_i, Z_i)$ using the generated experimental sample.[2]

The results are based on 500 Monte Carlo samples, N = 1,330, which is the sample size in the dataset, and $\mathcal{N}$ a fine grid from 500 to 4,000. All covariates, those in the dataset as well as the simulated ones, are studentized so that their variance is equal to one.

For the different specifications of $\bar{\gamma}$, Tables 1–3 report the selected sample size ($\hat{n}$), the selected number of covariates ($|\hat{I}|$), the ratio of costs for that selection divided by the budget B, the square root of the estimated MSE, $\sqrt{\widehat{MSE}_{\hat{n},N}(\hat{\mathbf{f}})}$, the bias and standard

[2]The method used here is not exactly the same as the method described in Step 4 of Section III. However, the difference would be minimal in the Monte Carlo experiments.

deviation of the estimated average treatment effect ($bias(\hat{\beta})$ and $sd(\hat{\beta})$), and the RMSE of $\hat{\beta}$ across the Monte Carlo samples of the experiment.

Overall, all three methods perform similarly well across different designs and the number of selected covariates tends to increase as κ becomes large. No single method dominates other methods, although POST-LASSO seems to perform slightly better than LASSO. In view of the Monte Carlo results, we argue that the empirical findings reported in Section V.A are likely to result from the lack of highly predictive covariates in the empirical example.

S7 Variables Selected in the School Grants Example

Table 4: School grants (outcome: math test): selected covariates in panel (a) of Table 7

OGA	LASSO	POST-LASSO
Child is male	Child is male	Child is male
Village pop.	Dist. to Dakar	Dist. to Dakar
Piped water	Dist. to city	Dist. to city
Teach-stud	Village pop.	Village pop.
No. computers	Piped water	Piped water
Req. (h) teach. qual.	No. computers	No. computers
Req. (h) teach. att.	Req. (h) teach-stud	Req. (h) teach-stud
Obs. (h) manuals	Hrs. tutoring	Hrs. tutoring
Books acq. last yr.	Books acq. last yr.	Books acq. last yr.
Any parent transfer	Provis. struct.	Provis. struct.
Teacher bacc. plus	NGO cash cont.	NGO cash cont.
Teach. train. math	Any parent transfer	Any parent transfer
Obst. (t) class size	NGO promised cash	NGO promised cash
Measure. equip.	Avg. teach. exp.	Avg. teach. exp.
	Teacher bacc. plus	Teacher bacc. plus
	Obs. (t) student will.	Obs. (t) student will.
	Obst. (t) class size	Obst. (t) class size
	Silence kids	Silence kids

Table 5: Definition of variables in Table 4

Variable	Definition
Child is male	Male student
Village pop.	Size of the population in the village
Piped water	School has access to piped water
Teach–stud	Teacher–student ratio in the school
No. computers	Number of computers in the school
Req. (h) teach. qual.	Principal believes teacher quality is a major requirement for school success
Req. (h) teach. att.	Principal believes teacher attendance is a major requirement for school success
Obs. (h) manuals	Principal believes the lack of manuals is a major obstacle to school success
Books acq. last yr.	Number of manuals acquired last year
Any parent transfer	Cash contributions from parents
Teacher bacc. plus	Teacher has at least a baccalaureate degree
Teach. train. math	Teacher received special training in math
Obst. (t) class size	Teacher believes class size is a major obstacle to school success
Measure. equip.	There is measurement equipment in the classroom
Dist. to Dakar	Distance to Dakar
Dist. to city	Distance to the nearest city
Req. (h) teach–stud	Principal believe teacher–student ratio is a major requirement for school success
Hrs. tutoring	Hours of tutoring by teachers
Provis. struct.	Number of provisional structures in school
NGO cash cont.	Cash contributions by NGO
NGO promised cash	Promised cash contributions by NGO
Avg. teach. exp.	Average experience of teachers in the school
Obst. (t) student will.	Teacher believes the lack of student willpower is one of the main obstacles to learning in the school
Obst. (t) class size	Teacher believes the lack of classroom size is one of the main obstacles to learning in the school
Silence kids	Teacher has to silence kids frequently

S8 Out-of-Sample Evaluations

In the empirical applications, we performed the covariate selection procedure as well as its evaluation (by RMSE and EQB) on the same pre-experimental sample. In this section, we study the sensitivity of our findings when the covariate selection and evaluation steps are performed on two separate samples.

We partition each of the datasets into five subsamples of equal size. Four of the five subsamples are merged to form the training set on which we perform the covariate selection procedure, and the remaining fifth subsample serves as evaluation sample on which we calculate the performance measures RMSE and EQB. Given the partition into five subsamples, there are five possible ways to combine them into training and evaluation samples. We perform the covariate selection on each of these five training samples using the same calibrated cost functions as in the main text, but adjusting the budget for the drop in sample size by letting the budget be the cost function $c(S, n)$ evaluated at the experimental selection $S = (1, \ldots, 1)'$ and n the length of the training sample. The output of the procedure consists of five sample size selections $\hat{n}$, five covariate selections, i.e. five values of $|\hat{I}|$, and five cost-to-budget ratios. Tables 6–8 show the averages of $\hat{n}$, $|\hat{I}|$, and "Cost/B" over those five different training samples. The RMSE is calculated using the estimate of γ from the training sample and data on Y and X from the evaluation sample. Similarly, the EQB is the budget necessary to achieve the RMSE on the evaluation sample equal to that of the experiment when the covariate selection procedures are applied to the training sample. Both RMSE and EQB are then averaged over the five subsamples.

Overall, the results of this out-of-sample evaluation exercise are similar to those reported in the full-sample analysis of the main text. Qualitatively, in both applications, the covariate selection procedures recommend larger sample sizes than the experiment. The recommended sample size may differ somewhat from those reported in the main text because the budget and training sample size is different, but the orders of magnitude are the same. In the school grants application, we notice that the recommended number of covariates selected tends to be smaller than in the full-sample evaluation of the main text, but if anything the covariate selection procedures manage to achieve even lower relative equivalent budgets and lower RMSE than the experiment.

Table 6: Day-care (outcome: cognitive test), 5-fold out-of-sample evaluation

method	$\hat{n}$	$\lvert\hat{I}\rvert$	cost/B	RMSE	EQB	relative EQB
experiment	1,330	36	1	0.029068	R$460,809.54	1
OGA	2,209	0.8	0.99503	0.020694	R$235,654.21	0.511
LASSO	2,260	0	0.99392	0.020777	R$237,425.38	0.515
POST-LASSO	2,146	1.8	0.99464	0.020647	R$234,494.95	0.509

Table 7: Day-care (outcome: health assessment), 5-fold out-of-sample evaluation

method	$\hat{n}$	$\lvert\hat{I}\rvert$	cost/B	RMSE	EQB	relative EQB
experiment	1,330	36	1	0.029313	R$460,809.54	1
OGA	2,221	0.6	0.99495	0.020708	R$232,066.95	0.504
LASSO	2,260	0	0.99392	0.020787	R$233,751.11	0.507
POST-LASSO	2,158	1.6	0.9949	0.020644	R$231,224.87	0.502

Table 8: School grants (outcome: math test), 5-fold out-of-sample evaluation

Method	$\hat{n}$	$\lvert\hat{I}\rvert$	Cost/B	RMSE	EQB	Relative EQB
			(a) Baseline outcome			
experiment	1,824	142	1	0.0082721	$27,523.74	1
OGA	2,618.4	1.2	0.99823	0.0044229	$16,609.53	0.603
LASSO	2,658	0	0.9991	0.004445	$16,621.60	0.604
POST-LASSO	2,638.2	1.2	0.99927	0.0044291	$16,651.77	0.605
			(b) Follow-up outcome			
experiment	609	143	1	0.0098756	$48,856.20	1
OGA	6,132	0	0.99885	0.0028432	$14,664.85	0.300
LASSO	6,132	0	0.99885	0.0028432	$14,664.85	0.300
POST-LASSO	6,132	0	0.99885	0.0028432	$14,664.85	0.300
			(c) Follow-up outcome, no high-cost covariates			
experiment	609	143	1	0.0098756	$48,856.20	1
OGA	6,092.8	0.8	0.99893	0.0027807	$14,571.10	0.298
LASSO	6,000.8	5.4	0.99905	0.002795	$14,651.75	0.300
POST-LASSO	6,040.4	2.4	0.99887	0.0027623	$14,532.12	0.297
			(d) Follow-up outcome, force baseline outcome			
experiment	609	143	1	0.0098756	$48,856.20	1
OGA	2,035.2	2.4	0.90783	0.0041647	$24,918.89	0.510
LASSO	2,494	1	0.99623	0.0046439	$25,522.72	0.522
POST-LASSO	2,494	1	0.99623	0.0034893	$23,651.72	0.484

S9 The Case of Multivariate Outcomes

In this section, we consider an extension to the case of multivariate outcomes. If data on a particular regressor is collected, then the regressor is automatically available for regressions involving any of the outcomes. Therefore, it is natural to select one common set of regressors for all outcomes. Hence, our regression problem corresponds to the special case of seemingly unrelated regressions (SUR) such that the vector of regressors is identical for each equation. In this case, it is well known that the OLS and GLS estimators are algebraically identical. In other words, there is no loss of efficiency in using the single-equation OLS estimator even if regression errors are correlated.

Suppose there are L outcome variables of interest, say $\{Y_{\ell,i} : \ell = 1, \ldots, L, \text{ and } i = 1, \ldots, N\}$. Then a multivariate analog of (3.8) can be written as

$$\min_{n \in \mathbb{N}_+, \boldsymbol{\gamma} = (\gamma_1', \ldots, \gamma_L')' \in \mathbb{R}^{ML}} \frac{1}{nNL} \sum_{\ell=1}^{L} \sum_{i=1}^{N} (Y_{\ell,i} - \gamma_\ell' X_i)^2 \qquad \text{s.t.} \qquad c(\mathcal{I}(\gamma), n) \leq B. \tag{S.13}$$

In other words, the stacked version of the OLS problem is equivalent to regressing $\boldsymbol{y} := (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_L')'$ on $I_L \otimes \boldsymbol{X}$ conditional on the budget constraint, where $\boldsymbol{y}_\ell = (Y_{\ell,1}, \ldots, Y_{\ell,L})'$, I_L is the L-dimensional identity matrix, and $\boldsymbol{X}$ is $N \times M$ dimensional matrix whose ith row is X_i'. Therefore, the OGA applies to this case as well with minor modifications. First, we need to redefine the outcome vector and the design matrix with the stacked $\boldsymbol{y}$ and the enlarged design matrix $I_L \otimes \boldsymbol{X}$. Suppose that a variable selection problem is on individual components of X_i. Then note that because of the nature of the stacked regressions, we need to apply a group OGA with each group consisting of L columns of $[I_L \otimes \boldsymbol{X}]_k$, where $k = (\ell - 1)M + m$ $(\ell = 1, \ldots, L)$ for each $m = 1, \ldots, M$.

S10 Counterfactual Increases of the Predictive Power of Covariates

In the second empirical application, we increase the correlation of the baseline outcome with the follow-up outcome as follows:

- First, we run a regression of follow-up outcome Y_{2i} on baseline outcome Y_{1i}, yielding regression coefficient $\hat{\rho}$ and residual $\hat{e}_i$.

- We then increase the predictive power of the baseline outcome by multiplying $\hat{\rho}$ by a factor w_1 and the residual by w_2 to define a new follow-up outcome

$$\widetilde{Y}_{2i} = w_1 \hat{\rho} Y_{1i} + w_2 \hat{e}_i.$$

- Then the variance of the original follow-up outcome can be decomposed into an "explained" and "unexplained" part as

$$Var(Y_{2i}) = \hat{\rho}^2 Var(Y_{1i}) + Var(\hat{e}_i),$$

and similarly for the new follow-up outcome

$$Var(\widetilde{Y}_{2i}) = w_1^2 \hat{\rho}^2 Var(Y_{1i}) + w_2^2 Var(\hat{e}_i).$$

- We choose w_2 so that the two outcomes have the same variance ($Var(Y_{2i}) = Var(\widetilde{Y}_{2i})$), i.e.

$$w_2 = \sqrt{\frac{(1 - w_1^2)\hat{\rho}^2 Var(Y_{1i})}{Var(\hat{e}_i)} + 1}.$$

- In panel (c) of Table 8, we set $w_1 = 1.2$, and in panel (d) of Table 8, $w_1 = 1.3$.

References

Barron, A. R., A. Cohen, W. Dahmen, and R. A. DeVore (2008): "Approximation and Learning by Greedy Algorithms," *The Annals of Statistics*, 36(1), 64–94.

Davis, G., S. Mallat, and M. Avellaneda (1997): "Adaptive greedy approximations," *Constructive Approximation*, 13(1), 57–98.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, ed. by T. P. Schultz, and J. A. Strauss, vol. 4, chap. 61, pp. 3895–3962. Elsevier.

Huang, J., T. Zhang, and D. Metaxas (2011): "Learning with Structured Sparsity," *Journal of Machine Learning Research*, 12, 3371–3412.

Ing, C.-K., and T. L. Lai (2011): "A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models," *Statistica Sinica*, 21(4), 1473–1513.

McConnell, B., and M. Vera-Hernandez (2015): "Going Beyond Simple Sample Size Calculations: a Practioner's Guide," Discussion paper.

Natarajan, B. (1995): "Sparse Approximate Solutions to Linear Systems," *SIAM Journal on Computing*, 24(2), 227–234.

Sancetta, A. (2016): "Greedy algorithms for prediction," *Bernoulli*, 22(2), 1227–1277.

Temlyakov, V. N. (2011): *Greedy Approximation*. Cambridge University Press, Cambridge.

Tropp, J. A. (2004): "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, 50(10), 2231–2242.

Tropp, J. A., and A. C. Gilbert (2007): "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, 53(12), 4655–4666.

Zhang, T. (2009): "On the Consistency of Feature Selection using Greedy Least Squares Regression.," *Journal of Machine Learning Research*, 10(3), 555 – 568.