Using Post-Double Selection Lasso in Field Experiments *

Jacobus Cilliers, *Georgetown University*

Nour Elashmawy, *Development Research Group, World Bank*

David McKenzie, *Development Research Group, World Bank*

July 17, 2024

Abstract

The post-double selection Lasso estimator has become a popular way of selecting control variables when analyzing randomized experiments. This is done to try to improve precision, and reduce bias from attrition or chance imbalances. We re-estimate 780 treatment effects from published papers to examine how much difference this approach makes in practice. We find it reduces standard errors by less than one percent compared to standard Ancova on average and does not select variables to model treatment in over half the cases. We discuss and provide evidence on the key practical decisions researchers face in using this method.

Keywords: Treatment effect; Randomized Experiment; Post-Double Selection Lasso; Attrition; Statistical Power.

JEL classification codes: C93; C21; O12.

*We thank authors of the different papers we replicated for answering queries on their code; Christian Hansen, Kaspar Wuthrich, Carolina Lopez, Anja Sautmann, Greg Lane, Erin Kelley, and participants in the World Bank half-baked seminar for useful comments. Elashmawy was funded by the Robert S. McNamara Fellowships Program.

1 Introduction

In a simple randomized experiment, regressing the outcome of interest on an indicator for treatment will give the difference-in-means estimator, which provides an unbiased estimate of the average impact of being assigned to treatment. However, although this is true in expectation, in any given random draw, the means of the treatment and control groups can differ from one another for many covariates, with the probability that these differences are large falling with sample size. One approach to improve balance and estimation precision has been through ex-ante design choices, such as the use of stratification or pairwise matching (Bai, 2022; Bruhn and McKenzie, 2009). A complimentary approach is ex-post adjustment of the difference in means, by controlling for different covariates in a regression. For example, adding the lagged outcome as a covariate in the regression for Ancova estimation can greatly improve power when the outcome is highly autocorrelated (McKenzie, 2012). However, once one moves beyond controlling for randomization strata and the lagged dependent variable to including other baseline variables as controls, the question arises of how many other covariates should be considered, and how they should be selected. Often the concern is raised that such covariate adjustments are ad hoc, involve substantial researcher degrees of freedom, and can give rise to p-hacking (Simmons et al., 2011).

The post-double selection lasso (PDS Lasso) estimator of Belloni et al. (2014b) has rapidly grown in popularity as a principled way of selecting control variables in field experiments. At first this may seem surprising, since PDS Lasso was originally developed for causal inference in observational studies. The method selects controls by using the least absolute shrinkage and selection operator (Lasso) twice: once to select covariates that help predict the outcome of interest, and once to select covariates that help predict treatment status. It then takes the union of these two sets as controls in the treatment regression. But if the goal is just to improve the efficiency of treatment estimation, then it is unclear why modeling treatment status is necessary, and why field experiment researchers do not instead use modern machine-learning approaches that just select covariates based on their ability to predict the outcome (Bloniarz et al., 2016; Guo et al., 2021; List et al., 2022; Wager et al., 2016; Wu and Gagnon-Bartsch, 2018).

This paper examines why and how PDS Lasso is being used and should be used in randomized field experiments. Two features that help distinguish field experiments from the large A/B online platform experiments studied in Guo et al. (2021) and List et al. (2022) are sample size and attrition. Field experiment researchers often work with relatively small sample sizes, often in the range of 100 to 1,000 observations, where adding sample size is expensive and statistical power a key concern. Large chance imbalances by treatment status may arise in small samples, leading to a desire to adjust differences in means for these baseline differences. Moreover, small sample sizes increase the importance of being able to improve precision

by adding covariates that help reduce the variance in the outcome of interest. But smaller samples limit the ability to benefit from non-linear machine learning approaches, and also raises questions about how well asymptotic results used to justify the choice of regularization parameters in Lasso work with the sample sizes common in field experiment applications. Wüthrich and Zhu (2023) recently showed that PDS Lasso can underselect variables in finite samples when the number of variables is reasonably large, potentially leading to omitted variable bias when these variables have moderate relationships with treatment. Second, the majority of field experiments in developing countries rely on survey data for many outcomes, which can be subject to attrition. In a survey of 96 field experiments, Ghanem et al. (2023) find an average attrition rate of 15 percent. This raises the concern that if the determinants of attrition differ with treatment status, then the sample for which data are available may no longer be comparable across treatment groups. This offers a further potential rationale for double selection methods, which can select covariates that predict treatment status in the attrited data. But the question remains as to whether this makes much difference in practice.

We replicate and re-analyze field experiment papers published in three economic journals between 2017 and 2022 which used PDS Lasso, resulting in 780 treatment estimates. We use this analysis to identify the key practical issues and performance of PDS Lasso, and compare the estimates and their standard errors to those that would be obtained using simple Ancova. We find that while authors typically include a long list of variables to be inputted as potential controls in PDS Lasso (a median of 182 controls), PDS Lasso typically ends up selecting very few control variables. The median is three controls, and in over half the cases, no variables at all are selected in the treatment regression step. When variables are selected for treatment, they are almost never those that are also selected for predicting the outcome of interest. As a result, PDS Lasso leads to minimal changes to treatment estimates and standard errors on average, with a median change in the coefficient of 0.01 standard deviations, and a median standard error that is 99.2% of that with Ancova. In over a quarter of the cases, standard errors are actually slightly larger than they would be just using Ancova. Researchers should therefore not anticipate significant power gains on average from using this method.

We combine this re-analysis with simulations to look at when PDS Lasso performs better or worse, and to help answer practical questions facing applied researchers. In our re-analysis, we do find the treatment regression step to be more likely to select control variables when there is attrition, but even then, typical changes in coefficients are small, reflecting that attrition in field experiments is often due to reasons uncorrelated with the outcome. We show that PDS Lasso sometimes ends up being less precise than Ancova (and having a higher mean-squared error in some simulations), since by inputting very many control variables, there is a risk that the Lasso penalization results in key variables such as the lagged dependent variable not being selected. We recommend included such a variable in the amelioration set that must be included.

We examine whether performance improves by using a different penalty parameter such as one selected by cross-validation. Although it gives slightly smaller standard errors on average, we find that it can overfit, sometimes resulting in substantially larger standard errors. Given these limited average gains, and risk of poor performance, using the standard plug-in penalty seems preferable. We suggest that researchers need to be considerably more judicious in their choice of control variables to input into this procedure, especially with the small samples typical in practice. We then conclude by discussing applications of this method with multiple outcomes, multiple treatments, and treatment interactions. We find two common mistakes arise in applied work, concerning how missing variables in potential controls are handled, and how the interacting variable is entered, and provide recommendations to avoid these errors. We conclude with a checklist for applying this approach in practice.

2 The Post-Double Selection Lasso Method

We summarize the most common existing approaches used to estimate treatment effects in field experiments (difference-in-means and Ancova), and then compare to the post-double selection lasso method of Belloni et al. (2014b).

Consider the following partial linear model for outcome y, for observations i=1,2,...,n:

$$y_i = \alpha + \gamma T_i + g(z_i) + \epsilon_i \tag{1}$$

Where T_i is a dummy variable which takes value one if unit i was assigned to treatment, and zero otherwise; z_i are a set of control variables, and ϵ_i is an unobservable that satisfies $E(\epsilon_i|T_i, z_i) = 0$.

2.1 Difference-in-means estimator

With pure random assignment, we have $E(T_i g(z_i)) = 0$, and so we can obtain an unbiased estimate of the average impact of being assigned to treatment, γ, through a simple difference-in-means equation:

$$y_i = \alpha + \gamma T_i + \omega_i \tag{2}$$

The variance of this difference-in-means estimator will then depend on the variance of the residual term ω_i.

With more complicated random assignment designs, treatment assignment may depend on control variables that are used to define randomization strata or matched pairs. Then one can add controls for these

strata or pairs to equation 2.

2.2 Ancova

While the difference-in-means estimator is unbiased, efficiency can be improved through the inclusion of control variables that help explain the outcome of interest y. One such variable that takes a special place in applied work is the baseline value of the outcome of interest, y_0. Approximating $g(z_i)$ with y_0 gives the Ancova estimator:

$$y_i = \alpha + \gamma T_i + \delta y_{i0} + \lambda' S_i + \upsilon_i \tag{3}$$

Where the S_i are a set of dummy variables for any randomization strata used. McKenzie (2012) shows this estimator improves power over the difference-in-means estimator, with the gain in power larger the more autocorrelated the outcome of interest is. This basic Ancova specification has become the default specification for many randomized field experiments. In addition to improving power, it adjusts the estimate for any baseline differences in the key outcome of interest that may have arisen from chance imbalances in the random draw, or from attrition, with the amount of adjustment data-driven and depending on how much this baseline difference predicts the future outcome.

In many applications we expect the baseline value of the outcome to have the strongest predictive power for future outcome values out of any baseline covariates. However, baseline measurement of the outcome may not always be available, or in some cases may be the same for everyone in the sample. For example, an experiment on young job-seekers may have employment and income as the main outcomes of interest, but at baseline everyone may be unemployed and earning zero income.

2.3 PDS Lasso in Theory

The baseline outcome is only one of many potential control variables that researchers could use. Belloni et al. (2014b) consider the problem of selecting controls from a set of p potential regressors $x_i = P(z_i)$, which can consist of z_i and different transformations and interactions of z_i that aim to approximate the function $g(z_i)$. They note that it is possible that $p > n$, that is, that the number of potential controls is high-dimensional and may even exceed the number of observations in the dataset. There are three reasons this could occur in a randomized field experiment. First, it is common for baseline surveys or application forms for programs to collect data on many characteristics of individuals or firms applying for a program. This may be augmented further by community or geographical characteristics. Second, since the functional form $g(.)$ is unknown, one may wish to consider interactions, polynomials, or other non-linear transformations of the baseline variables.

Third, as List et al. (2022) note, in some settings there may be many rounds of pre-treatment data, and then the question arises of how best to control for multiple lags of the variables.

Since their focus is on causal inference in observational studies, Belloni et al. (2014b) also use a partially linear model to model treatment:

$$T_i = m(z_i) + v_i \tag{4}$$

Where $E(v_i|z_i) = 0$.

The functions $g(.)$ and $m(.)$ are unknown. The key assumption Belloni et al. (2014b) make is that these models are approximately sparse, which means that there are linear approximations $x_i'\beta_{g0}$ to $g(z_i)$ and $x_i'\beta_{m0}$ to $m(z_i)$ that require only a small number of non-zero coefficients to approximate these functions up to a small approximation error. The vector x_i here includes a constant, as well as all the covariates to be considered.

The PDS Lasso method then follows a three-step procedure:

Step 1: Select control variables that predict the outcome y through a Lasso regression of y on the covariates x, not including an indicator for treatment:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_j x_{i,j} + \beta_p x_{i,p} + \varepsilon_i \tag{5}$$

The Lasso estimator solves the following equation:

$$\min_{\beta} \left\{ E\left[(y_i - x_i'\beta)^2\right] + \frac{\lambda}{n} \|\Psi\beta\| \right\} \tag{6}$$

Where Ψ is a diagonal matrix of penalty loadings,[1] and the key Lasso tuning parameter is given by λ. λ determines the penalty applied to selecting additional covariates: a small value of λ does not penalize additional covariates very much, and will result in more controls being selected, while a large value of λ penalizes each additional control a lot, and will result in fewer controls being selected. We discuss how to choose λ in section 2.4.

Let I_1 denote the set of control variables chosen in this step.

Step 2: Also select control variables that predict the treatment T through a Lasso regression of T on the covariates x, using the same λ as in Step 1:

$$T = \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \cdots + \alpha_j x_{i,j} + \alpha_p x_{i,p} + \varepsilon_i \tag{7}$$

[1]The penalty loadings incorporate the standard deviations of the regressors, which is equivalent to standardizing the regressors to have unit variance. The penalty parameters developed by Belloni et al. (2012) can also adjust account for heteroscedasticity and clustering.

Let I_2 denote the set of control variables chosen in this step.

Step 3: Then the PDS Lasso estimator of the treatment effect comes from regressing y on the treatment indicator, along with the union of the set of controls selected in the first two steps. That is from the union of I_1 and I_2. Belloni et al. (2014b) note that researchers may also wish to force the model to include additional variables from a set I_3, that he or she has theoretical reasons for wishing to include for robustness, even if they are not selected in either of the first two steps. They call this third set of variables the amelioration set I_3. The full set of controls selected is then the union of I_1, I_2 and I_3, which is denoted I. The treatment effect is estimated by the following least squares regression:

$$y_i = \alpha + \gamma T_i + x_i'\beta + \varepsilon_i \quad \text{with} \quad \beta_j = 0, \forall j \notin I \tag{8}$$

Inference then uses the standard heteroscedasticity robust standard errors from this linear regression. Belloni et al. (2014b) show that the resulting estimator allows for imperfect selection of the controls (does not assume that the approximation is exact), and provides confidence intervals that are valid uniformly across a large class of models, and achieves the semi-parametric efficiency bound under homoscedasticity.

The most common way of implementing this estimation in practice is through code developed by Ahrens et al. (2018), who developed the Stata program *pdslasso*. Stata versions 17 and up also include the built-in command *dsregress* which also implements double-selection Lasso. The two packages will give slightly different results when using their defaults for two reasons. First, the standard errors reported in *pdslasso* are slightly smaller because they do not adjust for the degrees of freedom from the selected controls.[2] Second, the default tuning parameter λ also differs.[3]

It is worth noting the similarities and differences with a common, but criticized, approach of deciding which control variables to include in a randomized experiment on the basis of a balance test. Here researchers do a t-test for difference in means, and then include as controls variables which are found to have a significant difference in means between the treatment and control groups. Permutt (1990) notes that this pre-testing can affect the size of subsequent tests, and be worse for power than simply randomly choosing covariates to control for. He instead argues for controlling for the covariates most highly correlated with the outcome. Step 2 of the PDS Lasso method is akin to a penalized balance test, while Step 1 does aim for find the variables most highly predictive of the outcome, irrespective of how balanced they are.

[2]To make these comparable, one needs to multiple the standard errors reported in *pdslasso* by $\sqrt{\frac{\frac{G}{G-1}}{\frac{N}{N-k-1}}}$, where G is the number of clusters, N the number of variables, and k the number of variables included in the regression.

[3]Both use a default plug-in λ described by equation (9), except that *dsregress* uses a version that allows for heteroskedastic errors.

2.4 How should the penalty parameter λ be chosen?

Belloni et al. (2012) provide a theoretically motivated choice of the tuning parameter by deriving an asymptotically optimal "plug-in" estimator for λ which sets:

$$\lambda = 2c\sqrt{n}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \tag{9}$$

where $\Phi^{-1}(.)$ is the inverse standard normal distribution, and the authors propose setting the constant $c = 1.1$, and $\alpha = 0.1/log(\max(p, n))$.[4]

Remark 1: we see from equation 9 that the larger is p, the number of covariates being considered, the larger is λ. That is, the penalty increases with the number of potential covariates. λ also increases with the sample size n, but λ/n in equation 6 will decrease with n, so that the larger the sample, the lower the penalty Lasso will place on adding more covariates.

Remark 2: with clustered randomized experiments, the effective amount of independent information in the data is reduced, and so more regularization is needed. The command *pdslasso* uses the cluster-lasso with default $\alpha = 0.1/log(n_{clusters})$ in this case, where $n_{clusters}$ is the number of clusters. All else equal, this means fewer controls will get selected in a clustered experiment than in an unclustered experiment of the same sample size.

Although this formula for λ ensures optimal asymptotic properties, it is less clear how well it performs in small samples. Wüthrich and Zhu (2023) provide simulation evidence and finite sample theory to show that a finite sample omitted variables bias can arise from the Lasso underselecting variables that have moderate explanatory power in cases where p is relatively large, and they suggest examining robustness of results to a 50% change in the plug-in λ. Note that the issue of omitted variable bias may be less of a concern in most experiments, and instead underselection may just result in less of the residual variance in the outcome being soaked up by selected regressors than is optimal.

An alternative is to choose the tuning parameter by cross-validation. Wager et al. (2016) suggest using cross-validation for single post-Lasso estimation in randomized experiments. Belloni et al. (2014a) note that while cross-validation is commonly used when the goal is prediction, it may not result in good performance when the goal is instead causal inference. Another practical issue with the use of cross-validation is that it can introduce further researcher discretion and complicate pre-specified replicability: the choice of cross-validated λ, and thus which controls are chosen, can depend on how many folds are used to split the data, the choice of random seed, and the stopping criteria used for selecting the optimal λ within a grid search.

[4] In practice, *pdslasso* appears to use $0.1/log(n)$, whereas *dsregress* uses $\alpha = 0.1/log(\max(p, n))$.

2.5 Why might researchers want to use double- rather than single-selection with an RCT?

PDS Lasso was developed for causal inference in observational studies, where the assumption is that treatment is only exogenous after conditioning on a set of controls. Equation 7 helps select which controls to condition treatment on. But if treatment is randomly assigned, then in expectation treatment is orthogonal to the controls. This raises the question of why we need double-selection, rather than just using Lasso (or other machine-learning approaches) for single-selection by choosing which covariates are most predictive of the outcome.

There are two main reasons why researchers might worry that treatment assignment is still correlated with baseline variables in a randomized field experiment. The first is the possibility of chance imbalances from random assignment with relatively small sample sizes. The second is attrition causing imbalances. However, even then, our intuition is that if either of these leads treatment to be correlated with a variable x_H, we would expect that this should not cause omitted variables bias unless x_H is also is correlated with y and so would be selected in equation 5 anyway. For example, suppose in a job training program we find that treatment assignment is correlated with hair color, but that hair color is unrelated to employment outcomes. Then a double-selection step which leads us to also add hair color to the regression could end up inflating standard errors and reducing precision by conditioning on what is effectively noise.

However, while it is true that we need not correct for variables which have no correlation with the error term in equation 1, Belloni et al. (2014a) explain why this logic can break down with regularization. The Lasso in step 1 will tend to select x variables that have large coefficients, but because of regularization, not select variables that have moderate associations with y. However, if these variables that are moderately associated with y are strong predictors of treatment status, then not including them can still lead to substantial omitted variables bias. The second step regression provides a second chance to select such variables, providing additional robustness against chance imbalances and selective attrition.

3 Empirical approach to examining how PDS Lasso performs in practice

How does PDS Lasso perform in practice when applied to the types of settings common in field experiments? We use two complimentary approaches to answer this question and to provide guidance on a number of practical and implementation questions which arise when using this method. The first is to replicate and re-examine a set of published papers which have used this method. The second is to use simulations.

3.1 Replications and Re-analysis of Published PDS Lasso papers

Our replication and re-analysis focuses on field experiment papers published from 2017 to 2022 in the following three journals: the American Economic Journal: Applied Economics, the American Economic Review, and the Journal of Development Economics. We chose these journals because the AEA journals have had data replication packages in place for this whole time, and because field experiments are commonly published in all three journals. Out of the 90 field experiment papers identified, 19 papers (22 percent) use the post-double selection approach to select their control variables. The remainder of the papers primarily used one or more of these other approaches such as difference-in-means, Ancova, and/or selection of covariates on an *ad hoc* basis or based on a balance test.[5]. A replication package was not available for one of these papers, giving a total of 18 papers for our analysis.

Table 1 provides some descriptive statistics of the selected papers. The median paper estimates treatment effects using PDS Lasso on 36 different outcomes, with the number of outcomes ranging from 3 to 189.[6] These include an extremely broad array of outcomes including employment, income, consumption, mental health, physical health, agriculture, environment, business, aspirations, soft skills, education, voting, livestock, and many other outcomes. They therefore encompass many of the key outcomes that appear in field experiments, and cover outcomes which we might expect to differ substantially in their ability to be predicted by baseline variables. In addition, 56% of the papers involve multiple treatments. This yields a total of 780 treatment estimates across the papers.

The median sample size is 1,913, and ranges from 418 to 13,987 observations. However, 15 of the 18 papers are clustered randomizations, with a median of 206 clusters and a minimum of 14 clusters. These sample sizes are further reduced for some outcomes by attrition. While the median attrition rate is 10 percent, with median differential attrition of 2 percent, there can be large variation in attrition across outcomes within the same study. This can occur due to item non-response, to some outcomes only be available conditional on another outcome (e.g. income may only be measured for those who work), or to a survey or administrative data source only being present for a subset of observations. As a consequence, one-quarter of the treatment estimates involve at least 24% attrition, and at least 7% differential attrition, and attrition rates exceed 50 percent in some cases. This provides us with many treatment estimates in which there may be concerns about potential imbalances between treatment and control due to small samples or attrition.

Authors of these studies tend to include a long list of variables to be inputted as potential control variables in PDS Lasso. The mean is 405 controls, and median is 182 controls. Notably these are both larger than the

[5]In particular, zero of the papers used the double-debiased machine learning approach of Chernozhukov et al. (2018), nor did we find papers using other machine learning approaches

[6]We include only the outcomes in which the authors use a PDS Lasso specification: some papers only do this for a subset of their outcomes as a robustness test.

median number of clusters in the study. The number of controls also shows substantial variation, ranging from six to 3,340 variables. We also look at two ratios: first, the ratio of inputted controls to the sample size of the experiment, which ranges from 0.003 to 1.73 with a median equal to 0.16. It is worth noting that the ratio exceeds 1 in fewer than 1 percent of the cases. Second, the ratio of inputted controls to the number of clusters which shows massive disparity with a minimum of 0.04 and a maximum of 17.58. We find that the mean is 2.27 and the median is 0.86.

For each of these 18 papers and 780 treatment effects, we re-estimate the treatment effect using three different specifications: (i) using PDS Lasso, with the control variable input set specified by the authors, using the *pdslasso* command and its default plug-in estimator for λ, but then adjusting the standard errors for a degrees of freedom correction; (ii) using Ancova, where we control for the lagged dependent variable (if available) and strata fixed effects; and (iii) using PDS Lasso, but with cross-validation to select the tuning parameter λ.[7] This enables us to examine how authors use PDS Lasso with experiments in practice, and what additional complications may arise that are not emphasized in theoretical work. We see how frequently PDS Lasso selects variables at each step of the estimation process, and how many variables typically are selected. We then compare the PDS Lasso and Ancova estimates to see how much of a change in standard errors (precision) is achieved by using this method, and how much of a change in the estimated treatment effect coefficients occurs. We then see how much these differences vary with features of the studies, such as the attrition rate.

3.2 Simulations

We supplement our re-analysis of real field experiments with additional evidence from simulations. In analyzing real experiments, we do not know the true treatment effect, and so while we can examine how much the coefficients and standard errors change, are not able to directly estimate bias or mean-squared error. The simulations also enable us to vary some design features such as the sample size, number of controls inputted, and degree of correlation between covariates and the outcome.

In our simulations, we use the following data generating process:

$$y_i = \gamma T_i + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i, \tag{10}$$

where $x_{i,j} \overset{\text{iid}}{\sim} N(0_p, I_p)$, $\epsilon_i \overset{\text{iid}}{\sim} (0,1)$, and each observation has an equal probability of being randomly assigned to treatment, T_i. Each new draw of the data randomly varies $x_{i,j}$, ϵ_i, and T_i. The object of interest is $\hat{\gamma}$, the coefficient on treatment, which is set equal to 0.5 in expectation. Standard errors are adjusted to account

[7] 10 folds, selecting the λ that gives the minimum of the CV-function.

for degrees of freedom. In all the simulations we allow for one variable, $x_{i,1}$, to be moderately or highly correlation with y_i. This variable can be thought of as the lagged dependent variable, which the researcher expects *ex ante* to be correlated with y_i, and would include in an Ancova estimation. In addition, four variables have a low level of correlation with y_i: $\beta_j = 0.05$ for $2 \leq j \leq 5$. The remaining x_i variables in the model are only correlated with the dependent variable by chance, and should not be included when estimating the treatment effect. This DGP thus satisfies the exact sparsity assumption, since only a few number of the coefficients are non-zero—a condition under which the PDS Lasso method is expected to perform well.

In Section 4 we compare the performance of PDS Lasso with the ANCOVA specification, defined as:

$$y_i = \gamma T_i + \beta_1 x_{i,1} + \epsilon_i, \tag{11}$$

Note that this is not the oracle estimator, since it does not include $x_{i,2}$ to $x_{i,5}$. The PDS lasso can thus out-perform a model using researcher-determined control variables, but it might also under-perform, if $x_{i,1}$ is not chosen. We then consider the performance of the PDS lasso under different features of the data generating process, such as the number of observations, n, and correlation between the regressors and the dependent variable, β_1. In Section 5 we vary different design decisions when implementing the PDS method, such as the number of potential regressors to include, p, and how to calculate the penalty parameter, λ.

4 How much difference does PDS Lasso make in practice?

PDS Lasso will improve on the base specification of Ancova and strata controls in a field experiment setting if the outcome of interest can be predicted well by other baseline covariates (improving precision), or if there are imbalances in treatment assignment related to controls that also are correlated with the outcome (reducing bias). It is unclear in the typical practical application how much of a gain this provides, and therefore how much researchers should plan on benefiting from the use of PDS Lasso when designing studies and doing power calculations. We therefore examine how the method operates in practice, using both our replication analysis and simulations.

4.1 How often are variables selected for Y and T in field experiments?

Recall that the field experiments in our sample included a mean (median) of 405 (182) potential control variables for PDS Lasso to select from. Figure 1 and Table 1 summarize how frequently the method selects variables from these input sets that help predict the outcome (step 1) and/or treatment status (step 2).

We see, despite the large number of potential control variables inputted, that PDS Lasso typically ends up selecting very few control variables: the mean is 3.62 and median is 2 variables, and the 75th percentile is only 5 variables. When we consider the two steps, we see that at least one variable is selected in the outcome regression (step 1) in 71% of cases, but the mean is for only 2.74 variables, and median for only 1 variable, to be selected. In over half (57%) of the cases, no variables are selected at all in the treatment regression (step 2), and when variables are selected, it is typically only one or two variables. Moreover, there is almost no overlap at all between the variables selected in step 1 to predict the outcome, and those selected in step 2 to predict the treatment. One interpretation is that any imbalances that occur in field experiments due to attrition or unlucky draws tend to be for idiosyncratic reasons, and do not result in selection into treatment on the basis of baseline variables that are strongly associated with future outcomes. Table A1 shows the results are not being driven by one or two papers with many outcomes, but are similar if we reweight the data at the paper level.

Given these results, we might expect to see PDS Lasso result in quite modest changes in precision and in coefficient estimates compared to Ancova, since in practice the method is not selecting many variables to control for, and the second step is not selecting variables that are very strongly correlated with the outcome. We examine this next.

4.2 How much does PDS Lasso change standard errors and treatment estimates relative to Ancova?

The bottom panel of Table 2 compares the estimated treatment effects and standard errors from PDS Lasso to those using Ancova[8]. Since we have a wide variety of outcomes measured in different units, for each outcome, we calculate the absolute change in the treatment coefficient, relative to the control group standard deviation. We see very minimal changes in treatment estimates on average, with a mean of 0.05 S.D. and a median of 0.01 S.D. The 75th percentile is 0.04 S.D. The upper tail is larger, with a 95th percentile of 0.15 S.D., but this and the maximum can reflect some outcomes where the control group standard deviation is close to zero, so that even a large standard deviation increase can be small in absolute terms.

To compare standard errors, we construct a ratio by dividing the standard error using PDS Lasso by the standard error using Ancova. A value smaller than one than indicates a smaller standard error, and hence an improvement in statistical power, from using PDS Lasso. The *pdslasso* command can also result in smaller standard errors even when the same variables as Ancova are used, due to not adjusting for degrees

[8]Recall that we are using the term Ancova here to describe a regression which just controls for randomization strata and the lagged dependent variable if available, but for outcomes where no lagged dependent variable is available, the comparison is just to a regression with strata fixed effects.

of freedom. We therefore show this ratio with and without adjusting for degrees of freedom. Table 2 shows that the standard error ratio is typically very close to unity, with a median of 0.992. While on average PDS Lasso yields marginally smaller standard errors, in about 25 percent of cases it results in larger standard errors than Ancova.

Figure 2 plots the ratio of PDS Lasso to Ancova coefficients, to this ratio of their standard errors. We see the mass is concentrated around 1 on both axes, suggesting that the majority of the time the two methods will give similar results. We examine the change in the statistical significance of outcomes when using PDS Lasso instead of Ancova and vice versa. Table A2 shows that using PDS Lasso instead of Ancova results in 49 (6.3%) and 42 (5.4%) insignificant outcomes that become significant at the 5 and 10 percent levels, respectively. There are a smaller number of treatment effects that are significant using Ancova that become insignificant when using PDS Lasso (15 (2%) at the 5 percent level, and 16 (2%) at the 10 percent level).

A key reason applied researchers use PDS Lasso is the hope that it will enable them to improve statistical power through selecting additional variables to control for in their regressions. In pre-analysis plans and grant applications they may then write that they plan on using this method and expect it to yield a reduction in the minimal detectable effect (MDE) size, or improvement in power. However, these results suggest that such improvements will be incredibly modest, and not guaranteed, relative to just using Ancova. Since the MDE with 80 percent power is 2.8 times the standard error, the implied MDE with PDS Lasso will only be 0.9% smaller than with Ancova. This will likewise result in very small changes in the sample size required to detect a given effect size. As an example, a sample size of 786 (393 treatment, 393 control) is needed to have 80 percent power to detect a 20 percent increase in an outcome like profits or income, in which the mean and standard deviation are the same size. Using PDS Lasso reduces this by 14 units to 772 at the median ratio.

4.3 When does PDS Lasso make more of a difference?

Although PDS Lasso does not select control variables for treatment in the majority of our cases, we might expect this to be more common when there is attrition. Panels (a) and (b) of Figure 3 shows this is the case. Panel (a) shows that the treatment regression selects at least one control variable only 12% of the time when there is zero attrition, and more often when there is attrition. However, the relationship is non-monotonic, which could indicate that many cases of high overall attrition still reflect balance on baseline variables by treatment status. In panel (b) we see a monotonic relationship between the differential attrition rate and likelihood of selecting control variables in the treatment regression. With no differential attrition, this occurs only 11% of the time, compared to 36% of the time with less than 5 percent differential attrition, and 67%

of the time with more than 5 percent differential attrition.

Panels (c) and (d) of Figure 3 examine how much the estimated treatment coefficient changes using PDS Lasso compared to Ancova as the attrition and differential attrition rates vary. We see a significantly higher mean percentage change in the treatment coefficient when differential attrition is greater than 5 percent, but even so, the mean percent change is only 0.08 standard deviations.

We would also expect PDS Lasso to make more of a difference to standard errors over Ancova when there is no lagged dependent variable available, and PDS Lasso is then potentially able to predict the outcome with other controls. We compared the standard error ratio (PDS Lasso/Ancova) for cases where no lagged dependent variable was available, to those where it was available and also included in PDS Lasso. There is not much difference at the median (0.996 when there is a lagged variable compared to 0.99 with no lagged variable available). However, at the 25th percentile the ratio is smaller (meaning more improvement) at 0.938 with no lag, compared to 0.974 with a lagged dependent variable available.

Finally, using the replication data, we also looked in more detail at the outliers, where PDS Lasso had resulted in the biggest change in standard errors compared to Ancova. Qualitatively we see these arise from cases where there was no lagged dependent variable available, and where just a few[9] controls are chosen by PDS Lasso that help predict the outcome well. For example, the largest reductions in standard errors compared to PDS Lasso occurs in Garbiras-Diaz and Montenegro (2022). They conduct an experiment to look at the effects of encouraging citizen oversight in Colombia. No lagged outcome is available, and the standard error after using PDS Lasso is 50-60% of that using just strata fixed effects. However, almost all of this improvement is coming from the inclusion of a single control: the number of candidates in the election strongly predicts the vote share (a regression of the outcome on this single regressor in the control group has a R^2 of 0.17). Hence in many cases, the gains from PDS Lasso are coming from adding a couple of controls that researchers might otherwise have used contextual knowledge to add.

4.4 Why is Ancova sometimes more precise than PDS Lasso?

We see that in one quarter of the empirical cases, PDS Lasso resulted in larger standard errors than Ancova. There are two potential reasons for this. The first is that by including many potential control variables for PDS Lasso to select among, it may fail to select the lagged dependent variable, or indeed any variables at all. In 23% of the replications in which a lagged dependent variable was available, but was not included by researchers in the amelioration set, the lagged variable was not selected by PDS Lasso.[10] Second, it could end up adding noise and reducing precision by selecting variables that are irrelevant.

[9] Of the 25 largest reductions in standard error, PDS Lasso selected a median of 3 controls to add.
[10] The lagged dependent variable was not selected in 21 out of 93 such cases.

We examine this in simulations reported in in Figure 4. In these simulations We set $p = 20$, vary the autocorrelation of the lagged dependent variable $\beta_1 \in (0.2, 0.4)$, have four weakly relevant controls, $\beta_j = 0.05$, for $2 \leq j \leq 5$, and 15 completely irrelevant controls $\beta_j = 0$ for $j > 5$, and then vary n. Figure 4(a) then shows that PDS Lasso fails to select the lagged dependent variable in moderately sized samples that are typical of many field experiments. When the correlation between $x_{i,1}$ and y is relatively strong ($\beta_1 = 0.4$) then the variable always gets selected in samples of 300 or larger. But when the correlation is relatively weak ($\beta_1 = 0.2$), then the variable only gets selected with 100 percent probability after the sample reaches 1,000 observations or more, and gets selected less than half the time with samples of 300. This problem becomes even more acute in clustered randomized trials, especially if the intra-cluster correlation is high. [11]

In contrast, we see in Figure 4(b) that PDS Lasso rarely selects the completely irrelevant variables included in the control set, and only starts selecting the weakly relevant controls ($\beta_j = 0.05$) with samples sizes of $1,000$ or larger. So the main issue for precision is failure to select the most important controls in small and moderate sample sizes. As a result, Figure 4(c) shows that the standard errors are larger compared to the ANCOVA specification in small samples, especially if the correlation between the lagged dependent variable dependent variable is relatively low. However, it is in small samples where statistical power is typically of most concern, and adding control variables to improve precision is of most interest. One solution is then to include the lagged dependent variable in the amelioration set and partial it out before PDS Lasso then selects other controls. Figure 4(d) shows that when this is done, PDS Lasso does at least as well as Ancova at all but the smallest sample sizes (where it occasionally selects an irrelevant variable), and starts to be more precise for samples of 300 or more in this simulation.

5 Practical Issues in Implementation

There are several questions that arise in practice when implementing PDS Lasso with field experiments. These include the appropriate choice of penalty parameter, how many and which controls to include, what to place in the amelioration set, whether the double step is needed at all, and how to deal with multiple treatments and treatment interactions. We discuss the theory and evidence for how researchers should decide on these questions.

[11] For example, Figure A1 in the appendix shows that with 50 clusters per treatment arm, 20 observations per cluster, an intra-cluster correlation (ICC) of 0.4, $\hat{\beta}_1 = 0.3$, the variable has about a 80 percent chance of being selected.

5.1 Should researchers use a penalty parameter other than the plug-in default?

As discussed above, the "plug-in" penalty parameter behaves well asymptotically, but Wüthrich and Zhu (2023) show that it can underselect variables in finite samples when the number of control variables in reasonably large. The alternative is cross-validation. Chernozhukov et al. (2024) suggest that cross-validation instead raises the risk of overfitting, and can perform poorly in moderately sized samples as a result. We examine how cross-validation performs in practice using both our replication data and simulations.

Figure 5 shows simulation results, varying the number of observations between 100 and $10,000$.[12] Panels (a) and (b) show that the cross-validation method tends to select far more variables compared to the plug-in approach, for both steps 1 and 2. For example, when $n = 500$ the median number of variables selected that are predictive of the dependent variable using the plug-in penalty parameter is one, compared to five for cross-validation. Since the oracle estimator includes five control variables, this suggest that there might be some power advantage of using cross-validation to select the penalty parameter in smaller samples. However, panels (a) and (b) also show that cross-validation often selects too many variables. The method is also unstable, with a high variation in the number of variables selected. This could reduce precision, due to reasons discussed above. Overall, panel (c) shows that the median standard error tends to be smaller for cross-validation in smaller samples, and equal to the plug-in parameter in very large samples. However, the variation in standard errors is larger, especially in smaller samples. That is, using cross-validation risks overfitting.[13]

Table 3 examines the performance of cross-validation in real field experiments by re-estimating each treatment effect using cross-validation instead of the plug-in penalty parameter. Comparing the number of variables selected to those presented in Table 1 we confirm that cross-validation tends to select substantially more variables: a mean (median) of 26.96 (13) controls compared to 3.7 (2) controls with the plug-in penalty parameter. The risk of overfitting is seen with a maximum of 264 controls selected. The median standard error is marginally smaller with cross-validation, at 98.4% of the plug-in standard error. However, there are some extreme cases where the standard error is substantially larger. Overall, given that cross-validation does not offer clear improvements, lacks theoretical justification, and has this risk of overfitting, there does not seem to be a compelling case for applied researchers to use cross-validation instead of the plug-in penalty.

Table A2 shows the change in significance when using cross-validation as compared to Ancova and PDS Lasso. Compared to the plug-in penalty parameter, the cross-validation approach leads to just as many

[12] We further set $p = 20$, $\beta_1 = 0.3$, and $\beta_j = 0.05$ for $j \in (2, 5)$.

[13] Two other approaches to reduce the risk of over-fitting from cross-validation are the (i) adaptive Lasso (Zou, 2006), or (ii) setting the penalty parameter such that cross-validated error is one standard error larger than the minimum. They can be specified in the *dsregress* command using the "*selection(adaptive)*" and "*selection(cv, serule)*" options, respectively. Our simulations showed that the adaptive lasso performed similar to simple cross-validation, although tended to still over-select in small samples. The 1 SE rule selected fewer variables than using the plug-in penalty parameter.

losses in significance as gains. In contrast, using cross-validation instead of Ancova is much more likely to result in a gain of statistical significance than a loss: there are 31 (4%) of the treatment estimates which lose significance at the 5 percent level, and 59 (7.8%) which would gain significance by using cross-validation instead of Ancova.

5.2 How many and which covariates should researchers include in the control set?

The theoretical treatments of PDS Lasso take the vector of potential covariates x, and its dimension p, as given, and provide a method for choosing which subset of controls from this larger set should be included. However, in practice, researchers typically have considerable choice over which variables they consider as potential covariates. In order to reduce further the p-hacking concerns of Simmons et al. (2011), this set should ideally be pre-specified. But then researchers need guidance on how many and which covariates to choose. One approach is the "kitchen sink" approach, of choosing a very large number of baseline covariates, potentially with p even larger than n. This approach seems particularly important in observational studies, where the assumption of exogeneity of treatment conditional on observables may be deemed more plausible if a very large set of observables is considered. Applied researchers then sometimes see this as showing robustness with no downside. For example, in introducing the Lasso for inference, StataCorp gives an example with 104 covariates and writes "we do not worry about overfitting the model, because the control variables that we specify are potential control variables. Lasso will select the relevant ones".[14] This kitchen sink approach describes a lot of the papers we replicate, given a mean (median) of 405 (182) inputted controls.

However, as equation 9 shows, the penalty parameter λ does increase with p, and so by starting with a very large set of potential controls, it is possible that none get selected, whereas some might get selected if PDS Lasso was run on a smaller subset. Belloni et al. (2014a, p. 40) note that distinguishing true predictive power from spurious associations becomes more difficult as more variables get considered, and so advise selection over a collection of variables that is "not overly extensive" and "where some persuasive economic intuition exists" for their inclusion. However, at the other extreme, if the number of controls being considered is not very large at all (p small relative to n), then there may be no advantage at all of running PDS Lasso. Belloni et al. (2014b) note that post-double-selection is first-order equivalent to a regression that just includes the full set of controls when p is small relative to n.

We illustrate this via simulations which vary the sample size, and number of inputted controls. Figure 6 shows the mean and median number of controls selected by PDS Lasso when we include 5, 40, and 1,600

[14] https://www.stata.com/features/overview/lasso-inferential-methods Accessed February 21, 2023.

control variables as inputs. Fewer variables get selected when more variables are included in the input set. This is particularly concerning with smaller samples, where by putting in many potential controls, we risk having none at all selected. But even with a sample size of 10,000 observations, inputting a large number of variables (1,600 in this example) leads to too few variables getting selected.

We have three additional pieces of advice for empirical practice. A first and important note is that researchers should ensure that their sample size n available for estimation does not change as one changes the set of variables in x. This means ensuring that there are no missing values in any of the variables included in x, which can be accomplished by dummying out missing values and including missing value indicator variables as additional covariates. We discovered this issue affects multiple published papers during our replications. In one extreme case, there was one paper where the variables were selected using a subsample that was a tenth the size of the overall sample due to this issue. One common issue that can then arise is that when a variable x_1 and a dummy variable indicating that x_1 is missing are both included in the set of controls inputted, PDS Lasso may only end up selecting one of the two. If researchers wish to include both, they can of course re-estimate the model by adding in the unselected variable. However, our view is that researchers should be comfortable including only one of the two. For example, age may not have any predictive power for the outcome nor be associated with treatment in an RCT, but refusing to tell your age (and hence having it missing), could be a strong predictor of differential attrition or of some outcomes of interest. Since we are not trying to interpret the coefficient on x_1, but just use it to improve precision or reduce imbalance, it does not matter that it may include some zeros for missings, or just be an indicator for missing responses on a variable not included in the regression.

A second practical recommendation is that, especially with small samples, researchers might want to discretize some continuous control variables that have highly skewed distributions or outliers (e.g., as a dummy variable for above or below the median), rather than include as continuous variables. Otherwise, there is a risk of overfitting based on just a few observations.

A third practical issue that comes up from applied researchers is how they should consider a set of fixed effects. For example, they may have fixed effects for each survey enumerator, or for different geographic regions, or different industries. They then wonder whether they should take an "all or nothing" approach, whereby either they include a full set of dummies for these different categories, or none at all. Inputting these fixed effects as control variables into PDS Lasso may instead select only one or two enumerator, industry, or region fixed effects to include. This may seem sensible if perhaps only one or two enumerators, or one or two regions may be predictive of the outcome or treatment status. However, recent work by Kolesár et al. (2024) notes that the approximate sparsity assumption can be problematic with categorical variables, and results are quite sensitive to the normalization used (e.g. to which category is set as the base category). Unless

researchers have domain-specific knowledge to help prune and/or combine many categories into a smaller subset, and to suggest sparseness, we therefore urge caution in using PDS Lasso to select only a subset of fixed effects, and instead suggest partialling out the full set of any fixed effects the authors wish to include.

5.3 What should go into the amelioration set?

Recall the amelioration set I_3 is the set of additional variables that researchers want included in the regression, even if they do not get selected in either of the two double-selection steps for predicting y or T. Wüthrich and Zhu (2023) recommend using variables suggested by theory and prior knowledge in this set to help mitigate concerns about finite sample omitted variable bias. This seems less of a concern in the experimental setting, if the goal is mostly to reduce residual variance rather than explain treatment, and it seems rare that researchers would have strong priors on variables that predict treatment given random assignment. Belloni et al. (2014b) note that one of the regularity conditions required for the approximate sparsity condition underlying their results is that the size of the amelioration set should not be substantially larger than the size of the set of variables selected by Lasso.

Our view is that the most likely variables to be chosen by researchers are those in the Ancova specification in equation (3): that is, the lagged dependent variable, and any randomization strata fixed effects. This guards against the underselection of the lagged dependent variable shown in section 4.4. As discussed by Bruhn and McKenzie (2009) it is also prudent to always include strata fixed effects in the regression. Belloni et al. (2016) consider the case of panel data and individual fixed effects, and note that approximate sparseness is likely to be inappropriate for dealing with individual specific heterogeneity, and that one should instead partial out individual fixed effects and assume approximate sparseness on demeaned data in the panel setting. Likewise, with strata fixed effects, one will usually want to include them all, and then require approximate sparseness conditional on the randomization strata. And as discussed in the prior section, given the recent work by Kolesár et al. (2024), any other fixed effects such as enumerator or regional fixed effects may also want to be included in this amelioration set rather than having PDS Lasso select only a subset of them.

There are then two ways of including this amelioration set. The first is to include them, but penalize the selection of other variables accordingly, so that p in equation (9) includes the variables in the amelioration set. This is implemented by the *aset()* option with the *pdslasso* command. The alternative is to apply zero penalty weight to them, which can be done by partialling out these variables before applying Lasso, in which case they will not be included as part of p. The *partial()* option does this. In practice, we find that out of the 18 replicated papers, only one paper used the aset() option and 12 papers used the partial() option. The 5 remaining papers used neither of the options, however, we partialled out the constant in order to get the

correct number of controls selected by PDS Lasso.[15]

We explore the difference in the total number of controls selected using the *partial()* option in comparison with the *aset()* option as shown in table A3 in the appendix. We notice that the overall number of variables selected is almost the same using either options. The only main distinction is in the maximum number of controls selected, which is 68 variables using *partial()* and 78 variables using *aset()*. We also check the difference in the number of variables selected between both options and we find that while in rare cases there can be a large difference in the number of control variables selected between the two approaches, the mean of the difference is only 1 variable, and the median is no difference.

It is not clear *ex ante* which method is most appropriate. On the one hand, since the plug-in penalty parameter can underselect in small samples, it seems appropriate not to further penalize the variables, especially in smaller samples. On the other hand, the increase in standard error due to degrees of freedom adjustment is particularly costly in small samples with many control variables. It could also depend on whether the strata dummies explain any variation in y_i, or how correlated it is with other potential control variables.

To answer this question we ran simulations for different numbers of strata (10 or 100), different number of observations (200 or 600), and varied whether the strata are correlated with the dependent variable or not as shown in table A4 in the appendix. Overall, we find almost no difference in standard errors. The decision whether to penalize the other variables or not therefore does not seem to matter much, at least for the DGP used in our simulations.

5.4 Is the double step even necessary?

We have seen that in over half the empirical cases, no variables at all are selected in the second step (predicting treatment). When variables are selected in this second step, Table 2 shows that they are almost entirely variables that are not selected in step one, so are not strongly correlated with the outcome. This raises the question of whether it is worth having this second step at all in a randomized experiment.

We examine this through simulations, reported in Figure 7. We hold the sample size constant, and introduce selective imbalance, by setting $x_{1,i} = \alpha_1 T_i + e_i$ before constructing equation 10, where $e_i \overset{\text{iid}}{\sim} N(0,1)$ and $\alpha \in (0.1, 0.2, 0.3)$. Such imbalance would occur if the attrition rates were different between the treatment and control groups and the probability of attrition is correlated with $x_{i,1}$.[16] These are the conditions under which the double selection would potentially be valuable.

The main take-away from Figure 7 is that including Step 2 reduces the Mean Square Error (MSE) when

[15] The *dsregress* program also allows one to specify a set of variables that must be included in the final regression, but there is no function to partial these variables in advance of performing variable selection.

[16] For example $\alpha \approxeq 0.2$ if 11 percent of observations in the treatment group, but those with with highest values of $x_{i,1}$, attrite.

there is a moderate degree of correlation between $x_{i,1}$ and y, but *increases* the MSE if the correlation is very low; it makes no difference when β_1 is high. When β_1 is very low, $x_{i,1}$ is not a source of omitted variable bias, so including the variable does not change the extent of bias (panel (b)), but it increases the standard errors when β_1 is small (panel (c)). In contrast, when β_1 is very high then the imbalanced variable gets selected in Step 1, so the double selection does not make any difference (panel (a)). Taken together, it is only in these intermediate cases—β_1 is sufficiently high, so that it is a source of omitted variable bias, but sufficiently small so it is not selected in the first step—that the second step reduces the MSE. The exact range of β_1 where Step 2 leads to a reduction in the MSE depends on the sample size and extent of imbalance.

This then suggests there can be value in the double-selection step, but that researchers should again be judicious in the choice of which variables to include in the control variables input set. Including a whole host of variables that might be linked to attrition, but that are at most very weakly correlated with the key outcome can increase mean-squared errors.

5.5 What if there are multiple outcomes and/or multiple treatments?

Our discussion so far has been limited to the canonical case of a single outcome y, and a single treatment. But all of the studies have multiple outcomes, and Table 1 shows, 56% of the studies employ multiple treatments. With multiple outcomes, PDS Lasso can simply be run separately for each outcome. However, there are two implications of this for empirical work that are worth noting. First, this does mean that a different set of control variables can be potentially chosen for each outcome, or for the same outcome measured at different points in time. For example, in one paper where the authors input 115 controls, PDS Lasso does not select any controls for one outcome but selects 13 controls for another outcome. In another paper with 3340 inputted controls, the range gets bigger and we find that while PDS Lasso does not select any controls for one of the outcomes, it selects 51 controls for another outcome. This is not a problem, and is indeed sensible, since, for example, the variables that best predict labor force participation may be different from those that predict earnings or fertility. But this does differ from the ad hoc "robustness check" approach to including covariates, which tends to add the same set of covariates for each outcome. Second, if the sample size differs across outcomes (as is often the case with item non-response in surveys), then the treatment selection regression in equation (7) might also select different variables for different outcomes. As a result, researchers should not simply run the treatment regression selection equation once and see whether any variables are selected, but ensure it is run for each sample being used for analysis.

With multiple treatments, we can extend equation 1 to include D treatments:

$$y_i = \alpha + \sum_{d=1}^{D} \gamma_D T_{D,i} + g(z_i) + \varepsilon_i \tag{12}$$

Then Urminsky et al. (2016) note that the method easily extends to incorporate these additional treatments by simply repeating step 2 (the Lasso treatment selection regression) for each treatment. Note here that each step 2 regression will be comparing a particular treatment to the combined group of the control and all other treatments. Then the post-double-selection regression will include the union of all covariates chosen to predict y, as well as those chosen to predict anyone of the different treatments. In the typical field experiment this may still select few, if any, variables, given random assignment sets each treatment orthogonal to these controls in expectation.

5.6 How should PDS Lasso be used with treatment interactions?

In addition to estimating the average effect of being assigned to treatment, it is common for researchers to examine treatment heterogeneity by interacting treatment with one or more covariates. For example, they may augment equation 1 to allow treatment to vary with variable $x_{i,1}$, to give the following partial linear model:

$$y_i = \alpha + \gamma T_i + \delta T_i x_{1,i} + \mu x_{1,i} + g(z_i) + \varepsilon_i \tag{13}$$

Where $x_{i,1}$ could also be part of the vector x used to approximate $g(z_i)$. Lasso can then be used to select variables that predict y and T as before, but now there is the additional term Tx_1 to also consider. Running step 2 (equation 7) to predict Tx_1 can be done to select additional covariates to add as controls, where x_1 should then be considered in the amelioration set, since we would always want to include the level of a variable when including its interaction with treatment.[17] If this is not done, then if x_1 is correlated with many of the other covariates, this step could end up selecting a large number of variables.

In practice it is common to see two different types of interactions. The first occurs when multiple follow-up periods are used, and treatment is interacted with time period to allow the effect of treatment to vary over time. We should then only expect to see different variables potentially chosen for the interaction terms if the attrition varies across rounds. The second type of interaction is when authors interact treatment with baseline variables like gender, education, or region to examine treatment heterogeneity. Then it is possible that the sample is balanced on observables overall, but once one looks within subgroups is imbalanced, or vice versa. We examine whether PDS Lasso selects different controls that predict T than the ones that predict Tx_1 in three papers with a total of 8 regressions and 9 treatments. We find varying results across

[17] We recommend partialling it out, so that the inclusion of this interacting variable is not penalized by lasso.

the three papers; while in one of them we find that PDS Lasso selects more controls that predict Tx_1 than the ones that predict T, we find the opposite in another paper. The third paper shows no difference between the set of controls chosen in both scenarios, as PDS Lasso did not select any controls except in few cases.

In the replications, one-third of the papers use interactions, resulting in 15 percent of the estimated coefficients in our replication sample being on an interaction. However, one common issue we encountered when examining replications of papers with interactions is that authors include the interacted variable in their *pdslasso* command as a treatment instead of partialling it out or adding it to the amelioration set. This led to the selection of more controls. For instance, including the interacted variable as a treatment in one paper resulted in having between 38 and 41 controls selected for several outcomes. However, after partialling out the interacted variable instead, we find that PDS Lasso selected between 8 and 11 controls. Appendix 6 shows an example of the incorrect syntax used, and the correct way to include interactions in this case.

6 A checklist for best empirical practice

Based on the above analysis, we offer the following guidelines for empirical researchers considering using PDS Lasso in their experimental analysis.

1. Be realistic as to how much power one is likely to gain. In the typical RCT we find very few variables get selected, and standard errors are less than one percent smaller than simply using Ancova. Power calculations should therefore not anticipate large improvements in power from using PDS Lasso.

2. Include the lagged dependent variable and any randomization strata in the amelioration set. This will help prevent underfitting where the lagged variable does not get selected.

3. Be judicious in the choice of the number of control variables inputted into PDS Lasso, and do not use a kitchen sink approach. Including very many variables can make it less likely the more important ones get chosen, and may worsen mean-squared error. For example, chance imbalances or attrition related to variables that are not correlated at all with the outcome need not be controlled for.

4. Be very careful about missing values, and ensure that all control variables being inputted have no missing values (by dummying out if necessary). Eight of the 18 papers we examined had the final sample using PDS Lasso smaller than the sample used for OLS because of this issue.

5. When including treatment interactions, ensure the interacting variable is included in the amelioration set, and not accidentally modelled as if it were another treatment. Multiple papers made coding syntax errors of this sort.

6. When using the *pdslasso* command, be aware of the lack of a degrees of freedom adjustment. Otherwise one can end up with smaller standard errors after using this command compared with using OLS even when no additional controls are selected. More generally, the implementation details can differ across software packages, and so it is useful to pre-specify exactly how the method will be implemented.

After following these steps, we see PDS Lasso as a useful robustness check for many field experiments, providing a less ad hoc way of selecting additional control variables on top of any lagged variable and randomization strata. Most of the time it should make very little difference to the estimated coefficients and standard errors. When it does make a sizeable change in the coefficients, this can provide a useful warning for researchers that they can no longer rely simply on random assignment to justify their results.

References

Ahrens, Achim, Christian Hansen, and Mark Schaffer, "Pdslasso and ivlasso: Progams for post-selection and post-regularization OLS or IV estimation and inference," http://ideas.repec.org/c/boc/bocode/s458459.html 2018. updated 2020.

Bai, Yuehao, "Optimality of matched-pair designs in Randomized Controlled Trials," *American Economic Review*, 2022, *112* (12), 3911–3940.

Belloni, Alexandre and Victor Chernozhukov, "High Dimensional Sparse Econometric Models: An Introduction," 2011. arXiv:1106.5242v2.

_ , Daniel Chen, Victor Chernozhukov, and Christian Hansen, "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 2012, *80*, 2369–2429.

_ , Victor Chernozhukov, and Christian Hansen, "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 2014, *28* (2), 29–50.

_ , _ , and _ , "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 2014, *81*, 608–650.

_ , _ , _ , and Damian Kozbur, "Inference in High-Dimensional Panel Models With an Application to Gun Control," *Journal of Business and Economic Statistics*, 2016, *34* (4), 590–605.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet Sekhon, and Bin Yu, "Lasso adjustments of treatment effect estimates in randomized experiments," *PNAS*, 2016, *113* (27), 7383–7390.

Bruhn, Miriam and David McKenzie, "In pursuit of balance: Randomization in practice in development field experiments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 200–232.

Chernozhukov, Victor, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis, "Applied Causal Inference Powered by ML and AI," 2024.

_ , Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, *Econometrics Journal*, 2018, *21*, C1–C68.

Garbiras-Diaz, Natalia and Mateo Montenegro, "All Eyes on Them: A Field Experiment on Citizen Oversight and Electoral Integrity," *American Economic Review*, 2022, *112* (8), 2631–2668.

Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Becerra, "Testing Attrition Bias in Field Experiments," *Journal of Human Resources, forthcoming*, 2023.

Guo, Yongyi, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman, "Machine Learning for Variance Reduction in Online Experiments," in "35th Conference on Neural Information Processing Systems" 2021.

Kolesár, Michal, Ulrich Müller, and Sebastian Roelsgaard, "The Fragility of Sparsity," 2024. arXiv:2311.02299v2.

List, John, Ian Muir, and Gregory Sun, "Using Machine Learning for Efficient Flexible Regression Adjustment in Economic Experiments," 2022. Working Paper.

McKenzie, David, "Beyond Baseline and Follow-up: The Case for more T in Experiments," *Journal of Development Economics*, 2012, *99* (2), 210–21.

Permutt, Thomas, "Testing for Imbalance of Covariates in Controlled Experiments," *Statistics in Medicine*, 1990, *9* (12), 1455–62.

Simmons, Joseph, Leif Nelson, and Uri Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, 2011, *22*, 1359–1366.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov, "Using Double-Lasso Regression for Principled Variable Selection," SSRN 2016.

Wager, Stefan, Wenfei Du, Jonathan Taylor, and Robert Tibshirani, "High-dimensional regression adjustments in randomized experiments," *PNAS*, 2016, *113* (45), 12673–12678.

Wu, Edward and Johann Gagnon-Bartsch, "The loop estimator: Adjusting for covariates in randomized experiments," *Evaluation Review*, 2018, *42* (4), 458–488.

Wüthrich, Kaspar and Ying Zhu, "Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis," *Review of Economics and Statistics*, 2023, *105* (4), 982–997.

Zou, Hui, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 2006, *101* (476), 1418–1429.

7 Tables

Table 1: Descriptive statistics

	Mean	SD	Min	P25	Median	P75	P95	Max	N
Study level									
Sample Size	3,374	3,800	418	1,038	1,913	4,079	13,987	13,987	18
Clusters	296	413	14	40	206	314	1,637	1,637	15
Number of Outcomes	43	46	3	9	37	54	189	189	18
Multiple treatments	0.56	0.51	0.00	0.00	1.00	1.00	1.00	1.00	18
Outcome × treatment level									
Inputted Controls	405	814	6	39	182	331	3,340	3,340	780
Inputted Controls to Sample Size	0.17	0.19	0.00	0.06	0.16	0.19	0.69	1.73	780
Inputted Controls to Clusters	2.27	4.40	0.04	0.20	0.86	2.00	16.06	17.58	694
Overall Attrition	0.14	0.16	0.00	0.04	0.10	0.23	0.34	0.93	780
Treatment Attrition	0.14	0.17	0.00	0.03	0.09	0.24	0.39	0.94	780
Control Attrition	0.16	0.16	0.00	0.04	0.11	0.24	0.35	0.92	780
Differential Attrition	0.05	0.08	0.00	0.00	0.02	0.07	0.17	0.56	780

Notes. Based on 780 outcome by treatment combinations from 18 studies.

Table 2: Empirical Performance of PDS Lasso and comparison with ANCOVA

	Mean	SD	Min	P25	Median	P75	P95	Max	N
Number of variables selected									
Overall	3.62	4.72	0	1	2	5	11	51	780
Selected for y	2.74	4.57	0	0	1	3	9	51	780
Selected for T	0.90	1.59	0	0	0	1	4	17	780
Selected for T, but not y	0.88	1.59	0	0	0	1	4	17	780
At least one variable selected									
Overall	0.83	0.37	0	1	1	1	1	1	780
Selected for y	0.71	0.45	0	0	1	1	1	1	780
Selected for T	0.43	0.50	0	0	0	1	1	1	780
Selected for T, but not y	0.43	0.50	0	0	0	1	1	1	780
Performance (vs ANCOVA)									
Change in stand. coef.[i]	0.05	0.16	0.00	0.00	0.01	0.04	0.15	2.20	779
SE ratio[ii]	0.970	0.103	0.502	0.949	0.992	1.004	1.045	2.438	776
SE ratio (no d.f. adj.)[iii]	0.950	0.104	0.501	0.932	0.964	0.989	1.021	2.426	776

Notes. Data is at an outcome × treatment level. The penalty parameter is calculated using the method proposed by Belloni and Chernozhukov (2011). The final panel compares the coefficient estimates and standard errors (SEs) with estimates using ANCOVA (equation 11).

[i] Absolute value of the difference in coefficient estimates, divided by the control group standard deviation

[ii] SE using PDS divided by the SE using ANCOVA

[iii] No adjustment to degrees of freedom are made to the SEs when using the PDS method.

Table 3: Comparing cross-validation with ANCOVA and with the plug-in penalty

	Mean	SD	Min	P25	Median	P75	P95	Max	N
Number of variables selected									
Overall	26.96	47.81	0	4	13	27	150	264	760
Selected for y	14.93	18.65	0	2	9	22	46	167	760
Selected for T	13.94	46.15	0	0	0	5	133	238	760
Selected for T, but not y	12.03	43.58	0	0	0	2	128	229	760
At least one variable selected									
Overall	0.90	0.30	0	1	1	1	1	1	780
Selected for y	0.84	0.37	0	1	1	1	1	1	780
Selected for T	0.43	0.50	0	0	0	1	1	1	780
Selected for T, but not y	0.34	0.48	0	0	0	1	1	1	780
Performance (vs ANCOVA)									
Change in stand. coef.	0.12	0.56	0.00	0.01	0.03	0.06	0.39	10.27	759
SE ratio[i]	0.956	0.142	0.500	0.906	0.965	1.000	1.180	2.182	760
Performance (vs plug-in)									
Change in stand. coef.	0.11	0.52	0.00	0.01	0.02	0.04	0.32	8.07	759
SE ratio[ii]	0.988	0.112	0.528	0.952	0.984	1.003	1.189	2.022	760

Notes. Performance of the PDS Lasso procedure when the penalty parameter is calculated using cross-validation (CV). Panels A and B shows number of variables selected. Panels C and D compares performance with ANCOVA, and when the penalty parameter is calculated using the method proposed by Belloni and Chernozhukov (2011) (referred to as the plug-in).

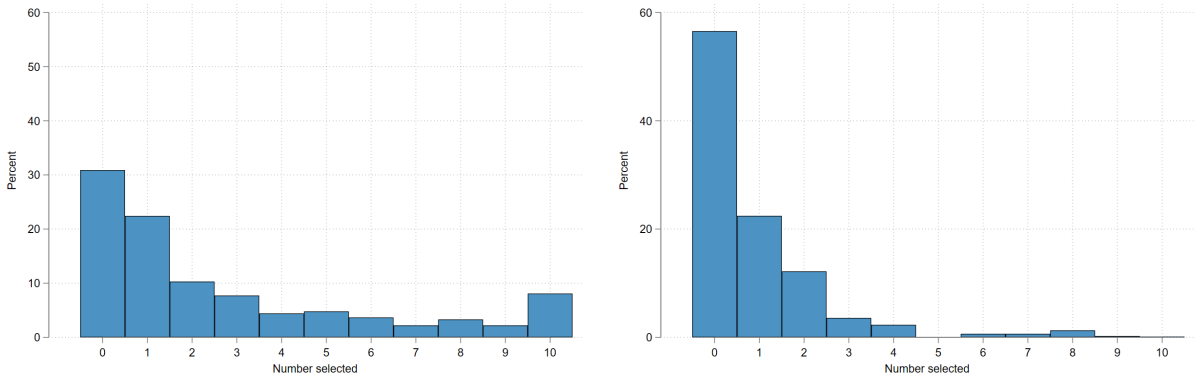
[i] SE using PDS Lasso with CV divided by SE using Ancova

[ii] SE using PDS Lasso with CV, divided by SE using PDS Lasso with the plug-in penalty parameter

Data is at an outcome × treatment level.

8 Figures

Figure 1: Number of variables selected in the first and second steps

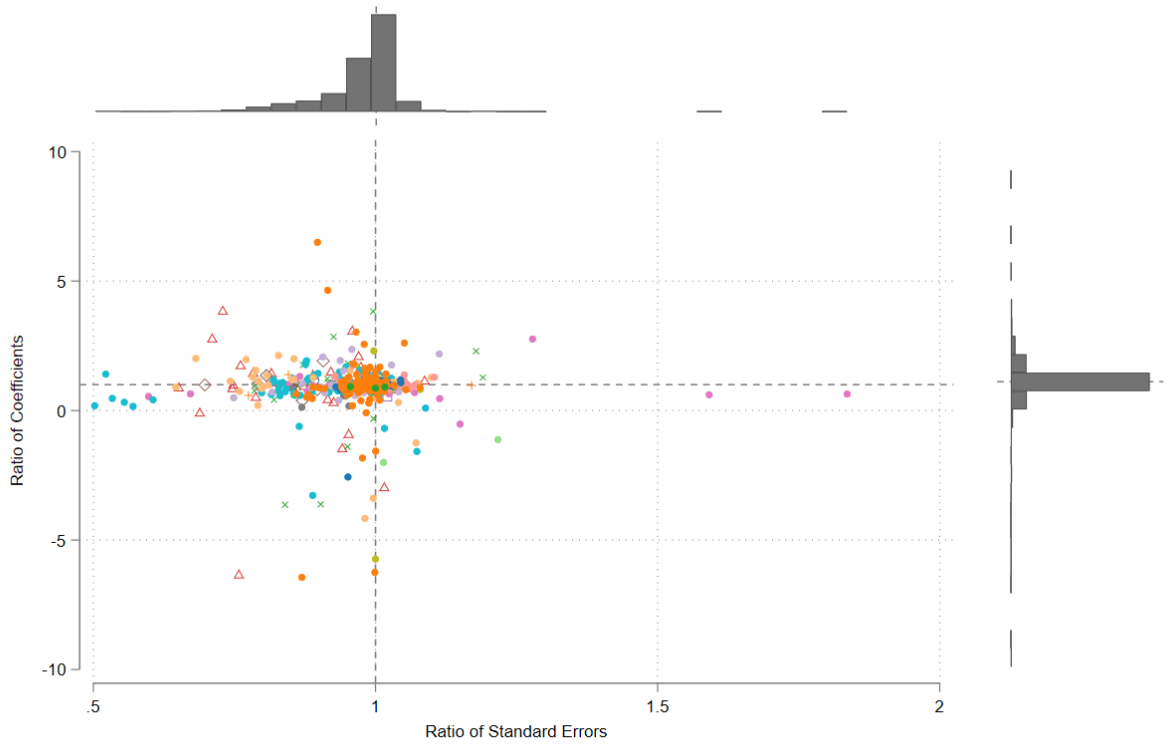


(a) Selected on dependent variable (step I)

(b) Selected on treatment (Step II)

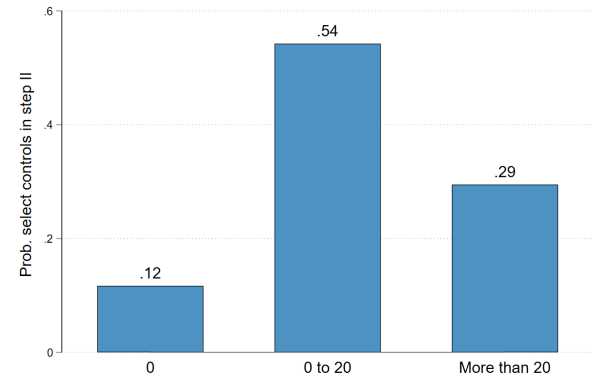
Notes: Distributions are top-coded at 10. Figure (a) shows a histogram of the number of variables selected that are predictive of the dependent variable, data is at an study-by-outcome level ($n = 271$). Figure (b) shows a histogram of the number of variables selected that are predictive of treatment assignment, and data is at an study-by-outcome-by-treatment level ($n = 780$).

Figure 2: Ratio of coefficients and standard errors: ANCOVA vs post-double selection method

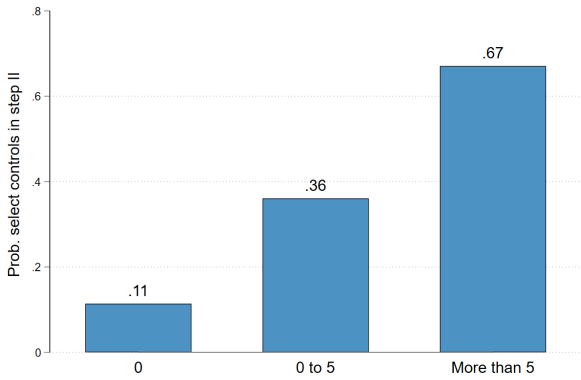


Notes. The ratio is equal to the estimated coefficient (or standard error) using the post-double selection method, divided by the estimated coefficient (or standard error) using ANCOVA. Each color represents a different paper. Distributions of the two variables are shown in the histograms to the above and right of the graph. 13 observations with coefficient ratios smaller than -10 or larger than 10 are omitted.

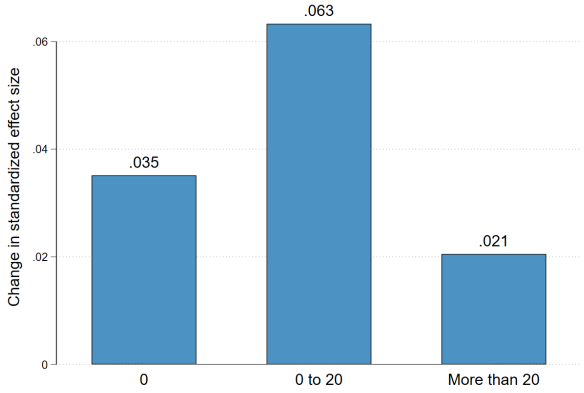
Figure 3: The probability of selecting a variable predictive of treatment, and change in effect size, varies with the attrition rate



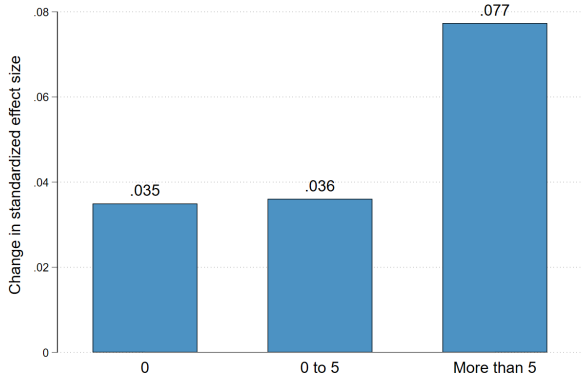
(a) Select controls, by attrition



(b) Select controls, by differential attrition



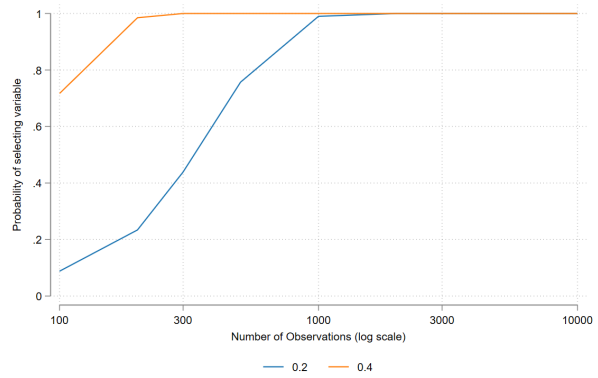
(c) Change in standardized β, by attrition



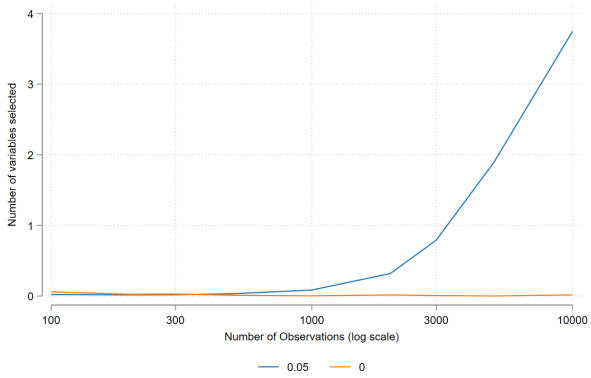
(d) Change in standardized β, by differential attrition

Notes. In panels (a) and (b), the y-axis is the probability that a variable is selected in step II. In panels (c) and (d), the y-axis is the absolute value of the change in the standardized effect size, comparing PDS lasso with ANCOVA. In panels (a) and (c), data is split by the overall attrition rate. In panels (b) and (d), data is split by the absolute value of the *difference* in attrition rates between the treatment group and control. The length of the bars designate the mean in each sub-group.

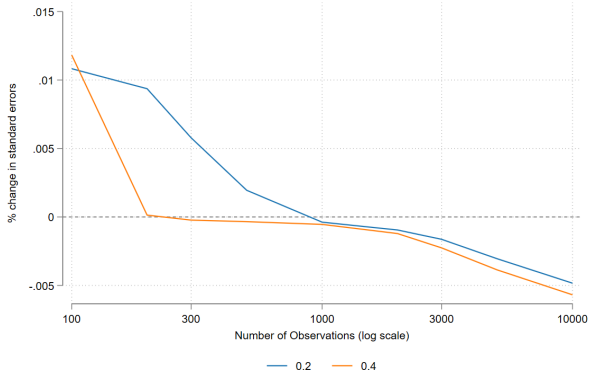
Figure 4: PDS Lasso sometimes is less precise than Ancova due to failure to select the lagged dependent variable



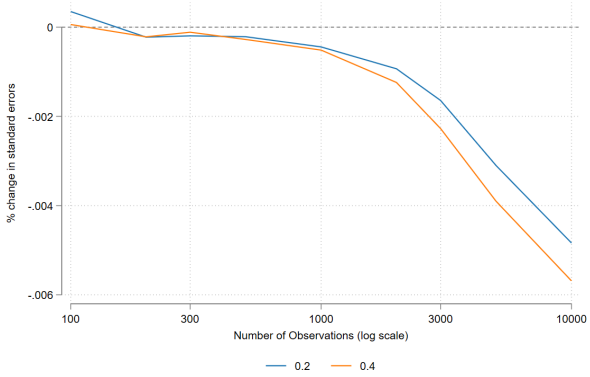
(a) Probability of selecting $x_{1,i}$



(b) No. of other variables selected



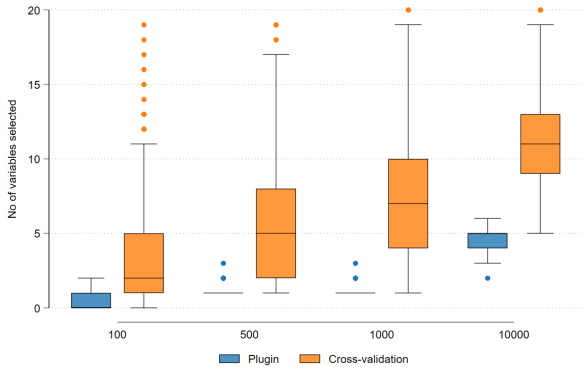
(c) Standard error, relative to ANCOVA



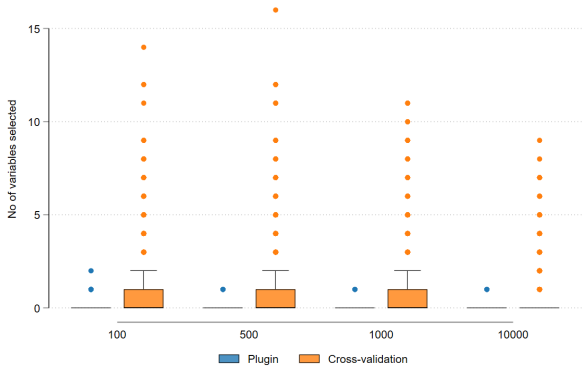
(d) Standard error (partial $x_{1,i}$), relative to ANCOVA

Notes: Data is generated using equation 10, where $x_{i,j} \overset{\text{iid}}{\sim} N(0_p, I_p)$, and $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$. $p = 20$. 1,000 simulation replications. Results shown for for different values of n and β_1. In each figure the x-axis is the natural log of n. Panel (a) shows the probability of selecting the lagged dependent variable, $x_{i,1}$, and different lines indicate different values for $\beta_1 \in (0.2, 0.4)$. Panel (b) shows the number of other variables selected, where $\beta_j = 0.05$ for $2 \leq j \leq 5$, and $\beta_j = 0$ for $i > 5$. Panel (c) shows the percentage change in the standard error, relative to ANCOVA: $y_i = \gamma T_i + \beta_0 + \beta_1 x_{i,1} + \epsilon_i$. Panel shows the percentage change in the standard error, relative to ANCOVA, when the lagged dependent variable is partial out in PDS lasso procedure.

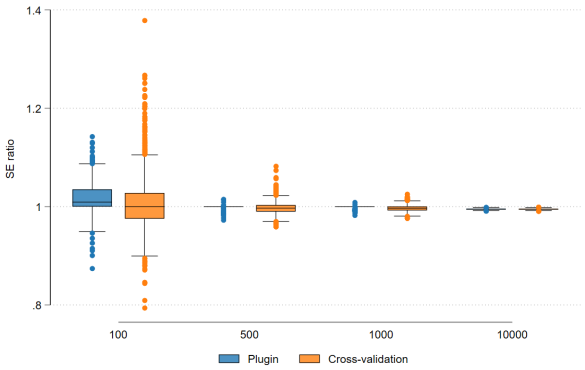
Figure 5: Cross-validation selects more variables than the plug-in penalty parameter and gives more volatile standard errors



(a) Variables selected (predict y_i)



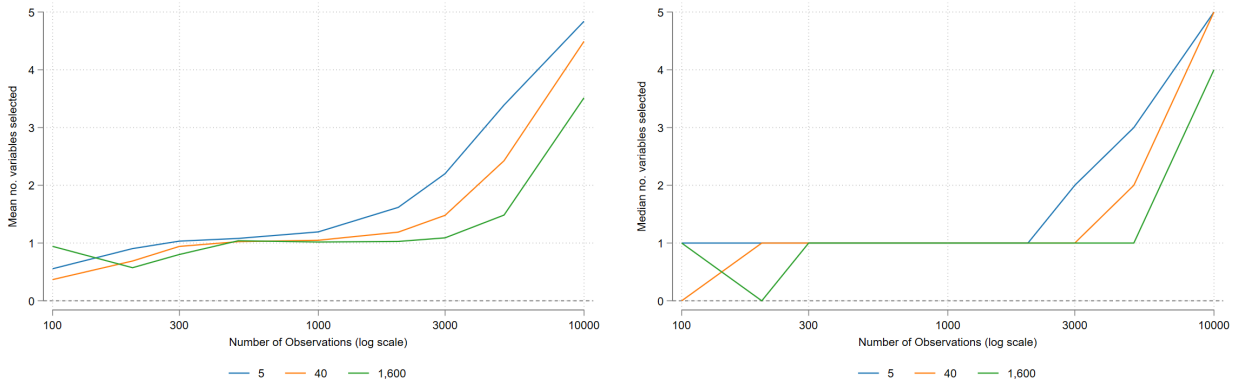
(b) Additional variables selected (predict T_i)



(c) Change in standard errors

Notes: Number of variables selected using the post-double selection (PDS) procedure, comparing two approaches to calculate the penalty parameter: the method proposed by Belloni et al. (2012), and 10-fold cross validation. See Section 3.2 for the data generating process. $p = 20$, $n \in (100, 500, 1000, 10,000)$. $\beta_1 = 0.3$, $\beta_j = 0.05$ for $j \in (2, 5)$. The remaining variables are only correlated with y_i by change. 1,000 simulation replications. Figure (a) shows box plots for the distributions of the number of variables selected in Step 1; (b) shows box plots for the number of additional variables selected using Step C. Panels C and D show the SE ratio: the SE using PDS Lasso, divided by the SE estimated using Ancova (i.e., only controlling for $x_{i,1}$).

Figure 6: Inputting many controls can result in fewer controls being selected by PDS Lasso

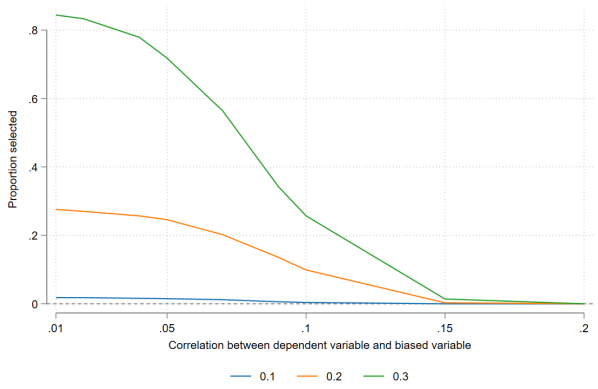


(a) Mean

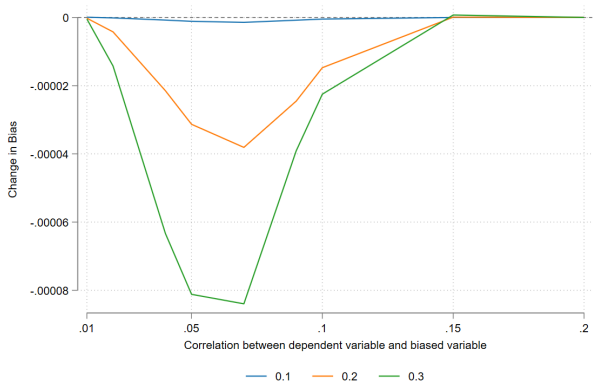
(b) Median

Notes: The mean and median number of variables selected, depending, depending on the number of predictors. Data is generated using equations (5) and (7), where $x_{i,j} \overset{\text{iid}}{\sim} N(0_p, I_p)$, and $\epsilon_i \overset{\text{iid}}{\sim} (0, 1)$. $\beta_1 = 0.3$; $\beta_j = 0.05$ for $2 \leq j \leq 4$; and $\beta_j = 0$ for $5 \leq j \leq 40$. $p \in (5, 40, 1600)$. In the first case (red line), variables $x_{i,1}$ to x_5 are inputed; in the second case (blue line), variables $x_{i,1}$ to x_{40} are inputed; in the final case, all 40 variables are interacted with each other and 1,600 are inputed. Variables are selected using the post double selection method proposed by Belloni et al (2014).

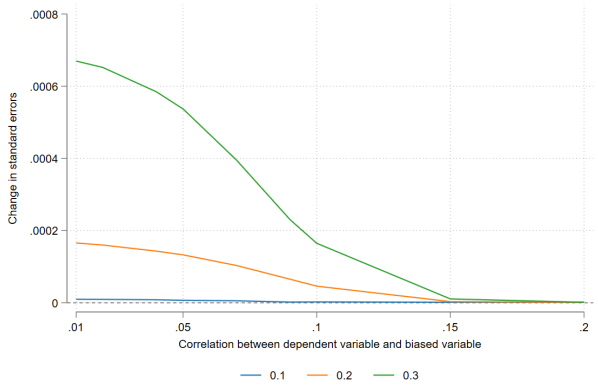
Figure 7: When does double-selection improve upon single-selection lasso?



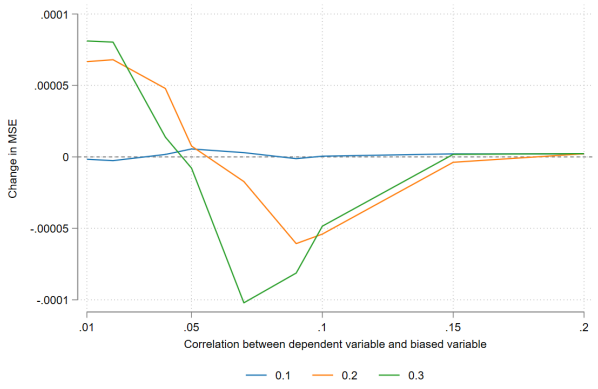
(a) Imbalanced variable selected in Step 2, but not Step 1



(b) Change in bias



(c) Change in standard errors

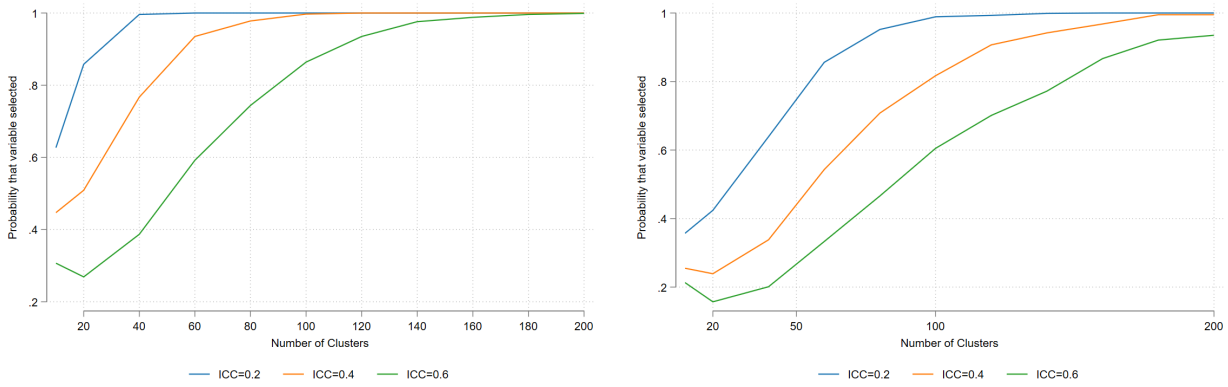


(d) Change in mean square error

Notes: Data is generated using equation 10, where $x_{i,j\neq 1} \overset{\text{iid}}{\sim} N(0_p, I_p)$, and $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$. We introduce imbalance by setting: $x_{i,1} = \alpha_1 T_i + e_i$, where $e_i \overset{\text{iid}}{\sim} N(0,1)$. Different lines in each panel show results for different levels levels of imbalance: $\alpha_1 \in (0.1, 0.2, 0.3)$. β_1 (on the x axis) ranges from 0.01 to 0.2. $(n,p) = (1,000, 20)$. All panels shows results of lasso regressions, using the penalty parameter of Belloni et al. (2012). *Double selection*=post-double selection method of Belloni et al. (2014b). *Single selection* only includes as controls $x_{i,j} \in I_1$. Panel (a) indicates the probability that $x_{i,1}$ is selected in Step 2 but not in Step 1. Panels (b) to (d) shows the difference between the Double and Single selection in terms of bias, the standard errors, and the mean square error, respectively. Bias=$(\frac{1}{N}\sum_{s=1}^{N}\hat{\beta}_1 - \beta_1)$ and MSE=$\frac{1}{N}\sum_{s=1}^{N}(\hat{\beta}_1 - \beta_1)$, where s indicates each random draw from the simulations. $N = 1,000$

A Appendix Tables and Figures

Figure A1: Variables selected with clustered randomized control trial



(a) Varies at an individual level

(b) Varies at a cluster level

Notes: The DGP is:

$$y_{i,c} = \gamma T_c + \sum_{j=1}^{10} \beta_j^I x_{i,c}^j + \sum_{j=1}^{10} \beta_j^C x_c^j + \mu_c + \epsilon_{i,j}, \tag{14}$$

We set 20 individuals per cluster and vary the number of clusters and the intra-cluster correlation, $\rho \in (0.2, 0.4, 0.6)$, where $\rho = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\epsilon^2}$ and σ_μ^2 and σ_ϵ^2 are the variance of μ_c and $\epsilon_{i,c}$, respectively. $p = 20$, $\beta_1^I = \beta_1^C = 0.3$, the other variables are correlated with $y_{i,c}$ by chance. Panel (a): probability that $x_{i,c}^1$ gets selected; Panel (b): probability that x_c^1 gets selected.

Table A1: Comparing post-double selection (PDS) method with ANCOVA: Weighted and Unweighted

	Unweighted		Weighted[i]	
	Mean	SD	Mean	SD
Number of variables selected				
Overall	3.62	4.72	3.97	4.71
Selected for y	2.74	4.57	3.26	4.68
Selected for T	0.90	1.59	0.73	1.36
Selected for T, but not y	0.88	1.59	0.71	1.35
At least one variable selected				
Overall	0.83	0.37	0.84	0.37
Selected for y	0.71	0.45	0.74	0.44
Selected for T	0.43	0.50	0.38	0.48
Selected for T, but not y	0.43	0.50	0.37	0.48
Performance (vs ANCOVA)				
Change in stand. coef.[ii]	0.05	0.16	0.05	0.15
SE ratio[iii]	0.970	0.103	0.975	0.153
SE ratio (no d.f. adj.)[iv]	0.950	0.104	0.952	0.160

Notes. Data is at an outcome × treatment level. Unweighted statistics give all 780 outcome by treatment estimates equal weight. Weighted estimates reweight at the paper level. The penalty parameter is calculated using the method proposed by Belloni and Chernozhukov (2011). The final panel compares the coefficient estimates and standard errors (SEs) with estimates using ANCOVA (equation 11).
[i]The weight of each paper is calculated by dividing 1 over the number of outcomes by treatment estimates per paper
[ii]Absolute value of the difference in coefficient estimates, divided by the control group standard deviation
[iii]SE using PDS divided by the SE using ANCOVA
[iv]No adjustment to degrees of freedom are made to the SEs when using the PDS method.

Table A2: Change of Significance when Using Different Approaches

	Ancova vs Pdslasso		Ancova vs Cvlasso		Pdslasso vs Cvlasso	
	insig to sig	sig to insig	insig to sig	sig to insig	insig to sig	sig to insig
Panel A. Change of Significance at 5 percent level						
Number of Observations	49	15	59	31	28	30
Percent	6.28	1.92	7.76	4.08	3.68	3.95
Panel B. Change of Significance at 10 percent level						
Number of Observations	42	16	52	31	28	28
Percent	5.38	2.05	6.84	4.08	3.68	3.68

Notes. In the first column, we show the number and percent of observations that are insignificant when using Ancova but significant when using PDS LASSO. The second column indicates the number and percent of observations that are significant when using Ancova but insignificant when using PDS LASSO. The same logic applies to the subsequent columns.

Table A3: Total Selected Controls using PDS Lasso When Using Partial vs Amelioration set

	Mean	SD	Min	P5	P25	Median	P75	P95	Max	N
Overall Number of Variables Selected										
Using Partial	19	14	1	2	9	17	24	45	68	604
Using Aset	20	15	1	2	8	18	26	48	78	604
Difference										
Overall Difference	-1	4	-46	-7	-2	0	0	3	9	604
Conditional on not being zero	-2	4	-46	-8	-4	-1	1	5	9	379

Notes. Some observations were dropped due to the difference in using partial or aset options to ensure comparability.

Table A4: Standard errors are almost identical partialling out variables compared to including them in the amelioration set

No. observations	200		600	
No. of strata	10	100	10	100
Panel A. Strata not correlated with y				
Amelioration set	0.135	0.135	0.077	0.077
Partial	0.136	0.136	0.077	0.078
Panel B. Strata correlated with y				
Amelioration set	0.135	0.135	0.078	0.078
Partial	0.135	0.135	0.078	0.078

Notes. Data is generated using equation 10, where $x_{i,j} \overset{\text{iid}}{\sim} N(0_p, I_p)$, and $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$. $p = 20$. $\beta_j = 0.05$ for $2 \leq j \leq 5$, $\beta_j = 0$ for $i > 5$. $n \in (200, 600)$. In Panel A strata fixed effects explain no variation in y_i. In Panel B strata fixed effects are quantiles of $x_{i,1}$, so explain some variation in y_i.

Table A5: List of Replicated Papers

Authors	Title	Journal	Year	Clustered trial	No. of Outcomes	No. of Treatments	No. controls inputted	No. of Estimates
Wheeler et al	LinkedIn (to) Job Opportunities: Experimental Evidence from Job Readiness Training	AEJ: Applied	2022	Yes	3	1	16	3
Lopez et al	Does Patient Demand Contribute to the Overuse of Prescription Drugs	AEJ: Applied	2020	Yes	11	2	331	42
Abel et al	The Value of Reference Letters: Experimental Evidence from South Africa	AEJ: Applied	2020	Yes	3	1	28	9
Bertrand & Crépon	Teaching Labor Laws: Evidence from a Randomized Control Trial in South Africa	AEJ: Applied	2021	No	12	1	115	12
Hussam et al (a)	The Psychosocial Value of Employment: Evidence from a Refugee Camp	AER	2022	Yes	14	2	6	35
Hussam et al (b)	Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field	AER	2022	Yes	4	1	91	8
Armand et al	Does Information Break the Political Resource Curse? Experimental Evidence from Mozambique	AER	2020	Yes	19	2	171	38
Andrabi et al	Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools	AER	2020	Yes	13	1	284	48
Alsan et al	Does Diversity Matter for Health? Experimental Evidence from Oakland	AER	2019	Yes	21	3	12	63
Carneiro et al	The Impacts of a Multifaceted Prenatal Intervention on Human Capital Accumulation in Early Life	AER	2021	Yes	27	1	3340	54
Garbiras-Díaz & Montenegro	All Eyes on Them: A Field Experiment on Citizen Oversight and Electoral Integrity	AER	2022	Yes	9	3	39	96
Dhar et al	Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India	AER	2022	Yes	3	1	61	10
Augsburg et al	When nature calls back: Sustaining behavioral change in rural Pakistan	JDE	2022	Yes	2	1	183	6
Heß et al	Environmental effects of development programs: Experimental evidence from West African dryland forests	JDE	2021	Yes	3	1	648	6
Islam et al	The Effects of Chess Instruction on Academic and Non-cognitive Outcomes: Field Experimental Evidence from a Developing Country	JDE	2021	Yes	30	1	14	30
Magnan et al	Information, technology, and market rewards: Incentivizing aflatoxin control in Ghana	JDE	2021	Yes	24	3	80	84
McIntosh & Zeitlin	Using household grants to benchmark the cost effectiveness of a USAID workforce readiness program	JDE	2022	Yes	18	2	332	189
Fernando	Seeking the treated: The impact of mobile extension on farmer information exchange in India	JDE	2021	No	21	2	409	47

Appendix 6: An example of *pdslasso* syntax to use when including interactions

Here *treatment* is the dummy variable for treatment, x is the interacting variable (e.g. female), and *interaction* is the interaction of *treatment* and x.

Commonly used incorrect syntax:

pdslasso y treatment interaction x ($controls), partial(fixedeffects laggedvariable)

Correct syntax:

pdslasso y treatment interaction ($controls x), partial(fixedeffects laggedvariable x)