NBER WORKING PAPER SERIES

MACHINE LABOR

Joshua Angrist
Brigham Frandsen

Working Paper 26584
http://www.nber.org/papers/w26584

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2019

This is a revised version of NBER Working Paper 26584, released December 2019. Many thanks to Ray Han and Shinnosuke Kikuchi for outstanding research assistance. Thanks also go to Alberto Abadie, Matias Cattaneo, Dean Eckles, Chris Hansen, Peter Hull, Guido Imbens, Simon Lee, Anna Mikusheva, Whitney Newey, Jose Olea, Parag Pathak, Bruce Sacerdote, Bernard Salanie, Stefan Wager, Chris Walters, and seminar participants at Columbia and MIT for helpful discussions and comments. They are not to blame for any of our mistakes or conclusions. This paper is dedicated to the memory of Alan Krueger. We would have liked to have his comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2019 by Joshua Angrist and Brigham Frandsen. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Machine Labor
Joshua Angrist and Brigham Frandsen
NBER Working Paper No. 26584
December 2019, Revised May 2020
JEL No. C21,C26,C52,C55,J01,J08

ABSTRACT

Machine learning (ML) is mostly a predictive enterprise, while the questions of interest to labor economists are mostly causal. In pursuit of causal effects, however, ML may be useful for automated selection of ordinary least squares (OLS) control variables. We illustrate the utility of ML for regression-based causal inference by using lasso to select control variables for estimates of effects of college characteristics on wages. ML also seems relevant for an instrumental variables (IV) first stage, since the bias of two-stage least squares can be said to be due to over-fitting. Our investigation shows, however, that while ML-based instrument selection can improve on conventional 2SLS estimates, split-sample IV, jackknife IV, and LIML estimators do better. In some scenarios, the performance of ML-augmented IV estimators is degraded by pretest bias. In others, nonlinear ML for covariate control creates artificial exclusion restrictions that generate spurious findings. ML does better at choosing control variables for models identified by conditional independence assumptions than at choosing instrumental variables for models identified by exclusion restrictions.

Joshua Angrist
Department of Economics, E52-436
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
angrist@mit.edu

Brigham Frandsen
Department of Economics
Brigham Young University
Provo, UT 84602
frandsen@byu.edu

1 Introduction

Many economic applications of ML originate in research on consumer choice. For instance, Bajari et al. (2015) uses ML to predict the demand for salty snacks. What does the ML toolkit offer empirical labor economics? Like marketing researchers, labor economists also benefit from more and bigger data sets. But most empirical labor questions turn on features of distributions, like conditional mean functions, rather than on the accuracy of individual prediction. Much of the applied labor agenda seeks to uncover causal effects, such as the effect of schooling on wages, using tools like regression and instrumental variables (IV). Other labor applications describe distributional shifts and trends, such as changes in income inequality. Whether causal or descriptive, labor questions involve few pure prediction problems.

The distinction between parameter estimation and individual prediction parallels that between slope coefficients and R^2 in regression analysis. Pursuing this analogy, Mullainathan and Spiess (2017) note that ML aims to improve the accuracy of fitted values ($\hat{y}$), rather than estimates of a regression slope coefficient or marginal effect. Empirical findings in labor economics rarely turn on $\hat{y}$. Yet, as Belloni, Chernozhukov and Hansen (2014*a*) observe, in any empirical application with many covariates, we'd like to guard against over-fitting and the vagaries of data mining. And two-stage least squares (2SLS) estimates of causal effects are made more precise by improving the first-stage R^2. Indeed, first-stage estimation is sometimes framed as a prediction problem, and the bias of 2SLS is arguably a consequence of over-fitting.

We consider three domains where ML might play a supporting role in pursuit of causal effects in labor economics. The first is data-driven selection of ordinary least squares (OLS) control variables. Hahn (1998) notes that efficient nonparametric matching estimators use controls to impute counterfactual outcomes. The fact that imputation is a form of prediction suggests ML is a good way to do it. We find empirical support for this idea in a replication and extension of the Dale and Krueger (2002) investigation of the causal effect of college characteristics on graduates' earnings (henceforth, DK02).

The DK02 research design conditions on the characteristics of colleges to which an applicant has applied and been admitted. The key identifying assumption here takes enrollment decisions conditional on application/admission sets to be as good as randomly assigned.

Graduates of highly selective and private colleges earn more, on average, than do those who attended less selective or public institutions. But this evidence of an elite school earnings advantage disappears after conditioning on 150 dummy variables indicating the selectivity of the schools in application/admissions sets. The cost of the DK02 dummy-variable control strategy is a two-thirds reduction in sample size. DK02 therefore deploys more parsimonious though also more parametric control strategies.

In the DK02 context, analysts seeking a smaller set of control variables must grapple with the fact that the college application process can be parameterized in many ways. This flexibility opens the door to potentially misleading specification searches. Regression analysis has long been subject to the concern that analysts cherry-pick regressors in service of an unscientific agenda (See, for example, Leamer's (1983) discussion of Ehrlich's (1975) influential analysis of the effects of capital punishment on homicide rates). The "post double selection" (PDS) lasso estimator introduced by Belloni, Chernozhukov and Hansen (2014*b*) can address this concern. Lasso (Tibshirani, 1996), which abbreviates the "least absolute shrinkage and selection operator," is a form of penalized regression that improves out-of-sample prediction by discarding some regressors and shrinking the coefficients on those retained. Estimators that use lasso solely for variable selection are said to be "post-lasso." The PDS procedure estimates causal effects in two steps. First, lasso is used to determine which covariates predict outcomes and which covariates predict treatment. The treatment effect is then estimated in a second step that includes the union of post-lasso controls selected for the outcome and treatment models as covariates in a conventional regression.

The value of ML for sensitivity analysis emerges when we use PDS to select the control variables characterizing sets of colleges to which members of the DK02 sample had applied and been accepted. Although the number and identity of lasso-chosen controls changes as we change the details required for lasso implementation, OLS estimates with ML-chosen controls robustly replicate earlier estimates showing null returns to elite or private college attendance. These encouraging findings should not be taken as suggesting ML creates valid conditional independence restrictions. Rather, ML tools seem helpful for choosing between alternative specifications that implement a common underlying conditional independence claim. The DK02 study, for example, is grounded in the idea that the characteristics of colleges to which applicants apply signal their ambition, while the set of schools to which they're admitted indexes their ability. This is a compelling but incomplete identifying assumption: in practice, schools can be described in many ways.[1]

Our second ML domain is the choice of instrumental variables. Use of ML for instrument selection is often motivated by the fact that 2SLS estimates in heavily over-identified models are biased. And 2SLS estimation is infeasible when the instrument set exceeds the sample size. ML would seem to provide a useful guide to instrument selection in the face of these problems, whittling a large set of potential instrumental variables down, keeping only those with a strong first stage. Motivated by this idea, theoretical work by Belloni et al. (2012); Carrasco (2012); Hansen and Kozbur (2014); Hartford et al. (2016) and others considers regularized models like lasso for first-stage estimation. We explore a pair of over-identified IV applications that would seem to have a role for ML-based instrument selection (though the settings considered here have far fewer instruments than observations). In contrast with encouraging findings on the utility of ML for selection of OLS control variables, our findings for instrument choice are mostly negative.

In simulations derived from the Angrist and Krueger (1991) (AK91) data, 2SLS estimation using a post-lasso first stage often improves on conventional 2SLS estimators using all available instruments, especially when lasso uses a plug-in rather than a cross-validated penalty. Lasso with a cross-validated penalty performs about like conventional 2SLS. However, the Angrist and Krueger (1995) split-sample IV (SSIV) estimator, an improved jackknife IV (IJIVE) estimator introduced by Ackerberg and Devereux (2009), and the usual limited information maximum likelihood (LIML) estimator are almost always better than 2SLS estimators using a post-lasso first stage, no matter how the lasso penalty is chosen.[2] These findings can be explained by the fact that approximate sparsity, a key lasso assumption, requires the unknown population first stage to have few parameters relative to sample size. In the applications we have in mind, the finite-sample behavior of IV estimators is better described by a Bekker (1994) many-instrument asymptotic sequence that fixes the ratio of

[1] Urminsky, Hansen and Chernozhukov (2016) discuss the value of PDS for principled variable selection. Empirical work using ML for the selection of controls includes Goller et al. (2019), which explores propensity score matching with an ML-based propensity score estimate. See also Lee, Lessler and Stuart (2010) for an earlier effort in the same vein. In a Monte Carlo study, Knaus, Lechner and Strittmatter (2018) compare ML-based estimates of individual average treatment effects, focusing on effect heterogeneity. We discuss a related paper by Wuthrich and Zhu (2019) below.

[2] In a scenario that includes some pure noise instruments, SSIV is itself exceptionally noisy. But IJIVE and LIML still out-perform ML procedures in this case.

sample size to first-stage parameters (see, for example, Angrist, Imbens and Krueger 1999). The Bekker sequence is not approximately sparse.

2SLS with an ML-chosen first stage also disappoints in a reexamination of the instrument selection strategy used by Gilchrist and Sands (2016). This study uses lasso to pick instrumental variables for the effect of a movie's opening-weekend viewership on subsequent ticket sales. ML is unimpressive here in spite of the fact that the first stage seems reasonably sparse. The potential drawbacks of ML for instrument choice are anticipated in part by Belloni et al. (2012), Belloni, Chernozhukov and Hansen (2013), and especially Hansen and Kozbur (2014), but our conclusions are less optimistic.[3] Even in models with a mix of strong and weak IVs, where an analyst might hope that lasso favors the strong, results using a post-lasso first stage exhibit substantial bias. Moreover, this bias is aggravated by the pre-testing of first-stage estimates implicit in lasso.[4]

Our exploration of the consequences of ML-based instrument selection shows that the disappointing performance of ML-augmented IV estimators is not unique to lasso. Random forest is an ML-inspired matching estimator that builds on the idea of regression trees. Ash et al. (2018) and Chernozhukov et al. (2018*a*) use random forest and related methods to select instruments for IV estimators of heavily over-identified models. When random forest methods are used to estimate first stage fitted values in the AK91 data, the resulting IV estimates are indistinguishable from 2SLS estimates with a saturated many-instrument first stage.

Our third domain concerns the selection of control variables in IV models with few instruments. This includes applications like Angrist and Evans (1998), which estimates causal effects of childbearing on mothers' labor supply using use twin births and sibling sex composition as a source of exogenous variable in family size. These just-identified IV estimates are made more plausible by conditioning on maternal characteristics (twin birth rates, for exam-

[3] Summarizing an analysis of the AK91 data, for example, Belloni, Chernozhukov and Hansen (2013) conclude that "The results in Table 5 are interesting and quite favorable to the idea of using lasso to perform variable selection for instrumental variables." Hansen and Kozbur (2014) note the poor performance of post-lasso IV in the absence of approximate sparsity, including the potential for pretest bias, but this work comments more on precision than bias. Hansen and Kozbur (2014) discuss a regularized jackknife IV estimator (JIVE) of the sort discussed by Angrist, Imbens and Krueger (1999), but omit LIML.

[4] Hall, Rudebusch and Wilcox (1996) appear to be the first to note the likelihood of bias from a pretested first stage. Andrews, Stock and Sun (2019) demonstrate the relevance of pretest bias in a simulation study based on articles appearing in the *American Economic Review*.

ple, are correlated with maternal age and schooling). Our exploration of this idea, inspired by Athey, Tibshirani and Wager's (2019) use of the Angrist and Evans (1998) data to select among high-dimensional controls and to model treatment effect heterogeneity, shows how random forest procedures founder when confronted with models that require an additive first stage for identification. The worst-case scenario here is an estimator with algorithmically-induced exclusion restrictions that yield meaningless yet statistically significant second-stage estimates.

2 Casting Regression in Two Roles

Regression uses linear models to describe conditional expectation functions (CEFs). The conditional expectation of a random variable, denoted Y_i for person i, as a function of data on a set of variables, X_i, can be written $E[Y_i|X_i = x]$. The symbol "E" in this expression denotes a population average, while the notation $E[Y_i|X_i = x]$ denotes the average of Y_i for everyone in a population of interest who has characteristics X_i equal to a particular value, x. For example, labor economists have long been interested in how much average (log) wages rise with each additional year of schooling. We compare $E[Y_i|X_i = 16]$, the average wages of on-time college graduates, with $E[Y_i|X_i = 12]$, the earnings of high school graduates. The notation $E[Y_i|X_i]$ represents population mean Y_i for any value in the support of random variable X_i. Written this way, the CEF is random because X_i is random.

Because $E[Y_i|X_i = x]$ takes on as many values as there are choices of x, labor economists and others doing applied econometrics often aspire to simplify or approximate the CEF, so as to highlight or summarize important features of it. The regression of Y_i on X_i does this by providing the best linear approximation to the CEF. Formally, given a set of $k = 1, ..., K$ explanatory variables, X_i, the $K \times 1$ regression slope vector, β, can be defined as the minimum mean squared error (MMSE) linear approximation:

$$\beta = \arg\min_b E\left[\{E[Y_i|X_i] - X_i'b\}^2\right] = E[X_iX_i']^{-1}E[X_iY_i]. \tag{1}$$

Moreover, if the CEF is indeed linear, then regression finds it.

The contemporary ML agenda is more likely to use data on schooling to predict *individual* earnings than to approximate the CEF. But the law of iterated expectations implies that the best (MMSE) linear predictor of Y_i coincides with the best linear approximation to $E[Y_i|X_i]$. That is,

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i] = \arg\min_b E\left[\{Y_i - X_i' b\}^2\right]. \tag{2}$$

The distinction between CEF approximation and individual prediction is therefore of no consequence for *parameters*: the regression slope vector that approximates the CEF also provides the best linear predictor of Y_i given X_i. The ordinary least squares (OLS) estimator, denoted here by $\hat{\beta}_{LS}$, replaces expectations with sums in (2) and provides the best linear predictor in the sample in which it's fit.

There seems to be little daylight between predictive regression and econometric regression models motivated by an interest in conditional distributions. A gap opens, however, when an analyst aspires to use regression to generate predictions in *new* data. Assuming $\hat{\beta}_{LS}$ is computing using data on the first n observations only, the regression prediction of Y_{n+1} given X_{n+1} is $\hat{y}_{n+1} = X_{n+1}' \hat{\beta}_{LS}$. Even in the realm of linear models, $\hat{y}_{n+1}$ is not the last word in *out-of-sample* prediction.

A better out-of-sample predictor augments the least squares minimand, (2), with a regularization term that favors smaller coefficients and lower-dimensional models over an unrestricted OLS fit. Much of the ML toolkit can be said to consist of prediction augmented by regularization. Ridge regression, introduced by Hoerl and Kennard (1970), is an early version of this idea: the ridge regularization term is the sum of squared regression coefficients. Lasso, a method associated with contemporary ML, regularizes by including the sum of the absolute value of coefficients in the estimation minimand. The family of widely-used regularized regression estimators can be defined as solving

$$\min_b E\{Y_i - X_i' b\}^2 + \lambda ||b||_q^q, \tag{3}$$

where λ is a user-chosen tuning parameter and $||b||_q = (\sum_k |b_k|^q)^{1/q}$. Ridge sets $q = 2$; lasso sets $q = 1$. A best subset estimator is obtained by letting $q \to 0$ (since the resulting estimand penalizes the number of non-zero coefficients).[5]

A second gap between the econometric and predictive ML frameworks arises from the asymmetry with which most empirical Labor views regressors. The modern empirical paradigm usually distinguishes between the components of X_i: one is a causal variable of interest, the

[5] Abadie and Kasy (2018) compare the Bayes risk of alternative estimators in this family.

rest, a set of supporting controls whose coefficients are of little interest. An empirical example highlights the significance of this distinction.

When Regression Reveals Causal Effects

Adapting the pioneering study by Dale and Krueger (2002), Angrist and Pischke (2015) ask whether it pays to attend a private university like Duke instead of a state school like the University of North Carolina. Is the money spent on private college tuition justified by future earnings gains? The causal regressor here is a dummy variable that indicates graduate i attended private college, denoted by D_i. Control variables are represented by a vector, A_i. The outcome of interest, Y_i, is a measure of earnings roughly 20 years post-enrollment. Our sample consists of the College and Beyond survey data analyzed in Dale and Krueger (2002).

The causal relationship between private college attendance and earnings can be described in terms of potential outcomes: Y_{1i} represents the earnings of individual i were he or she to go private ($D_i = 1$), while Y_{0i} represents i's earnings after a public education ($D_i = 0$). The causal effect of attending a private college is the difference, $Y_{1i} - Y_{0i}$. We see only Y_{1i} or Y_{0i}, depending on the value of D_i. The analyst therefore aspires to measure an average causal effect like $E\left[Y_{1i} - Y_{0i}\right]$, or an effect conditional on treatment, $E\left[Y_{1i} - Y_{0i} | D_i = 1\right]$.

The link between causal inference and regression is facilitated by a constant-effects framework that highlights the problem of selection bias, glossing over the distinction between different sorts of causal averages. The constant-effects causal model can be written:

$$\begin{aligned} Y_{0i} &= \alpha + \eta_i && (4) \\ Y_{1i} &= Y_{0i} + \rho, && (5) \end{aligned}$$

where the first equation defines α to be the mean of Y_{0i} and the individual deviation from this mean to be η_i. The second line says that the causal effect, $Y_{1i} - Y_{0i}$, is a constant, ρ. Using the fact that observed outcomes are related to counterfactual outcomes by

$$Y_i = Y_{0i} + \left(Y_{1i} - Y_{0i}\right) D_i,$$

we can use the constant-effects model to write

$$Y_i = \alpha + \rho D_i + \eta_i. \qquad (6)$$

Equation (6) casts the problem of selection bias in terms of η_i, which looks like a regression error term. Unlike regression residuals, however, which are uncorrelated with regressors by definition, η_i is correlated with D_i.

Regression-based solutions to the problem of selection bias begin with a key conditional independence assumption. Specifically, causal claims for regression are founded on the assumption that

$$E\left(\eta_i|D_i = 1, A_i = a\right) = E\left(\eta_i|D_i = 0, A_i = a\right), \tag{7}$$

where A_i is a vector of m control variables and a is a particular value of A_i. In other words, in the population with $A_i = a$, the private and public earnings comparison is an apples-to-apples contrast. This *ceteris paribus* claim can be written compactly as:

$$E\left(\eta_i|D_i, A_i\right) = E\left(\eta_i|A_i\right). \tag{8}$$

Controls satisfying 8 must be "pre-treatment variables," that is, they cannot themselves be outcomes. In the Dale and Krueger (2002) empirical strategy, the control vector A_i identifies the sets of schools to which the college graduates in the sample had applied and were admitted. Equations (7) and (8) say that, conditional on having applied to Duke and UNC, and having been admitted to both, those who chose Duke have the same average *potential* earnings as those who went to UNC. Dale and Krueger (2002) and Angrist and Pischke (2015) provide evidence in support of this claim.

The final element of the causal regression story is the assumption that the conditional mean of Y_{0i} is a linear function of A_i :

$$E\left(\eta_i|A_i\right) = \gamma' A_i. \tag{9}$$

This implies

$$\eta_i = \gamma' A_i + \varepsilon_i,$$

where it's surely true that

$$E\left(\varepsilon_i|A_i\right) = 0. \tag{10}$$

Combining these pieces generates a linear CEF with a causal interpretation:

$$\begin{aligned} E\left(Y_i|A_i, D_i\right) &= \alpha + \rho D_i + E\left(\eta_i|A_i\right) \\ &= \alpha + \rho D_i + \gamma' A_i. \end{aligned}$$

The regression model,

$$Y_i = \alpha + \rho D_i + \gamma' A_i + \varepsilon_i, \tag{11}$$

can therefore be used to construct unbiased estimates of the causal effect of interest, ρ. The control coefficient vector, γ, need not be economically interesting, but may provide diagnostic information useful for assessing the plausibility of (8).[6]

Private college alumni earn more than those who went public. Remarkably, however, a set of well-chosen controls serves to eliminate evidence of an elite college premium based on uncontrolled comparisons. This can be seen in Panel A of Table 1. A private schooling earnings premium of around 21 log points estimated with no controls (reported in the first column of the table) falls to a still-substantial 14 points (reported in the second column) when estimated with ten controls for applicant ability, like SAT scores and class rank, and family background in the form of parents' income. In contrast with the substantial private college premia reported in the first two columns, however, estimates in columns 3-4 show that, conditional on controls for the selectivity of schools to which graduates had applied and been admitted, the private premium falls to zero.

The choice of selectivity controls used to compute the estimates reported in columns 3-4 of Table 1 is motivated by the idea that, within each selectivity group, students are likely to have similar educational and career ambitions, while they were also judged similarly capable by college admissions staff. Within-group comparisons should therefore be considerably more apples-to-apples than uncontrolled comparisons involving all students. Because there are many unique combinations of application and admissions choices, it's helpful to group similarly selective schools like Princeton and Yale together. The models used to construct the estimates in columns 3-4 therefore control for sets of schools grouped by their Barron's selectivity (Barron's magazine groups schools into 6 selectivity groups). This model can be written

$$Y_i = \alpha + \rho D_i + \underbrace{\delta_0' C_i + \sum_{j=1}^{150} \delta_j GROUP_{ji}}_{\gamma' A_i} + \varepsilon_i, \tag{12}$$

where $\{GROUP_{ji}; j = 1, ..., 150\}$ is a set of dummy variables indicating application and admission to a particular configuration of Barron's selectivity groups, with group coefficients denoted δ_j (for 151 groups with variation in private college attendance). The vector C_i

[6] The utility of regression for causal inference is not limited to models with constant effects. Provided the parameterization of A_i is suitably flexible (as with sets of dummy variables for categorical controls), the OLS estimand is a weighted average of control-specific average causal effects (see, for example, Angrist and Krueger (1999) for details).

contains the additional controls used to construct the estimates reported in column 2. The full set of controls in A_i includes C_i and the selectivity group dummies, though, conditional on the latter, the former may be unnecessary.[7]

Estimates of equation (12) suggest that the earnings premium enjoyed by private college grads reflects the high Y_{0i}'s of those who aim higher and are more attractive to admissions officers, rather than capturing a causal effect of private attendance. The similarity of the estimates in columns 3 and 4 also shows this conclusion to be unaffected by adjustment for further controls. Since applicant characteristics like SAT score and family background are strongly predictive of earnings, the similarity of the estimated private school effects across these two columns has the implication that, after conditioning on the selectivity groups indicated by $\{GROUP_{ji}; j = 1, ..., 150\}$, private attendance must be uncorrelated with applicant characteristics.[8] Following a brief digression, we use ML to explore the robustness and sensitivity of this finding.

2.1 It's Only Fitting

Students of econometrics learn to distinguish between slope coefficients and goodness of fit. The latter is usually measured by

$$R^2 = 1 - \frac{S(Y_i - X_i'\hat{\beta}_{LS})}{S(Y_i - m_y)},$$

where $S(\cdot)$ denotes a sample sum of squares and m_y is the sample mean of Y_i. By minimizing $S(Y_i - X_i'\hat{\beta}_{LS})$, the OLS estimator maximizes in-sample R^2, while many ML methods aim to boost out-of-sample R^2.

How important is fit as a *research* goal? The quality of the individual predictions generated by models linking private college attendance with future wages is summarized by the R^2

[7] College selectivity categories are determined by Barron's Profiles of American Colleges 1978. For example, one of the selectivity groups coded by $\{GROUP_{ji}; j = 1, ..., 150\}$ indicates those who applied to one highly competitive school and two competitive schools, and were admitted to one of each. Our sample consists of people from the 1976 college-entering cohort who appear in the College and Beyond survey and who were full-time workers in 1995. The analysis excludes graduates of historically black schools and is further restricted to applicant selectivity groups containing some students who attended public universities and some students who attended private universities. The dependent variable is the log of pre-tax annual earnings in 1995. Regressions are weighted to make the sample representative of the population of graduates of 30 College and Beyond institutions. 68.6 percent of the sample with Barron's matches attended a private school.

[8] Angrist and Pischke (2015) estimate this piece of the omitted variables bias (OVB) formula directly.

statistics reported in Table 1. The proportion of earnings variance explained by covariates ranges from tiny (just under 2%) to modest (almost 14%). Models that control for variables like GPA and family background yield the best fit, while those controlling for selectivity groups but omitting these personal characteristics (reported in column 3) have an R^2 of only around 6%. Still, the estimates generated by the latter likely provide a more useful guide to the economic returns to private education than does the estimate associated with a higher R^2 in column 2. This is is evident in that fact that, once we control for selectivity groups, remaining covariates are uncorrelated with elite school attendance (a result detailed in Angrist and Pischke 2015).

Beyond the DK02 study that we expand on here, Alan Krueger's many path-breaking empirical contributions testify to the primacy of causality over fit in empirical labor economics. His work shows, for example, that company-owned fast-food franchises pay their workers more than franchisee-owned stores (Krueger, 1991), suggesting a role for efficiency wages in the low-wage labor market; that workers with computer skills earn substantially more and receive a higher rate of return to their schooling than other workers (Krueger, 1993), illuminating the theory of skill-biased technical change; that IV estimates of the effect of schooling on wages are close to the corresponding OLS estimates (Angrist and Krueger, 1991), suggesting there's little OVB in Mincerian wage equations; that minimum wages don't appear to depress employment (Card and Krueger, 1994; Katz and Krueger, 1992), prompting a re-think of the competitive labor market paradigm; and that workers who attended schools with more resources earn more (Krueger, 1999; Card and Krueger, 1992*a*,*b*).

These influential empirical findings provide evidence on causal relationships that are central to labor economics, yet unrelated to regression R^2. Whether measured in-sample or out-of-sample, labor R^2s are mostly pitiable. Replacing OLS with a penalized estimator like ridge or lasso will almost certainly improve the out-of-sample fit associated with estimates like those in Table 1. An analyst who seeks only to predict the wages of college graduates might be led (by lasso, say) to dispense with variables describing college characteristics, individual computer skills, and high school resources. Other variables, such as where sample respondents live and work, are likely more important. Yet, this misses the point of any causal inquiry.

3 Welcome to the Machine

3.1 ML Picks OLS Controls

Predictive ML fails to discriminate between causal and control variables, but economists using ML are free to draw such distinctions. In an important econometric extension of the ML toolkit, Belloni, Chernozhukov and Hansen (2014*b*) introduce the method of *post double selection* (PDS), an empirical strategy that uses lasso to pick regression control variables like A_i, given interest in a particular causal effect, like that of private college attendance. PDS builds on Robinson (1988)'s partially linear model, which seeks to identify an additive causal effect using flexible and possibly nonlinear controls.[9] The DK02 research design, potentially involving hundreds of control variables, seems like a promising test bed for the PDS framework.

Returning to the simple causal structure embodied in (4) and (5), the identifying assumption motivating PDS can be stated as

$$E\left(\eta_i|D_i, A_i\right) = E\left(\eta_i|A_i\right) = \gamma_0' A_i + r_{0i}, \tag{13}$$

where A_i is now a vector of m controls with m potentially larger than the sample size, n, and r_{0i} is an approximation error that is small in a sense made precise in Belloni, Chernozhukov and Hansen (2014*b*). The full set of m covariates is sometimes said to make up a "dictionary" of possible controls. Importantly, the Belloni, Chernozhukov and Hansen (2014*b*) framework maintains the hypothesis that we observe any variables needed to support identification under equation (13). The fact that γ_0 is non-zero for only a subset of controls is the key PDS assumption of approximate sparsity: the number of *possible* controls may exceed the sample size, but the model of interest is well-approximated by $s_0 < n$ non-zero elements. ML uses approximate sparsity for estimation when the identity of the *specific* controls needed to support a conditional independence assumption is unknown.

PDS also includes a linear model for the propensity score, that is, for the conditional probability of treatment given A_i. We write this as

$$E[D_i|A_i] = \gamma_1' A_i + r_{1i}, \tag{14}$$

[9] Hill (2011) uses high-dimensional Bayesian regression trees to separately impute $E\left(Y_{1i}|A_i\right)$ and $E\left(Y_{0i}|A_i\right)$ when estimating average treatment effects.

where γ_1 likewise has $s_1 < n$ non-zero elements and r_{1i} is approximation error. PDS regularity conditions also require the propensity score to lie strictly between zero and one, at least for some values of A_i, so that treatment status varies conditional on A_i. Finally, the union of the sets of covariates with non-zero coefficients in either (13) or (14) is assumed to contain a total of $s < n$ unique variables (as a theoretical matter, s may be an increasing function of sample size).

Faced with an abundance of candidate controls, PDS finds a list of variables adequate to control OVB, while rendering causal inference feasible. This search can be formulated as a model selection problem in the context of a two-equation system,

$$Y_i = \gamma_0' A_i + \rho D_i + r_{0i} + \varepsilon_i \tag{15}$$

$$D_i = \gamma_1' A_i + r_{1i} + \nu_i, \tag{16}$$

where error terms ε_i and υ_i are mean-independent of regressors (implied by (13) and (14)) and ρ in the first equation is a constant additive causal effect.[10]

Substituting equation (16) into equation (15) generates a reduced form regression of outcomes on controls,

$$Y_i = \gamma' A_i + r_i + \zeta_i, \tag{17}$$

where $\gamma = \gamma_0 + \rho\gamma_1$,$r_i = r_{0i} + \rho r_{1i}$, and ζ_i is a residual that's mean-independent of A_i. The PDS procedure starts by fitting lasso regression models to both (16) and (17). Lasso is a penalized regression estimator that minimizes equation (3) with $q = 1$. Lasso deletes some variables from the covariate dictionary, while shrinking the coefficients on those retained towards zero. PDS ignores lasso shrinkage, using lasso only as a model selection device. OLS estimation of a model including only the variables retained by lasso is called "post-lasso" estimation. Let M_i denote the union of control variables selected by lasso estimation of the propensity score and reduced form. The PDS estimator of ρ is the coefficient on D_i in

$$Y_i = \pi' M_i + \rho D_i + \xi_i, \tag{18}$$

where ξ_i is regression residual. Belloni, Chernozhukov and Hansen (2014*b*) give conditions under which the resulting estimates of ρ are consistent and asymptotically normal.

[10] Belloni, Chernozhukov and Hansen (2014*b*) link PDS with Robinson (1988), which specifies potentially nonlinear functions $m_0(A_i)$ and $m_1(A_i)$ in place of $\gamma_0' A_i$ and $\gamma_1' A_i$. The linear model described by (15) and (16) is then seen as approximating these functions in an asymptotic sequence that shrinks approximation error as the sample size grows.

3.2 Elite College Effects

The estimates in Table 1 suggest that, conditional on the Barron's categories of the colleges to which graduates had applied and been admitted, private college attendance is unrelated to earnings. But control for OVB using 150 Barron's dummies leaves around 5,600 observations, down from over 14,000 observations in the full College and Beyond sample of graduates with earnings. The problem here is that many selectivity groups are populated by sets of graduates that uniformly attended a public or private college. OLS estimation of a model including the set of Barron's dummies is implicitly a panel-data-style "within estimator" that drops observations on regressors in covariate cells where the regressor D_i is fixed at zero or one, that is, cells with a degenerate propensity score. The sample that can be used to estimate (12) need not be representative of the population covered by the College and Beyond survey.

PDS winnows a large dictionary of potential control variables, retaining variables that seem likely to mitigate omitted variables bias. It's worth noting, however, that while a post-lasso algorithm applied to the set of Barron's dummies included in equation (12) may drop some of these dummies, it won't *combine* them. Rather, by dropping dummies that are deemed unnecessary, post-lasso estimators expand the reference group for the set of dummies retained. Suppose, for example, that applicants apply to and are admitted to one of three sets of schools. This scenario generates two Barron's dummies, plus a reference group. Omitting one dummy pools this group with the original reference group. Likewise, lasso won't pool groups of applicants with a degenerate probability of assignment in a manner that makes such groups informative about treatment effects. Recognizing these problems, analysts have proposed lasso-type strategies that penalize *differences* in coefficients like the δ_js in (12). But such methods (known as fused lasso) seem unlikely to be attractive for models with categorical control variables indicating many categories.[11]

In addition to saturated control for 151 Barron's selectivity groups, the DK02 study

[11] The cardinality of the set of all possible subsets of a set with J elements is 2^J. For variables like the Barron's categories used to compute the estimates in Table 1, the number of parameters required to model all possible dummy coefficient differences far exceeds the rate required for approximate sparsity (as described by Belloni, Chernozhukov and Hansen (2014*b*)) and is arguably also larger than that required for the "many covariates" asymptotic sequence discussed by Cattaneo, Jansson and Newey (2018*a*). Note also that lasso estimators, like ridge estimators, are sensitive both to the choice of omitted group when categorical variables are coded with dummies, and to regressor scale. The lasso routines used here standardize regressors, in some cases employing the regressor-specific penalty loadings detailed in Belloni, Chernozhukov and Hansen (2014*b*).

explores a parsimonious control strategy that conditions only on the average SAT score of the schools to which graduates applied, plus dummies for the number of schools applied to (specifically, three dummies indicating those who applied to two, three, and four or more schools). The DK02 paper labels this specification, which can be estimated in the full C&B sample, a "self-revelation model." The self-revelation model is motivated by the hypothesis that college applicants have a pretty good idea of the sort of schools within their reach, and of the set of schools where they're likely to be well-matched. An applicant's self-assessment is reflected in the average selectivity of the schools they've targeted, while the number of applications submitted is a gauge of academic ambition.

The small set of DK02 self-revelation controls yields a model suitable for estimation in the full sample. But the set of controls used by this strategy could just as well have been something else, perhaps characteristics of the most or least selective school to which applicants applied instead of the average. This motivates an exploration of alternative parsimonious control schemes, reported in Table IV of DK02. ML methods–and PDS in particular–seem useful for a more systematic exploration of the sensitivity of causal conclusions when many equally plausible control variables are at hand.

Columns 5 and 6 in Panel A of Table 1 report estimates of private school effects from the self-revelation model. When estimated using the sample for which we can control for Barron's matches, the self-revelation model likewise generates a small and statistically insignificant private school effect (specifically, 0.036 with a standard error of 0.029). Moreover, as can be seen in column 6, the estimated private attendance effect is almost unchanged when the self-revelation model is estimated using the full College and Beyond sample rather than the sample with Barron's matches.

The DK02 study focuses on a continuous measure of college selectivity–the average SAT score of students enrolled at the college attended–rather than a dichotomous private attendance variable. Although PDS is motivated as a strategy for estimation of dichotomous treatment effects, the logic behind it applies to models with continuous causal regressors (Belloni, Chernozhukov and Hansen (2014*b*) evaluate PDS in a simulation study involving a normally distributed regressor). As a benchmark for ML estimates of average SAT effects, Panel B of Table 1 reports estimates of the earnings gain generated by attendance at a more selective school (columns 2 and 4 of this panel replicate results reported in DK02). Without controls, each 100 point increment in *alma mater* selectivity is associated with around 11% higher earnings among graduates, a substantial gap that falls to a still-significant 7.6% when estimated with controls for individual characteristics like SAT scores and class rank. As with the private earnings premium, however, the estimates reported in columns 3 and 4 of Panel B suggest that college selectivity is unrelated to earnings when SAT effects are estimated with ability and ambition controls in the form of dummies for Barron's selectivity groups. Likewise, as can be seen in column 5, self-revelation controls serve to eliminate college selectivity effects. Finally, the estimates in column 6 show that this conclusion holds in the full College and Beyond sample.

We also consider effects of a third treatment variable, dichotomous like the private attendance dummy, but measuring college selectivity like average SAT scores. This is a dummy for schools that Barron's ranks as being "highly competitive" or better (HC+). Roughly 73% of the full College and Beyond sample attended HC+ schools, close to the 72% who attended a private school (The private and HC+ dummies differ for 13% of the College and Beyond sample). As can be seen in Panel C of Table 1, the premium associated with HC+ attendance is close to that associated with private attendance. Moreover, like the estimated private college effects reported in Panel A, the HC+ effect falls but remains substantial when estimated with controls for a few individual characteristics. Finally, as with the private and selective college estimates in Panels A and B, the HC+ effect disappears conditional on dummies for Barron's selectivity groups and when estimated with self-revelation controls in the Barron's-group sample. Interestingly, however, self-revelation estimates computed using the full sample fail to replicate the statistical zeroes reported in columns 3 and 4. Rather, the estimated premium for HC+ attendance reported at the bottom of column 6 is a marginally statistically significant 0.068.

PDS-Supported Sensitivity Analysis

Our PDS estimator for private college effects begins with a dictionary containing 384 possible control variables, including the personal characteristics used for column 2 and the self-revelation controls used for column 4. This dictionary omits dummies for Barron's selectivity groups, relying on coarser summary statistics to describe the colleges to which graduates applied and were accepted. Specifically, the dictionary adds the number of colleges applied to; indicators for being accepted to one, two, three or four or more colleges; indicators for being rejected from one, two, three or four or more colleges; mean SAT scores at the most selective school, least selective school, and for all schools where the applicant was accepted; mean SAT scores at the most selective school, least selective school, and for all schools where the applicant was rejected, and all two-way interactions and squared terms associated with the underlying list of possible controls, except for squares of dummy variables, which are redundant. We see this dictionary as encompassing a wide range of alternatives to the self-revelation model.

PDS estimates of private college effects, reported in the first three columns of Table 2, are mostly similar to the corresponding estimates computed in models with Barron's dummies and self-revelation controls. For example, using a plug-in penalty computed by Stata 16's `lasso linear` command, the PDS-estimated private attendance effect is 0.038 with a standard error of 0.04. This is generated by a model that retains 18 controls. The plug-in penalty used to compute this estimate, based on a formula in Belloni, Chernozhukov and Hansen (2014*b*), is data-driven, though not cross-validated. As can be seen in column 2, a cross-validated penalty retains far more controls (100), but yields similar estimates. An alternate procedure, `cvlasso`, part of a set of Stata routines called Lassopack (Ahrens, Hansen and Schaffer 2019) adds a few more controls (for a total of 112), but again yields similar estimated private school effects. These results appear in column 3.[12]

The tendency for cross-validation to produce smaller penalties (and hence to include more controls) also surfaces in results reported by Chetverikov, Liao and Chernozhukov (2019). This is a caution for practitioners: implementation details may matter in some applications, even if not in our Table 2. Other relevant computational considerations include the use of regressor-specific penalty loadings, choice of software, and options affecting cross-validation. In view of the increasingly wide variety of lasso estimation routines, an Appendix details our choices in this regard. Appendix Table A1 also compares elite college effects computed using alternative implementations beyond those used for the estimates reported in Table 2. With one important exception, these are qualitatively similar to the estimates reported in Table 2.

[12] The Belloni et al. (2012) and Belloni, Chernozhukov and Hansen (2014*b*) plug-in penalties generalize the penalty formula proposed by Bickel et al. (2009). The plug-in penalty requires two user-specified constants, c and γ, which we set at the rlasso (Ahrens, Hansen and Schaffer (2019)) defaults ($c = 1.1$ and $\gamma = .1/log(n)$). Belloni, Chernozhukov and Hansen (2014*b*) suggest using $c = 1.1$ and $\gamma = .05$. These choices may also be consequential for the number of controls retained by lasso.

PDS estimates of the effect of elite college attendance as measured by school average SAT scores are reported in the first three columns of Panel B in Table 2. These are close to zero and about as precise as the benchmark full-sample estimates of average SAT effects reported in Table 1. In this case, PDS estimation with plug-in, cross-validated and `cvlasso` tuning parameters retains 24, 151, and 58 controls, respectively. The wide variation in control variable choice induced by changes in tuning parameters is an important caution for researchers looking to interpret coefficients on the control variables themselves. But this variation also suggests that the findings in Table 1 should not be seen as the product of a judicious specification search. Finally, as with the benchmark self-revelation estimates of HC+ attendance effects reported in column 5 of Table 1, PDS estimates of HC+ effects reported in Panel C of Table 2 show positive effects, two of which are marginally significant. Again, tuning parameter choice generates considerable variation in controls, but this variation is not reflected in estimates of the causal effect of interest.

How important is *double* selection in this context? Results in columns 4-6 of Table 2 are from a procedure applying lasso to the reduced-form regression of the outcome variable on controls (the reduced form excludes the treatment variable). This single-selection estimator naturally relies on fewer controls than does PDS. Outcome-only selection of controls also generates somewhat larger HC+ effects. In contrast with the rest of Table 2, the impression left by columns 4-6 of Panel C is one of significant effects on the order of 0.08. For theoretical reasons detailed by Belloni, Chernozhukov and Hansen (2014*b*), the smaller PDS estimates (with similar standard errors) are likely to be more reliable. Reinforcing this conclusion, outcome selection using an alternative plug-in penalty yields a model with only a single control and an outlying estimated HC+ effect of 0.22. At the same time, single selection applied to the propensity score with this penalty generates an estimate with a standard error almost 50% larger than that of the corresponding PDS estimate. These results appear in Appendix Table A1 (see columns 1, 5, and 9 in Panel C of this table).

A conventional, ML-free approach to probing the sensitivity of regression estimates simply widens the set of controls. Column 7 of Table 2 reports estimates and standard errors of the effect of elite college attendance from models that include the full set of controls in the dictionary underlying lasso. Because some controls are linearly dependent, the model used to construct these estimates retains 303 of 384 controls in the dictionary. Full-dictionary control is feasible here because the dictionary of controls is not truly high-dimensional in the sense of containing as many variables as there are observations. As it turns out, the full-control estimates in column 7 are similar to those generated by PDS.

An empirical example does not make a theorem, of course. Wuthrich and Zhu (2019) use a mix of simulation evidence and theory to show that PDS bias mitigation depends on design features like regressor variance and the extent of OVB. Moreover, we've examined a scenario in which OLS with full-dictionary control is feasible and effectively removes OVB. Even so, PDS seems a useful tool for sensitivity analysis in a regression context, where analysts may choose from an abundance of possible control variables. The fact that the target causal estimate remains reasonably stable even while the list of selected controls varies widely from one routine to another reinforces claims of robustness. It's worth emphasizing, however, that our causal interpretation of the ML estimates in Table 2 turns on a maintained conditional independence assumption. ML methods do not create quasi-experimental variation. Rather, ML uses data to pick from among a large set of modeling options founded on a common identifying assumption.

A further PDS plus is that in the DK02 application, the models selected save degrees of freedom, possibly increasing precision and external validity. The DK02 strategy that controls for 160 variables, including 150 application/admissions group dummies, trims a starting sample of over 14,000 for the estimation of private school effects to around 5,600. This changes the precision of estimated elite college effects little. But it still seems noteworthy that replacing group dummies with a larger set of potential controls ultimately yields a far more parsimonious setup that makes use of the entire data set. The resulting gains in sample coverage are analogous to those yielded by propensity score matching over full covariate matching when covariates are discrete and high-dimensional (as, for example, in Abdulkadiroğlu et al. 2017).

4 ML Picks Instruments

The asymptotic sampling variance of 2SLS estimates is inversely proportional to the first stage R^2, a statistic that summarizes the quality of the first stage fit to the CEF of endogenous D_i given an instrument vector, Z_i. This fact encourages the use of many instruments. On the other hand, 2SLS estimates are biased, with a finite-sample distribution shifted towards the mean of the corresponding OLS estimates. Additional instruments aggravate this bias when their explanatory power is low (see, e.g., Angrist and Krueger, 1999). This bias-variance trade-off appears to open the door to a fruitful empirical strategy that uses machine learning to select instruments. Use of ML for instrument selection is discussed and explored in work by Belloni, Chernozhukov and Hansen (2011), Belloni et al. (2012), and Mullainathan and Spiess (2017), among others.[13]

4.1 Machining the AK91 First Stage

How valuable is a machine-specified first stage for labor IV? We explore this question by revisiting Angrist and Krueger (1991), an influential IV study that uses quarter of birth (QOB) dummies as instruments to estimate the economic returns to schooling (henceforth, AK91). The QOB identification strategy is motivated by the fact that children who start school at an older age attain the minimum dropout age after having completed less schooling than those who enter school younger. Because most children start school in the year they turn six, those born later in the year are younger when school starts, and are therefore constrained by compulsory attendance laws to spend more time in school before reaching the dropout age. AK91 documents a strong QOB first stage, showing that highest grade completed increases with QOB for American men born in the 1920s and 1930s.

The AK91 endogenous variable is highest grade completed; the dependent variable is the log weekly wage in a sample of 329,509 men born between 1930 and 1939 from the 1980 Census public use files. A regression of schooling on three QOB dummies and 9 year of birth (YOB) dummies generates an F statistic for the three excluded QOB instruments of around 36. 2SLS estimates using three QOB dummies therefore seem unlikely to suffer substantial weak-instrument bias. The many-weak instrument angle surfaces when QOB dummies are interacted with dummies for year of birth (YOB) and place (state) of birth (POB). These interactions are motivated by the fact that the relationship between QOB and schooling varies

[13] Okui (2011) and Carrasco (2012) appear to be the first explorations of ridge-type regularized IV as a solution to the weak instruments problem. Carrasco and Tchuente (2015) discuss regularized LIML. Hansen and Kozbur (2014) regularize the Angrist, Imbens and Krueger (1999) jackknife IV estimator. Donald and Newey (2001) truncate an instrument list based on approximate mean squared error. Chamberlain and Imbens (2004) introduce a random effects procedure for models with many weak instruments that is closely related to LIML

both across cohorts, as compulsory attendance laws have grown less important, and across states, since states set school attendance policy. Interacting three QOB dummies with nine YOB and 50 POB dummies generates 180 excluded instruments.[14] The first-stage F statistic in this case (controlling for additive YOB and POB main effects) falls to around 2.6. As first noted by Bound, Jaeger and Baker (1995), this many-weak first stage may generate estimates of the economic returns to schooling that are close to OLS solely by virtue of finite-sample bias. A fully interacted QOB-YOB-POB first stage has 1530 instruments. The first stage F statistic in this case falls below 2, so the potential bias of 2SLS here is even larger.

As in the previous section, our framework for instrument selection maintains the underlying identifying assumptions that motivate IV estimation. In particular, we do not aspire to find valid instruments, but rather to choose among them. We assess the consequences of instrument choice for the bias and dispersion of the resulting IV estimates; problems of statistical inference are left for future work. This investigation begins by examining an ML strategy in which a conventional 2SLS second stage is estimated using the instrument set retained by a post-lasso first stage, as suggested by Belloni et al., 2012.

Lasso for instrument selection is evaluated in a simulation experiment calibrated so that OLS estimates are misleading. In the absence of omitted variables or endogeneity bias in OLS estimates, it's hard to gauge the potential for finite sample bias in 2SLS estimates. For example, with a single fourth-quarter dummy as the instrument, 2SLS in the AK91 sample generates an estimated return to schooling of 0.074. The corresponding OLS estimate is 0.071.[15] This just-identified IV estimate (which, like LIML, is approximately median-unbiased) suggests OLS is a good guide to the causal effect of schooling on wages. But then we should expect OLS and 2SLS estimates to be close regardless of instrument strength (see also Cruz and Moreira (2005), which argues that even heavily over-identified AK91 estimates have little bias). This leads us to craft a simulation design that preserves the structure of the AK91 sample and IV estimates, but introduces substantial omitted variables bias in the corresponding OLS estimates.

Starting with the full AK91 1980 census sample, we computed average highest grade

[14] Washington, DC is treated as a state.

[15] These estimates, which include no controls, are from Table 6.5 in Angrist and Pischke (2015). The corresponding standard error is 0.028. The 2SLS estimate with three QOB dummies as instruments and YOB dummies included as controls is 0.105, with a standard error of 0.02.

completed in each QOB-YOB-POB cell (a total of 2040 means). Call these cell averages $\bar{s}(q, c, p)$ where $q = 1, ..., 4; c = 1930, ...1939; p = 1, ..., 51$. Simulated schooling, $\tilde{s}_i$, is a Poisson draw with mean μ_i, where

$$\mu_i = \max[1, \bar{s}(Q_i, C_i, P_i) + \kappa_1 \nu_i], \tag{19}$$

and variables Q_i, C_i, and P_i are i's quarter, cohort (year), and place (state) of birth. This mean is censored below at 1. The standard normal variable ν_i is multiplied by a scale parameter, κ_1 chosen to generate a first stage R^2 and partial F statistic matching those from a 2SLS procedure that uses 180 excluded instruments in the original data. This benchmark specification uses three QOB dummies interacted with 10 YOB dummies and 50 POB dummies as instruments, controlling for a full set of POB-by-YOB interactions. The original-data F-statistic in this 180-instrument model is 2.56.

Our simulated dependent variable builds on the conditional mean function generated by 2SLS estimation with 180 instruments in the AK91 sample. Specifically, let $\hat{y}(C_i, P_i)$ be the second-stage fitted value this model generates after subtracting $\hat{\rho}_{2SLS} S_i$, where $\hat{\rho}_{2SLS}$ is the 2SLS estimate of the returns to schooling and S_i is the endogenous schooling variable. The notation here reflects the fact that this estimated fitted value varies only by YOB and POB. The simulated dependent variable is then constructed as

$$\tilde{y}_i = \hat{y}(C_i, P_i) + 0.1\tilde{s}_i + \omega(Q_i, C_i, P_i)(\nu_i + \kappa_2 \epsilon_i), \tag{20}$$

where $\tilde{s}_i$ is simulated schooling drawn according to(19). The causal effect of schooling on wages is fixed at 0.1. Error term ϵ_i is standard normal, while weight $\omega(Q_i, C_i, P_i)$ is set to generate a conditional variance of residual wages in each QOB-YOB-POB cell proportional to the variance of 2SLS residuals in the original data (again, using the 180 instrument model). Finally, setting the scale parameter $\kappa_2 = .1$ generates an OLS estimand equal to 0.207, or roughly double the causal effect of interest. Each simulation sample begins with a bootstrap sample of $\{Q_i, C_i, P_i\}$ from the original data. Simulated schooling and wages are then constructed for this draw as described by equations (19) and (20).

Across 999 simulation draws, 2SLS estimates have bias around 0.04, while the bias of OLS is 0.107 by construction. Using the full set of $QOB \times YOB \times POB$ dummies as instruments (for a total of 1530 excluded instruments) increases 2SLS bias by about 50%, to 0.061. These results appear in the first two rows of Table 3, which also shows the average first-stage F-static across simulations above column headings. The Monte Carlo standard deviation of 0.011 is close to the (robust) standard error estimated for this model using the original data. Not surprisingly, moving from 180 to 1530 instruments increases precision, at the price of increased bias when instruments are added. In both models, the median absolute deviation of the 2SLS estimates (MAD, defined as the median of the absolute value of the difference between simulated estimates and the median simulation estimate) is a little lower than the corresponding standard deviation. The Monte Carlo median absolute error (MAE, defined as the median of the absolute value of the difference between simulated estimates and 0.1) is close to the bias.

The bias reduction yielded by a post-lasso first stage depends strongly on the manner in which the penalty term is chosen. On average, a cross-validated (CV) penalty retains 74 of 180 and 99 of 1530 instruments. As can be seen in the row immediately below the 2SLS estimates, however, post-lasso estimation using CV-chosen penalties yields almost no bias reduction over 2SLS, while slightly reducing precision.[16]

Swapping cross-validation for a plug-in penalty leaves far fewer instruments. This is because the modified plug-in penalty proposed by Belloni et al. (2012) is much larger than the corresponding CV penalty. Starting with a dictionary of 180 instruments, the plug-in penalty retains only 2 instruments, on average, and even fewer when starting with 1530 instruments (in a few simulation runs, the plug-in estimator retains no instruments). Our findings here are also consistent with simulation results comparing lasso estimated with CV and plug-in penalties reported by Belloni et al. (2012) and Chetverikov, Liao and Chernozhukov (2019). Use of a much smaller instrument set reduces bias to around 0.015 in the two models.

The bias reduction yielded by a plug-in penalty comes at the cost of reduced precision. With so few instruments retained, the standard deviation of estimated schooling coefficients is about 0.035, while the median absolute error of these estimates is about 0.028. This is a considerable improvement on the bias and dispersion of 2SLS estimates. But three non-ML IV estimators that are often used in many-weak instrumental variables scenarios, SSIV, IJIVE, and LIML, do better. LIML is an approximately (median) unbiased maximum likelihood

[16] For the estimates in Table 3, cross-validated lasso penalty terms are chosen once using the original data. Plug-in penalties are recalculated in each simulation draw. Conditional on covariates, the original data and simulation draws are independent. Lasso is re-estimated for each draw.

analog alternative to 2SLS (see, for example, Davidson and MacKinnon 1993). SSIV, a split-sample version of 2SLS introduced by Angrist and Krueger (1995), estimates first-stage parameters in half the sample, carrying these over to the other half to compute fitted values. SSIV uses these "cross-sample fitted values" as instruments.[17] SSIV is consistent under a Bekker (1994) many-instrument asymptotic sequence and is therefore also approximately unbiased. The improved jackknifed instrumental variables (IJIVE) estimator suggested by Ackerberg and Devereux (2009) constructs a first-stage fitted value for each observation in a leave-out sample omitting that observation, after partialing out covariates using the full sample. Ackerberg and Devereux (2009)show that IJIVE is superior to the JIVE estimators discussed in Angrist, Imbens and Krueger (1999). Like SSIV and LIML, JIVE-type estimators are Bekker-unbiased.

The results in Table 3 suggest SSIV, IJIVE, and LIML estimates using both 180 and 1530 instruments are indeed virtually unbiased, though LIML and IJIVE are more precise than SSIV (compare, for example, Monte Carlo standard deviations of 0.012 for LIML and 0.016 for SSIV using 1530 instruments). The standard deviation of these estimators mostly lies between that of the lasso estimators computed using plug-in and CV penalties. LIML, IJIVE, and SSIV out-perform the best of the lasso estimators on MAE grounds. This reflects the fact that even with a relatively severe plug-in penalty, lasso-based estimates remain biased. The near-unbiasedness of LIML is perhaps surprising since LIML is often viewed as having no finite-sample moments (see, e.g., Hausman et al., 2012). The median-unbiasedness of LIML, IJIVE, and SSIV is apparent from the fact that MAE for these estimators is almost indistinguishable from MAD.[18]

As can also be seen in the rows of Table 3 grouped under the split-sample IV heading, SSIV estimates computed with an instrument list chosen by lasso with a CV penalty are unbiased. But there would seem to be little reason to prefer lassoed SSIV over full-dictionary SSIV, since the latter is more precise and has smaller MAE. At the same time, use of a plug-in penalty in a post-lasso SSIV procedure yields a first-stage that mostly chooses no instruments. Specifically, post-lasso SSIV with a plug-in penalty picks zero instruments in

[17] Angrist and Krueger (1995) call this version of split-sample IV an "unbiased split-sample estimator."

[18] LIML is the maximum likelihood estimator of a linear equation with an endogenous regressor under normality, but the GMM justification for LIML requires only conditional homoskedasticity (Hausman et al. 2012). Our simulation errors are normal but realistically heteroskedastic, so it seems fair to say the simulation design does not stack the deck in favor of LIML.

670 out of 999 iterations for the 180-instrument case, and in 893 out of 999 iterations for the 1530-instrument case. The IV estimates computed when instruments are selected are biased and much less precise than conventional SSIV estimates. Finally, using a sample split just to choose instruments (though not for first stage estimation) yields estimates only marginally better than 2SLS when applied to the 1530-instrument model (compare MAEs of 0.046 and 0.056) and a little worse for the 180-instrument model (compare MAEs of 0.043 and 0.040).[19]

4.2 Theoretical considerations

Lasso for instrument selection faces two challenges. First is the fact that any over-identified 2SLS estimator is biased. The second is a pretesting problem.

The "Bekker sequence" (named for Bekker (1994) and used by Angrist and Krueger (1995)) describes the bias of IV estimators using an asymptotic sequence that fixes the (limiting) number of observations per instrument as the sample size grows. This sequence shows that with many weak instruments we should expect 2SLS estimates to be biased towards the corresponding OLS estimates in inverse proportion to the first-stage F statistic for excluded instruments. By contrast, LIML, SSIV, and IJIVE are Bekker-unbiased. As documented repeatedly (see, e.g., Angrist, Imbens and Krueger 1999), the Bekker sequence describes the finite-sample behavior of alternative IV estimators extraordinarily well.[20]

ML methods are often motivated by the desire to fit relationships when the number of predictors is very large, perhaps even of the same order of magnitude as the sample size. In an IV context, this sounds like a many-weak scenario. But the asymptotic sequence that justifies use of lasso for first stage estimation relies on the sample size increasing relative to the number of parameters estimated. In such a sequence, the dictionary of possible instruments may be much larger than the sample size, but the number of parameters in an ML-engineered first

[19] This estimator, reported in the row labeled "Post-lasso (IV choice split only, CV penalty)", splits the sample, using one half-sample and the cross-validation penalty chosen in the original data to select instruments via lasso. This instrument set is then used for conventional 2SLS estimation in the other half sample. All of split-sample procedures enforce an equal split and average results from complementary splits. Chernozhukov et al. (2018*a*) discuss IV strategies that use lasso or other ML estimators in combination with SSIV-type sample splitting.

[20] The Bekker sequence has antecedents in Kunitomo (1980) and Morimune (1983), though Bekker (1994) appears to be the first motivated by quasi-experimental applications like AK91. Hansen, Hausman and Newey (2008) generalize the Bekker sequence to approximate the behavior of a wider class of estimators under weaker conditions.

stage is still limited. In particular, the Belloni et al. (2012) approximate sparsity condition implies $\lim_{n\to\infty} \frac{s}{n} = 0$, where s is the number of instruments needed to approximate the first stage CEF. By contrast, the Bekker sequence allows the limit of $^{s}/_{n}$ to be fixed at a number strictly between zero and one. Under Bekker, the fact that lasso truncates the instrument list reduces the bias of 2SLS estimates, but does not eliminate it.

Perhaps the AK91 application is an unfair test of the lasso idea. The number of AK91 instruments is at least two orders of magnitude below sample size, while the first stage is arguably more cloned than sparse. Even so, AK91 is often seen as representative of empirical labor applications in which many weak instruments are a concern (e.g., Staiger and Stock (1997), Chamberlain and Imbens (2004), and Hansen, Hausman and Newey (2008)). Belloni, Chernozhukov and Hansen (2011), Belloni and Chernozhukov (2011), Belloni, Chernozhukov and Hansen (2013) and Hansen and Kozbur (2014) use AK91 data as a testbed for machine-chosen first stages. We note, however, that other IV scenarios may indeed favor lasso. For example, Belloni et al. (2012) report simulation results for a Monte Carlo experiment with 100 possible instruments, a sample size of 100, and a sparse first stage with exponentially or discontinuously declining first-stage coefficients. In this experiment, lasso-based IV estimation outperforms 2SLS, LIML, and a modification of LIML due to Fuller (1977). We note also that Belloni et al. (2012) consider procedures for weak-instrument-robust hypothesis testing in combination with lasso, extending an approach in Staiger and Stock (1997). But this may be unnecessary: Bekker (1994) gives standard-error formulas consistent under a many-IV sequence; Kolesár et al. (2015) show this provides good confidence interval coverage, while Hansen, Hausman and Newey (2008) generalize Bekker standard errors to allow for heteroskedasticity.

Post-Lasso as Pre-Test

As first noted by Hall, Rudebusch and Wilcox (1996), estimation after screening instruments on the basis of the statistical significance of first stage coefficients need not improve, and may even aggravate, the bias of IV estimates. Pretesting estimated first-stage coefficients aggravates bias because when population first stage coefficients are truly zero or close to it, high in-sample correlation with an endogenous regressor is associated with a high in-sample correlation with omitted variables (or structural error terms).[21]

The theoretical link between post-lasso IV and pretesting is most visible in the case where the instruments are a set of orthonormalized variables (say, mutually exclusive dummies normalized by group size). In this case, post-lasso selects an instrument when the associated first-stage coefficient exceeds a constant. In particular, letting $\hat{\pi}_j$ denote the coefficient on the j-th instrument from an OLS first stage using orthonormalized instruments, post-lasso estimators retain the$j$th instrument when

$$|\hat{\pi}_j| > c_n, \tag{21}$$

where c_n, is determined by the lasso penalty and sample size (see, e.g., Hastie, Tibshirani and Wainwright 2015). The analogy with pretesting arises because pretest estimators retain $\hat{\pi}_j$ using a rule like (21), where the threshold is proportional to the estimated standard error of $\hat{\pi}_j$, which depends on sample size. Lasso regularity conditions imply that lasso and pretest thresholds converge at different rates. In the data at hand, however, lasso and 2SLS with a pretested first stage can be operationally similar.

Evidence of pretest bias emerges when LIML is computed with a post-lasso instrument list. This can be seen in the rows in Table 3 labeled “post-lasso LIML”. When lasso penalties are cross-validated, the otherwise unbiased LIML estimator exhibits bias of 0.022 in the 180-instrument model (with 74 instruments retained) and 0.048 when using 1530 instruments (with an average of 99 retained). Lasso with a plug-in penalty retains only two instruments, but here too, we see evidence of bias. With a plug-in penalty, the combination of bias and reduced precision yields an MAE of around 0.026 using both instrument lists, two to three times as large as conventional LIML. Not surprisingly, with only two instruments retained, the behavior of plug-in lassoed LIML is close to that of post-lasso 2SLS using the same small instrument set. By way of comparison, the table also shows an explicitly pretested LIML estimator, which retains instruments with a first-stage t-statistic in the upper decile of t-statistics for the full set of instruments. The point here is not to recommend this for empirical practice. Rather, the similarity of bias and MAE for pretested LIML and lassoed LIML using a cross-validated penalty highlight the pretesting analogy.

For the same reason that the SSIV estimator is essentially unbiased, sample splitting elim-

[21] Andrews, Stock and Sun (2019) survey and assess the pretesting problem in modern empirical work.

inates the risk of pretest bias. In particular, when estimated in a separate sample, estimated first stage coefficients are independent of second stage residuals. As we've seen, however, SSIV (and IJIVE) with all instruments tends to do better than a lassoed SSIV estimator that uses half the sample to pick instruments as well as estimate first stage coefficients.

A Walk in the Woods

The bias engendered by an ML-chosen instrument list is not unique to lasso. This is evident in results from an IV procedure that uses regression trees to estimate first-stage conditional mean functions. A "tree" in this context is the mean of the variable to be instrumented, conditional on a sequence of splits into subsamples. Given predictors like YOB and QOB, a tree-based first stage might split schooling into older and younger workers, and then split the older group by QOB, while leaving all of the younger group pooled (perhaps because compulsory attendance laws matter little for those born later). Splits are chosen or skipped so as to minimize MSE or some other measure of fit. "Leaves" on the resulting trees are endpoints in each sequence of splits. Athey and Imbens (2019) note that regression trees can be interpreted as a nearest-neighbor-type matching procedure where an observation's neighbors are those found on the same leaf. Random forests, introduced by Breiman 2001, elaborate on regression trees by using bootstrap samples (or subsampling) to decide where to split, and by looking only at randomly selected subsets of predictors when deciding where to split. Random forest predictions average the predictions yielded from each of these smaller samples.

Building on methods described in Hartford et al. (2017), Ash et al. (2018) explore a procedure using random forest first stage fitted values to compute IV estimates of the effects of appellate court decisions on the length of sentences handed down in district courts. The characteristics of appellate court judges, who are selected by random assignment, play the role of (high dimensional) instruments. In a related application of ML to IV, Athey, Tibshirani and Wager (2019) and Chernozhukov et al. (2018*a*) use a random forest procedure to select and partial out non-excluded (exogenous) covariates.[22]

We explore the utility of these estimation strategies for IV by using random forest first

[22] Athey, Tibshirani and Wager (2019) and Chernozhukov et al. (2018*b*) also use random forests to model treatment effect heterogeneity.

stage fitted values as instruments for education, controlling for a full set of YOB-by-POB fixed effects. In a random forest procedure that mirrors the 1530 instrument model we've estimated by 2SLS, the fitted values from a random forest fit of schooling to QOB, YOB, and POB are all but indistinguishable from the fitted values generated by a saturated regression model. This result is for a random-forest fit computed using a minimum leaf size of 1, the default for Stata's `rforest` command (a constraint not often binding in the AK91 data set). Not surprisingly, therefore, the 2SLS estimates generated by 2SLS with random forest fits as instrument, reported at the bottom of Table 3, are indistinguishable from the conventional 2SLS estimates, reported at the top. Likewise, an SSIV procedure based on random forest first-stage fitted values, the results of which appear in the last row in this table, replicates the relatively good performance of SSIV.

Increasing minimum leaf size to 800, a value used for IV estimation by Athey, Tibshirani and Wager (2019), leads to estimates that differ slightly from the corresponding conventional 2SLS results. The larger leaf size combines some QOB-YOB-POB cells, reducing bias from 0.061 for conventional 2SLS to 0.057 using random forest fits. MAE for this procedure is about like lasso with a cross-validated penalty, but far higher than lasso using a plug-in penalty. Changing leaf size leaves the performance of random forest SSIV essentially unchanged because this constraint is largely irrelevant (as we confirmed for minimum leaf sizes of 10 and 100). The estimates in Table 3 offer little reason to favor IV using a random forest first stage over conventional 2SLS, or, for that matter, over plug-in-penalized lasso.

4.3 IV at the Movies

Gilchrist and Sands (2016) uses lasso to select instruments for a 2SLS procedure in which the ratio of the number of instruments to sample size is an order of magnitude higher than in the 1530 instrument version of AK91. We might expect the relative performance of post-lasso instrument selection to improve in this setting. The Gilchrist and Sands (2016) study is motivated by an inquiry into social spillovers from movie viewership: filmgoers discuss movies they've seen with friends and coworkers, perhaps increasing viewership. Weather induces quasi-experimental variation in opening weekend viewership that identifies this social effect.

The Gilchrist and Sands (2016) sample contains information on the total dollar value of ticket sales for 1,381 movies over the course of 1,671 weekend days (the unit of observation for econometric analysis). The instruments for opening weekend viewership are nationally aggregated weather measures that summarize conditions near the nation's movie theaters on a given day. Theater weather conditions are proxied by conditions measured at weather stations in the same zip code. This identification strategy is motivated by the idea that the weather is randomly assigned and that good weather reduces viewership. The instrument dictionary includes 52 weather variables, such as the proportion of theaters experiencing 75-80 degree temperatures, indicators of snow and rain, and average hourly precipitation. Exogenous covariates in the model include dummies for the timing of opening weekend days. Additional exogenous controls include summary measures of weather conditions in the periods for which subsequent viewership is measured. These variables control for possible serial correlation in the weather. There are a total of 142 (mostly dummy) controls.

The IV estimates reported by Gilchrist and Sands (2016) are the result of a "manual 2SLS procedure" in which exogenous covariates are first partialed (using OLS) from opening weekend and subsequent viewership, and from the excluded instruments. Specifically, the paper reports estimates from a model regressing residual subsequent viewership on first stage fitted values using "weather shocks" as instruments. The latter are residuals from regressions of weather variables on covariates. Subject to the requirement that controls and samples be identical in all three partialing steps, this procedure is the same as 2SLS estimation of a model that includes exogenous covariates as controls instead of partialing them out (though manual 2SLS standard errors are incorrect). We therefore focus on the 2SLS equivalent of the Gilchrist and Sands (2016) estimates, and lasso versions thereof.

The full-dictionary 2SLS estimate of the effects of opening weekend viewership on viewership a week later is $0.5(SE = .022)$; the manual 2SLS estimate reported in Gilchrist and Sands (2016) is $0.475(SE = 0.024)$, a result we replicate using data posted by the authors. These estimates use all 52 excluded instruments. Our corresponding OLS estimate is $.449(SE = 0.016)$, while the original OLS estimate is $0.423(SE = 0.015)$. Small differences between our estimates and the originals arise because the original procedure partials both dummy variable and weather controls from the outcome variable, while partialing only contemporaneous weather variables out of the endogenous variable.

Using a single lasso-selected instrument, Gilchrist and Sands (2016) report an opening weekend effect of $0.474(SE = 0.047)$. The instrument in this case, a dummy for pervasive good weather, has a strong first stage, with a t-statistic over 6 (and hence a first-stage F close to 40). The fact that this differs little from 2SLS estimates using all 52 excluded weather instruments and from the corresponding OLS estimates points to limited scope for bias in the IV estimates. As we noted in the discussion of AK91, when OLS is indistinguishable from low-dimensional, strongly identified IV estimates, finite-sample concerns usually evaporate. This leads us to explore a simulation design built from the Gilchrist and Sands (2016) model and data, but with more OLS bias. IV procedures can then be evaluated on the basis of their ability to get closer than OLS to the truth.

Second-weekend attendance is our outcome variable of interest (the original study looks at opening weekend effects on viewership in weeks 2 through 6, finding declining effects). The simulation design starts by regressing opening weekend attendance (the endogenous regressor of interest) on exogenous covariates and the full set of excluded instruments to obtain first-stage fitted values. The list of exogenous covariates includes indicators for calendar year, day of the week, week of the year, holidays, and measures of weather conditions during the movie's second weekend. Call these first-stage fitted values $\hat{a}\left(X_{dt}, Z_t\right)$, where X_{dt} is the vector of exogenous covariates on second-weekend day d (Friday, Saturday, Sunday) among movies opening in week t, and Z_t is the vector of excluded instruments for movies opening in week t (the weather instruments vary only by week). Simulated opening-weekend attendance, $\tilde{a}_{dt}$, is drawn from a standard gamma distribution with shape parameter $\mu_{dt} = \max\left\{\delta, \hat{a}\left(X_{dt}, Z_t\right)\right\}/k_1$. This yields a skewed, non-negative continuous distribution. Scaling by $k_1 = 1.35$ approximates the first-stage R^2 and partial F-statistic in the original data. Because the gamma functional form requires a shape parameter bounded away from zero, the shape parameter is censored below by setting $\delta = .01$. Simulated attendance is then drawn by first generating a uniform random variable, v_{dt}, and evaluating the inverse gamma conditional distribution function with shape parameter μ_{dt} at v_{dt}. The appearance of v_{dt} in the simulated outcome residual is our source of endogeneity.

Our simulated outcome builds on LIML estimates in the original data using all 52 instruments. Specifically, let $\hat{y}\left(X_{dt}\right)$ be the dependent variable fitted value, after subtracting $\hat{\rho}_{LIML} a_{dt}$, where $\hat{\rho}_{LIML}$ is a LIML estimate of the effects of opening weekend attendance on second-weekend attendance, and a_{dt} is observed opening weekend attendance on day d for movies opening in week t. The notation here reflects the fact that this estimated fitted value varies only by X_{dt}. The simulated dependent variable is then constructed as

$$\tilde{y}_{dt} = \hat{y}\left(X_{dt}\right) + .6\tilde{a}_{dt} + \omega\left(X_{dt}, Z_t\right)\left(k_2\Phi^{-1}\left(v_{dt}\right) + \varepsilon_{dt}\right), \tag{22}$$

so the causal effect of opening weekend attendance is fixed at .6. Error component ε_{dt} is standard normal, while $\omega\left(X_{dt}, Z_t\right)$ is set to generate a conditional variance of residual second-weekend attendance given exogenous covariates and excluded instruments proportional to the variance of second-stage LIML residuals in the original data.[23] Finally, $k_2 = -1.5$ generates OLS estimates around 0.23, biased in the same direction as OLS in the original data, but much more so. Each simulation draw begins with a bootstrap sample of (X_{dt}, Z_t), with simulated opening-weekend and second-weekend attendance constructed as described above.[24]

The coefficients of interest are OLS, 2SLS, and post-lasso 2SLS estimates of parameter ρ in an IV setup modeling ticket sales on day d of the second weekend, y_{dt}, as a function of ticket sales on day d of the opening weekend, a_{dt}, among movies opening in week t. This model can be written:

$$\begin{aligned} y_{dt} &= \rho a_{dt} + X_{dt}'\gamma_0 + \varepsilon_{dt} \\ a_{dt} &= Z_t'\pi + X_{dt}'\gamma_1 + v_{dt}, \end{aligned}$$

where π is the vector of first-stage coefficients.

The bias of OLS estimates across simulations is -0.37, while 2SLS is about half as bad, with a negative bias of 0.17. In other words, both procedures yield estimated effects of opening weekend sales on second weekend sales that are much reduced from the causal effect of 0.6. Adding 52 worthless (standard uniform) instruments to the original dictionary of 52 weather instruments raises 2SLS bias by almost 50%, to 0.24. These benchmark estimates appear in the first two rows of Table 4, which also show that the MAE of 2SLS is indistinguishable from the bias of estimates using either 52 or 104 instruments. First-stage F statistics fall from close to 3 with 52 instruments to around 2 with 104 instruments.

As can be seen in the third and fourth rows of the table, 2SLS with a post-lasso first stage shortens the instrument list considerably but does little to reduce bias. Specifically,

[23] Specifically, we regressed the squared residuals generated by full-dictionary LIML on all instruments and covariates and used the square root of the predicted values to scale the simulated error term.

[24] The sample includes 557 opening weekends. The bootstrap sample draws individual days independently rather than by weekend. This is consistent with Gilchrist and Sands (2016), which reports standard errors described as clustered, but with clusters equal to the unit of observation.

using the larger plug-in penalty reduces the 52-instrument list to a list of 12, while the 104 instrument list falls to around 23. Lasso with a cross-validated penalty generates 37 and 58 instruments, respectively. But post-lasso 2SLS estimates remain substantially biased with both instrument lists and either penalty choice. Using the larger plug-in penalty, for example, yields second-stage estimates with a bias of -0.132. The bias-reduction payoff to post-lasso is larger when the instrument dictionary includes 52 noise variables and lasso is tuned with a plug-in penalty. In particular, the bias of 2SLS falls from around -0.24 to -0.15. On the other hand, post-lasso instrument selection using cross-validated penalties leaves bias in the 104 instrument model almost unchanged from that of 2SLS.

The middle rows of Table 4, which describe the behavior of SSIV, IJIVE, and LIML estimates, show that SSIV estimates are less biased than 2SLS estimates computed using post-lasso to choose instruments, but more biased than the SSIV results in the AK91 simulation. SSIV also suffers here from low precision, with Monte Carlo standard deviations ranging from around 0.6 to 6.6. This dispersion reflects a few extreme SSIV realizations. MAD for SSIV is far below variance, however. Remarkably, SSIV still beats post-lasso for both models on MAE grounds, with the SSIV advantage most impressive when the instrument list includes 52 real instruments only. Moreover, IJIVE improves markedly on SSIV. The bias of IJIVE is small (though not zero) and the IJIVE standard deviation is less than a quarter of SSIV's. The MAD and MAE for IJIVE are indistinguishable (and well below that of SSIV), indicating that IJIVE is median unbiased.

As in the AK91 results, LIML shines. Specifically, LIML estimates are virtually unbiased using both instrument sets and about as precise as post-lasso 2SLS estimates computed with a plug-in penalty. The upshot is that MAE for LIML is less than half of the MAE for post-lasso IV estimates constructed using a plug-in penalty. The robustly good performance of LIML in this case may be surprising given that the simulation residuals are heteroskedastic and the sample size is modest. But this finding is consistent with simulation evidence in favor of LIML reported in Angrist, Imbens and Krueger (1999).

Our description of post-lasso IV estimators concludes with a brief account of results generated by the Stata 16 `poivregress` command (documented in Stata 2019). Motivated by Chernozhukov, Hansen and Spindler (2015), `poivregress` allows the list of instruments and the list of exogenous covariates to be modeled as high-dimensional, applying lasso to the selection of variables in both. As with the other estimators described in Tables 3 and 4, we focus first on the consequences of lasso for instrument selection. The `poivregress` estimates discussed here use a plug-in penalty (the appendix gives other implementation details).

Simulation results using `poivregress` to choose instruments, reported at the bottom of Table 4, look much like those using Lassopack commands with a plug-in penalty. The resulting estimates exhibit bias on the order of 0.13 for models estimated with 52 instruments and 0.14 for models estimated with 104 instruments, only slightly below the bias of estimates computed using Lassopack commands. This in spite of the fact that `poivregress` retains an average of 1.3 instruments conditional on retaining any. It's noteworthy that `poivregress` fails to select *any* instruments in about two-thirds of the runs when starting with a dictionary of 52, while no instruments are selected in about three-fourths of the runs starting with a dictionary of 104. Surprisingly, `poivregress` reports second stage estimates even for these runs. Unsurprisingly, IV estimates generated without excluded instruments are biased and imprecise, with an MAE approaching that of OLS.

The bottom two rows of Table 4 show `poivregress` estimates computed for a procedure in which control variable coefficients are penalized along with first stage coefficients. That is, controls are treated as high-dimensional. Of 142 possible controls, 15-16 are retained in models that also retain instruments. Again, the average number of instruments retained is close to 1, conditional on having any retained. These estimates are less biased than post-lasso estimates computed with a plug-in penalty. Compare, for example, a bias of -0.096 using `poivregress` with a bias for plug-in post-lasso of -0.132 in the 52-instrument model. The better performance of `poivregress` with high-dimensional controls appears to reflect a higher first stage F when redundant controls are dropped. Even so, SSIV, IJIVE, and LIML estimates of this model are better on MAE grounds. For the model with 104 instruments, which perhaps hews closest to the idea of sparsity, `poivregress` MAE beats that of SSIV (but not LIML or IJIVE) in runs for which instruments are selected.

Application of `poivregress` to models that select controls as well as instruments also yields results without instruments. In this case, instruments are retained in about half of 999 simulations runs, while no instruments are selected and an estimate of zero reported for 372 and 409 runs in the 52- and 104-instrument models, respectively. Remaining runs generate non-zero IV estimates even with no instruments chosen (though instrument-free IV estimates are widely dispersed). On balance, the estimates in Table 4 make a weak case for ML-augmented IV. But other scenarios, perhaps with as many instruments as observations, might favor ML-based instrument selection more clearly.

5 ML Picks IV Controls

Identification in instrumental variables models may turn on control for covariates as well as on the choice of instruments. For example, in a study of the effects of family size on parents' labor supply, Angrist and Evans (1998) use the occurrence of multiple second births and samesex sibships as a source of quasi-experimental variation in the probability of having a third child. Because twin birth rates increase with maternal age and education, estimators exploiting the twins experiment are made more credible by conditioning on these covariates. The Angrist and Evans (1998) samesex instrument exploits the fact that, among women with two children, the probability of a third birth increases when the first two are both boys or both girls. But parents may care about the sex of their first and second born for reasons other than homogeneity. Ananat and Michaels (2008), for example, argue that male firstborns reduce divorce. Male births also generate more samesex sibships (because they're more likely), so the samesex identification strategy may be improved by allowing for additive male birth effects.

2SLS estimators include control variables as exogenous covariates in linear models. But ML methods can control for covariates without functional form assumptions. We briefly explore the ability of random forest routines to capture covariate effects in IV identification strategies that require some degree of control. This investigation is inspired by Athey, Tibshirani and Wager (2019), which uses random forest methods to model heterogeneous causal effects of family size when these are identified by sibling sex composition in the Angrist and Evans (1998) data.

Although motivated by the possibility of heterogeneous causal effects, the Athey, Tibshirani and Wager (2019) random forest IV procedure also generates an unconditional IV estimate of the form

$$\hat{\rho}_{ATW} = \frac{\sum_i \{[Y_i - \hat{g}_Y(X_i)][(Z_i - \hat{g}_Z(X_i)]\}}{\sum_i \{[D_i - \hat{g}_D(X_i)][(Z_i - \hat{g}_Z(X_i)\}},$$

where the function $\hat{g}_Z(X_i)$ is the (leave-out) fitted value from a random forest estimate of $E[Z_i|X_i]$ and functions $\hat{g}_Y$ and $\hat{g}_D$ are defined similarly for dependent and endogenous variables denoted by Y_i and D_i, respectively. Random forest centering of the moments underlying $\hat{\rho}_{ATW}$ adjusts for covariate effects in a manner analogous to linear control for exogenous covariates in 2SLS. Indeed, replacing random forest routines with linear regression generates a version of $\hat{\rho}_{ATW}$ equal to 2SLS.[25]

The behavior of $\hat{\rho}_{ATW}$ is explored here in a setting that requires additive controls. As a benchmark, column 1 in Panel A of Table 5 reports conventional 2SLS estimates of effects of childbearing on labor supply using a dummy variable indicating samesex sibships to instrument a variable indicating mothers with three or more children (everyone in the sample has at least two). As in Angrist and Evans (1998), these estimates show a first-stage effect of samesex sibships on the probability of having more than two children equal to about 0.07. 2SLS using the samesex instrument generates substantial and precisely estimated negative labor supply effects of a third birth. Specifically, the birth of a third child reduces employment rates by about 12 points, with a concomitant decline of about 5 weeks worked and a 5 hour reduction in the work week. These effects are smaller than the corresponding OLS estimates (not reported here), suggesting a high degree of selection bias in the latter.

Columns 2-4 report estimates of $\hat{\rho}_{ATW}$ computed using the Stata random forest routine `rforest`, implemented with minimum leaf sizes 10, 100, and 800. The estimates in columns 5-7 were computed using the `regression_forest` command contained in the Generalized Random Forest (GRF) software package distributed by the authors of Athey, Tibshirani and Wager (2019).[26] The `rforest`-based estimates in columns 2-4 are remarkably imprecise, with standard errors over 10 times larger than the 2SLS standard errors for minimum leaf size of 10 and over three times larger for a minimum leaf size of 800. This imprecision reflects the fact that samesex is deterministically related to the additive male birth indicators included

[25] A related procedure outlined in Chernozhukov et al. (2018*a*) uses random forest and other ML methods to partial covariates from instruments, dependent variables, and endogenous variables, in combination with a sample splitting strategy similar to SSIV. The moment conditions motivating this procedure (equations 4.4 and 4.8 in Chernozhukov et al. (2018*a*)) appear to be the same as those motivating the estimators considered by Athey, Tibshirani and Wager (2019). The Athey, Tibshirani and Wager (2019) procedure uses jackknifed random forest fits rather than sample splitting.

[26] The appendix gives computational details. Athey, Tibshirani and Wager (2019) use a leaf size of 800 for IV estimation.

as covariates. Specifically, the samesex instrument can be written

$$ss_i = m_{1i}m_{2i} + (1 - m_{1i})(1 - m_{2i}),$$

where ss_i indicates mothers of a samesex sibship and m_{ji} indicates mothers with a male child at birth j. 2SLS uses additivity to accommodate this dependence, distinguishing interaction terms from additive effects. The `rforest` routine struggles with this.

As can be seen in columns 5-7, the `regression_forest` program does better with the samesex instrument than does `rforest`. In particular, estimated labor supply effects in these columns are similar to those generated by 2SLS. But these too are considerably less precise than the corresponding 2SLS estimates, with standard errors as much as 50% larger (compare for example, the 2SLS estimate of the effect on hours equal to -4.8 with a standard error of 1 to the `regression_forest` estimate of -4.6 with a standard of about 1.5 in column 5). A second noteworthy feature of this set of results is a set of large first stage estimates, ranging from 0.24 in column 7 to 0.57 in column 5. These estimates presumably reflect the underlying parameterization of m_{ji} effects on ss_i implicit in the random forest fit. While this parameterization yields samesex residuals with enough variance to generate informative second stage estimates, it renders the first stage uninterpretable.

Randomly Excluded

Random forest partialing may have undesirable consequences beyond second-stage imprecision or a reparameterized first stage. Since random forest is not regression, random forest residuals may be correlated with the covariates that made them. In an IV context, failure to orthogonalize covariates and instruments risks the creation of unintended exclusion restrictions that lead to misleading second stage estimates. This phenomenon is analogous to the risk of spurious identification when a probit or logit first stage is used to instrument a dummy endogenous variable (see, e.g., Angrist 2001).

We illustrate this point using an "artificial instruments" experiment of the sort that inspired the Bound, Jaeger and Baker (1995) critique of Angrist and Krueger (1991). This experiment (originally suggested by Alan Krueger) used randomly generated instruments to reveal the bias of heavily over-identified 2SLS estimates in cases where the instruments are uninformative. Our version constructs a single just-identifying instrument that is highly correlated with covariates, but unrelated to treatment conditional on covariates.

The covariates used for IV estimation in Angrist and Evans (1998) and Athey, Tibshirani and Wager (2019) include mother's age ($agem_i$) and mother's education ($educm_i$). Our first artificial instrument is a function of these two variables plus random noise:

$$h_{1i} = agem_i + educm_i + u_i \equiv x_i + u_i,$$

where u_i is standard uniform, drawn independently of covariates. We refer to $x_i = agem_i + educm_i$ as a "covariate index". Conditional on the covariates used to construct the index, instrument h_{1i} should have no identifying power. 2SLS estimates computed using h_{1i} as an instrument in a model including mothers' education, age, and other covariates appear in the first column of Panel B in Table 5. These estimates have large standard errors and indeed appear uninformative. For example, where the samesex instrument generates an estimated reduction of 5.3 weeks worked with a standard error of 1.2, instrument h_{1i} generates an estimate with a standard error around 22.

Random forest partialing of covariates from h_{1i} yields a residual that remains correlated with x_i. Figure 6 documents this by plotting residuals from a random forest fit of h_{1i} to the covariates used to construct the 2SLS estimates report in column 1 of Table 5. The figure shows average residuals conditional on x_i, along with the conditional mean of OLS fitted values and OLS residuals given x_i. Not surprisingly, mean OLS fitted values are linear in x_i, while mean OLS residuals are flat. Mean random forest residuals, by contrast, turn up or down for values at the ends of the support of x_i. Smaller leaf size reduces but does not eliminate this correlation.[27]

The risks posed by Figure 6 for IV are apparent in the IV estimates reported in columns 2-7 of Panel B in Table 5. These estimates generate a misleading impression of large and (for the most part) statistically significant effects. The problem is especially severe for estimates computed with a larger minimum leaf size. The spurious identification conjured by random forest partialing stems from the failure to fit an additive linear model (repeated draws of h_{1i} generate similar findings). Some of the artificial IV estimates shown in columns 2-4 of the table (computed using `rforest`) are implausibly large, implying, for example, a fall in

[27] The figure plots residuals computed by `rforest`. A plot constructed using the `regression_forest` routine looks similar.

employment rates in excess of one. An attentive analyst might not be fooled here. But the estimates computed using `regression_forest` and reported in columns 5-7 are both small enough and precise enough to give the impression of a meaningful finding.

The failure to fit (or "learn", in ML vernacular) the relationship between h_{1i} and covariates may seem at odds with results using random forest to estimate the AK91 first stage, reported in Table 3. In the AK91 simulations, random forest fits a 1530 instrument first stage perfectly, recovering the empirical CEF. Random forest does worse with the artificial Angrist and Evans (1998) first stage because the number of covariate cells in this case is much larger. While the AK91 design has roughly 2000 cells and around 200 observations per cell, the Angrist and Evans (1998) first stage has around 161,000 cells, with 1.6 observations per cell. This necessitates some smoothing, which random forest delivers as promised. Yet this flexible ML routine misses important features of the CEF that it's been tasked to model.

Figure 6 suggests reducing minimum leaf size moderates the correlation between random forest residuals and covariates. The estimates reported in column 5 of Panel B show that partialing controls from h_{1i} with a leaf size of 10 (and therefore little regularization) generates no statistically significant second stage estimates. This offers an interesting contrast with the lasso-IV estimates reported in Table 3, where larger tuning parameters induce more regularization, mitigating bias. But the small-leaf strategy is a double-edged sword. The results in Panel A using the real samesex instrument with a minimum leaf size of 10 are either so imprecise as to be useless (i.e., those in column 2, computed using `rforest`), or generated by a first stage that fails to describe the causal effect of the samesex experiment on fertility (i.e., an estimate of 0.572 in column 5, computed using `regression_forest`).

The risk of spurious identification using random forest partialing arises even when instruments have some signal. Consider, for instance,

$$h_{2i} = agem_i + educm_i + (u_i \cdot ss_i) = x_i + (u_i \cdot ss_i).$$

Identification using h_{2i} hinges on control for the covariates that go into x_i. Unlike h_{1i}, however, artificial instrument h_{2i} has a strong and precisely estimated first stage effect on fertility of about 0.082, reported in the first column of Panel A in Table 6. A 2SLS estimator has no trouble extracting the signal in h_{2i} while successfully purging covariate effects. For example, the estimated employment reduction due to a third child is about -0.11 whether computed using ss_i or h_{2i}, while the standard error increases by about a third when using the latter. 2SLS estimates for other outcomes are similarly close.

In contrast with the good performance of 2SLS using h_{2i}, random forest partialing mostly yields estimates that seem as distorted as those computed using an instrument with no information. Estimates computed using `rforest`, reported in columns 2-4 of Panel A, are too large to be coherent, yet worryingly precise given their magnitudes (though many are outside the bounds of dependent variable support). Part of the problem here is the extraordinary sensitivity of the random forest first stage, which falls to near zero and negative in columns 2-4 and 7, and shrinks to 0.048 and 0.014 in columns 5-6. With a minimum leaf size of 100, estimates computed using `regression_forest` are noisy and therefore mostly not significantly different from zero. Estimates with a minimum leaf size of 10 are in the ballpark of the corresponding 2SLS estimates, but, as in Panel A of Table 5, markedly less precise. And, `regression_forest` estimates using h_{2i} as an instrument with a minimum leaf size of 800, shown in column 7, are arguably more troubling than the corresponding `rforest` estimates because they're way off base, statistically significant, and small enough to imply effects within the bounds of dependent variable support.

Finally, it's noteworthy that random forest partialing can yield misleading estimates even in a scenario without signal-extraction concerns. This is highlighted by estimation using an artificial instrument without noise. Specifically, we use

$$h_{3i} = agem_i + educm_i + ss_i$$

to compute $\hat{\rho}_{ATW}$. These estimates are reported in Panel B of Table 6, along with a 2SLS benchmark. As expected, 2SLS estimates using h_{3i} as an instrument, reported in the first column 1 of Panel B, are identical to those using the original samesex instrument. But most of the estimates of $\hat{\rho}_{ATW}$ computed using h_{3i}, shown in the rest of Panel B, are little better than those computed using h_{2i}. The exception is the set of estimates computed using `regression_forest` with a minimum leaf size of 10. Again, however, estimates of $\hat{\rho}_{ATW}$ are considerably less precise than the corresponding 2SLS estimates. Interestingly, the addition of x_i to ss_i to construct h_{3i} resolves the first stage parameterization issue highlighted in Panel A of Table 5: the first stage effect of h_{3i} in column 5 (Panel B) of Table 6 is 0.065, down from 0.57 in column 5 (Panel A) of Table 5. This sensitivity to parameterization hardly seems reassuring.

6 Summary and Conclusions

The Belloni, Chernozhukov and Hansen (2014*b*) PDS procedure provides a partially automated scheme for regression sensitivity analysis. Application of PDS to estimation of effects of elite college attendance shows how this approach can support causal conclusions in linear models. The identity and length of the list of PDS-included controls varies with changes in lasso tuning parameters and software. But the resulting estimates of causal effects are stable, consistently showing little evidence of an elite college advantage. In this application, PDS appears to offer a coherent data-driven complement to ad hoc robustness checks.

Our findings on ML in IV applications are less encouraging. In simulations modeled after Angrist and Krueger (1991) and Gilchrist and Sands (2016), 2SLS estimates with a post-lasso first stage sometimes improve on 2SLS with all available instruments. But SSIV, IJIVE, and LIML do better than 2SLS procedures that use lasso for instrument selection. In the AK91 design, estimates with a random forest first stage simply replicate 2SLS and SSIV, so this application of ML seems gratuitous. We note that our analysis focuses on estimator bias and dispersion, rather than procedures for inference. As has been shown elsewhere, however, many-instrument (Bekker, 1994) standard error formulas for LIML appear to provide good coverage. It seems likely that similar formulas can be obtained for SSIV (perhaps along the lines of those for JIVE in Chao et al. 2012).

The simulation results reported here show LIML to be surprisingly robust to heteroskedasticity. While some types of heteroskedasticity can confound LIML, this need not be true. Our results hint at the empirical relevance of heteroskedastic scenarios discussed by Bekker and Van Der Ploeg (2005) and Hausman et al. (2012). For example, LIML remains Bekker-unbiased under heteroskedasticity with dummy instruments and equal group sizes. More general conditions for this result are constant instrument leverage and an orthogonality condition given by Hausman et al. (2012). Scenarios involving a far higher ratio of instruments to observations than we've considered might favor ML-based instrument selection over SSIV and LIML. Such scenarios are, as yet, rarely seen in applied microeconometrics.

As a theoretical matter, our divergent conclusions on the utility of ML for control variable selection and for instrumental variable selection may be related to results in Cattaneo, Jansson and Newey (2018*a*,*b*). This work shows that regression estimators relying on high-dimensional controls to identify causal effects are consistent under a many-covariate sequence analogous to the Bekker sequence for IV. Unlike 2SLS, which is biased in all but just-identified models, the high-dimensional regression estimator at the heart of the PDS procedure is Bekker-unbiased (maintaining the conditional independence assumption motivating the procedure). Moreover, double selection works to mitigate the consequences of selection errors. With IV, pretest bias can be avoided by versions of lasso that use sample splitting to separate first-stage and second-stage estimates. In our applications, however, SSIV mostly does better without the complications of lasso.

Beyond matters of bias and precision, our analysis highlights the potential risks of nonlinear IV. In models with dummy endogenous variables, for example, a probit first stage equation includes nonlinear terms that may create unintended identifying restrictions (Angrist 2001). Random forest partialing of covariates in a just-identified model may likewise create artificial exclusion restrictions that lead to misleading second-stage estimates. This would seem to be a caution for applications relying on other nonlinear ML routines to pick controls in an IV setting. The bias documented in our experimental scenarios is not integral to the ML methods we've explored. In some applications, regression trees, random forests, elastic and neural networks, and ensemble methods that combine these may indeed be harmless. As far as empirical Labor goes, however, the payoff to these baroque procedures does not yet appear to justify the risk.

Table 1: OLS Estimates of Elite College Effects

	Basic Controls		DK02 Selection controls			
		Personal	Barron's	Barron's	Self-revelation	
		charac-	matches	matches	Barron's	
	None	teristics	only	w/pers. char.	sample	Full sample
	(1)	(2)	(3)	(4)	(5)	(6)
A. Private School Effects						
Estimate	0.212	0.139	0.007	0.013	0.036	0.037
	(0.060)	(0.043)	(0.038)	(0.025)	(0.029)	(0.039)
R-squared	0.019	0.107	0.058	0.138	0.111	0.114
No. of controls	0	10	150	160	13	14
N	14238			5583		14238
B. Effects of School Average SAT/100						
Estimate	0.109	0.076	0.008	0.004	0.004	0.000
	(0.026)	(0.016)	(0.029)	(0.016)	(0.017)	(0.018)
R-squared	0.019	0.107	0.066	0.140	0.107	0.113
No. of controls	0	10	334	344	13	14
N	14238			9166		14238
C. Effects of Attending Schools Rated Highly Competitive +						
Estimate	0.225	0.153	0.018	0.022	0.031	0.068
	(0.046)	(0.030)	(0.047)	(0.035)	(0.032)	(0.029)
R-squared	0.020	0.108	0.048	0.129	0.106	0.114
No. of controls	0	10	128	138	13	14
N	14238			4945		14238

Notes: This table reports OLS estimates of the effect of college characteristics on graduate earnings, estimated with various sets of controls. Estimates use College and Beyond sampling weights and cluster standard errors by institution. Controls used for column 2 include graduates' SAT scores, log parental income, indicators for female, black, Hispanic, Asian, other/missing race, high school top 10 percent, high school rank missing, and athlete. Controls for estimates reported in Panel A, column 3 include 150 dummies (for 151 categories) indicating the Barron's selectivity mix of schools to which graduates applied and were admitted. Controls for column 4 include Barron's dummies and the personal characteristics used for column 2. The Barron's model in Panel B includes 334 dummies; the Barron's model in Panel C includes 128 dummies. Columns 5-6 models replace dummies for Barron's selectivity groups with the average SAT score of schools applied to, along with indicators for applying to two, three, and four or more schools.

Table 2: Post-Lasso Estimates of Elite College Effects

	Double-selection (PDS)			Outcome selection			All controls
	plugin (16)	C.V. λ	cvlasso	plugin (16)	C.V. λ	cvlasso	OLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			A. Private School Effects				
Estimate	0.038	0.020	0.040	0.046	0.043	0.042	0.017
	(0.040)	(0.039)	(0.041)	(0.041)	(0.043)	(0.043)	(0.039)
No. of controls	18	100	112	10	35	50	303
			B. Effects of School Average SAT/100				
Estimate	-0.009	-0.013	-0.009	-0.008	-0.009	-0.008	-0.012
	(0.020)	(0.018)	(0.019)	(0.020)	(0.019)	(0.019)	(0.018)
No. of controls	24	151	58	10	34	43	303
			C. Effects of Attending Schools Rated Highly Competitive +				
Estimate	0.068	0.051	0.073	0.076	0.080	0.082	0.053
	(0.033)	(0.033)	(0.033)	(0.031)	(0.032)	(0.032)	(0.033)
No. of controls	17	185	106	10	34	43	303

Notes: The sample size is 14,238. Estimates in columns 1-3 are from a post-double-selection (PDS) lasso procedure. Results in columns 4-6 are from a procedure applying lasso to a reduced-form regression of the the outcome on the dictionary of controls. Columns 1 and 4 show results using the Stata 16 lasso linear command to select controls with a plug-in penalty, and OLS to compute the estimates. Columns 2 and 5 use lasso linear with 10-fold cross validation to select the penalty. Columns 3 and 6 use Stata 15 (Lassopack) cvlasso to select the penalty, rlasso to select controls, and OLS to compute estimates. See the appendix for detals. Column 7 reports OLS estimates including the entire set of controls. Controls include those used for column 5 of the previous table plus the following: indicators for being accepted to two colleges, three colleges, and four or more colleges; indicators for being rejected from one college, two colleges, three colleges, and four or more colleges; the number of schools applied to; the average SAT score among schools at which the applicant was accepted; the average SAT score among schools from which the applicant was rejected; the highest average SAT score across schools at which the applicant was accepted; the highest average SAT score across schools from which the applicant was rejected; the lowest average SAT score among schools at which the applicant was accepted; the lowest average SAT score among schools from which the applicant was rejected, and all two-way interactions of the above variables. The control dictionary contains 384 variables. OLS estimates use weights, and are reported with robust standard errors clustered by institution. All lasso commands use regressor-specific penalty loadings.

Table 3: Angrist and Krueger (1991) Simulation Results

	180 Instruments (QOB*YOB; POB*YOB; Average F=2.5)					1530 Instruments (QOB*YOB*POB; Average F=1.7)				
Estimator	Avg. IVs retained	Bias	Standard deviation	Median abs. dev.	Median abs. error	Avg. IVs retained	Bias	Standard deviation	Median abs. dev.	Median abs. error
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS		0.107	0.0004	0.0003	0.1070					
2SLS	180	0.0403	0.0108	0.0075	0.0397	1530	0.0611	0.0046	0.0032	0.0611
Post-lasso IV (CV penalty)	74.0	0.0390	0.0120	0.0082	0.0384	99.0	0.0559	0.0084	0.0059	0.0560
Post-lasso IV (plug-in penalty, IVs selected)*	2.1	0.0143	0.0346	0.0218	0.0279	1.6	0.0149	0.0367	0.0224	0.0271
Split-Sample IV	180	-0.0009	0.0237	0.0158	0.0158	1530	-0.0001	0.0164	0.0112	0.0115
Post-lasso SSIV (CV penalty)	63.1	-0.0015	0.0258	0.0172	0.0173	63.0	-0.0013	0.0280	0.0183	0.0183
Post-lasso SSIV (plug-in penalty, IVs selected)**	2.1	-0.0724	1.3168	0.0274	0.0287	3.4	0.0197	0.0504	0.0228	0.0292
Post-lasso (IV choice split only, CV penalty)	63.1	0.0429	0.0144	0.0097	0.0431	63.0	0.0460	0.0141	0.0093	0.0459
IJIVE***	180	-0.0011	0.0194	0.0130	0.0131	1530.0	0.0001	0.0123	0.0088	0.0087
LIML	180	-0.0016	0.0185	0.0123	0.0124	1530	-0.0034	0.0117	0.0079	0.0083
Post-lasso LIML (CV penalty)	74.0	0.0222	0.0152	0.0102	0.0220	99.0	0.0484	0.0094	0.0066	0.0483
Post-lasso LIML (plug-in penalty, IVs selected)*	2.1	0.0126	0.0347	0.0221	0.0273	1.6	0.0138	0.0366	0.0221	0.0257
Pretested LIML (t => 3.12 for 180, t=>2.3 for 1530)	18	0.0222	0.0236	0.0148	0.0238	153	0.0385	0.0163	0.0111	0.0393
Random forest first stage, 2SLS using RF fits as instruments (min leaf size=1)							0.0611	0.0047	0.0030	0.0612
Random forest 2SLS, min leaf size = 800							0.0567	0.0065	0.0045	0.0567
Random forest first stage, SSIV using RF fits as instruments (min leaf size =1)							-0.0003	0.0158	0.0109	0.0108
Random forest SSIV, min leaf size = 800							-0.0005	0.0158	0.0104	0.0103

Notes: The table describes simulation results for 999 Monte Carlo estimates of the economic returns to schooling using simulated samples constructed from the Angrist and Krueger (1991) census sample of men born 1930-39 (N=329,509). The causal effect of schooling is calibrated to 0.1; the OLS estimand is 0.207. The instruments used to compute the estimates described by columns 1-5 consist of 30 quarter-of-birth-by-year-of-birth and 150 quarter-of-birth-by-state-of-birth interactions (average F-stat = 2.5, average concentration parameter = 270). The instruments used to compute the estimates described by columns 6-10 are quarter-of-birth-by-year-of-birth-by-state-of-birth interactions (average F-stat = 1.7, average concentration parameter = 1050). All models include saturated year of birth by state of birth controls. Columns 1 and 6 report the average number of instruments retained by lasso. Post-lasso estimates are computed as described in the appendix. Split-Sample IV uses first stage coefficients estimated in one half-sample to construct a cross-sample fitted value used for IV in the other. Sample-splitting procedures average results from complementary splits. Post-lasso with an IV-choice split only uses post-lasso in half the sample to pick instruments, doing 2SLS with these and own-sample fitted values in the other half. IJIVE implements Ackerberg and Devereaux's (2009) improved jack-knifed instrumental variables procedure. "Post-lasso LIML" is LIML using the instrument set selected by a post-lasso first stage. "Pretested LIML" estimates are computed using conventional LIML, retaining only instruments with a first-stage t-statistic in the upper decile of t-statistics for the full set of instruments. Simulation sets choose lasso penalties once, using the original AK91 data. Random forest routines are described in the appendix.

*The plug-in penalty generates a lasso first stage that includes no instruments in 11 simulation runs with 180 instruments and in 57 simulation runs with 1530 instruments. Statistics reported in these rows are for runs completed.

**Post-lasso SSIV with a plug-in penalty picks zero instruments in 670 of 180-instrument runs, and in 893 of 1530-instrument runs. Statistics reported in these rows are for runs completed.

***IJIVE results based on 599 simulation iterations.

Table 4: Simulation Results for Opening Weekend Effects

	Original Instruments (F=2.85)					Original plus 52 noise instruments (F=2.06)				
	Avg. IVs retained	Bias	Standard deviation	Median abs. dev.	Median abs. error	Avg. IVs retained	Bias	Standard deviation	Median abs. dev.	Median abs. error
Estimator	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS		-0.374	0.015	0.010	0.374					
2SLS	52	-0.165	0.042	0.027	0.165	104	-0.239	0.034	0.022	0.238
Post-lasso IV (CV penalty)	36.6	-0.160	0.054	0.029	0.163	58.2	-0.219	0.063	0.029	0.228
Post-lasso IV (plug-in penalty)	12.2	-0.132	0.092	0.053	0.142	22.5	-0.150	0.095	0.067	0.159
Split-sample IV	52	0.053	0.568	0.095	0.093	104	-0.109	6.610	0.134	0.134
IJIVE	52	0.019	0.133	0.067	0.067	104	0.004	0.166	0.082	0.087
LIML	52	0.007	0.089	0.057	0.057	104	0.009	0.104	0.064	0.065
poivregress (fixed # of controls)										
Some IVs selected	1.32	-0.133	0.084	0.053	0.137	1.26	-0.145	0.077	0.050	0.143
No IVs selected*	0	-0.287	0.360	0.198	0.316	0	-0.291	0.352	0.189	0.325
poivregress (high-dim controls)										
Some IVs selected (15-16 ctls retained)	1.22	-0.096	0.108	0.069	0.107	1.22	-0.098	0.104	0.068	0.106
No IVs selected** (49-50 ctls retained)	0	-1.58	18.5	0.354	0.477	0	-1.90	19.5	0.375	0.477

Notes: The table reports simulation results for 999 Monte Carlo estimates of the effect of opening weekend ticket sales on second weekend ticket sales using simulated samples constructed from the data used by Gilchrist and Sands (2016) (N=1,671). The causal effect of interest is calibrated to 0.6. Columns 1-5 show results using the original instruments. Columns 6-10 report the results of adding 52 randomly generated (standard uniform) instruments to the original 52-instrument dictionary. Lasso estimates are computed after partialing out included exogenous covariates. Post-lasso IV estimates are computed as described in the appendix. Split-Sample IV uses first stage coefficients estimated in one half-sample to construct a cross-sample fitted value used for IV in the other. Sample-splitting procedures average results from complementary splits.

*poivregress reports estimates with zero instruments selected in 640 of 999 runs for the original 52-instrument set and in 741 of 999 runs using the 104 instrument set.

**poivregress with high-dimensional controls reports estimates with zero instruments selected in 141 of 999 runs for the original 52-instrument set and in 132 of 999 runs using the 104 instrument set. In 372 runs with 52 instruments and 409 runs with 104 instruments, this version of poivregress selects zero instruments and reports an estimate of zero.

Table 5: IV Estimates After Random Forest Partialing

		Random Forest					
		`rforest`			`regression_forest`		
	2SLS	Leaf 10	Leaf 100	Leaf 800	Leaf 10	Leaf 100	Leaf 800
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			A. Samesex Instrument				
More than 2 children	0.0676 (0.0018)	0.0804 (0.0234)	0.084 (0.0123)	0.087 (0.0067)	0.572 (0.020)	0.370 (0.0107)	0.239 (0.0066)
Employment	-0.118	0.046	-0.077	-0.115	-0.110	-0.123	-0.118
	(0.028)	(0.342)	(0.168)	(0.086)	(0.0402)	(0.0323)	(0.0314)
Weeks worked	-5.28	0.153	-2.04	-3.59	-4.48	-5.34	-5.27
	(1.23)	(14.9)	(7.35)	(3.74)	(1.73)	(1.41)	(1.35)
Hours/week	-4.82	0.792	-2.79	-4.97	-4.60	-5.26	-4.95
	(1.04)	(13.2)	(6.49)	(3.26)	(1.49)	(1.19)	(1.16)
			B. Artificial Instrument (covariate index + uniform noise)				
More than 2 children	-0.0066 (0.0031)	-0.0051 (0.0012)	-0.0096 (0.0009)	-0.0144 (0.0007)	-0.0108 (0.0028)	-0.0173 (0.0017)	-0.0169 (0.0007)
Employment	-0.035	-1.24	-1.07	-1.08	-0.0824	-0.451	-0.655
	(0.504)	(0.35)	(0.14)	(0.07)	(0.286)	(0.112)	(0.050)
Weeks worked	-4.94	-51.5	-43.4	-43.2	-8.01	-23.7	-28.0
	(21.8)	(14.6)	(5.58)	(2.67)	(12.3)	(4.86)	(2.10)
Hours/week	-22.5	-49.2	-38.6	-32.7	-19.4	-19.8	-21.8
	(19.7)	(13.7)	(5.08)	(2.29)	(11.1)	(4.38)	(1.87)

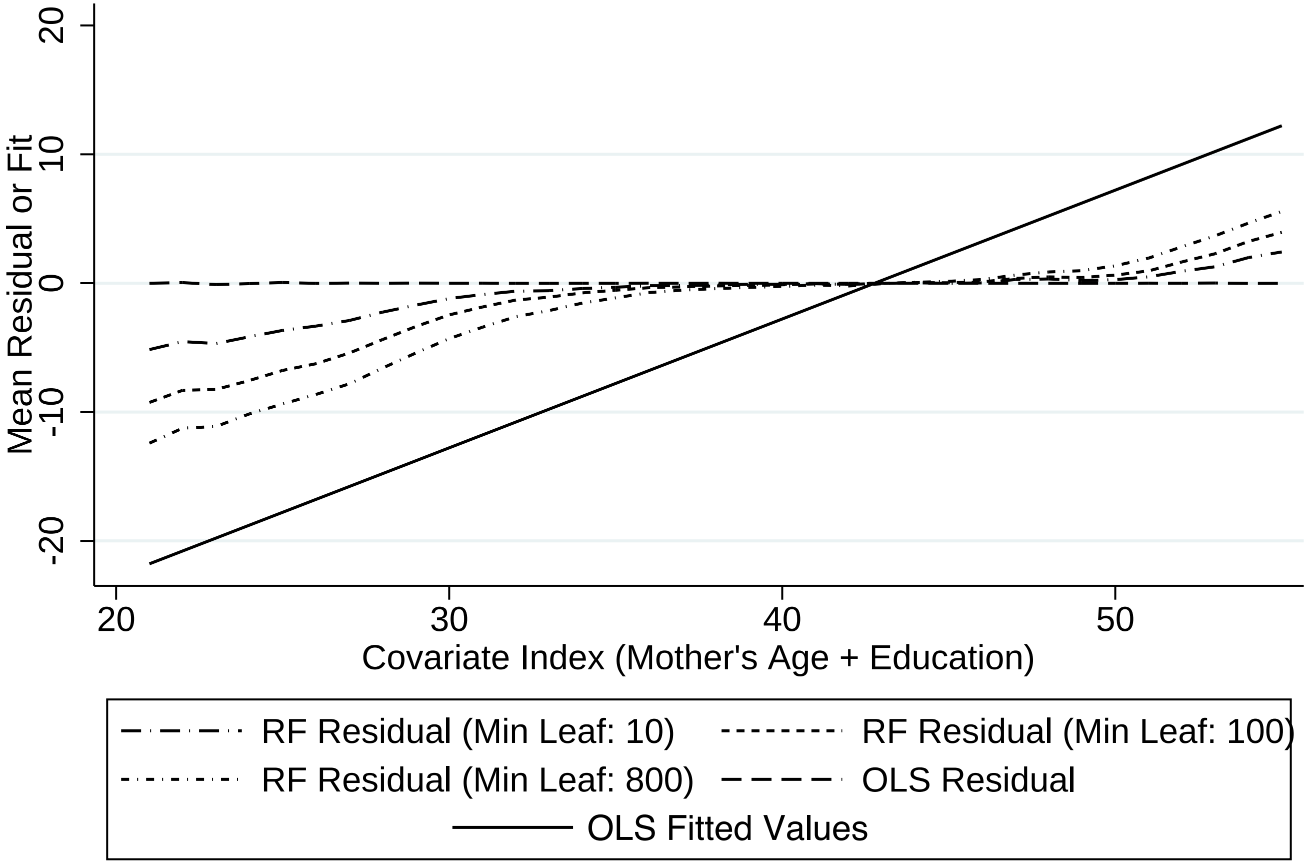
Notes: This table reports 2SLS and random forest IV estimates of the effect of having more than two children on the outcome variables listed at left. The estimates in this table differ from those in the previous table in that they use different instruments. Estimates in Panel A use an artificial instrument contructed from the sum of mother's age, education, and the product of a samesex dummy and uniform random noise. This instrument yields 2SLS estimates similar to the 2SLS estimates reported in Panel A of the previous table. Estimates in Panel B use an instrument contructed from the sum of mother's age, education, and the samesex dummy. This instrument generates 2SLS estimates **identical** to the 2SLS estimates reported in Panel A of the previous table. The sample includes married women from the 1980 PUMS aged 21-35 with two or more children. The sample size is 254,652.

Table 6: Random Forest IV Experiments With a Signal

		Random Forest					
		`rforest`			`regression_forest`		
	2SLS	Leaf 10	Leaf 100	Leaf 800	Leaf 10	Leaf 100	Leaf 800
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Instrument with some signal (covariate index + samesex × uniform noise)							
More than 2 children	0.0819 (0.0027)	-0.0038 (0.0012)	-0.0082 (0.001)	-0.0130 (0.0007)	0.0475 (0.003)	0.0136 (0.0017)	-0.0100 (0.0007)
Employment	-0.111	-1.83	-1.30	-1.19	-0.0633	0.310	-0.974
	(0.036)	(0.64)	(0.18)	(0.078)	(0.069)	(0.146)	(0.095)
Weeks worked	-5.04	-80.0	-55.9	-51.3	-3.38	15.3	-45.2
	(1.57)	(27.8)	(7.8)	(3.3)	(2.93)	(6.3)	(4.2)
Hours/week	-4.31	-73.2	-47.3	-37.3	-0.965	12.3	-29.6
	(1.32)	(25.6)	(6.9)	(2.8)	(2.55)	(5.6)	(3.4)
B. Instrument with full-strength signal (covariate index + samesex)							
More than 2 children	0.0676 (0.0018)	-0.0038 (0.0013)	-0.0086 (0.0010)	-0.0123 (0.0007)	0.0648 (0.0024)	0.0287 (0.0015)	-0.0051 (0.0007)
Employment	-0.118	-1.87	-1.26	-1.23	-0.0922	0.0018	-1.66
	(0.028)	(0.68)	(0.18)	(0.08)	(0.041)	(0.0541)	(0.249)
Weeks worked	-5.28	-82.4	-54.5	-53.3	-5.17	1.55	-73.5
	(1.23)	(29.3)	(7.5)	(3.6)	(1.74)	(2.32)	(10.8)
Hours/week	-4.82	-73.6	-45.8	-38.6	-3.58	-0.344	-47.8
	(1.04)	(26.6)	(6.6)	(3.0)	(1.52)	(2.04)	(7.9)

Notes: This table reports 2SLS and random forest IV estimates of the effect of having more than two children on the outcome variables listed at left. The estimates in this table differ from those in the previous table in that they use different instruments. Estimates in Panel A use an artificial instrument contructed from the sum of mother's age, education, and the product of a samesex dummy and uniform random noise. This instrument yields 2SLS estimates similar to the 2SLS estimates reported in Panel A of the previous table. Estimates in Panel B use an instrument contructed from the sum of mother's age, education, and the samesex dummy. This instrument generates 2SLS estimates identical to the 2SLS estimates reported in Panel A of the previous table. The sample includes married women from the 1980 PUMS aged 21-35 with two or more children. The sample size is 254,652.

Figure 1: Random Forest Residuals are Correlated with Covariates



Note: The figure plots residuals from random forest and least squares fits of an artificial instrument on a linear function of covariates. The covariate list contains mother's education, mother's age, mother's age at first birth, an indicator for the sex of each of the first two children, ages of the first two (in quarters), and three race indicators (black, Hispanic and other race). The artificial instrument is the sum of mother's age and education plus uniform (0,1) noise. Plotted points are averages conditional on the value of mother's age + education.

References

Abadie, Alberto, and Maximilian Kasy. 2018. "Choosing among Regularized Estimators in Empirical Economics: The Risk of Machine Learning." *Review of Economics and Statistics*, 101(5): 743–762.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak. 2017. "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation." *Econometrica*, 85(5): 1373–1432.

Ackerberg, Daniel A, and Paul J Devereux. 2009. "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity." *The Review of Economics and Statistics*, 91(2): 351–362.

Ahrens, Achim, Christian B Hansen, and Mark E Schaffer. 2019. "Lassopack: Model Selection and Prediction with Regularized Regression in Stata." *arXiv preprint arXiv:1901.05397*.

Ananat, Elizabeth O, and Guy Michaels. 2008. "The Effect of Marital Breakup on the Income Distribution of Women with Children." *Journal of Human Resources*, 43(3): 611–629.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics*, 11(1): 727–753.

Angrist, Joshua, and William Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review*, 88(3): 450–77.

Angrist, Joshua D. 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of business & Economic Statistics*, 19(1): 2–28.

Angrist, Joshua D, and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106(4): 979–1014.

Angrist, Joshua D., and Alan B. Krueger. 1995. "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business & Economic Statistics*, 13: 225–35.

Angrist, Joshua D, and Alan B Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*. Vol. 3, 1277–1366. New Holland:Elsevier.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton:Princeton University Press.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics*, 14: 57–67.

Ash, Elliott, Daniel Chen, Xinyue Zhang, Zhe Huang, and Ruofan Wang. 2018. "Deep IV in Law: Analysis of Appellate Impacts on Sentencing Using High-Dimensional Instrumental Variables." http://users.nber.org/ dlchen/papers/Deep_IV_in_Law.pdf (Working Paper).

Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics*, 11: 685–725.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics*, 47(2): 1148–1178.

Bajari, Patrick, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. 2015. "Machine Learning Methods for Demand Estimation." *American Economic Review*, 105(5): 481–85.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica*, 62(3): 657.

Bekker, Paul A, and Jan Van Der Ploeg. 2005. "Instrumental Variable Estimation Based on Grouped Data." *Statistica Neerlandica*, 59(3): 239–267.

Belloni, Alexandre, and Victor Chernozhukov. 2011. "High Dimensional Sparse Econometric Models: An Introduction." In *Inverse Problems and High-Dimensional Estimation.* 121–156. New York:Springer.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica*, 80(6): 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian B Hansen. 2013. "Inference for High-Dimensional Sparse Econometric Models." In *Advances in Economics and Econometrics: Tenth World Congress of Econometric Society, Volume III.* , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel, Chapter 7, 245–295. Cambridge:Cambridge University Press.

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

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014*b*. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies*, 81(2): 608–650.

Belloni, A, V Chernozhukov, and C Hansen. 2011. "Lasso Methods For Gaussian Instrumental Variables Models." *arXiv preprint arXiv:1012.1297.*

Bickel, Peter J, Yaacov Ritov, Alexandre B Tsybakov, et al. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector." *The Annals of Statistics*, 37(4): 1705–1732.

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association*, 90(430): 443–450.

Breiman, Leo. 2001. "Random Forests." *Machine Learning*, 45(1): 5–32.

Card, David, and Alan B Krueger. 1992*a*. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy*, 100(1): 1–40.

Card, David, and Alan B Krueger. 1992*b*. "School Quality and Black-White Relative Earnings: A Direct Assessment." *The Quarterly Journal of Economics*, 107(1): 151–200.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *The American Economic Review*, 84(4): 772–793.

Carrasco, Marine. 2012. "A Regularization Approach to the Many Instruments Problem." *Journal of Econometrics*, 170(2): 383–398.

Carrasco, Marine, and Guy Tchuente. 2015. "Regularized LIML for Many Instruments." *Journal of Econometrics*, 186(2): 427 – 442.

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey. 2018*a*. "Alternative Asymptotics and the Partially Linear Model with Many Regressors." *Econometric Theory*, 34(2): 277–301.

Cattaneo, Matias D, Michael Jansson, and Whitney K Newey. 2018*b*. "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity." *Journal of the American Statistical Association*, 113(523): 1350–1361.

Chamberlain, Gary, and Guido Imbens. 2004. "Random Effects Estimators with Many Instrumental Variables." *Econometrica*, 72(1): 295–306.

Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen. 2012. "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments." *Econometric Theory*, 28(1): 42–86.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments." *American Economic Review*, 105(5): 486–90.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018*a*. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal*, 21(1): C1–C68.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018*b*. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." NBER Working Paper No. 24678.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov. 2019. "On Cross-Validated Lasso." *arXiv preprint arXiv:1605.02214*.

Cruz, Luiz M, and Marcelo J Moreira. 2005. "On the Validity of Econometric Techniques with Weak Instruments Inference on Returns to Education using Compulsory School Attendance Laws." *Journal of Human Resources*, 40(2): 393–410.

Dale, Stacy Berg, and Alan B Krueger. 2002. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." *The Quarterly Journal of Economics*, 117(4): 1491–1527.

Davidson, Russell, and James G MacKinnon. 1993. *Estimation and Inference in Econometrics.* Oxford:Oxford University Press.

Donald, Stephen G, and Whitney K Newey. 2001. "Choosing the Number of Instruments." *Econometrica*, 69(5): 1161–1191.

Ehrlich, Isaac. 1975. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *American Economic Review*, 65(3): 397–417.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1): 1–22.

Fuller, Wayne A. 1977. "Some Properties of a Modification of the Limited Information Estimator." *Econometrica*, 45(4): 939–953.

Gilchrist, Duncan Sheppard, and Emily Glassberg Sands. 2016. "Something to Talk About: Social Spillovers in Movie Consumption." *Journal of Political Economy*, 124(5): 1339–1382.

Goller, Daniel, Michael Lechner, Andreas Moczall, and Joachim Wolff. 2019. "Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed." IZA Working Paper 12526.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica*, 66: 315–31.

Hall, Alastair, Glenn Rudebusch, and David Wilcox. 1996. "Judging Instrument Relevance in Instrumental Variables Estimation." *International Economic Review*, 37(2): 283–98.

Hansen, Christian, and Damian Kozbur. 2014. "Instrumental Variables Estimation with Many Weak Instruments using Regularized JIVE." *Journal of Econometrics*, 182(2): 290–308.

Hansen, Christian, Jerry Hausman, and Whitney Newey. 2008. "Estimation With Many Instrumental Variables." *Journal of Business & Economic Statistics*, 26(4): 398–422.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2016. "Counterfactual Prediction with Deep Instrumental Variables Networks." *arXiv preprint arXiv:1612.09596.*

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. "Deep IV: A Flexible Approach for Counterfactual Prediction." *Proceedings of the 34th International Conference on Machine Learning-Volume 70.*

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: the Lasso and Generalizations.* London:Chapman and Hall/CRC.

Hausman, Jerry A, Whitney K Newey, Tiemen Woutersen, John C Chao, and Norman R Swanson. 2012. "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments." *Quantitative Economics*, 3(2): 211–255.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67.

Katz, Lawrence F, and Alan B Krueger. 1992. "The Effect of the Minimum Wage on the Fast-Food Industry." *Industrial and Labor Relations Review*, 46(1): 6–21.

Knaus, Michael, Michael Lechner, and Anthony Strittmatter. 2018. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence." CEPR Discussion Paper No. DP13402.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens. 2015. "Identification and Inference with Many Invalid Instruments." *Journal of Business & Economic Statistics*, 33(4): 474–484.

Krueger, Alan B. 1991. "Ownership, Agency, and Wages: An Examination of Franchising in the Fast Food Industry." *The Quarterly Journal of Economics*, 106(1): 75–101.

Krueger, Alan B. 1993. "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989." *The Quarterly Journal of Economics*, 108(1): 33–60.

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

Kunitomo, Naoto. 1980. "Asymptotic Expansions of the Distributions of Estimators in a Linear Functional Relationship and Simultaneous Equations." *Journal of the American Statistical Association*, 75(371): 693–700.

Leamer, Edward. 1983. "Let's Take the Con out of Econometrics." *American Economic Review*, 73(1): 31–43.

Lee, Brian K, Justin Lessler, and Elizabeth A Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine*, 29(3): 337–346.

Morimune, Kimio. 1983. "Approximate Distributions of k-Class Estimators when the Degree of Overidentifiability is Large Compared with the Sample Size." *Econometrica*, 521(3): 821–841.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, 31(2): 87–106.

Okui, Ryo. 2011. "Instrumental Variable Estimation in the Presence of Many Moment Conditions." *Journal of Econometrics*, 165(1): 70–86.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. 2012. "How Many Trees in a Random Forest?" 154–168, Springer. New York.

Robinson, Peter M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica*, 56(4): 931–954.

Schonlau, Matthias. 2019. "RFOREST: Stata module to implement Random Forest algorithm." Boston College Department of Economics Working Paper S458614.

Staiger, Douglas, and James H Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65(3): 557–586.

Stata. 2019. *Stata Lasso Reference Manual Version 16.* College Station, Texas:Stata Press.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B*, 58(1): 267–288.

Townsend, Wilbur. 2017. "ELASTICREGRESS: Stata module to perform elastic net regression, lasso regression, ridge regression." Boston College Department of Economics Working Paper S458397.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2016. "Using Double-Lasso Regression for Principled Variable Selection." SSRN Working Paper No. 273374.

Wuthrich, Kaspar, and Ying Zhu. 2019. "Omitted variable bias of Lasso-based inference methods: A finite sample analysis." SSRN Working Paper No. 3379123.

Appendix

We experimented with lasso commands in Stata 16, documented in Stata (2019), the Lassopack routines for Stata 15 documented in Ahrens, Hansen and Schaffer (2019), and the Elasticregress routines for Stata 15 documented in Townsend (2017). Different routines default to or allow the user to select different sorts of penalties. As in Belloni, Chernozhukov and Hansen (2013) and related work, our estimates use two types of penalty terms, one a Bickel et al. (2009)-type (BRT) plug-in penalty, described in Belloni, Chernozhukov and Hansen (2014*b*), another using cross-validation. Each lasso estimation procedure has 3 parts: penalty selection; lasso estimation (to select controls or instruments); final estimation. Stata 16 lasso commands applied to problems other than estimation of elite college effects proved slow and, in some cases, numerically unstable, as documented in Table 3. This led us to Lassopack and Elasticregress for lasso estimation with large samples and/or high dimensional controls and instruments. Random forest estimates reported in Table 3 were computed in Stata; those in Table 5 use Stata and R.

Table 2: Post-Lasso Estimates of Elite College Effects

The sample size is 14,238. Estimates in columns 1-3 are from a post-double-selection (PDS) lasso procedure. Results in columns 4-6 are from a procedure applying lasso to a reduced-form regression of the outcome on the dictionary of controls. Controls include those used for the "self-revelation" model in Table 1 plus the following: indicators for being accepted to two colleges, three colleges, and four or more colleges; indicators for being rejected from one college, two colleges, three colleges, and four or more colleges; the number of schools applied to; the average SAT score among schools at which the applicant was accepted; the average SAT score among schools from which the applicant was rejected; the highest average SAT score across schools at which the applicant was accepted; the highest average SAT score across schools from which the applicant was rejected; the lowest average SAT score across schools at which the applicant was accepted; the lowest average SAT score across schools from which the applicant was rejected, and all two-way interactions and squared terms associated with the underlying list of possible controls. The dictionary of controls includes 384 variables, of which 303 are linearly independent. Penalties are computed once, using the original AK91 sample. Computational details are as follows:

- Columns 1 and 4 (Plug-in penalties)
  - Penalty: Uses Stata 16 `lasso linear` command to select controls, specifying a plug-in penalty, specifically,

```
lasso linear log-income 'dictionary-of-controls' [iw=weight],selection(plugin,het)
lasso linear elitetreatment 'dictionary-of-controls' [iw=weight],selection(plugin,het)
```

  - Lasso (control selection): Uses `lasso linear` as specified above.

- Final estimates: Least squares regression of the outcome on the elite school variable, controlling for the variables selected by lasso (as needed for PDS and single selection), using College and Beyond sampling weights, with standard errors clustered at the institution level.

- Columns 2 and 5 (Cross-validated penalties)

  - Penalty: Computed using `lasso linear` to select controls using 10-fold cross-validation, which is the default in this case,

    ```
    lasso linear log-income 'dictionary-of-controls' [iw=weight]
    lasso linear elitetreatment 'dictionary-of-controls' [iw=weight]
    ```

  - Lasso (control selection): Uses `lasso linear` as specified above.
  - Final estimates: Least squares regression of the outcome on elite school variables, controlling for the variables selected by lasso (as needed for PDS and single selection), using College and Beyond sampling weights, with standard errors clustered at the institution level.

- Columns 3 and 6 (Cross-validated penalties using `cvlasso`)

  - Penalty: Computed using the Lassopack command `cvlasso`. This yields a cross-validated MSE-minimizing penalty level, λ^{CV}. Lassopack `rlasso` was also used to compute the default plug-in penalty, $\lambda^{default}$, specifying institutional weights and clustering (clustering induces robust, covariate-specific penalty loadings following Belloni, Chernozhukov and Hansen 2014*b*). A cross-validated penalty scaling factor is then computed as $c^{CV} = 1.1\lambda^{CV}/\lambda^{default}$. The factor 1.1 arises because the default scaling factor in `rlasso` is 1.1.
  - Lasso (control selection): Uses `rlasso`, with c^{CV} replacing $c = 1.1$, specifying weights and clustering at the institution level.
  - Final estimates: Least squares regression of the outcome on elite school variables, controlling for the variables selected by lasso (as needed for PDS and single selection), using College and Beyond sampling weights, with standard errors clustered at the institution level.

Appendix Table A1: Alternative Post-Lasso Estimates of Elite College Effects

This table reports estimates using Lassopack `command rlasso` and Elasticregress command `lassoregress.`

- Columns 1, 5, and 9 use`rlasso` and a plug-in penalty

  - Penalty: Computed using Lassopack `rlasso` with default plug-in penalty, specifying weights and clustering by institution.
  - Lasso (control selection): Uses `rlasso` with regressor-specific penalty loadings.
  - Final estimates: Least squares regression of the outcome on elite school variables, controlling for the variables selected by Lasso, using College and Beyond sampling weights, with standard errors clustered at the institution level.

- Remaining columns use the Elasticregress command `lassoregress` and cross-validated penalties done three ways

  - Penalty: Computed via 10-fold cross validation as implemented in `lassoregress`, specifying institutional weights. Columns 2, 6, and 10 use the default cross-validated penalty, which minimizes cross-validated MSE. Columns 3, 7, and 11 specify the option`, lambda1se`, which uses the largest penalty such that the cross-validated MSE is within one standard deviation of the minimum. Columns 4, 8, and 12 calculate the penalty as 10 times the default.

  - Lasso (control selection): Estimated by the `lassoregress` command call that computes penalties.

  - Final estimates: Least squares regression of the outcome on elite school variables, controlling for the variables selected by Lasso, using College and Beyond sampling weights, with standard errors clustered at the institution level.

Elite college effects computed using the R-based package glmnet (Friedman, Hastie and Tibshirani, 2010) are similar to those reported in Tables A1 and 2. When applied to estimate the propensity score, however, the number of controls retained under glmnet-determined cross-validated penalties is generally much larger than the number of controls retained by Lassopack and Elasticregress. This is the result of a smaller penalty chosen for equation (16); glmnet lasso with cvlasso-determined penalties behaves like the Stata lasso routines, as does glmnet lasso on the PDS reduced form, equation (17).

Table 3: Angrist and Krueger (1991) Simulation Results

For estimates using the large AK91 sample with many fixed effects, Lassopack was faster and appeared to be more stable than Stata 16's `poivregress (`used for Table 4 and described below)

- Penalty: Lasso estimates computed with a plug-in penalty use Lassopack `ivlasso` with default parameters. Estimates using cross-validated penalties were computed using Lassopack `cvlasso` as applied to a first-stage equation, specifying the option`,fe` to control for a full set of state-of-birth and year-of-birth interactions. A scaling factor for `ivlasso` is then computed as described for Table 2, above.

- Lasso (instrument selection): Computed using Lassopack routine `ivlasso` controlling for state-of-birth and year-of-birth interactions via the`,fe` option. Estimates using plug-in penalties use Lassopack `ivlasso` defaults. Estimates using cross-validated penalties employ the scaling adjustment described for the `cvlasso` estimates reported in Table 2. Note that `ivlasso` computes first stage estimates by calling `rlasso`.

- Final estimates: Computed via post-lasso 2SLS using `ivlasso.`

LIML estimates were computed using Stata `ivregress liml`. Lassoed versions of LIML use the instrument lists chosen for post-lasso 2SLS. Pretested LIML estimates use the instrument list described in the text. SSIV estimates split the sample in equal-sized halves randomly. One half-sample is used to estimate first stage parameters by OLS; these are carried over to the second half to compute cross-sample fitted values. Cross-sample fitted values and covariates are used to compute second-stage parameters using `ivregress`. The sample halves are then swapped, and the two resulting estimates averaged.

Post-lasso LIML estimates use the instruments chosen for post-lasso 2SLS. Post-lasso SSIV recomputes the lasso first stage in each half sample.

IJIVE estimates use the improved jackknifed instrumental variables procedure described in Ackerberg and Devereux (2009). We implement this by first partialing the covariates (state-of-birth and year-of-birth interactions) out of the outcome, endogenous regressor, and instruments, and then applying the `jive` Stata command to the residualized variables.

Random forest IV estimates in this table use Stata's `rforest` command (documented in Schonlau (2019)) to fit the first stage, with predictors YOB, POB, and QOB. Random forest estimates were computed with a minimum leaf size of 1 and 800, averaging results from 100 trees with no maximum depth (these are `rforest` defaults; Oshiro, Perez and Baranauskas 2012 finds little payoff to more trees). The number of variables randomly investigated is equal to the square root of the number of right-hand-side variables (also a default setting). Random forest 2SLS uses random forest fitted values as excluded instruments, in a model with saturated control for year of birth and state of birth. Random forest SSIV fits the first stage with Stata command `rforest` in half the sample, assigning cross-sample fitted values to the relevant cells in the other half. Second-stage estimates are then obtained using these cross-fitted fitted values as instruments with saturated YOB-by-POB controls. As with the other SSIV estimates in the table, random forest SSIV swaps half samples and averages the resulting second-stage estimates from each.

Table 4: Simulation Results for Opening Weekend Effects

These simulations link a film's opening weekend ticket sales to subsequent ticket sales. OLS estimates are from a regression of second-weekend ticket sales on opening weekend ticket sales, controlling for year, week-of-year, day-of-week, and holiday dummies, as well as a set of second-weekend weather controls that includes indicators for the maximum temperature in 10-degree increments and indicators for rain, snow, and indicators for average precipitation in quarter inches per hour. 2SLS estimates (computed by `ivregress 2sls`) in columns 1-5 include the same exogenous controls used for OLS, with a set of 52 opening-weekend weather indicators used as excluded instruments. Estimates in columns 6-10 these instruments plus an additional 52 uniform noise instruments. Lasso estimates were computed as follows:

- Penalty: Lasso estimates computed with a plug-in penalty use Lassopack `ivlasso` with default parameters, partialing controls using the Lassopack`,partial()` option. Estimates using cross-validated penalties were computed using Lassopack `cvlasso` with opening-weekend ticket sales as the dependent variable and the controls and instruments as explanatory variables, partialing controls using the`,partial()` option. This yields a cross-validated MSE-minimizing penalty level, λ^{CV}. A scaling factor, c^{CV}, is then computed as described for Table 3, above. Penalties are recomputed for each simulation draw.

- Lasso (instrument selection): Computed using Lassopack `ivlasso`, with controls partialed via the`,partial()` option, using the scaling factor computed as described for Table 3, above.

- Final estimates: Computed via post-lasso 2SLS using `ivlasso.`

LIML estimates were computed using `ivregress liml`. SSIV estimates split the sample in equal-sized halves randomly. One half-sample is used to estimate first stage parameters by OLS; these are carried over to the second half to compute cross-sample fitted values. Cross-sample fitted values and covariates are used to compute second-stage parameters using `ivregress`. The sample halves are then swapped, and the two resulting estimates are averaged.

The bottom four rows show results generated using Stata 16's `poivregress` command with the default plug-in penalty. This command is described on page 5 of Stata (2019) as

> ....partialing-out lasso instrumental-variables linear regression. This command estimates coefficients, standard errors, and confidence intervals and performs tests for variables of interest, both exogenous and endogenous, while using lassos to select from among potential control variables and instruments.

In models with a fixed number of controls, we computed `poivregress` estimates using the command

`poivregress` *week2tickets* ‘*low-dim exogenous vars*’ `(`*week1tickets* =‘*high-dim instruments*’`),vce(robust).`

For `poivregress` estimates treating controls as high dimensional, we used

`poivregress` *week2tickets* `(`*week1tickets* =‘*high-dim instruments*’`), controls(`‘*high-dim exogenous vars*’`) vce(robust).`

Pages 267-8 of Stata (2019) describe the multi-step sequence of regression and post-lasso partialing implemented by this command. The fact that `poivregress` reports IV estimates with no instruments retained appears to be an artifact of numerical imprecision in the construction of first-stage residuals computed at the final partialing step.

Figure 6 and Table 5: IV Estimates After Random Forest Partialing

These exhibits use the Angrist and Evans (1998) sample of married women from the 1980 Census. Random forest partialing for Figure 6 uses the Stata `rforest` command discussed in the context of Table 3. Estimates in Table 5 use `rforest` and the `regression_forest` command contained in the Generalized Random Forest (GRF) software package referenced by Athey, Tibshirani and Wager (2019). Residuals plotted in the figure and used as instruments were computed using leaf sizes indicated in legends and column headings. GRF parameter settings mostly equal to those use by Athey, Tibshirani and Wager (2019). The number of variables randomly investigated is equal to the square root of the number of right-hand-side variables plus 20, with a subsample rate of 5 percent. Our implementation computes 100 trees. We obtained similar estimates using much larger numbers of trees. GRF `regression_forest` reports leave-out fitted values, as suggested by Athey, Tibshirani and Wager (2019).

Table A1: Alternative Post-Lasso Estimates of Elite College Effects

	Double-selection (PDS)				Treatment (score) selection				Outcome selection			
	plug-in (15)	C.V. λ	1 S.E. λ	10xC.V. λ	plug-in (15)	C.V. λ	1 S.E. λ	10xC.V. λ	plug-in (15)	C.V. λ	1 S.E. λ	10xC.V. λ
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. Private School Effects												
Estimate	0.055	0.021	0.023	0.037	0.058	0.021	0.021	0.048	0.205	0.039	0.040	0.035
	(0.041)	(0.039)	(0.038)	(0.041)	(0.062)	(0.040)	(0.039)	(0.054)	(0.050)	(0.042)	(0.041)	(0.041)
Treatment residual s.d	0.339	0.304	0.308	0.330	0.339	0.305	0.308	0.332	0.492	0.333	0.335	0.356
No. of controls	6	107	91	27	5	93	86	25	1	34	10	2
B. Effects of School Average SAT/100												
Estimate	-0.014	-0.014	-0.015	-0.014	-0.028	-0.009	-0.009	-0.015	0.111	-0.013	-0.009	-0.018
	(0.023)	(0.018)	(0.019)	(0.020)	(0.027)	(0.020)	(0.021)	(0.021)	(0.018)	(0.020)	(0.021)	(0.017)
Treatment residual s.d	0.414	0.390	0.391	0.409	0.415	0.393	0.393	0.409	0.943	0.411	0.414	0.528
No. of controls	15	90	64	15	14	71	61	14	1	34	6	2
C. Effects of Attending Schools Rated Highly Competitive +												
Estimate	0.079	0.057	0.050	0.073	0.076	0.054	0.052	0.075	0.220	0.076	0.079	0.063
	(0.029)	(0.034)	(0.034)	(0.032)	(0.043)	(0.034)	(0.035)	(0.037)	(0.036)	(0.033)	(0.031)	(0.030)
Treatment residual s.d	0.342	0.289	0.288	0.330	0.342	0.289	0.288	0.332	0.470	0.335	0.337	0.357
No. of controls	8	106	82	30	7	88	79	28	1	34	6	2

Notes: This table reports estimates computed using alternative lasso routines. Estimators are as described in the note to Table 2. Columns 1, 5, and 9 use the default plug-in penalty implemented in Lassopack rlasso; these estimates (like those in Table 2) use regressor-specific penalty loadings. Columns 2, 6, and 10 use a cross-validated penalty, as implemented in the Elasticregress lassoregress command (see Townsend, 2018); these estimates standardize regressors but omit regressor-specific penalty loadings. Columns 3, 7, and 11 use the largest penalty such that the cross-validated mean squared error is within one standard error of the minimum cross-validated mean squared error, a variation implemented in lassoregress. Hastie, Tibishrani, and Wainwright (2016) suggest this modification. Columns 4, 8, and 12 use a penalty equal to 10 times the cross-validated penalty used for columns 2, 6, and 10.