Machine Labor

Joshua D. Angrist, *Massachusetts Institute of Technology and National Bureau of Economic Research*

Brigham Frandsen, *Brigham Young University*

The utility of machine learning (ML) for regression-based causal inference is illustrated by using lasso to select control variables for estimates of college characteristics' wage effects. Post-double-selection lasso offers a path to data-driven sensitivity analysis. ML also seems useful for an instrumental variables (IV) first stage, since two-stage least squares (2SLS) bias reflects overfitting. While ML-based instrument selection can improve on 2SLS, split-sample IV and limited information maximum likelihood do better. Finally, we use ML to choose IV controls. Here, ML creates artificial exclusion restrictions, generating spurious findings. On balance, ML seems ill-suited to IV applications in labor economics.

I. Introduction

Many economic applications of machine learning (ML) originate in research on consumer choice. For instance, Bajari et al. (2015) use ML to predict

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, Serena Ng, Jose Olea, Parag Pathak, Bruce Sacerdote, Bernard Salanie, Stefan Wager, and Chris Walters as well as seminar participants at Columbia, Massachusetts Institute of Technology, and the *JOLE* Conference in Honor of Alan Krueger 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. Contact the corresponding author, Brigham Frandsen, at frandsen@byu.edu.

Submitted December 3, 2020; Accepted October 29, 2021.

Journal of Labor Economics, volume 40, number S1, April 2022.
 https://doi.org/10.1086/717933

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 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). Causal inference focuses on parameters, such as average causal effects, rather than the prediction of individual choices or outcomes.

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 estimate a regression slope coefficient or marginal effect. Empirical findings in labor economics rarely turn on $\hat{y}$. Yet as Belloni, Chernozhukov, and Hansen (2014a) observe, in any empirical application with many covariates, we would like to guard against overfitting and the vagaries of data mining. These concerns extend to causal models with many control variables or many instruments.

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 that 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.

The Dale and Krueger (2002) 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 Dale and Krueger (2002) dummy-variable control strategy is a two-thirds reduction in sample size. We would like to have a more parsimonious control strategy.

In the Dale and Krueger (2002) 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 (Leamer 1983). The post-double-selection (PDS) lasso estimator introduced by Belloni, Chernozhukov, and Hansen (2014b) can addresses this concern. Lasso (Tibshirani 1996), which abbreviates the "least absolute shrinkage and selection operator," is a form of regularized regression that improves out-of-sample prediction by discarding some regressors and shrinking the coefficients on those retained. Post-lasso

estimators use lasso solely for variable selection. 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 model.

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 Dale and Krueger (2002) sample had applied and been accepted. Although the number and identity of lasso-chosen controls change 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 that ML creates valid conditional independence restrictions. Rather, ML tools seem helpful for choosing between alternative specifications that implement a common underlying conditional independence claim.[1]

Our second ML domain is the choice of instruments for IV estimation. Use of ML for instrument selection is motivated by the fact that two-stage least squares (2SLS) estimates in heavily overidentified 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 instruments 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 consider regularized models like lasso for first-stage estimation. We explore a pair of overidentified IV applications that would seem to have a role for ML-based instrument selection (although 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) 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. And the Angrist and Krueger (1995) split-sample IV (SSIV) estimator, an improved jackknife IV (IJIVE)

[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) in sec. III.

estimator introduced by Ackerberg and Devereux (2009), and the limited information maximum likelihood (LIML) estimator are almost always better (i.e., less biased and have lower median absolute error) than 2SLS estimators using a post-lasso first stage, no matter how the lasso penalty is chosen. 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 effectively characterized by a Bekker (1994) many-instrument asymptotic sequence that fixes the ratio of the sample size to the number of first-stage parameters (Angrist, Imbens, and Krueger [1999] demonstrate the empirical relevance of the Bekker sequence for a leading example).

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 instruments for the effect of a movie's opening-weekend viewership on subsequent ticket sales. ML is unimpressive here in spite of the fact that a small number of terms appear to approximate the first-stage conditional expectation function (CEF) well. 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.[2] Even in models with a mix of strong and weak instruments, 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 pretesting of first-stage estimates implicit in lasso.[3]

Our third domain concerns the selection of control variables in IV models with many (possible) control variables but few instruments. This includes applications like that of Angrist and Evans (1998), who estimate causal effects of childbearing on mothers' labor supply using twin births and sibling sex composition as a source of exogenous variation in family size. These just-identified IV estimates are made more plausible by conditioning on maternal characteristics (twin birth rates, e.g., are correlated with maternal age and schooling). Our exploration of this idea, inspired by Athey, Tibshirani, and Wager (2019), shows how random forest procedures founder when confronted with models that require a low-dimensional additive first stage for identification.

[2] Summarizing an analysis of the Angrist and Krueger (1991) data, e.g., Belloni, Chernozhukov, and Hansen (2013, 281) conclude that "the results in Table 5 are interesting and 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.

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

The worst-case scenario here is an estimator with algorithmically induced spurious exclusion restrictions that yield meaningless, yet statistically significant, second-stage estimates.

II. Casting Regression in Two Roles

Regression uses linear models to describe conditional CEFs. The conditional expectation of a random variable, denoted Y_i for person i, given a set of variables, X_i, can be written $\mathrm{E}[Y_i|X_i = x]$. Because $\mathrm{E}[Y_i|X_i = x]$ 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, assuming X_i includes a set of K explanatory variables indexed by k, the $K \times 1$ regression slope vector, β, can be defined as the minimum mean squared error (MMSE) linear approximation:

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

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 $\mathrm{E}[Y_i|X_i]$. That is,

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

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 OLS estimator, denoted here by $\hat{\beta}_{\mathrm{LS}}$, replaces expectations with sums in equation (2) and provides the best linear predictor in the sample in which it is fit.

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

A better out-of-sample predictor augments the least squares minimand, equation (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:

$$\min_b \frac{1}{n}\sum_{i=1}^{n}\{Y_i - X_i'b\}^2 + \lambda\sum_k |b_k|, \tag{3}$$

where λ is a user-chosen tuning parameter.

A second gap between the econometric and predictive ML frameworks arises from the asymmetry with which most empirical labor economists view regressors. The modern empirical paradigm usually distinguishes between the components of X_i: one is a causal variable of interest, the rest a set of supporting controls whose coefficients are of secondary interest. An empirical example highlights the significance of this distinction.

A. 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 (UNC). Is the money spent on private college tuition justified by future earnings gains? The causal regressor here is a dummy variable, D_i, that indicates graduate i attended private college. The outcome of interest, Y_i, is a measure of earnings roughly 20 years after 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 to a private college ($D_i = 1$), 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 $\mathrm{E}[Y_{1i} - Y_{0i}]$, or an effect conditional on treatment, $\mathrm{E}[Y_{1i} - Y_{0i}|D_i = 1]$.

The link between causal inference and regression is easiest to make in 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

$$Y_{0i} = \alpha + \eta_i, \tag{4}$$

$$Y_{1i} = Y_{0i} + \rho, \tag{5}$$

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

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

we can use the constant-effects model to write

$$Y_i = \alpha + \rho D_i + \eta_i. \tag{6}$$

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 estimates are founded on the assumption that

$$\mathrm{E}[\eta_i|D_i = 1, A_i = a] = \mathrm{E}[\eta_i|D_i = 0, A_i = a], \tag{7}$$

where A_i is a vector of 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

$$\mathrm{E}[\eta_i|D_i, A_i] = \mathrm{E}[\eta_i|A_i]. \tag{8}$$

In the Dale and Krueger (2002) empirical strategy, the control vector A_i identifies the sets of schools to which 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. Angrist and Pischke (2015) provide support for this claim, showing that applicants matched on A_i have similar family income and SAT scores.

The last element of the causal regression story is the assumption that the conditional mean of η_i is a linear function of A_i:

$$\mathrm{E}[\eta_i|A_i] = \gamma' A_i. \tag{9}$$

This implies

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

where it is surely true that

$$\mathrm{E}[\varepsilon_i|A_i] = 0. \tag{10}$$

Combining equations (8) and (9) generates a linear CEF with a causal interpretation:

$$\begin{aligned}\mathrm{E}[Y_i|A_i, D_i] &= \alpha + \rho D_i + \mathrm{E}[\eta_i|A_i] \\ &= \alpha + \rho D_i + \gamma' A_i.\end{aligned}$$

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 equation (8).[4]

Regression produces the best (in-sample) linear predictor of individual outcomes, but our interest in equation (11) does not derive from this fact. Indeed, the predictive accuracy of most labor regressions as measured by R^2 is pitiable. Rather, causal regressions like equation (11) are valuable if and when they generate unbiased estimates of causal effects. An analyst who seeks only to predict the wages of college graduates would likely do well to dispense with elite college measures and application and admission indicators altogether, focusing instead on where graduates live and work, since these variables are highly predictive of earnings. Yet this sort of analysis misses the point of causal inquiry.

B. Benchmark Private College Premia

Private college alumni earn more than those who attended public schools. 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 10 controls for applicant ability, like SAT scores and class rank, and for 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 and 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 and 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 is helpful to group similarly selective schools like Princeton

[4] 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 (this interpretation is detailed in Angrist and Krueger 1999).

Table 1
OLS Estimates of Elite College Effects

	Basic Controls		DK02 Selection Controls			
					Self-Revelation	
	None (1)	Personal Characteristics (2)	Barron's Matches Only (3)	Barron's Matches with Personal Characteristics (4)	Barron's Sample (5)	Full Sample (6)
	A. Private School Effects					
Estimate	.212	.139	.007	.013	.036	.037
	(.060)	(.043)	(.038)	(.025)	(.029)	(.039)
R^2	.019	.107	.058	.138	.111	.114
Number of controls	0	10	150	160	13	14
N	14,238			5,583		14,238
	B. Effects of School-Average SAT Score/100					
Estimate	.109	.076	.008	.004	.004	.000
	(.026)	(.016)	(.029)	(.016)	(.017)	(.018)
R^2	.019	.107	.066	.140	.107	.113
Number of controls	0	10	334	344	13	14
N	14,238			9,166		14,238
	C. Effects of Attending Schools Rated Highly Competitive or Better					
Estimate	.225	.153	.018	.022	.031	.068
	(.046)	(.030)	(.047)	(.035)	(.032)	(.029)
R^2	.020	.108	.048	.129	.106	.114
Number of controls	0	10	128	138	13	14
N	14,238			4,945		14,238

NOTE.—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 col. 2 include graduates' SAT scores, log parental income, and indicators for female, black, Hispanic, Asian, other/missing race, high school top 10%, high school rank missing, and athlete. Controls for estimates reported in panel A, col. 3, include 150 dummies (for 151 categories) indicating the Barron's selectivity mix of schools to which graduates applied and were admitted. Controls for col. 4 include Barron's dummies and the personal characteristics used for col. 2. The Barron's model in panel B includes 334 dummies; the Barron's model in panel C includes 128 dummies. Columns 5 and 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. DK02 = Dale and Krueger (2002).

and Yale together. The models used to construct the estimates in columns 3 and 4 therefore control for sets of schools grouped by their Barron's selectivity (Barron's magazine groups schools into six selectivity groups). This model can be written

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

where $\{\text{GROUP}_{ji}; j = 1, \ldots, 150\}$ 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 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, although conditional on the latter, the former may be unnecessary.[5]

Estimates of equation (12) suggest that the earnings premium enjoyed by private college graduates 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, like SAT score and family background, that are strongly predictive of earnings. Is this lack of omitted-variable bias (OVB) a robust result, or did we just get lucky? ML helps answer this.

III. Welcome to the Machine

A. 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 econometric extension of the ML toolkit, Belloni, Chernozhukov, and Hansen (2014b) introduce the method of PDS, an empirical strategy that uses lasso to pick regression control variables. The Dale and Krueger (2002) 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 equations (4) and (5), the identifying assumption motivating PDS can be stated as

$$\mathrm{E}[\eta_i|D_i, A_i] = \mathrm{E}[\eta_i|A_i] = g(A_i), \tag{13}$$

where A_i is a vector of control variables as before. Function $g(A_i)$ is shorthand for an unknown, possibly nonlinear model capturing the dependence of potential outcomes on controls. The Belloni, Chernozhukov, and Hansen

[5] College selectivity categories are determined by Barron's *Profiles of American Colleges 1978*. For example, one of the selectivity groups coded by $\{\text{GROUP}_{ji}; j = 1, \ldots, 150\}$ 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 colleges 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 pretax annual earnings in 1995. Regressions are weighted to make the sample representative of the population of graduates of 30 College and Beyond institutions. A total of 68.6% of the sample with Barron's matches attended a private school.

(2014b) framework maintains the hypothesis that the conditioning variables, A_i, are observed.

Faced with an abundance of candidate controls, PDS finds a list of variables adequate to control OVB while rendering causal inference feasible. This strategy has an analog in a partially linear model in the style of Robinson (1988). The partially linear model of interest here can be written

$$Y_i = g(A_i) + \rho D_i + \varepsilon_i, \tag{14}$$

where condition (13) ensures that the error term, ε_i, is mean independent of A_i and D_i, and ρ is defined as a constant additive causal effect.

PDS adds to this structure a model for the propensity score—that is, for the conditional probability of treatment given A_i—and a model for the reduced form, defined as $\mathrm{E}[Y_i|A_i]$. These can be written as follows:

$$\mathrm{E}[D_i|A_i] = m_1(A_i), \tag{15}$$

$$\mathrm{E}[Y_i|A_i] = m_0(A_i) = g(A_i) + \rho m_1(A_i). \tag{16}$$

A partially linear estimator of ρ regresses $Y_i - m_0(A_i)$ on $D_i - m_1(A_i)$, replacing unknown functions with consistent nonparametric estimates thereof. This yields consistent estimates of ρ because both $Y_i - m_0(A_i)$ and $D_i - m_1(A_i)$ are uncorrelated with the unknown control function, $g(A_i)$. In fact, consistency requires only one of m_0 or m_1 to be specified correctly. In the Robinson model, A_i is low dimensional, while the PDS scenario envisions a need for dimension reduction.

PDS uses the assumption of (approximate) sparsity to approximate the unknown, possibly nonlinear regression functions $m_0(A_i)$ and $m_1(A_i)$. Let $\tilde{A}_i$ denote A_i augmented with transformations, adding, for example, polynomial terms, dummy variables, and interaction terms. Approximations of m_1 and m_0 can then be written

$$m_1(A_i) = \gamma_1' \tilde{A}_i + r_{1i},$$

$$m_0(A_i) = \gamma_0' \tilde{A}_i + r_{0i},$$

where r_{0i} and r_{1i} are approximation errors. The assumption of approximate sparsity means that all but a few of the elements of γ_1 and γ_0 are zero and that the approximation errors are small in a sense made precise in Belloni, Chernozhukov, and Hansen (2014b). In other words, selection bias can be eliminated using only a subset of the elements of $\tilde{A}_i$.

The augmented control vector, $\tilde{A}_i$, is sometimes said to make up a "dictionary" of possible controls. The dictionary is presumed to be of dimension p, where p may exceed sample size. Even where the underlying set of controls is of modest dimension, the fact that functions m_0 and m_1 are left unspecified can lead to a a high-dimensional set of controls. Importantly, however, $m_0(A_i)$ is presumed to be well approximated by $s_0 < n$ regressors, while $m_1(A_i)$ is likewise well approximated by $s_1 < n$ regressors.

PDS implements sparsity via lasso, a regularized regression estimator that minimizes expression (3). Lasso deletes some variables from the covariate dictionary while shrinking the coefficients on those retained toward 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 m_0 and m_1. Exploiting the properties of multivariate regression models, PDS obtains the bivariate regression of $Y_i - m_0(A_i)$ on $D_i - m_1(A_i)$ by OLS estimation of

$$Y_i = \pi' M_i + \rho D_i + \xi_i, \tag{17}$$

where ξ_i is a regression residual.

The key approximate sparsity condition supporting PDS, formalized by Belloni, Chernozhukov, and Hansen (2014b), puts limits on $s = \max\{s_0, s_1\}$, the maximum number of nonzero coefficients in the linear approximation to $m_0(A_i)$ and $m_1(A_i)$. This condition requires that $s^2\log^2(\max\{p, n\})/n$ converges to zero and that the root mean squares of the approximation error terms, r_{0i}, r_{1i}, be no more than $C\sqrt{s/n}$ for some fixed constant C.[6]

How important is *double* selection, that is, the fact that M_i contains the union of post-lasso-chosen controls for the outcome reduced form and the treatment propensity score? Given conditional independence assumption (8), a regression of Y_i on D_i and reduced-form controls yields unbiased estimates. In practice, the reduced form is unknown, and the post-lasso approximation to it is inherently imperfect. Inclusion of lasso-retained covariates from the propensity score mitigates the bias arising from such specification errors.[7]

B. Estimating Elite College Effects

The estimates at the top of 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 more than 14,000 observations in the full College and Beyond sample of graduates with earnings. A key problem here is that many selectivity groups are populated by sets of graduates in which D_i equals 0 or 1 for everyone. OLS estimation of a model including the full set of Barron's dummies is implicitly a panel-data-style within-group estimator that drops observations in

[6] Also needed for asymptotic normality are a sparse eigenvalue condition on the Gram matrix, $E[A_iA_i']$, and the existence of bounds on various moments of the data; see Belloni, Chernozhukov, and Hansen (2014b) for details.

[7] More formally, Belloni, Chernozhukov, and Hansen (2014b) establish the uniform consistency of a PDS procedure given regularity conditions under which single-selection estimators are not uniformly consistent. Doubly-robust propensity score estimators offer similar bias mitigation in a semiparametric context (see, e.g., Bang and Robins 2005).

covariate cells where the regressor D_i is fixed at 0 or 1—that is, cells with a degenerate propensity score. The sample that can be used to estimate equation (12) need not be representative of the population covered by the College and Beyond survey.

PDS offers an alternative to full selectivity-group control while retaining variables that seem likely to mitigate OVB. It is worth noting, however, that although a post-lasso algorithm applied to the set of Barron's dummies included in equation (12) may drop some of these dummies, it will not 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 those in the omitted group with the original reference group. Likewise, lasso will not pool groups of applicants with a degenerate probability of assignment in a manner that makes such groups informative about treatment effects. Recognizing problems with lasso applied to dummy variables, analysts have proposed lasso-type strategies that penalize differences in coefficients like the δ_js in equation (12). But such methods (known as fused lasso) seem unlikely to be attractive for models with categorical control variables indicating many categories. Our PDS implementation therefore builds on an alternative control strategy that does not rely on a large set of dummy controls.[8]

In addition to saturated control for 151 Barron's selectivity groups, the Dale and Krueger (2002) study 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 Dale and Krueger (2002) paper labels this specification a "self-revelation model." The 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 are likely to be well matched. An applicant's self-assessment is reflected in the average selectivity of the schools they have targeted, while the number of applications submitted is a gauge of academic ambition.

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

[8] 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 parameter constraints required for approximate sparsity. It also seems worth noting that lasso estimators, like ridge estimators, are sensitive to the choice of omitted group when categorical variables are coded with dummies.

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 Dale and Krueger (2002) 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 [2014b] evaluate PDS in a simulation study involving a normally distributed regressor). With continuous treatments, function $m_1(A_i)$ becomes a model for the conditional mean of treatment rather than the conditional probability of treatment.

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 (cols. 2 and 4 of this panel replicate results reported in Dale and Krueger 2002). 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 (denoted 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 zeros 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.

C. PDS-Supported Sensitivity Analysis

The small set of Dale and Krueger (2002) self-revelation controls yields a model estimable in the full College and Beyond 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 average selectivity. ML methods—and PDS in particular—seem useful for a systematic exploration of the sensitivity of causal conclusions when many equally plausible specifications are available.

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 columns 5 and 6 of table 1. The 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, at the least selective school, and for all schools where the applicant was accepted; mean SAT scores at the most selective school, at the 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. This dictionary encompasses a wide range of alternatives to the self-revelation model.

PDS estimates of private college effects, reported in the first three columns of panel A in 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 (2014b), is data driven although 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.[9]

[9] The Belloni et al. (2012) and Belloni, Chernozhukov, and Hansen (2014b) plug-in penalties generalize the penalty formula proposed by Bickel, Ritov, and Tsybakov (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 $\gamma = 0.1/\log(n)$). Belloni, Chernozhukov, and Hansen (2014b) suggest using $c = 1.1$ and $\gamma = 0.05$.

Table 2
Post-lasso Estimates of Elite College Effects

	Double Selection (PDS)			Outcome Selection			All Controls
	Plug-In (1)	CV λ (2)	cvlasso (3)	Plug-In (7)	CV λ (8)	cvlasso (9)	OLS (7)
	A. Private School Effects						
Estimate	.038	.020	.040	.046	.043	.042	.017
	(.040)	(.039)	(.041)	(.041)	(.043)	(.043)	.039
Number of controls	18	100	112	10	35	50	303
	B. Effects of School-Average SAT Score/100						
Estimate	−.009	−.013	−.009	−.008	−.009	−.008	−.012
	(.020)	(.018)	(.019)	(.020)	(.019)	(.019)	(.018)
Number of controls	24	151	58	10	34	43	303
	C. Effects of Attending Schools Rated Highly Competitive or Better						
Estimate	.068	.051	.073	.076	.080	.082	.053
	(.033)	(.033)	(.033)	(.031)	(.032)	(.032)	.033
Number of controls	17	185	106	10	34	43	303

NOTE.—The sample size is 14,238. Estimates in cols. 1–3 are from PDS lasso procedures. Results in cols. 4–6 are from a procedure applying lasso to a reduced-form regression of 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 tenfold cross validation (CV) 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 details. Column 7 reports OLS estimates including the entire set of controls. Controls include those used for col. 5 of 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 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.

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 an important caution for practitioners: implementation details are likely to 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 (available online) details our choices in this regard further. Table A1 (tables A1–A3 are available online) compares estimates of elite college effects computed using additional cross-validation schemes. With one exception, these are qualitatively similar to the estimates reported in table 2.

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+ 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.

It is interesting to contrast PDS with single-selection lasso. 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. The argument for double selection given in section III.A implies that 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.[10]

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 is not truly high dimensional in the sense of containing more variables than 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 the

[10] 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 estimates. These results appear in table A1.

quality of PDS bias mitigation depends on design features like regressor variance and the extent of OVB. Moreover, we have 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. Findings where the target causal estimate remains reasonably stable while the list of selected controls varies from one routine to another reinforce claims of robustness.

It is worth emphasizing that a 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. This facilitates estimation in high-dimensional control scenarios and may increase precision (although that is not the finding here). We have also noted considerable sensitivity to implementation details, specifically to software choice and lasso penalty determination.

IV. ML Picks Instruments

The sampling variance of a 2SLS estimate is inversely proportional to the first-stage R^2. This fact encourages the use of many instruments. On the other hand, 2SLS estimates are biased, with a finite-sample distribution shifted toward 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 suggest a fruitful empirical strategy that uses ML 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.[11]

A. Machining the Angrist and Krueger (1991) 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. The QOB identification strategy is motivated by the fact that children who start school at an older age attain the minimum school dropout age after having completed less schooling than

[11] 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.

those who enter school younger. Because most children start school in the year they turn 6 years old, 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. Angrist and Krueger (1991) document a strong QOB first stage, showing that highest grade completed increases with QOB for US men born in the 1920s and 1930s.

The Angrist and Krueger (1991) endogenous variable is highest grade completed; the dependent variable is the log weekly wage. Our replication focuses on a sample of 329,509 men born between 1930 and 1939 from the 1980 census public use files. In this sample, a regression of schooling on three QOB dummies and nine year of birth (YOB) dummies generates an *F*-statistic for the three excluded QOB instruments of around 36. This strong relationship reflects the fact that fourth-quarter births complete around 0.12 more years of schooling than first-quarter births and are around 2 percentage points more likely to graduate high school (*t*-statistics for these effects exceed 7). 2SLS estimates using three QOB dummies as instruments therefore seem unlikely to suffer substantial weak-instrument bias.

The many-weak-instrument angle surfaces when QOB dummies are interacted with dummies for YOB and place (state) of birth (POB). These interactions are motivated by the fact that the relationship between QOB and schooling varies both across cohorts (as compulsory attendance laws have grown less important) and across states (since states set school attendance policy).[12] Interacting three QOB dummies with nine YOB and 50 POB dummies (including one for the District of Columbia) generates 180 excluded instruments. 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-instrument first stage may generate estimates of the economic returns to schooling that are close to the corresponding OLS estimates solely by virtue of finite-sample bias. A fully interacted QOB-YOB-POB first stage has 1,530 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 aspire not 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. Our investigation begins by examining an ML strategy in which conventional 2SLS is carried out using the instrument set retained by a lasso preliminary stage, an approach suggested by Belloni et al. (2012).

[12] Angrist and Krueger (1992) explore this heterogeneity.

Lasso for instrument selection is evaluated here in a simulation experiment calibrated so that OLS estimates are misleading. In the absence of omitted variables or endogeneity bias in OLS estimates, it is 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 Angrist and Krueger (1991) sample generates an estimated return to schooling of 0.074. The corresponding OLS estimate is 0.071.[13] This just-identified IV estimate (which, like LIML, is approximately median unbiased) suggests that 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 overidentified Angrist and Krueger [1991] estimates have little bias). This leads us to craft a simulation design that preserves the structure of the Angrist and Krueger (1991) sample and IV estimates but introduces substantial OVB in the corresponding OLS estimates.

Our simulated endogenous school variable is constructed from average highest grade completed in each QOB-YOB-POB cell, computed in the Angrist and Krueger (1991) 1980 sample (for a total of 2,040 means). Call these cell averages $\bar{s}(q, c, p)$, where $q = 1, \ldots, 4$; $c = 1{,}930, \ldots, 1{,}939$; and $p = 1, \ldots, 51$. Simulated schooling, $\tilde{s}_i$, is a Poisson draw with mean μ_i, where

$$\mu_i = \max[1, \bar{s}(Q_i, C_i, P_i) + \kappa_1 \nu_i] \tag{18}$$

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. Mean μ_i is generated with the aid of a standard normal variable, denoted ν_i, multiplied by a scale parameter, κ_1. Scale is 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 Angrist and Krueger (1991) 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 first-stage 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 Angrist and Krueger (1991) 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 schooling. 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

[13] 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.

$$\tilde{y}_i = \hat{y}(C_i, P_i) + 0.1\tilde{s}_i + \omega(Q_i, C_i, P_i)(\nu_i + \kappa_2\epsilon_i), \quad (19)$$

where $\tilde{s}_i$ is schooling simulated using equation (18). 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 $\kappa_2 = 0.1$ generates an OLS estimand equal to 0.207, or roughly double the causal effect of interest. Each simulation 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 (18) and (19).

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 × YOB × POB dummies as instruments (for a total of 1,530 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*-statistic across simulations above column headings. The Monte Carlo standard deviation of 0.011 is close to the (robust) standard error estimated for the 180-instrument model using the original data. Not surprisingly, moving from 180 to 1,530 instruments increases precision, at the price of increased bias. As can be seen in columns 4 and 9, the median absolute deviation (MAD, defined as the median of the absolute value of the difference between simulated estimates and the median simulation estimate) of the 2SLS estimates is somewhat below the corresponding standard deviation. The Monte Carlo median absolute error (MAE, reported in columns 5 and 10 and 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 heavily on the manner in which the penalty term is chosen. On average, a cross-validated penalty retains 74 of 180 and 99 of 1,530 instruments. As can be seen in the row immediately below the 2SLS estimates, post-lasso estimation using cross-validation-chosen penalties yields almost no bias reduction over 2SLS while slightly increasing sampling variance (as reflected in the Monte Carlo standard deviation).[14]

Swapping the cross-validated penalty 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 cross-validated penalty. Starting with a dictionary of 180 instruments, the plug-in penalty retains only two instruments, on average, and even fewer when starting with

[14] Cross-validated lasso penalty terms for the estimates in table 3 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 reestimated for each draw.

Table 3
Angrist and Krueger (1991) Simulation Results

Estimator	180 Instruments (QOB × YOB; POB × YOB; Average F = 2.5)					1,530 Instruments (QOB × YOB × POB; Average F = 1.7)				
	Average IV Retained (1)	Bias (2)	SD (3)	MAD (4)	MAE (5)	Average IV Retained (6)	Bias (7)	SD (8)	MAD (9)	MAE (10)
OLS		.107	.0004	.0003	.1070					
2SLS	180	.0403	.0108	.0075	.0397	1,530	.0611	.0046	.0032	.0611
Post-lasso IV (cross-validated penalty)	74.0	.0390	.0120	.0082	.0384	99.0	.0559	.0084	.0059	.0560
Post-lasso IV (plug-in penalty, IV selected)[a]	2.1	.0143	.0346	.0218	.0279	1.6	.0149	.0367	.0224	.0271
Split-sample IV	180	−.0009	.0237	.0158	.0158	1,530	−.0001	.0164	.0112	.0115
Post-lasso SSIV (cross-validated penalty)	63.1	−.0015	.0258	.0172	.0173	63.0	−.0013	.0280	.0183	.0183
Post-lasso SSIV (plug-in penalty, IV selected)[b]	2.1	−.0724	1.3168	.0274	.0287	3.4	.0197	.0504	.0228	.0292
Post-lasso (IV choice split only, cross-validated penalty)	63.1	.0429	.0144	.0097	.0431	63.0	.0460	.0141	.0093	.0459
IJIVE[c]	180	−.0011	.0194	.0130	.0131	1,530.0	.0001	.0123	.0088	.0087
LIML	180	−.0016	.0185	.0123	.0124	1,530	−.0034	.0117	.0079	.0083
Post-lasso LIML (cross-validated penalty)	74.0	.0222	.0152	.0102	.0220	99.0	.0484	.0094	.0066	.0483
Post-lasso LIML (plug-in penalty, IV selected)[a]	2.1	.0126	.0347	.0221	.0273	1.6	.0138	.0366	.0221	.0257
Pretested LIML ($t \geq 3.12$ for 180, $t \geq 2.3$ for 1,530)	18	.0222	.0236	.0148	.0238	153	.0385	.0163	.0111	.0393

NOTE.—This 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 between 1930 and 1939 (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 cols. 1–5 consist of 30 quarter-of-birth-by-year-of-birth and 150 quarter-of-birth-by-state-of-birth interactions (average F-statistic = 2.5, average concentration parameter = 270). The instruments used to compute the estimates described by cols. 6–10 are quarter-of-birth-by-year-of-birth-by-state-of-birth interactions (average F-statistic = 1.7, average concentration parameter = 1,050). 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 jackknifed 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 Angrist and Krueger (1991) data.

[a] 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 1,530 instruments. Statistics reported in these rows are for runs completed.

[b] Post-lasso SSIV with a plug-in penalty picks zero instruments in 670 of the 180-instrument runs and in 893 of the 1,530-instrument runs. Statistics reported in these rows are for runs completed.

[c] IJIVE results are based on 599 simulation iterations.

1,530 instruments (in a few simulation runs, the plug-in estimator retains no instruments). Our findings here are consistent with simulation results comparing lasso estimates computed with cross-validated 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 when starting with either instrument set.

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 MAE of these estimates is about 0.028. This is a considerable improvement on the bias of 2SLS estimates. But three non-ML IV estimators that are often used in many-weak-IV scenarios—LIML, SSIV, and IJIVE—do better. LIML is an approximately (median) unbiased maximum likelihood alternative to 2SLS (see, e.g., 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.[15] SSIV is consistent under a Bekker (1994) many-instrument asymptotic sequence and is therefore also approximately unbiased. The 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 that SSIV, IJIVE, and LIML estimates using both 180 and 1,530 instruments are indeed virtually median unbiased, although LIML and IJIVE are more precise than SSIV (compare, e.g., Monte Carlo standard deviations of 0.012 for LIML and 0.016 for SSIV using 1,530 instruments).[16] The standard deviation of these estimators mostly lies between that of the lasso estimators computed using plug-in and cross-validated penalties. LIML, IJIVE, and SSIV outperform 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 median unbiasedness of LIML, IJIVE, and SSIV is apparent from the fact that MAE for these estimators is almost indistinguishable from MAD.[17]

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

[16] Blomquist and Dahlberg (1999) also find that LIML and JIVE perform better than SSIV.

[17] LIML is the maximum likelihood estimator of a linear equation with an endogenous regressor under normality, but the generalized-method-of-moments justification for LIML requires only conditional homoskedasticity (Hausman et al. 2012). Our

Grouped under the split-sample IV heading, table 3 reports SSIV estimates computed with an instrument list chosen by lasso. Using a cross-validated penalty, these are virtually 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 no instruments in 670 out of 999 iterations for the 180-instrument case and in 893 out of 999 iterations for the 1,530-instrument case. The post-lasso SSIV estimates computed when instruments are retained are biased and much less precise than conventional SSIV estimates. Finally, using a sample split just to choose instruments (although not for first-stage estimation) yields estimates only marginally better than 2SLS when applied to the 1,530-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).[18]

B. Theoretical Considerations

Lasso for instrument selection faces two challenges. First is the fact that any overidentified 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 toward the corresponding OLS estimates in inverse proportion to the first-stage F-statistic for excluded instruments. By contrast, LIML, SSIV, and JIVE estimators are Bekker unbiased. Angrist, Imbens, and Krueger (1999), among others, show that the Bekker sequence describes the finite-sample behavior of alternative IV estimators extraordinarily well.[19]

simulation errors are normal but realistically heteroskedastic, so it seems fair to say that the simulation design does not stack the deck in favor of LIML.

[18] This estimator, reported in the row labeled "Post-lasso (IV choice split only, cross-validated 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 our split-sample procedures enforce an equal split and average results from complementary splits. Chernozhukov et al. (2018) discuss IV strategies that use lasso or other ML estimators in combination with SSIV-type sample splitting.

[19] The Bekker sequence has antecedents in Kunitomo (1980) and Morimune (1983), although Bekker (1994) appears to be the first motivated by quasi-experimental applications like Angrist and Krueger (1991). Hansen, Hausman, and Newey (2008) generalize the Bekker sequence to approximate the behavior of a wider class of estimators under weaker conditions.

ML methods are often motivated by prediction problems in which 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-instrument scenario. But the asymptotic sequence that justifies use of lasso for first-stage estimation has 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 stage is still limited. In particular, the Belloni et al. (2012) approximate sparsity condition implies $\lim_{n\to\infty} s/n = 0$, 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 0 and 1, a scenario Ng (2013) refers to as "dense." In the Bekker sequence, the fact that lasso truncates a dense instrument list reduces the bias of 2SLS estimates but does not eliminate it.

Perhaps the Angrist and Krueger (1991) application is an unfair test of the lasso idea. The number of Angrist and Krueger (1991) instruments is at least two orders of magnitude below sample size, so Angrist and Krueger (1991) is not a true high-dimensional scenario. Even so, Angrist and Krueger (1991) 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; Hansen, Hausman, and Newey 2008). Likewise, Belloni, Chernozhukov, and Hansen (2011), Belloni and Chernozhukov (2011), Belloni, Chernozhukov, and Hansen (2013), and Hansen and Kozbur (2014) use Angrist and Krueger (1991) data as a test bed for machine-chosen first stages.

Other IV scenarios may indeed favor lasso. Belloni et al. (2012) report simulation results for a sample size of 100 and a sparse first stage with exponentially or discontinuously declining first-stage coefficients. In this experiment, lasso-based IV outperforms 2SLS, LIML, and the Fuller (1977) modification of LIML. Belloni et al. (2012) also consider 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 that these provide good confidence interval coverage, and Hansen, Hausman, and Newey (2008) generalize Bekker standard errors to allow for heteroskedasticity.

1. Post-lasso as Pretest

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 with weak instruments. Pretesting estimated first-stage coefficients aggravates bias because high in-sample correlation with an endogenous

regressor is associated with a high in-sample correlation with omitted variables (or structural error terms) when instruments are weak.[20]

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 the size of the groups they indicate). 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 jth instrument from an OLS first stage using orthonormalized instruments, post-lasso estimators retain the jth instrument when

$$\left|\hat{\pi}_j\right| > c_n, \tag{20}$$

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 inequality (20), 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 median-unbiased LIML estimator exhibits bias of 0.022 in the 180-instrument model (with 74 instruments retained) and 0.048 when using 1,530 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 MAE for all-instrument LIML. Not surprisingly, with only two instruments retained, the behavior of plug-in lassoed LIML is close to that of post-lasso (plug-in) 2SLS using a similarily 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 problem with the latter.

2. *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. Given predictors

[20] Andrews, Stock, and Sun (2019) survey and assess the pretesting problem in modern applications of IV.

like YOB and QOB, a tree-based first stage might split schooling into branches containing 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 mean squared error or some other measure of fit. "Leaves" on the resulting trees are end points in each sequence of splits. Regression tree predictions are given by average outcomes for observations on a leaf (see, e.g., Athey and Imbens 2019). Random forests, introduced by Breiman (2001), elaborate on regression trees by using bootstrap samples and looking only at randomly selected subsets of predictors when deciding where to split. Random forest predictions average the predictions from each of these sampled splits.

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) use a random forest procedure to select instruments and to select and partial out nonexcluded (exogenous) covariates, an ML application discussed in section V.

We explore the utility of IV with random forest first-stage fitted values using 1,530 QOB-based instruments for education, controlling for a full set of YOB-by-POB fixed effects. In this application, random forest reproduces the first-stage fitted values from conventional 2SLS with a saturated first stage, so the second-stage estimates are indistinguishable from conventional 2SLS estimates. Similarly, when combined with sample splitting, the random forest results are indistinguishable from SSIV. These results, reported in the appendix, offer little reason to favor IV using a random forest first stage over conventional non-ML estimators.

C. 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 1,530-instrument version of Angrist and Krueger (1991). We might therefore 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 have 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 1,671 weekend days (the unit of observation for econometric analysis). The instruments for opening-weekend viewership are average weather conditions near theaters on a film's opening weekend. This identification strategy is motivated by the idea that the weather is randomly assigned and that good weather reduces movie attendance. The instrument dictionary includes 52 weather variables, such as the proportion of theaters experiencing 75°F–80°F 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 residualized "weather shocks" as instruments. 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 (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 = 0.022); 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 from the regressor of interest.

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 more than 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 noted in the discussion of Angrist and Krueger (1991), 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, with more OLS bias. IV procedures are then 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 2–6, finding declining effects). The simulation design starts by regressing opening-weekend attendance 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 $\hat{a}(X_{dt}, Z_t)$, 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 $\mu_{dt} = \max\{\delta, \hat{a}(X_{dt}, Z_t)\}/k_1$. This yields a skewed, nonnegative 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 distribution requires a shape parameter bounded away from zero, μ_{dt} is censored below by setting $\delta = .01$. Simulated attendance is then drawn by first generating a uniform random variable, ν_{dt}, and evaluating the inverse gamma conditional distribution function with shape parameter μ_{dt} at ν_{dt}. The appearance of ν_{dt} in the simulated outcome residual is our source of endogeneity.

The simulated outcome builds on LIML estimates in the original data using all 52 instruments. Specifically, let $\hat{y}(X_{dt})$ be the dependent variable fitted value, after subtracting $\hat{\rho}_{\text{LIML}} a_{dt}$, where $\hat{\rho}_{\text{LIML}}$ 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

$$\tilde{y}_{dt} = \hat{y}(X_{dt}) + 0.6\tilde{a}_{dt} + \omega(X_{dt}, Z_t)(k_2\Phi^{-1}(v_{dt}) + \varepsilon_{dt}), \tag{21}$$

so the causal effect of opening-weekend attendance is fixed at 0.6. Error component ε_{dt} is standard normal, while $\omega(X_{dt}, Z_t)$ 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.[21] 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

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

(X_{dt}, Z_t), with simulated opening-weekend and second-weekend attendance constructed as described above.[22]

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 as

$$y_{dt} = \rho a_{dt} + X'_{dt}\gamma_0 + \varepsilon_{dt},$$
$$a_{dt} = Z'_t\pi + X'_{dt}\gamma_1 + v_{dt},$$

where π is the vector of first-stage coefficients. We follow Gilchrist and Sands (2016) in specifying a day-specific causal model that links, say, Friday attendance 1 week after opening with Friday attendance on opening weekend.

The bias of OLS estimates across simulations is −0.37, while 2SLS is about half as bad, with a bias around −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 the 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, using the larger plug-in penalty yields a list of 12 instruments, while the 104 instrument list falls to around 23. Lasso with a cross-validated penalty retains 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

[22] 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), who report standard errors described as clustered, but with clusters equal to the unit of observation.

Table 4
Simulation Results for Opening-Weekend Effects

	Original Instruments ($F = 2.85$)					Original Plus 52 Noise Instruments ($F = 2.06$)				
Estimator	Average IV Retained (1)	Bias (2)	SD (3)	MAD (4)	MAE (5)	Average IV Retained (6)	Bias (7)	SD (8)	MAD (9)	MAE (10)
OLS		−.374	.015	.010	.374					
2SLS	52	−.165	.042	.027	.165	104	−.239	.034	.022	.238
Post-lasso IV (CV penalty)	36.6	−.160	.054	.029	.163	58.2	−.219	.063	.029	.228
Post-lasso IV (plug-in)	12.2	−.132	.092	.053	.142	22.5	−.150	.095	.067	.159
Split-sample IV	52	.053	.568	.095	.093	104	−.109	6.610	.134	.134
IJIVE	52	.019	.133	.067	.067	104	.004	.166	.082	.087
LIML	52	.007	.089	.057	.057	104	.009	.104	.064	.065

NOTE.—This 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 installments. Columns 6–10 report the results of adding 52 randomly generated (standard uniform) installments 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.

than the SSIV results in the Angrist and Krueger (1991) 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 (although not zero), and the IJIVE standard deviation is less than a quarter of that for SSIV. The MAD and MAE for IJIVE are indistinguishable (and well below that of SSIV), indicating that IJIVE is median unbiased.

As in the Angrist and Krueger (1991) results, LIML estimates are virtually median unbiased using both instrument sets. Here, LIML is 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). 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 yield good confidence interval 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 appendix includes a brief account of simulation 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. These results, computed using a plug-in penalty, are mostly similar to those shown in table 4 when the procedure is constrained to use a fixed number of controls and when some instruments are retained. It is noteworthy, however, that poivregress fails to select any instruments in about two-thirds or more of these simulation runs. Worryingly, poivregress reports second-stage estimates for many no-instrument runs. IV estimates generated without excluded instruments are biased and imprecise. Simulations using poivregress to select control variables as well as instrumental variables look a little better in that bias is reduced and more runs finish with at least some instruments retained. Here too, however, the procedure reports second-stage estimates with no instruments excluded. And in both variations, LIML and IJIVE estimates exhibit less bias and lower MAE than the corresponding poivregress estimates.

V. ML Picks IV Controls

Identification in IV 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 same-sex 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 variables. The Angrist and Evans (1998) same-sex 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 secondborn for many reasons. Ananat and Michaels (2008), for example, argue that male firstborns reduce divorce. The same-sex identification strategy may therefore be improved by allowing for additive male birth effects.

2SLS estimators incorporate 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 model 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.

The Athey, Tibshirani, and Wager (2019) random forest IV procedure generates an unconditional IV estimate of the form

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

where Z_i indicates families, X_i is a vector of controls, the function $\hat{g}_Z(X_i)$ is the (leave-out) fitted value from a random forest estimate of $\mathrm{E}[Z_i|X_i]$, 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. Replacing random forest centering (i.e., subtracting fitted values) with linear regression residuals generates a version of $\hat{\rho}_{\text{ATW}}$ equal to 2SLS. But random forest centering may be more flexible and, hence, more robust.[23]

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 same-sex 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 same-sex sibships on the probability of having more than two children equal to about 0.07. 2SLS using the same-sex 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. 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 $\hat{\rho}_{\text{ATW}}$ 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).[24] The rforest-based estimates in columns 2–4 are remarkably imprecise, with

[23] A related "double/debiased" procedure outlined in Chernozhukov et al. (2018) 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 (eqq. [4.4] and [4.8] in Chernozhukov et al. 2018) 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.

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

Table 5
IV Estimates after Random Forest Partialing

		Random Forest					
		rforest			regression_forest		
	2SLS (1)	Leaf 10 (2)	Leaf 100 (3)	Leaf 800 (4)	Leaf 10 (5)	Leaf 100 (6)	Leaf 800 (7)
	A. Same-Sex Instrument						
First stage: >2 children	.0676	.0804	.084	.087	.572	.370	.239
	(.0018)	(.0234)	(.0123)	(.0067)	(.020)	(.0107)	(.0066)
Employment	−.118	.046	−.077	−.115	−.110	−.123	−.118
	(.028)	(.342)	(.168)	(.086)	(.0402)	(.0323)	(.0314)
Weeks worked	−5.28	.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)
	B. Artificial Instrument (Covariate Index + Uniform Noise)						
First stage: >2 children	−.0066	−.0051	−.0096	−.0144	−.0108	−.0173	−.0169
	(.0031)	(.0012)	(.0009)	(.0007)	(.0028)	(.0017)	(.0007)
Employment	−.035	−1.24	−1.07	−1.08	−.0824	−.451	−.655
	(.504)	(.35)	(.14)	(.07)	(.286)	(.112)	(.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)
	C. Instrument with Some Signal (Covariate Index + Same Sex × Uniform Noise)						
First stage: >2 children	.0819	−.0038	−.0082	−.0130	.0475	.0136	−.0100
	(.0027)	(.0012)	(.001)	(.0007)	(.003)	(.0017)	(.0007)
Employment	−.111	−1.83	−1.30	−1.19	−.0633	.310	−.974
	(.036)	(.64)	(.18)	(.078)	(.069)	(.146)	(.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)

NOTE.—This table reports 2SLS and random forest IV estimates of the effect of having more than two children on employment and weeks worked. Estimates in panel A use a same-sex instrument. Estimates in panel B use an artificial instrument constructed from the sum of mother's age, mother's education, and uniform random noise. Estimates in panel C use an instrument constructed from the sum of mother's age, mother's education, and the product of the same-sex dummy with uniform random noise. The sample includes married women from the 1980 Public Use Microdata Sample aged 21–35 with two or more children. The sample size is 254,652.

standard errors more than 10 times larger than the 2SLS standard errors for minimum leaf size of 10 and more than three times larger for a minimum leaf size of 800. This imprecision reflects the fact that same sex is deterministically related to the additive male birth indicators included as covariates. Specifically, the same-sex instrument can be written as

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

where ss_i indicates mothers of a same-sex 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 to make this distinction.

As can be seen in columns 5–7, the regression_forest program does better with the same-sex 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 40% larger (compare, e.g., the 2SLS estimate of the effect on employment equal to −0.12 with a standard error of 0.028 to a similar regression_forest estimate with a standard of about 0.04 in col. 5). A second noteworthy feature of this set of results is the 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 same-sex 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) uses randomly generated instruments to reveal the bias of heavily overidentified 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 randomly drawn 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 mother's education, age, and the other covariates used to construct the real-instrument 2SLS estimates reported in panel A 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 same-sex 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 1 documents this by plotting residuals from a random forest fit of h_{1i} to the covariates used to construct the 2SLS estimates reported 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, conditional mean OLS fitted values are linear in x_i, while conditional mean OLS residuals are flat. Conditional 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.[25]

The risks posed by figure 1 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 employment rates in excess of 1. An attentive analyst might not be fooled here. But the estimates computed using regression_forest, 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 Angrist and Krueger (1991) first stage. In the Angrist and Krueger (1991) simulations, random forest fits a 1,530-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

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

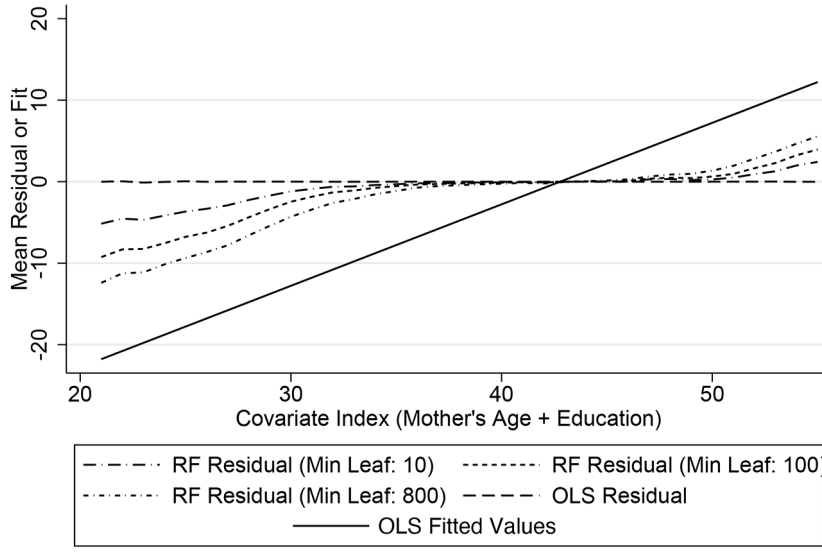

FIG. 1.—Random forest residuals are correlated with covariates. This figure plots residuals from random forest and least squares fits of an artificial instrument on a linear function of covariates. The covariate list contains mother's education, mother's age, mother's age at first birth, an indicator for the sex of each of the first two children, ages of the first two children (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. RF = random forest.

covariate cells in this case is much larger. While the Angrist and Krueger (1991) design has roughly 2,000 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 has been tasked to model.

Figure 1 suggests that 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 same-sex instrument with a minimum leaf size of 10 are either so imprecise as to be useless (i.e., those in col. 2, computed using rforest) or generate a first stage farthest from the causal effect of the same-sex experiment on fertility (i.e., an estimate of 0.572 in col. 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 C in table 5. 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}, although the standard error increases by about 30% when using the latter. 2SLS estimates for weeks worked are similarly close.

In contrast with the good performance of 2SLS using h_{2i}, random forest partialing mostly yields estimates as distorted or misleading as those computed using an instrument with no information. Estimates computed using rforest, reported in columns 2–4 of panel C, are too large to be coherent yet too precise for their magnitudes to be put down to sampling variance. These estimates reflect a small and unstable random forest first stage, which falls to near zero and negative (although significantly different from zero) in columns 2–4 and 7 and shrinks to 0.048 and 0.014 in columns 5 and 6. With a minimum leaf size of 100, IV estimates computed using regression_forest and instrument h_{2i} are noisy and of the wrong sign. The regression_forest estimates with a minimum leaf size of 10 are in the ballpark of the corresponding 2SLS estimates but, as in panel A, are 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 are way off base, statistically significant, and small enough to imply effects within the bounds of dependent variable support.

VI. Summary and Conclusions

The Belloni, Chernozhukov, and Hansen (2014b) PDS procedure provides a partially automated scheme for regression sensitivity analysis. Application of PDS to the estimation of effects of elite college attendance shows how this approach can support causal conclusions in a regression context. The identity and length of the list of PDS-included controls varies with changes in lasso tuning parameters and software. But in the application studied here, 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.

The evidence on ML in IV applications is less encouraging. In simulations modeled on 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. The asymptotic sparsity condition required by lasso-based methods appears to poorly approximate this empirical setting. The simulation results reported here also 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), in which LIML remains consistent.

Our divergent conclusions on the utility of ML for control variable selection and for IV selection can be related to the differing consequences of high dimensionality for regression and IV. 2SLS estimates are inconsistent in a many-instrument asymptotic sequence that reproduces the finite-sample behavior of IV estimators with many weak instruments. Lasso mitigates but does not fix this inconsistency while also risking pretest bias. This contrasts with the behavior of many-covariate regression models where the number of control variables increases in proportion to the sample size. Cattaneo, Jansson, and Newey (2018a, 2018b) show that OLS estimates of causal effects in a partially linear model like equation (14) are consistent and asymptotically normal in a Bekker-like sequence.

ML methods aim to improve out-of-sample fit, while the empirical labor applications we highlight seek to estimate causal effects. Beyond the Dale and Krueger (2002) and Angrist and Krueger (1991) studies expanded 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 establishments (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 minimum wages do not depress employment (Katz and Krueger 1992; Card and Krueger 1994), prompting a rethink of the competitive labor market paradigm; and that workers who attended schools with more resources earn more as a result (Card and Krueger 1992a, 1992b; Krueger 1999). This work generates evidence on causal relationships central to labor economics yet unrelated to goodness of fit.

In some contexts, pursuit of better fit is harmless. But our analysis also highlights the potential risks of fit-focused IV. Random forest partialing of covariates in a just-identified model generates 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.

References

Ackerberg, Daniel A., and Paul J. Devereux. 2009. Improved JIVE estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics* 91, no. 2:351–62. https://doi.org/10.1162/rest.91.2.351.

Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer. 2019. Lassopack: Model selection and prediction with regularized regression in Stata. 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, no. 3:611–29.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics* 11, no. 1:727–53. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua D. 2001. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics* 19, no. 1:2–28.

Angrist, Joshua D., and William N. Evans. 1998. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review* 88, no. 3:450–77.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. Jackknife instrumental variables estimation. *Journal of Applied Econometrics* 14:57–67.

Angrist, Joshua D., and Alan B. Krueger. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, no. 4:979–1014.

———. 1992. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association* 87, no. 418:328–36.

———. 1995. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and Economic Statistics* 13:225–35.

———. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, vol. 3, 1277–1366. Amsterdam: Elsevier.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'metrics: The path from cause to effect*. Princeton, NJ: Princeton University Press.

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. Unpublished manuscript. http://users.nber.org/dlchen/papers/Deep_IV_in_Law.pdf.

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. *Annals of Statistics* 47, no. 2:1148–78. https://doi.org/10.1214/18-AOS1709.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. Machine learning methods for demand estimation. *American Economic Review* 105, no. 5:481–85.

Bang, Heejung, and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, no. 4:962–73.

Bekker, Paul A. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62, no. 3:657–681.

Bekker, Paul A., and Jan Van Der Ploeg. 2005. Instrumental variable estimation based on grouped data. *Statistica Neerlandica* 59, no. 3:239–67.

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, no. 6:2369–429.

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, Victor Chernozhukov, and Christian Hansen. 2011. Lasso methods for Gaussian instrumental variables models. arXiv:1012.1297.

———. 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, chap. 7, 245–295. Cambridge: Cambridge University Press.

———. 2014a. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives* 28, no. 2:29–50.

———. 2014b. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81, no. 2:608–50.

Bickel, Peter J., Yaacov Ritov, and Alexandre B. Tsybakov. 2009. Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics* 37, no. 4:1705–32.

Blomquist, Soren, and Matz Dahlberg. 1999. Small sample properties of LIML and jackknife IV estimators: Experiments with weak instruments. *Journal of Applied Econometrics* 14, no. 1:69–88.

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, no. 430:443–50.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45, no. 1:5–32.

Card, David, and Alan B. Krueger. 1992a. Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy* 100, no. 1:1–40.

———. 1992b. School quality and black-white relative earnings: A direct assessment. *Quarterly Journal of Economics* 107, no. 1:151–200.
———. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84, no. 4:772–93.
Carrasco, Marine. 2012. A regularization approach to the many instruments problem. *Journal of Econometrics* 170, no. 2:383–98.
Carrasco, Marine, and Guy Tchuente. 2015. Regularized LIML for many instruments. *Journal of Econometrics* 186, no. 2:427–42.
Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey. 2018a. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory* 34, no. 2:277–301.
———. 2018b. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association* 113, no. 523:1350–61.
Chamberlain, Gary, and Guido Imbens. 2004. Random effects estimators with many instrumental variables. *Econometrica* 72, no. 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, no. 1:42–86.
Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21, no. 1:C1–C68.
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, no. 5:486–90.
Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov. 2019. On cross-validated lasso. 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, no. 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. *Quarterly Journal of Economics* 117, no. 4:1491–527.
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, no. 5:1161–91.
Fuller, Wayne A. 1977. Some properties of a modification of the limited information estimator. *Econometrica* 45, no. 4:939–53.

Gilchrist, Duncan Sheppard, and Emily Glassberg Sands. 2016. Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy* 124, no. 5:1339–82.

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 no. 12526, Institute of Labor Economics, Bonn. https://ideas.repec.org/p/iza/izadps/dp12526.html.

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, no. 2:283–98.

Hansen, Christian, Jerry Hausman, and Whitney Newey. 2008. Estimation with many instrumental variables. *Journal of Business and Economic Statistics* 26, no. 4:398–422. https://doi.org/10.1198/073500108000000024.

Hansen, Christian, and Damian Kozbur. 2014. Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics* 182, no. 2:290–308.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2016. Counterfactual prediction with deep instrumental variables networks. arXiv:1612.09596.

———. 2017. Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: The lasso and generalizations*. London: Chapman & 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, no. 2:211–55.

Hoerl, Arthur E., and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, no. 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, no. 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, Center for Economic and Policy Research, Washington, DC.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. 2015. Identification and inference with many invalid instruments. *Journal of Business and Economic Statistics* 33, no. 4:474–84.

Krueger, Alan B. 1991. Ownership, agency, and wages: An examination of franchising in the fast food industry. *Quarterly Journal of Economics* 106, no. 1:75–101.

———. 1993. How computers have changed the wage structure: Evidence from microdata, 1984–1989. *Quarterly Journal of Economics* 108, no. 1:33–60.

———. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114, no. 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, no. 371:693–700.

Leamer, Edward. 1983. Let's take the con out of econometrics. *American Economic Review* 73, no. 1:31–43.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine* 29, no. 3:337–46.

Morimune, Kimio. 1983. Approximate distributions of *k*-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica* 521, no. 3:821–41.

Mullainathan, Sendhil, and Jann Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31, no. 2:87–106.

Ng, Serena. 2013. Variable selection in predictive regressions. In *Handbook of economic forecasting*, vol. 2, ed. G. Elliott, C. Granger, and A. Timmermann, 752–789. Amsterdam: Elsevier.

Okui, Ryo. 2011. Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics* 165, no. 1:70–86.

Robinson, Peter M. 1988. Root-*n*-consistent semiparametric regression. *Econometrica* 56, no. 4:931–54.

Staiger, Douglas, and James H. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, no. 3:557–86.

Stata. 2019. *Stata lasso reference manual*. Version 16. College Station, TX: Stata Press.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58, no. 1:267–88.

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.