Stockholm School of Economics
Department of Economics
659 Degree project in economics
Spring 2018

Needles in a haystack: a machine learning approach to instrumental variables selection

Filip Mellgren (23644) and Vera Lindén (23611)

Abstract

This paper explores the comparative merits of two different machine learning algorithms for variable selection in an instrumental variables (IV) setting with many weak instruments. We apply a new method, post-L_2boosting, to Angrist and Krueger's (1991) classical paper about the effect of schooling on earnings. We compare the performance of the post-L_2boosting with another recently suggested method, post-LASSO estimation, on the same data. Among the methods used in this paper, post-LASSO is superior for increasing the first stage F-statistic of the IV estimation, implying that it more effectively reduces finite sample bias. However, our findings are not conclusive as further research is needed regarding the effects of different tuning techniques for the hyper parameters.

Keywords: instrumental variables, LASSO, L_2boosting,

JEL: C18, C26, C52, C55

Supervisor: Mark Bernard

Date submitted: 15 May 2018

Date examined: 28 May 2018

Discussants: Tim Lundqvist and Jakob Östgren

Examiner: Johanna Wallenius

Acknowledgements

We would like to express our sincerest gratitude toward our supervisor Mark Bernard for his invaluable support and encouragement. We also want to extend a special thank you to Abhijeet Singh for his insightful advice.

To Anurag Dey and Sofia Mellgren, thank you for believing in us and for always being there.

Contents

1 Introduction

One of the major conundrums of economic research is the distinction between correlation and causation. Unlike the natural sciences, the social sciences are often unable to abstract away complexities of reality by laboratory experiments. Instead, various techniques such as instrumental variables estimation provide an avenue to establish causal effects from observational data.

The developments in machine learning may at a first glance not appear to be directly related to the pursuit of causality however, machine learning algorithms can be useful tools in an economist's toolbox. While the goal of inference remains the same, in an age of big data, researches may need to use unconventional methods to handle the size and complexity of modern economic data (Varian 2014, Athey 2017).

In the context of instrumental variables, 'the finite-sample biases in instrumental variables are a consequence of overfitting'(Mullainathan & Spiess 2017, p. 100). The problem of overfitting is a central concept in machine learning and relates to how a model of sufficient complexity can perfectly estimate a set of data points in a sample but fail to hold any predictive abilities when applied to data out of sample. One specific way in which machine learning could be used in conjunction with econometrics is by mitigating this finite sample bias.

In this paper, we investigate if some prominent machine learning algorithms can be helpful in resolving econometric problems arising from weak instruments. These methods could be useful when there is an abundance of potential instruments and the researcher needs to make a choice as to which instruments to include when constructing an IV estimator (Belloni et al. 2011, p. 1). The machine learning algorithms we investigate include a variant of gradient boosting, known as post-L_2boosting and post-LASSO. We also consider a practice commonly used by machine learning practitioners which involves testing a model's predictive abilities out of sample. This is useful for understanding if too many variables have been used that does little in capturing the true relationship.

Under the assumption of constant causal effects, our results show that both post-LASSO and post-L_2boosting can substantially increase the first stage F-statistic that tests for weak instruments over naïvely including all of the instruments. We posit that post-LASSO is the more suitable method, as it produces a larger increase in the F-statistic.

This paper is organised as follows: Section 2 provides some background to machine learning and its relevance for economics. We begin with an overview of what machine learning is and how its applications differ from the search for causality, before providing the inspiration for this thesis as well as some examples of where machine learning has been used by economists in the past.

Section 3 introduces relevant theory and Section 4 provides a brief overview of previous research related to machine learning approaches to instrumental variables selection. Here we also introduce the uninitialised reader to Angrist's and Krueger's paper *Does Compulsory Schooling Attendance Affect Schooling and Earnings?*.

Section 5 outlines the research focus of this paper; comparing variable selection methods. In Section 6 and 7 we describe the the method and the data used, and in Section 8 we present our results.

In the final Sections 9 and 10 we discuss how our results in relation to previous research findings. We also provide an avenue for how the analysis in this paper may be improved by future research.

2 Background

Machine learning generally refers to some computer programme that improves its performance - learning - of a given task with experience (Mitchell 1997). This learning can be supervised, meaning that both predictors, (x), and outcomes, (y), are pre specified, or unsupervised, in which case the computer programme is fed an input and then independently produces an output. Supervised and unsupervised learning can be applied to a range of tasks such as prediction, classification, clustering, and dimensionality reduction.

While econometrics is mostly concerned with establishing causal relationships, not simply predictions, there are elements of inference suitable for application of machine learning (Ludwig et al. 2017, Athey & Imbens 2016, Hartford et al. 2017). Furthermore, machine learning has the potential to influence econometrics by automating model selection through a data-driven and systematic approach. 'This approach constrasts with economics, where (in principle, though rarely in reality) the researcher picks a model based on principles and estimates it once' (Athey 2017, p. 2).

Inspired by Athey (2017), Belloni et al. (2011) and Luo & Spindler (2017), we examine machine learning algorithms that can be used for instrument selection assuming there is no *a priori* information of what instruments to include in the first stage of an IV estimation. We investigate an empirical paper, Angrist & Krueger (1991), which has been scrutinised and declared to be suffering from bias related to weak instruments. It is worth noting that for datasets such as this: 'even when there appears to be only a few instruments, the problem is effectively high-dimensional because there are many degrees of freedom in how instruments are actually constructed' (Mullainathan & Spiess 2017, p. 101).

One could argue that the application of machine learning methods in econometrics is not new. Within the field of machine learning it is best practice to split the data into different samples: one that is used to train the model and the other to evaluate its performance. Angrist & Krueger (1995) and Angrist et al. (1999) use similar techniques to remedy bias in instrumental variables estimation. Indeed the first stage of a two-stage least squares (2SLS) is essentially a prediction task (Mullainathan & Spiess 2017) and machine learning can be used for instrument selection (Belloni et al. 2011, Luo & Spindler 2017).

3 Theory

In this section we review relevant theory related to instrumental variables and machine learning. We begin by examining the problems associated with weak instruments before providing a brief overview of the fundamentals of machine learning and some central concepts. We then describe the specific machine learning methods used in this thesis: post-LASSO and postL_2boosting.

3.1 Instrumental variables estimation

Instrumental variables (IV) estimation is a method for establishing causal inference that exploits exogenous variation in the instrument(s) to enable measurement of the effects of an, otherwise endogenous, explanatory variable on some outcome. It was first used by Wright (1928) and was later developed by Theil (1953), who introduced the two-stage least squares (2SLS).

Using this method, Angrist & Krueger (1991) explored the relationship between educational attainment and earnings. Equation (1), the first stage equation, establishes the relationship between the instrument and the endogenous variable of interest and is a regression of educational attainment on quarter of birth. Equation (2), the structural equation, describes the relationship of interest, the effect of education on earnings, and therefore regresses the log weekly wage on educational attainment.

$$E_i = X_i\pi + \sum_c Y_{ic}\delta_c + \sum_c \sum_j Y_{ic}Q_{ij}\theta_{jc} + \epsilon_i \tag{1}$$

$$lnW_i = X_i\beta + \sum_c Y_{ic}\xi_c + \rho E_i + \mu_i \tag{2}$$

E_i is the individual i's educational attainment, measured as years of schooling, X_i is a vector of covariates, and Y_{ic} and Q_{ij} are dummy variables indicating whether the individual was born in a specific year c or quarter j, respectively. W_i is weekly wage and ρ is the return on education.

In order to obtain valid instruments for the endogenous variable which can be used to construct consistent estimators, two conditions need to be met.

- Instrument exogeneity: $cov(z, u) = 0$

  The instrument (z) has no partial effect on the dependent variable, conditional on the endogenous variable (x) and the controls (K).

- Relevance: $cov(z, x) \neq 0$

  The instrument explains part of the variance in the endogenous variable of interest.

In practice, it is difficult to find instruments that perfectly satisfy these criteria, and some deviations may be permitted. Yet, even small violations of the above stated criteria can cause significant problems. These problems were largely overlooked by researchers until Bound et al. (1995) drew attention to them.

3.2 Weak instruments

There are two problems with weak instruments: inconsistency and finite sample bias. These arise from the instruments being only mildly correlated with the endogenous variable. This does not violate the assumption of relevance, but is nevertheless cause for concern. Bound et al. (1995) showed that the results in Angrist & Krueger (1991) may suffer from problems related to weak

instruments by demonstrating how similar results could be obtained by replacing the instruments with simulated, random data reflecting the basic irrelevance of the instruments.

Bound et al. (1995) showed that any inconsistency arising from imperfectly exogenous instruments was amplified by weak instruments. Furthermore, they considered the relative inconsistency of the 2SLS estimate to the OLS estimate which is given by Equation (3):

$$\frac{plim\hat{\beta}_{IV} - \beta}{plim\hat{\beta}_{OLS} - \beta} = \frac{\sigma_{\hat{x},u}/\sigma_{x,u}}{R^2_{x,z}} \tag{3}$$

In Equation (3), $R^2_{x,z}$ refers to the population partial R^2 from regressing x on the instruments z after the controls have been partialled out. A low $R^2_{x,z}$ means the relative inconsistency of the 2SLS is sensitive to any deviations from the assumption of instrument exogeneity, $cov(z, u) = 0$. Therefore, a small violation of the exclusion restriction may cause a large asymptotic bias.

Building on Sawa (1969), Bound et al. (1995) further posited that 2SLS estimates are biased in the direction of OLS estimates in finite samples. Furthermore, the probability distribution of an IV estimation is the same as the corresponding biased OLS estimate, except for the degrees of freedom. This means that the bias of the IV estimator is close to the bias of OLS, when the sample size is small. The bias of the IV estimator occurs despite having perfectly exogenous instruments and is of the same magnitude as the OLS estimate in the limit as $R^2_{x,z}$ goes to zero. Consider the IV estimator for a single instrument that is given by the Equation (4):

$$\beta_{IV} = \frac{cov(z_i, y_i)}{cov(z_i, x_i)} \tag{4}$$

When $cov(z_i, x_i) = 0$, the IV estimator is undefined. In a finite sample, however, there will inevitably exist some random correlation between z_i and x_i which is not helpful in detecting any causal relationship from x to y.

Through their simulation, Bound et al. (1995) showed that the estimates for returns to schooling remain similar, despite exchanging the quarter of birth variables to a randomised quarter of birth (that is: population $R^2_{x,z} = 0$, but not the finite sample partial correlation). The results demonstrate how the seemingly enormous sample size of $n > 300,000$ in Angrist & Krueger (1991) is not enough to evade the finite sample bias, as a result of weak instruments.

Bound et al. (1995), Buse (1992), Staiger & Stock (1994) conclude that the finite sample bias is inversely proportional to the number of instruments when the instruments are weak. In addition, Staiger & Stock (1994) find that $1/F$ approximates the magnitude of the finite sample bias of IV estimates relative to OLS estimates. Where F is the F-statistic that tests for joint significance of the instruments from the first stage regression of the endogenous variable on the instruments and the covariates. This F-statistic is given by:

$$F = \frac{(\sum(\hat{x}_r - x)^2 - \sum(\hat{x}_{ur} - x)^2)/q}{\sum(\hat{x}_{ur} - x)^2/(n - k - 1)} \tag{5}$$

Where the restricted model (subscript r) only consists of controls and the unrestricted model (subscript ur) consists of both excluded instruments and controls, q is the number of excluded instruments and k is the number of excluded instruments and controls. In essence, the F-statistic indicates whether adding the instruments provide any additional explanatory power by testing for joint significance. Practically, a higher F-statistic indicate a stronger first stage and less problems related to having weak variable bias. As Bound et al. (1995) point out, the finite sample F-statistic tends to be an upward biased estimator of the population F-statistic. For this reason, Stock & Yogo (2005) suggested that the F-statistic should be above 10 (well above what is normally required

of an F-statistic to indicate significance). Having a first stage F-statistic exceeding 10 is now a widely used rule of thumb. However, in general, a larger F-statistic indicates less concerns for bias.

3.3 Supervised machine learning

The underlying principles of supervised machine learning is to obtain high predictive accuracy by modelling a function of independent variables that explains dependent variables. In order to obtain high predictive accuracy, supervised learning modelling takes a data driven approach that does not require rigid assumptions of the underlying relationships of the data. What is important is that there exist correlations, or other non-linear relationships, within the data which can serve to predict an outcome of interest. This approach is fundamentally different from statistical inference and econometrics (Athey & Imbens 2017, p. 22). For instance, when performing prediction tasks, omitted variables are not necessarily a problem in machine learning. In fact, it might be beneficial to exclude variables and obtain bias in order to achieve better predictions. This is not the case for econometric purposes as biased parameters do not lend themselves to interpretation.

We next discuss some more specific concepts and introduce machine learning algorithms that we use in our analysis.

3.4 Bias-variance trade-off

As discussed by Wooldridge (2009, p. 91) the bias of an estimator refers to the difference between the expected value of the estimator and the true value of what is being estimated. Often, this is induced by expressing a complex real world phenomenon with a too simple model that do not take into account crucial aspects. Likewise, a supervised learning model can be biased if the expected predicted values from the model (depending on the data used to estimate the model) are not the same as the actual values. The richer, or the more complex, a model is made, the less biased it becomes. However, simpler models have the virtue of being more robust to the set of data they were estimated with. This means that such models are less dependent on the random sample and will be more general in the sense that they will appear similar in a broader set of contexts. Such models are said to have a low variance.

A model that exhibits low bias but high variance is referred to as being overfitted. For example, if too many variables are used to explain a small set of observations, the explanatory power *in sample* rise as a result of adjusting or interpolating the model to noise present in the data. Such a model will fail to generalise to other data sets and exhibits large discrepancies between model accuracy in sample versus out of sample. Likewise, a model that includes too few variables will not be able to capture important variance and result in a simple model with biased estimators. A model that lacks crucial variables is referred to as being underfitted. Due to the conflicting natures of over- and underfitted models, there is an inherent bias-variance trade-off. Bias increases whenever there are missing variables from the analysis whereas variance increase when too many variables are used on too little data. In other words, the more situation specific a model is made, the more accurate it will be on a particular data set, without necessarily extending its precision to other samples from similar populations.

It can be shown that the expected mean squared error (MSE) for a given data point x_0 out of sample is given by:

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon). \tag{6}$$

Unlike economists who often strive for zero bias of the estimated parameters at a tolerable level

of variance (i.e. statistical significance), machine learning would be concerned with minimising the quantity in Equation (6). The way a machine learning model is evaluated is typically done by fitting a model, $\hat{f}(x)$, on a training set (typically an 80% sub sample of the data at hand). This model fit is then tested on a remaining, held out, test set by plugging in independent observational data, (x), to obtain estimated values $\hat{f}(x)$ that is then compared to the actual data (y) using a loss function, typically the MSE.

3.5 Cross validation

Econometric practice favours simpler models over complex ones (given that bias has been dealt with) to prevent overfitting. Supervised machine learning, on the other hand, often relies on a data driven approach to balance bias against variance. One of the more common methods include k fold cross validation.[1] Cross validation works by splitting the data systematically into k folds of equal size. It is an iterative process where all folds but one, $(k-1)$, are used to fit a model that is then validated against the one remaining fold. The process is repeated until each fold has been left out and tested against once for all different model specifications.[2] This results in k different estimates of the loss function for each model specification (E_i in the figure below).

These k estimates are then averaged to obtain an averaged score for each model specification. The model specification that is associated with the lowest average value of the loss function is the model specification deemed to be best by the cross validation.

Both bias and variance of the test error depend on the choice of k. Small values of k are typically associated with a high bias, especially for small sample sizes. Large values of k become less biased but induce more variance. Following a simulation study, Kohavi et al. (1995) recommend setting $k = 10$, on the basis that it achieves a good trade-off between bias and variance.

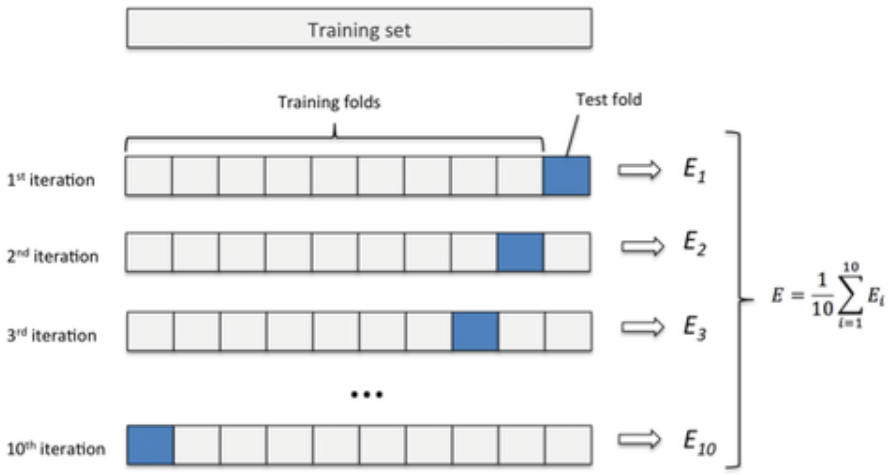

Figure 1: Illustration of 10-fold cross validation. Image from Buhagiar (2017, p. 6).

3.6 LASSO

The least absolute shrinkage and selection operator (LASSO) was first introduced by Tibshirani (1996) and is an instance of *regularisation*, which refers to a process of reducing overfitting. The

[1] cross validation has previously been used in economics, in the context of kernel regressions (Athey 2017).

[2] An example of what is meant by different model specifications could be a model that can take on different values of a hyper parameter λ. When the hyper parameter λ is varied, different model specifications are obtained. The optimal level of λ, say λ^*, can be determined by cross validation.

LASSO can thereby reduce model variance by inducing more bias. It is also useful for variable selection. Another method that is closely related to LASSO is the ridge regression.

The way LASSO and ridge regression perform regularisation is by adding a penalty term to the loss function, usually the residual sum of squares (RSS). The penalty term added to the LASSO, $\lambda \geq 0$ is proportional to the absolute value of the estimated parameters and LASSO thereby pulls any coefficients toward 0. Ridge regression works in a similar way but adds λ times the square of the coefficients as a penalty.

The ridge regression minimises the following equation:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2, \lambda \geq 0. \tag{7}$$

The LASSO only differs from the ridge regression by using the absolute values of the coefficients instead of the squared terms:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|, \lambda \geq 0. \tag{8}$$

For the minimisation problems (7) and (8), any reduction in the RSS must be large enough to offset the increase induced by the penalty term. The λ in the formulas is a hyper parameter that can be tuned using cross validation, and it determines the sensitivity of the penalty term.

An advantage of LASSO over ridge regression is the fact that it can set some coefficients to 0 whereas the ridge regression will not. Therefore, the LASSO can be a useful tool for selecting variables. To illustrate this feature of LASSO in comparison to the ridge regression, we consider the case of a model containing two coefficients, β_1 and β_2. The ridge regression and the LASSO solve the following minimisation problems, respectively (James et al. 2013, p. 220):

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j^2| \leq s \tag{9}$$

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s \tag{10}$$

Where there exists an s for each λ such that the problems in Equations (7) and (8) yield the same solutions as Equations (9) and (10) respectively. In Figure 2, $\hat{\beta}$ represents the least squares estimates that is subject to the LASSO's constraint and the ridge regression's constraint. The LASSO's constraint is illustrated by the blue diamond in the figure and defined by the L_1 norm: $s \leq |\beta_1 + \beta_2|$. The blue circle is the ridge regression's constraint s, defined by the L_2 norm: $s \leq \sqrt{\beta_1^2 + \beta_2^2}$. The ellipses around the $\hat{\beta}$ represent different levels of RSS, the function we intend to minimise in Equations (9) and (10). The solution will be the intersection between the constraint and the ellipse that represents the lowest RSS. Note that with sufficiently relaxed constraints (represented by a larger diamond and a larger circle), both the LASSO and the ridge regressions equal the least squares estimate $\hat{\beta}$.

The shape of the constraints is what gives LASSO its variable selection property. Because the LASSO's loss function has vertices due to its constraint being subject to the L_1 norm, coefficients that result from a LASSO estimation will sometimes be set to zero. In Figure 2, the LASSO's estimate of $\beta_1 = 0$ whereas the ridge regressions estimate of $\beta_1 > 0$. The figure shows the intuition for two dimensions, but the same logic applies as the dimensionality is increased.

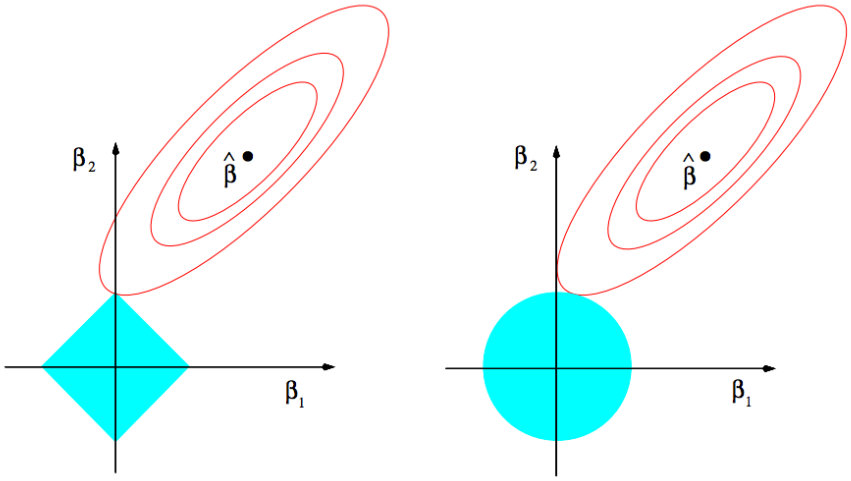

Figure 2: Illustration of the error and constraint functions for LASSO (to the left) and the ridge regression (to the right). Image from Tibshirani (1996, p. 6).

The parameter λ in Equations (9) and (10) is a hyper-parameter. Intuitively, when $\lambda = 0$, the ridge and LASSO estimations become the same as the OLS estimates and a high values of λ are associated with more strict regularisation. The value of λ is ideally set to a value that minimises the loss function out of sample. For this reason, it can be determined by cross validation.

3.6.1 Post-LASSO

The post-LASSO estimator uses LASSO for variable selection, and then proceeds by adding the selected variables to an OLS estimation with the full set of pre-specified controls. The point of using post-LASSO instead of using the coefficients directly obtained from the LASSO estimation is that the LASSO does not guarantee that all desired control variables are included. However, the post-LASSO guarantees that all controls are included in the final model. In this way, post-LASSO avoids potential bias caused by not including all the control variables while exploiting a subset of highly predictive variables, thereby limiting both bias and variance. Belloni et al. (2011) take advantage of this bias-variance trade-off by using the LASSO to select variables before the first stage of an IV regression in order to select relevant instruments from a long list of candidate instruments. This is will be explored further in subsection 6.1.1.

3.7 L_2boosting

Boosting algorithms refer to a family of algorithms for sequential model building that was developed by Friedman (2001). Boosting combines many weak predictors, called *base learners*, into a single, stronger, predictor. These algorithms learn *slowly* and have proven successful in reducing both bias and variance, which has made them popular statistical learning algorithms.

Further developments in boosting algorithms have added the L_2 (least squares) penalty function (Bühlmann & Yu 2003), and showed its usefulness in economic applications by using it for variable selection in an IV settling (Luo & Spindler 2017).

This thesis uses the following specification of the boosting algorithm for variable selection:

1. Start/ Initialization: $\beta^0 = 0$ (p-dimensional vector), $f^0 = 0$, set maximum number of iterations m_{stop} and set iteration index to 0.

2. At the $(m+1)^{th}$ step, calculate the residuals $U_i^m = y_i - x_i^{'}\beta^m$.

3. For each predictor variable $j = 1, ..., p$ calculate the correlation with the residuals:

$$\gamma_j^m := \frac{\sum_{i=1}^n U_i^m x_{i,j}}{\sum_{i=1}^n x_{i,j}^2}$$

Select the variable j^m this is the most correlated with the residuals, i.e., $max_{1 \leq j \leq p} |corr(U^m, x_j)|$[1].

4. Update the estimator: $\beta^{m+1} := \beta^m + \eta\gamma_{j^m}^m e_{j^m}$ where e_{j^m} is the j^mth index vector and $f^{m+1} := f^m + \eta\gamma_{j^m}^m x_{j^m}$. $0 < \eta \leq 1$.

5. Increase m by one. If $m < m_{stop}$, continue with 2; otherwise stop.

(Luo & Spindler 2017, p. 2)

Where m_{stop} is the number of iterations the algorithm runs to reduce the residuals. Akin to LASSO, the L_2boosting in a high-dimensional setting, can be prevented from over-fitting by introducing regularisation. In this case, the regularisation parameter takes the form of stopping the algorithm early by limiting the number of iterations to something less than m_{stop}, e.g. m^*. Obtaining an appropriate value for m^*, can be achieved by using k fold cross validation or a data dependent stopping rule. If k fold cross validation is used, the value of m^* is set to the m that minimises the loss function, such as the RSS, over the k folds.

A data dependent stopping rule proposed by Luo & Spindler (2016) works by finding the first m for which the following inequality holds:

$$\frac{||U^m||_{2,n}^2}{||U^{m-1}||_{2,n}^2} = \frac{\hat{\sigma}_{m,n}^2}{\hat{\sigma}_{m-1,n}^2} > (1 - C * log(p)/n) \tag{11}$$

Where $||U^m||_{2,n}^2$ and $||U^{m-1}||_{2,n}^2$ are the RSS at iteration m and $m-1$ respectively. The number of variables (or predictors) are denoted by p and the number of observations is captured by the variable n. The interpretation of this rule is that optimal stopping is reached when the estimated variance of the residuals ($\hat{\sigma}_{m,n}^2$) in one step, relative to the variance of the residuals in the previous step ($\hat{\sigma}_{m-1,n}^2$), has not been reduced more than the specified threshold (to the right of the inequality sign).

Every boosting algorithm has a base learner, which refers to the function used to reduce the loss function. The L_2boosting described in the algorithm above uses a linear base learner, but it is more common to use a tree based learner that reduces the loss function by fitting a regression tree on the residuals at every step. The benefits of using a linear base learner over a tree based learner is that the final model at m^* takes the form of a linear regression. In contrast, the tree based method becomes more of a black box as the sequential tree building leaves no such simple formula. However, trees will often be better at detecting non linear relationships.

The rate of convergence for L_2boosting is determined by the shrinkage parameter $0 < \eta \leq 1$. Luo & Spindler (2017) investigated the simple case where $\eta = 1$. By only updating the residuals with a fraction $\eta < 1$ of the estimated model at every step, the algorithm was able to potentially discern more information from the data set. To see why, consider the mth step where the algorithm uses ηx_j to explain the residuals. If we continue to find that x_j is best at explaining the residuals for the following $1/\eta$ steps for each variable, the estimation with the smaller η become the same as the estimation when $\eta = 1$. If not, the $0 < \eta < 1$ estimation was able to find a pattern that the $\eta = 1$ estimation was unable to find and have hence built a more predictive model. In general,

a smaller value of η will require more iterations (a higher m^*), but yields better predictions in return.

3.7.1 Post-L_2Boosting

The post-L_2Boosting follows the same principle as the post-LASSO but uses L_2Boosting instead of LASSO for variable selection. The selected variables are those with non-zero coefficients in the L_2boosting model.

3.8 Contrasts to other dimension reduction methods

Using LASSO and gradient boosting is not the only way to reduce the dimensionality of a data set. Other methods include factor analysis and Principal Components Analysis (PCA). Such methods aim to reduce dimensionality by identifying *factors* or *components* that summarise the original set of variables using fewer dimensions by exploiting correlations in the data set, i.e. the variables used in the regression estimation is a smaller set of variables that are defined as linear combinations of the original variables. As a result, some dimensions that otherwise would have been included in a regular estimation are excluded if they contain only a small amount of proper variance.

A crucial contrast between dimension reduction models and variable selection using post-LASSO or post-L_2boosting is that the reduction of dimensions by a PCA may result in non interpretable models where the dimensions are few but inexplicable. In contrast, the variables selected using LASSO or L_2boosting will be the same variables that enter the analysis.

There are other data driven model selection procedures such as forward and backward selection. Such algorithms perform model selection by adding/withdrawing variables to/from a model. It should be noted that a forward selection procedure that builds a model by taking the RSS as a loss function in the setting of a multiple linear regression proceed in a similar way to L_2boosting with a linear base learner having $\eta = 1$.

4 Previous research

This section is dedicated to the previous research this thesis builds upon. First, we review Angrist and Krueger's seminal paper on the returns to education. Next we cover the work by Belloni et al. (2011) who demonstrate how the LASSO can be applied for variable selection on the example of returns to education. Lastly, we cover the work by Luo & Spindler (2017), which brought L_2boosting to econometrics.

4.1 Angrist and Krueger (1991)

In *Does Compulsory School Attendance Affect Schooling and Earnings?* Angrist & Krueger (1991) estimate returns to schooling using exogenously determined variability in compulsory school attendance as an instrument for education. In the first-stage of the 2SLS the authors predict years of schooling for an individual using the following equation:

$$E_i = X_i\pi + \sum_c Y_{ic}\delta_c + \sum_c \sum_j Y_{ic}Q_{ij}\theta_{jc} + \epsilon_i$$

Where variable E_i is education of individual i in number of years, X_i is a vector of covariates, Q_{ij} is a dummy variable indicating if the ith individual was born in quarter j, $j \in \{1, 2, 3\}$ and Y_{ic} indicates whether the ith individual was born in year c, $c \in \{0, 1, 2...9\}$

They motivate the first stage by establishing that individuals born relatively early in a year will be older when they start school and hence reach the legal drop out age with less schooling than individuals born later in the year. In addition, they argue that quarter of birth is determined exogenously, and does not influence earnings in any other way than through schooling.

In their original paper, Angrist and Krueger used two sets of instruments. The first estimate used 27 instruments, three quarter of birth dummies interacted with nine year of birth dummies. The second set includes additional interactions of quarter of birth with state of birth, producing a total of 177 instruments. They also alluded to the possibility of introducing instruments composed of interactions between all three variables; quarter of birth, year of birth, and state of birth. This implied that the total number of possible instruments was 1530 (including quarter of birth without interactions).

There is now wide consensus that the use of this many instruments is not good practice and can be misleading due to inconsistency and finite sample bias (Bound et al. 1995). Bound et al. (1995) provide that the F-statistics and R^2 from the first stage regression can be informative when determining the validity of the instruments.

The results in Angrist & Krueger (1991) have since been re-examined on several occasions. A few examples, which do not concern variable selection but are nevertheless worth mentioning as they are of interest in the context of machine learning include Angrist & Krueger (1995), Carrasco (2012), Hansen & Kozbur (2014). The first study used split-sample instrumental variables (SSIV), which exhibits a characteristic of machine learning: sample splitting. The other two studies employed the ridge regression, explained in Subsection 3.6, to improve the IV estimation.

4.2 Belloni, Chernozhukov & Hansen (2011)

Belloni et al. (2011) looked at a high-dimensional, sparse setting to employ a LASSO based method called post-LASSO to eliminate weak and redundant instruments. High-dimensionality refers to a setting where the number of observations n can potentially be larger than the number of available predictors p. The sparsity condition implies that only a few of the available predictors are useful in predicting the endogenous variable, i.e. only a few variables have coefficients significantly different from zero. The post-LASSO estimator is an ordinary least squares (OLS) estimator but with an additional step where the LASSO has been used for variable selection, beforehand. For detailed description of the LASSO please refer to Subsection 3.6.

Belloni et al. (2011) demonstrated the merits of the post-LASSO first on simulated data and on application to Angrist & Krueger (1991). The post-LASSO, with cross validation, successfully selected 12 instruments among the full set of 1530 potential instruments. The most important instruments (quarter of birth, without interactions) were included. This is reassuring as the algorithm independently selected these instruments without any previous 'knowledge' of which instruments would be preferable from a theoretical point of view.

4.3 Luo & Spindler (2017)

Luo & Spindler (2017) presented the post-L_2boosting as an underutilised and competitive alternative to the post-LASSO for variable selection among many variables in a sparse scenario. Here, 'post' refers to the same variable selection function as in the LASSO case and 'L_2' refers to the penalty term, which is the squared errors. Luo & Spindler (2017) illustrate their findings on both simulated data, and an empirical example. In both cases, the data is relatively small and the number of variables is large relative to the number of observations.

The simulated data has 100 observations and includes 200 variables with diminishing predictive power. The post-L_2boosting yields a slightly lower bias than estimates from the post-LASSO algorithm. The empirical example is an IV estimation of the the relationship between GDP and appellate court decisions. Their dataset included 90 countries and 60 variables. The post-L_2boosting estimates 'replicate the Lasso estimates but with smaller standard errors' (Luo & Spindler 2017, p. 3). The two findings seem to suggest that post-L_2boosting outperforms post-LASSO in terms of precision, whilst not inducing any larger bias.

Belloni et al. (2011) and Luo & Spindler (2017) propose methods for application in similar, sparse situations. Luo & Spindler (2017) wrote their paper after Belloni et al. (2011) and claimed that post-L_2boosting match the performance of the post-LASSO for IV estimation. Thus, it seems reasonable to expect that the proposed post-L_2boosting algorithm should do equally well on the empirical example chosen by Belloni et al. (2011). To the best of our knowledge, no such comparison on this particular dataset has been made. It is also of interest to see how well the L_2boosting performs with increased observations and variables. Furthermore, we are curious to study the behaviour of the F-statistic, as neither study reports them. This paper will explore this in Sections 8 and 9.

5 Specification of detailed research focus

When the choice of appropriate instruments is not obvious, the machine learning algorithms previously suggested by Belloni et al. (2011) and Luo & Spindler (2017) can guide the extraction of instruments with a high predictive power. But which algorithm is the better of the two?

To test the usefulness of these methods we use data from Angrist & Krueger (1991). This is akin to Belloni et al. (2011), but we extend the analysis to include the the post-L_2boosting algorithm proposed by Luo & Spindler (2017). This serves two purposes. First, we explore if post-L_2boosting is helpful in selecting among many variables when the data set is moderately big and includes only a weak relationship between the target variable and the potential instruments. Secondly, we bridge existing research by collectively linking LASSO and L_2boosting to a classical data set and thereby obtain useful comparisons on a well known benchmark. We hope to contribute to establishing the merits (or lack thereof) of a data driven approach to variable selection.

In addition, we investigate how well the different approaches predict data out of sample. The point of doing this is to see if more restrictive sets of variables chosen by the LASSO and the L_2boosting may be able to predict the data better than the more complex specification used in Angrist & Krueger (1991).

It should be emphasised that we perform our analyses under the assumption of constant causal effects, implying that those assigned the treatment (possibility of dropping out of school earlier) will also comply with this, regardless of their unobserved individual characteristics. Reality, however, may exhibit heterogeneous treatment effects; Angrist & Krueger (2001, p. 77) observed 'the quarter-of-birth instrument is most relevant for those who are at high probability of quitting school as soon as possible, with little or no effect on those who are likely to proceed on to college'.

Our research question:

In the case of Angrist and Krueger (1991) under the assumption of constant causal effects, which variable selection method is better, post-LASSO or post-L_2boosting, in terms of mitigating finite sample bias and inconsistency arising from many weak instruments?

6 Method

In this section we describe the methods applied to the data in order to answer the research question. By applying machine learning algorithms for variable selection we strive to reduce the number of excluded instruments in an IV estimation which suffers from weak instruments.

6.1 Variable selection methods

Two variable selection algorithms from the machine learning literature are used: post-LASSO and post-L_2boosting. Both have been proposed as useful algorithms to filter out weak instruments and were introduced in Section 3. For the post-LASSO, we use two variants for selecting the hyper parameter λ. For the post-L_2boosting, we use four variants based on two different criteria for selecting the hyper parameter m.

6.1.1 Variable selection using post-LASSO

The first algorithm we implement for variable selection is the post-LASSO. As described above, the post-LASSO consists of two step: one LASSO regression that shrinks the number of variables and one subsequent OLS regression that represent the first stage of an IV estimation. Had we used only the LASSO, there would be a risk of not all controls being selected. We need all the controls in order to compare the different first stages of the IV regression produced by the different methods. Had we used fewer control variables for some regressions, more would differ than just the set of instruments. Hence, we use the post-LASSO, which allows us to add unselected control variables from the first (LASSO) step to the second (OLS) step.

In the first step of the post-LASSO, a regular LASSO is used to predict the endogenous variable x using variables in $K \cup M$, where K denotes the full set of control variables and M the full set of instruments. We regress x on K and M using LASSO for our model specifications, allowing the LASSO to set some coefficients to zero, thereby reducing the dimensionality of the data. The instruments in M having non zero coefficients are then included in a set of qualified instruments, Z, that we use to regress x on K and Z using OLS. The equations below illustrate how these two steps of the post-LASSO relate to the two stages of the IV regression. The second stage of the post-LASSO is the one later used for evaluation.

$$\text{First step post-LASSO: } x_i = \alpha_1' M_i + \alpha_2' K_i + w_i$$
$$\text{Second step post-LASSO} = \text{First stage IV: } x_i = \pi_1' Z_i + \pi_2' K_i + v_i$$
$$\text{Second stage IV: } y_i = \beta_1 \hat{x}_i + \beta_2' K_i + u_i$$

Like Belloni et al. (2011) we use 10-fold cross validation to determine the regularisation hyper parameter λ, *see Equation* (8). The 10-fold cross validation then selects a λ_{min}, which produces the lowest value of the sum of squares loss function.

As can be seen in Equation (12), minimising the squared error loss function $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is intrinsically linked to keeping the R^2 high by penalising coefficients toward 0 for variables that do not predict well out of sample.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

Keeping the overall R^2 high while selecting among instruments is analogous to keeping the partial $R^2_{x,z}$ high, as only the weakest instruments are discarded. Thereby the issue of inconsistency

arising from the partial $R^2_{x,z}$ in Equation (3) being reduced by the variable selection is considered. As a result the variable selection has the potential to decrease finite sample bias while not worsening inconsistency.

Belloni et al. (2011) include an additional, more conservative, plug-in rule that we do not implement. Instead, we apply another method that induces more regularisation than λ_{min} (Friedman et al. 2001, p. 80). This method uses the same cross validation, but rather than choosing λ_{min}, it elects the most regularised λ that yields an error within one standard error from the minimum of observed values in the cross validated loss-function, call this value λ_{1se}. As λ_{1se} is chosen to regularise more, this value is expected to yield a more sparse solution to the variable selection problem than λ_{min} while remaining reasonably close to the estimated lowest level of the loss function.

6.1.2 Variable selection using post-L_2boosting

As with the post-LASSO, the post-L_2boosting algorithm consists of two steps, where the first step involves selecting the variables and the second step involves OLS estimation. The variable selection step in this case is performed by gradient boosting with a linear base learner. The linear base learner assures that the output from the boosting is a linear function of a subset of variables used to explain x. This subset of instruments Z, which have been selected by the boosting, enter the last part of the post-L_2boosting estimation. This is akin to the OLS estimation in the last part of the post-LASSO, the only difference being what Z were chosen by the different algorithms.

We let the gradient boosting have an L_2 loss function which gives a model optimised to reduce the RSS. Each iteration seeks to minimize the loss function by finding the variable most correlated with the residuals. We set the slowing parameter $\eta = 1$. A lower η would be helpful in discerning a relationship between x, M, and K that does well in predicting x by allowing more iteration steps without overfitting. However, the task is to do variable selection and tolerate that not all explanatory power is accounted for. This is also the implementation used by Luo & Spindler (2017).

Special care needs to be taken when setting the maximum number of iterations (m_{stop}) performed in the first step of the post-L_2boosting. Ideally, this number is as small as possible for computing reasons. Yet, it needs to be larger than the optimal number of iterations (m^*).[3] As m_{stop} is set before knowing m^*, it has to be guessed or set by calibration. We reason that, if all variables had been independent, the algorithm would have had a total of $m = K + Z$ iterations. This occurs because each variable can only be chosen once, as all variance related to that variable is accounted for in the first inclusion and not updated by subsequent contributions. Setting all variables to be independent before applying gradient boosting is called orthogonal gradient boosting. We do not use orthogonal boosting, but it is helpful for understanding the regular L_2boosting (which we do use) and it helps us select m_{stop}.

In order to prevent the first part of the post-L_2boosting from overfitting, the optimal number of iterations, m^*, needs to be determined. We first apply a k-fold cross validation to determine the optimal number of iterations m^*. We let $k = 10$ as this is recommended in the literature (Kohavi et al. 1995), and use the cross validation to find the number of iterations that minimises the loss function. Specifically, data from 9 folds are used to train a model that is then evaluated against data in the remaining fold by calculating the squared error loss function for each integer m, $0 < m \leq m_{stop}$. This process is repeated 10 times (until each fold has been left out once). Hence, this process is demanding in terms of hardware requirements.

[3]Note the difference between the maximum number of iterations m_{stop}, which is an arbitrary number selected by the researcher and the optimal number of iterations m^*, which is selected through by the algorithm.

As with the LASSO, a cross validated m^* has the benefit of keeping the first stage R^2 high while discarding variables that do not predict well out of sample. Since at most one additional variable is chosen at each iteration step, the stopping criteria results in a variable selection that excludes variables that are unable to further reduce the $\sum_{i=1}^{n}(y_i - \hat{y})^2$, assuming not all variables predict well out of sample.

However, minimising the loss function will not penalise based on the amount of variables in the model. As long as the variables hold predictive power, the variables will be included. Since the F-statistic is inversely related to the number of variables, we want the number of variables to be reduced. Hence, minimising the loss function might therefore not be optimal in order to increase the F statistic. Therefore, we also implement the data driven rule (Luo & Spindler 2016, p. 18). According to this rule, optimal stopping is reached at m^*, which is the first iteration when the following inequality holds:

$$\frac{\hat{\sigma}^2_{m,n}}{\hat{\sigma}^2_{m-1,n}} > 1 - C\frac{log(p)}{n} \tag{13}$$

Where $\hat{\sigma}^2$ is the estimated variance, m denotes the mth iteration, n is the sample size and C is some constant and p is the number of variables. We report results for $C = 0.25$, $C = 0.5$ and $C = 0.75$. In general, a larger value of C will be associated with earlier stopping.

6.2 Evaluation of chosen models

Mitigating finite sample bias, in the presence of weak instruments, is obtained by strengthening the relationship between the instruments and the endogenous variable of interest. This thesis focuses on how to strengthen this relationship by selecting only the strongest instruments. As suggested by Bound et al. (1995), to evaluate the strength of the first stage, and therefore our variable selection methods, two metrics are considered: the first stage R^2, and the F-statistic. Both measures relate to the relevance condition that is necessary for valid instruments. The F-statistic has a direct relationship with the finite sample bias, as the bias is proportional to $1/F$, while the R^2 is more related to inconsistency. By reporting both, we may appreciate how the overall bias may be affected by the variable selection procedures.

6.2.1 Comparison based on the weak instruments F-statistics

The F-statistic is the joint significance test of all the instruments in the first stage. It has the null hypothesis that the variables do not explain the endogenous variable x. Model specifications with higher F-statistics are better in the sense that the risk of finite sample bias is lower (Bound et al. 1995) and may also lead to less biased estimates of the effect of x explaining y. Therefore, higher values are preferable to lower values when interpreting this statistic. For an absolute comparison of the merits of a single F-statistic, the '$F > 10$' rule of thumb gives a useful indication. The formula for the statistic testing for weak instruments was given in Equation (5).

Removing weak instruments could increase the F-statistic, especially if the removed instruments are poor predictors of x, as the number of instruments, q, enter the denominator of Equation (5) directly.

6.2.2 Comparison based on the first stage R^2

Due to its relatedness with inconsistency, Bound et al. (1995) suggested looking at the partial R^2 for an indication of the strength of the excluded instruments in the first stage. We investigate this in addition to the overall first stage R^2 with the motivation that, given the same control variables K, it is straightforward to make comparisons between different selections of instruments as the controls do not change between different specifications.

To get a sense of the population first stage R^2, we investigate how well the methods perform out of sample. We expect more restrictive approaches to have a higher out of sample R^2 if the model is overfitted. A benefit of comparing out of sample R^2 is that we get around the problem of R^2 always increasing with more variables, even if these variables do not have any true explanatory power. We use the out of sample R^2 instead of reporting the adjusted R^2 on the basis that the adjusted R^2 is a biased estimator of the population R^2. The second stage R^2 is not reported as it is irrelevant for our analysis.

If there are problems with overfitting, we would expect the out of sample R^2 to increase for models with a more restrictive choice of instruments. As mentioned before, increasing the R^2 decreases inconsistency, as it enters the denominator of Equation (3). We note that this does not evaluate the violation of instrumental exogeneity directly but does remedy the problems of inconsistency if the R^2 increases.

To estimate the out of sample R^2, the data is first randomly split into an 80% training set that is used to train the post-LASSO and post-L_2boosting model fits. These trained models are then used to make predictions on the withheld 20% of the data, the test set. The out of sample R^2 is calculated using:

$$R^2 = \frac{\sum(\widehat{x}_{i,test} - \overline{x}_{train})^2}{\sum(x_{i,test} - \overline{x}_{train})^2} \tag{14}$$

Equation (14) is the proportion of explained variance to the proportion of total variance. The reason for using $\overline{x}_{train}$ rather than $\overline{x}_{test}$ is because both are estimators of the same quantity, namely the population mean of x, and $\overline{x}_{train}$ is a better estimator because it was estimated with more data.

7 Data

For implementing our method, we work with the same data as Angrist & Krueger (1991). This is a publicly available sample from the US population census conducted on 1 April 1980. The data set includes 329,509 observations, with data on the log of weekly wages, years of education, year of birth, quarter of birth, and state of residence at the time of the census. The studied population is white males, born 1930–1939.

7.1 Potential instruments

In their original paper, Angrist & Krueger (1991) used quarter of birth `(QOB)` interacted with year of birth `(YOB)` to identify variations of educational attainment within a given year. They also used quarter of birth interacted with state of birth to allow for seasonal patterns within states.

We chose to include pure quarter of birth dummy variables in addition to the variables used by Angrist & Krueger (1991). This, is in accordance with Belloni et al. (2011). Note that including these three dummy variables means that we need to exclude three `(QOB)*(YOB)` dummies.

Belloni et al. (2011) use post-LASSO to select amongst a total of 1530 potential instruments. Due to the computational complexities of cross validating the L_2boosting algorithm, we restrict ourselves to a set of 180 potential instruments. Selection among these variables also constitute a more direct comparison to the original set of instruments used by Angrist & Krueger (1991).

Type of instrument	Number of instruments
Quarter of birth	3
Quarter of birth * Year of birth	27
Quarter of birth * State of birth	150
Total	180

Table 1: Overview of potential instruments

We analyse two different model specifications, found in columns 1 through 4 of Table VII in Angrist & Krueger (1991). These specifications were chosen because they include the maximum number of instruments, for which we have available data. The first model, hereafter referred to as *age excluded*, is an IV estimate of the returns to schooling and includes control variables for year of birth and state of birth. The equation for *age excluded* is:

$$E_i = X_i\pi + \sum_{j=1}^{3} Q_{ij}\omega_j + \sum_{c=0}^{8}\sum_{j=1}^{3} Y_{ic}Q_{ij}\theta_{jc} + \sum_{s=1}^{50}\sum_{j=1}^{3} S_{is}Q_{ij}\tau_{sj} + \epsilon_i \tag{15}$$

Where X_i is a vector of covariates, with π being the vector coefficients related to the variables. The three summations refer to the different instruments; quarter of birth Q, quarter of birth interacted with year of birth Y, and quarter of birth interacted with state of birth S, one for each of the 50 states.

The second model, hereafter referred to as *age included*, includes covariates for *age* and age^2, in addition to the covariates in the previous model. Since *age* and age^2 are measured in quarters, two instruments must be removed from *age included* to avoid perfect multicollinearity. This leaves us with 178 potential instruments for the *age included* replication.

7.2 Data preparation

We load and prepare the data in the statistical computations program R (R Core Team 2017). The quarter of birth, year of birth, and state of birth variables are redefined as dummy variables, with

one dummy per category. We create the interaction terms for quarter of birth (`QOB`) multiplied by year of birth (`YOB`), and quarter of birth multiplied by state of birth (`State`).

We remove all variables for being born in 1939 or in the fourth quarter, the dummy variables and their interaction terms. For the *age included* specification, which includes two age variables, we remove the dummy variable for being born in the third quarter of 1938 (`QOBYOB.38`) and the dummy variable for being born in the third quarter and the state of Wisconsin (`QOBState.3_55`).

To facilitate the model specifications, we name the total set of potential instruments `M`, the set of control variables for replicating *age excluded* `K.2`, and the set of control variables for replicating *age included* as `K.4`. Table 2 serves as a control to verify that everything is prepared in accordance with the study we replicate.

Table 2: Replication of table VII in Angrist & Krueger (1991)

	Dependent variable:			
	Return to education	Return to education	Return to education	Return to education
	OLS	*instrumental variable*	*OLS*	*instrumental variable*
	(1)	(2)	(3)	(4)
Years of education	0.067^{***}	0.093^{***}	0.067^{***}	0.091^{***}
	(0.0003)	(0.009)	(0.0003)	(0.011)
age and age^2	*Excluded*	*Excluded*	*Included*	*Included*
Observations	329,509	329,509	329,509	329,509
R^2	0.129	0.114	0.129	0.117
Adjusted R^2	0.129	0.114	0.129	0.117
Residual Std. Error	0.634 (df = 329268)	0.639 (df = 329448)	0.634 (df = 329268)	0.638 (df = 329446)
F Statistic (df = 240; 329268)	203.631^{***}		203.631^{***}	
Note:				*p<0.1; **p<0.05; ***p<0.01

7.2.1 Parallelisation and random seed

Our calculations involve parallelisation for cross validating the L_2boosting in order to facilitate faster calculations. This makes use of a method called *forking* which is only possible on Unix based computers. In order to reproduce the results on a computer running on Windows, the parameter `papply` (in the function `cvrisk` in the package `mboost` by Hothorn et al. (2017)) needs to be changed from `mclapply` to `lapply`. Because the involvement of several computing cores means the standard seed setting in R will not suffice to reproduce any results, we use the random number generator developed in L'ecuyer et al. (2002) by setting the R random number generator `RNGkind` to `'L'Ecuyer-CMRG'`.

7.3 Programming the LASSO

In order to perform the LASSO, we use the R package `glmnet` and the function `cv.glmnet`. The function does k-fold cross validation (we set $k = 10$) over 100 automatically generated values of λ and returns both λ_{min} and λ_{1se} (*see Subsection* 6.1.1) based on the MSE loss function. λ_{min} and λ_{1se} returned by the cross validation are indicated in Figure 3 by the dashed lines. The left dashed line in each graph indicate λ_{min} and the right dashed line in each graph indicate λ_{1se}. The left graph represents the *age excluded* specification and the graph to the right represents the *age included* specification. Values above the graph show how many variables are associated with each value of $log(\lambda)$, measured on the bottom axis.

In `glmnet`, we set the parameter `alpha` to 1, meaning that the penalty term should only include the absolute value of the parameters, and no squared functions thereof. (If we had set `alpha`=0 we would have had a ridge regression, and for any value 0< `alpha` <1 we have a combination of LASSO and the ridge regression (an *elastic net*).)

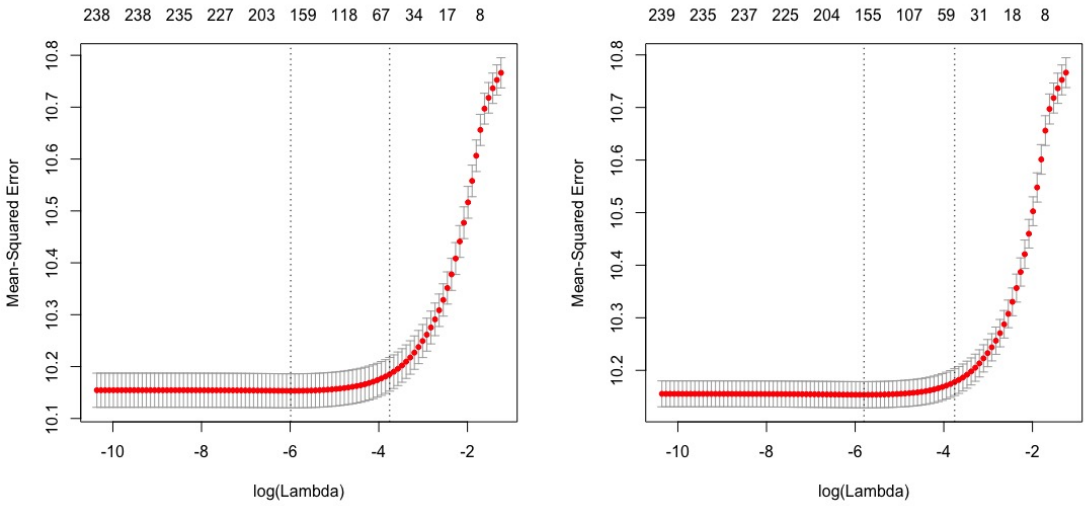

Figure 3: Cross validated λ

The left and the right graphs show the cross validated λ with corresponding MSE for model specifications *age excluded* and *age included* respectively. The values above the graphs indicate the number of chosen variables. Note that the right dashed line intersects the red curve at the same height as the upper standard error by the left dashed line.

7.4 Programming the boosting

We use the R-package `mboost` by Hothorn et al. (2017) to implement the L_2boosting and use the function `glmboost` which we give a least squares (L_2) loss function by setting the `family` argument to `GaussReg()`. We also allow the function to centre the variables for quicker convergence.

Despite the variables in our data set being correlated (and we do not implement orthogonal gradient boosting) the idea of keeping the number of iterations in the region of $K + M$ is used and we set the maximum number of iterations at `250` (*see* 6.1.2).

Figure 4 shows the 10-fold cross validated estimations of out of sample MSE in thin grey, and the average estimated out of sample MSE in black produced by the function `cvrisk`. The point where the cross validated average MSE is the lowest occurs for the optimal number of iterations (m^*). 250 turned out to be sufficiently large maximum number of iterations, surpassing the optimal number of iterations; $m^*_{excl} = 213$ and $m^*_{incl} = 247$. These parameters are relatively high, meaning that many instruments were selected. We also see that the curve showing the squared errors is relatively flat after about 40 iterations. This suggests that reducing the number of instruments by earlier stopping would not increase the MSE substantially. We should also note the spread between the different cross validation curves (in thin grey). The spread indicates a high variance and that the stopping could have occurred after much fewer iterations, depending on chance.

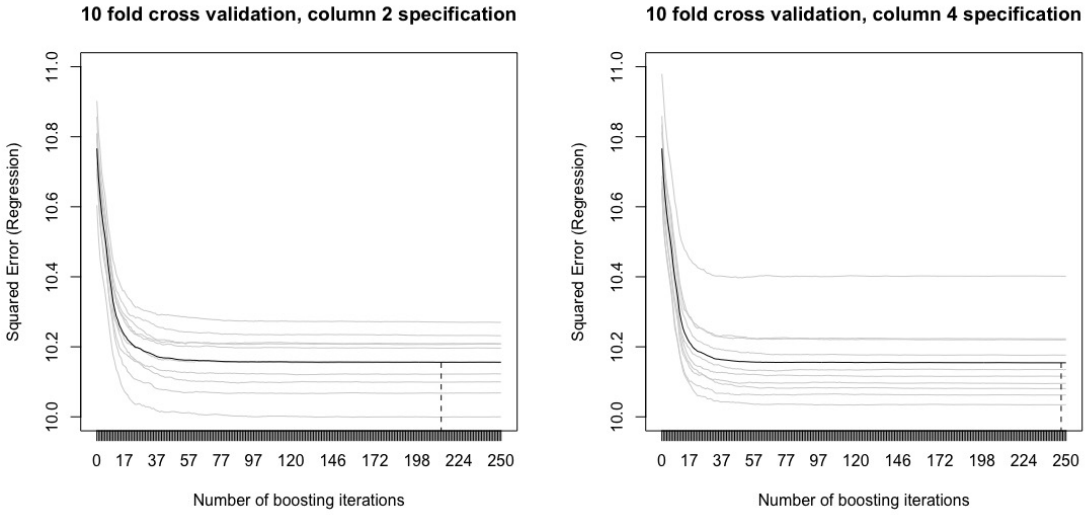

Figure 4: Cross validated m^*

The figure shows the 10-fold cross validation (one value of the squared error for each fold and every value of m). The black line is the average squared error associated with each $0 < m \leq 250$. The cross validation finds that $m^*_{excl} = 213$ and $m^*_{incl} = 247$.

We report the optimal number of iterations associated with different values of C in Table 3. The data dependent stopping rule stops earlier for higher values of C, and earlier than the cross validated (cv) stopping criterion for any value of C.

Table 3: Optimal stopping for cross validation and different values of C

	Post-L_2boosting							
age and age^2	Excluded				Included			
C	CV	0.25	0.5	0.75	CV	0.25	0.5	0.75
m^*	213	133	99	89	247	124	86	73

7.5 Programming the IV estimations

The IV estimations are implemented using the function `ivreg` from the R package `AER` by Kleiber & Zeileis (2008). The output provides diagnostics such as the test for weak instruments that is the first stage F-statistic. The first stage R^2 is obtained by regressing years of education (`EDUC`) on chosen instruments Z and appropriate controls K, using the standard R function `lm`. The Tables 8 and 9 in the appendix report what instruments were chosen for each model specification.

8 Results

Table 4 shows the number of instruments selected by the LASSO and L_2boosting for the different stopping criteria; 10-fold cross validation and the data dependent stopping rules specified in Subsection 3.7. The full tables of chosen instruments can be found in Table 8 and 9 in Appendix. LASSO with λ_{1se} is the most conservative variable selection method and select the fewest instruments. On the other end of the spectrum, we have methods with cross validated stopping criteria. This indicates that many instruments do contain some information and contribute to explain educational attainment.

Number of instruments selected		
age and age^2	Excluded	Included
Variable selection method		
post-LASSO, λ_{1se}	11	9
post-LASSO, λ_{min}	119	113
post-Boosting, CV	111	122
post-Boosting, $C = 0.25$	61	65
post-Boosting, $C = 0.5$	40	35
post-Boosting, $C = 0.75$	32	25

Table 4: Number of instruments selected by the various algorithm specifications.

Tables 5 and 6 show 2SLS estimates on the whole data set for our two model specifications *age excluded* and *age included* respectively. Column 1 uses the full set of 180, or 178[4] instruments. Columns 2 and 3 show estimates for the two versions of the post-LASSO. Column 4 through 7 show estimates for post-L_2boosting with different tuning techniques for determining the stopping rule; 10-fold cross validation, $C = 0.25$, $C = 0.50$, and $C = 0.75$.

In the bottom rows of Tables 5 and 6 we find the first stage 'F-statistic'. Post-LASSO, with the conservative stopping rule (λ_{1se}), results in the highest F-statistic for both specifications and exceeds the $F = 10$ threshold for the *age included* specification. The F-statistic differs substantially for the different versions of the post-L_2boosting algorithm, depending on the stopping criteria. We find that the most conservative stopping rule for post-L_2boosting results in the highest F-statistics. However, no stopping rule produces an F-statistic exceeding the $F = 10$ threshold.

In comparing the two algorithms tuned by the same technique (cross validation), post-L_2boosting yields slightly higher F-statistics. However, the difference is small considering both algorithms exhibit large variances in the cross validated values of the loss function (*see figures 3 and 4*). The large variability between the MSEs for the L_2boosting, the broad confidence intervals of the cross validated errors of the LASSO, and the flat shapes of the estimated MSEs all suggest that the number of iterations, and selected λ, is sensitive to random seed.

Standard errors are fairly similar across the different methods, and all have a significance level of 1%. This suggests that there is no significant trade-off in precision between the different methods.

[4]Recall that since two control variables, *age* and *age*^{2}, are added in *age included* two instruments are removed.

Table 5: *Age excluded* specification.

	Dependent variable:						
	Log weekly wage						
	$AK91$	post-LASSO		post-L_2boost			
		λ_{1se}	λ_{min}	$m_{cv} = 213$	$m_{.25} = 133$	$m_{.5} = 99$	$m_{.75} = 89$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Years of education	0.093***	0.105***	0.092***	0.094***	0.086***	0.082***	0.083***
	(0.009)	(0.015)	(0.010)	(0.010)	(0.010)	(0.011)	(0.012)
Controls							
9 Year of birth dummies	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
50 State of birth dummies	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Age and age squared	*No*	*No*	*No*	*No*	*No*	*No*	*No*
Statistics							
1st stage R^2	0.0582	0.0574	0.0582	0.0581	0.0579	0.0578	0.0577
Partial R^2	0.00141	0.00055	0.00133	0.00132	0.00111	0.00097	0.00088
F-statistic	2.582	16.589	3.692	3.911	5.983	7.96	9.02

Note: *p<0.1; **p<0.05; ***p<0.01

$n = 329509$. Standard error in parentheses. 'AK91' refers to a model with all instruments, 'λ_{1se}' denotes post-LASSO with regularisation parameter set using the 1 standard error rule, 'λ_{min}' denotes post-LASSO with regularisation parameter set using cross validation. m_{cv} denotes L_2boosting with number of iterations set using cross validation. $m_{.25}$, $m_{.5}$, and $m_{.75}$ denotes L_2boosting with number of iterations set by the inequality rule for different values of C indicated by the subscript.

Table 6: *Age included* specification.

	Dependent variable:						
	Log weekly wage						
	$AK91$	post-LASSO		post-L_2boost			
		λ_{1se}	λ_{min}	$m_{cv} = 247$	$m_{.25} = 124$	$m_{.5} = 86$	$m_{.75} = 73$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Years of education	0.091***	0.111***	0.089***	0.092***	0.086***	0.075***	0.078***
	(0.011)	(0.024)	(0.012)	(0.011)	(0.012)	(0.014)	(0.016)
Controls							
9 Year of birth dummies	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
50 State of birth dummies	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Age and age squared	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Statistics							
1st stage R^2	0.0582	0.0574	0.058	0.0582	0.058	0.0578	0.0577
Partial R^2	0.00106	0.00023	0.00079	0.00101	0.00081	0.00058	0.00046
F-statistic	1.972	8.247	2.292	2.72	4.09	5.424	6.056

Note: *p<0.1; **p<0.05; ***p<0.01

$n = 329509$. Standard error in parentheses. 'AK91' refers to a model with all instruments, 'λ_{1se}' denotes post-LASSO with regularisation parameter set using the 1 standard error rule, 'λ_{min}' denotes post-LASSO with regularisation parameter set using cross validation. m_{cv} denotes L_2boosting with number of iterations set using cross validation. $m_{.25}$, $m_{.5}$, and $m_{.75}$ denotes L_2boosting with number of iterations set by the inequality rule for different values of C indicated by the subscript.

Table 7 shows the R^2 from the first stage regression, estimated on a test set. The differences between the different techniques are on the order of less than 10^{-3} in magnitude. The original specification obtains the highest out of sample first stage R^2. However, any small decrease may have a large effect on consistency, given that the $R^2_{x,z}$ is already low in the original study (0.00141 and 0.00106). The $R^2_{x,z}$ seems to decrease faster for the model specification *age included* and is at its lowest for post-LASSO with λ_{1se}. Furthermore, Table 7 does not indicate that the original specification in Angrist & Krueger (1991) (AK91) was overfitted, since even the cross validated tunings of the algorithms had lower out of sample R^2 than the original specification. Hence, when the algorithms select fewer variables we obtain simpler models with higher F-statistics at the expense of out of sample performance (R^2).

Table 7: R^2 on a 20 % test set

age and age^2	AK91	post-L_2boosting				post-LASSO	
		m_{cv}	$m_{.25}$	$m_{.5}$	$m_{.75}$	λ_{1se}	λ_{min}
Excluded	0.05804	0.05788	0.05768	0.05752	0.05744	0.05718	0.05790
Included	0.05803	0.05791	0.05770	0.05747	0.05735	0.05715	0.05780

This table shows out of sample R^2 for the different variable selection methods and their variations, compared to original model specification in Angrist & Krueger (1991). m_{cv} indicates m^* selected using cross validation while $m_{.25}$, $m_{.5}$, and $m_{.75}$ indicate m^* for different values of C denoted in the subscript. λ_{1se} indicates λ chosen by the one standard error stopping rule, and λ_{min} indicates the cross validated λ.

9 Discussion

Both LASSO and boosting are successful to the sense that they select quarter of birth without any interaction terms, which is 'the variable that most cleanly satisfies Angrist's and Krueger's argument for the validity of the instrument set' (Belloni et al. 2011, p. 20). In addition, there are apparent similarities between the variables chosen by the different variable selection methods. This is reassuring because if there were no pattern, it would either mean that no instrument was superior in predicting educational attainment (which we know not to be true); or there would have been something wrong with one or several of the methods used. Not surprisingly, as shown in Tables 8 and 9, more restrictive versions of the same algorithm never select instruments not selected by more relaxed versions.

L_2boosting yields promising results when we compare columns 3 and 4 in Tables 8 and 9 where both hyper parameters were chosen by cross validation. Comparing the F-statistics for post-L_2boosting and post-LASSO show that post-L_2boosting achieves slightly higher F-statistics. In addition, we note that the $R^2_{x,z}$ in Tables 8 and 9 do not differ substantially between the different models. This means that finite sample bias is further reduced when exchanging post-LASSO for post-L_2boosting (since F increases). Meanwhile, the inconsistency remains roughly constant (since the $R^2_{x,z}$ differs little). The overall effect is therefore a smaller bias for the L_2boosting compared to the post-LASSO. On the other hand, Figures 3 and 4 indicate high standard errors for the estimated MSE, suggesting that the results may be sensitive to random seed.

More restrictive criteria than tuning by 10-fold cross validation seem to generally yield more significant F-statistics. Consequently, using the 1 standard error rule with the L_2boosting algorithm to select an $m_{1se} < m^*$ could be beneficial as such a rule could yield a more restrictive variable selection. On the other hand, higher F-statistics are also associated with a lower $R^2_{x,z}$ and test set R^2. Therefore, if there is cause for concern regarding the exogeneity assumption, deviating from the cross validated hyper parameters may be undesirable as this is likely to increase the inconsistency of the IV estimate. On the other hand, if the concern regards finite sample bias, being more restrictive in the variable selection process may be preferable as the F-statistics seem likely to increase.

Concretely, we see that post-LASSO with λ_{1se} is superior in terms of increasing the F-statistic. The quantitative difference between the best F-statistics from the two algorithms is 16.589 compared to 9.02 for model specification *age excluded*. This is a substantial, and qualitatively significant, difference as the boosting algorithm does not produce an F-statistic above the $F = 10$ threshold. In practice, therefore, using the wrong method could lead the researcher to incorrectly conclude that the IV-estimate is insignificant.

10 Conclusion

We have tried to determine which variable selection method, post-LASSO or post-L_2boosting, works better by tying existing research to a single data set and adding theoretically motivated evaluation metrics. In doing so, we wish to contribute to bridging the gap between machine learning and econometrics and show that the two disciplines can be thought of as complementary.

Our approach provides comparable results indicating that variable selection *per se* is helpful before an IV estimation when there are many weak instruments to choose from, and the researcher has no *a priori* knowledge of what instruments should be included. This is consistent with previous findings. However, we are unable to declare any algorithm as strictly better than the other, especially if there is cause for concern regarding instrument endogeneity. If we can assume instrument exogeneity, post-LASSO with λ_{1se} is the better choice, as it yields the highest F-statistic.

There seems to be a trade-off between inconsistency and finite sample bias that is determined by the choice of hyper parameters λ and m. However, the fact that our analysis was only made for a single data set means we only get an indication for how the results might extend to other data sets.

In conclusion, we recommend the applied researcher to think carefully about the data at hand before selecting the regularisation hyper parameters. Regarding what algorithm to use, both post-LASSO and post-L_2boosting seem helpful, with the former being easier to implement, faster to compute, and producing the highest F-statistic in our study. An avenue for further research could be to apply more homogeneous hyper parameter tuning techniques, such as investigating the 1 standard error rule for L_2boosting, and to perform similar analyses on simulated and other empirical data sets.

References

Angrist, J. D., Imbens, G. W. & Krueger, A. B. (1999), 'Jackknife instrumental variables estimation', *Journal of Applied Econometrics* **14**(1), 57–67.

Angrist, J. D. & Krueger, A. B. (1991), 'Does compulsory school attendance affect schooling and earnings?', *The Quarterly Journal of Economics* **106**(4), 979–1014.

Angrist, J. D. & Krueger, A. B. (1995), 'Split-sample instrumental variables estimates of the return to schooling', *Journal of Business & Economic Statistics* **13**(2), 225–235.

Angrist, J. D. & Krueger, A. B. (2001), 'Instrumental variables and the search for identification: from supply and demand to natural experiments', *Journal of Economic perspectives* **15**(4), 69–85.

Athey, S. (2017), The impact of machine learning on economics, *in* 'Economics of Artificial Intelligence', University of Chicago Press.

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360.

Athey, S. & Imbens, G. W. (2017), 'The state of applied econometrics: causality and policy evaluation', *Journal of Economic Perspectives* **31**(2), 3–32.

Belloni, A., Chernozhukov, V. & Hansen, C. (2011), 'Lasso methods for gaussian instrumental variables models'.

Bound, J., Jaeger, D. A. & Baker, R. M. (1995), 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak', *Journal of the American statistical association* **90**(430), 443–450.

Buhagiar, J. (2017), Automatic segmentation of indoor and outdoor scenes from visual lifelogging, PhD thesis, University of Malta.

Bühlmann, P. & Yu, B. (2003), 'Boosting with the l_2 loss: regression and classification', *Journal of the American Statistical Association* **98**(462), 324–339.

Buse, A. (1992), 'The bias of instrumental variable estimators', *Econometrica: Journal of the Econometric Society* pp. 173–180.

Carrasco, M. (2012), 'A regularization approach to the many instruments problem', *Journal of Econometrics* **170**(2), 383–398.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* pp. 1189–1232.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Hansen, C. & Kozbur, D. (2014), 'Instrumental variables estimation with many weak instruments using regularized jive', *Journal of Econometrics* **182**(2), 290–308.

Hartford, J., Lewis, G., Leyton-Brown, K. & Taddy, M. (2017), Deep iv: a flexible approach for counterfactual prediction, *in* 'International Conference on Machine Learning', pp. 1414–1423.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. (2017), *Mboost: model-based boosting*. R package version 2.8-1.
URL: *https://CRAN.R-project.org/package=mboost*

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

Kleiber, C. & Zeileis, A. (2008), *Applied Econometrics with R*, Springer-Verlag, New York. ISBN 978-0-387-77316-2.
URL: *https://CRAN.R-project.org/package=AER*

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Ijcai', Vol. 14, Montreal, Canada, pp. 1137–1145.

L'ecuyer, P., Simard, R., Chen, E. J. & Kelton, W. D. (2002), 'An object-oriented random-number package with many long streams and substreams', *Operations research* **50**(6), 1073–1075.

Ludwig, J., Mullainathan, S. & Spiess, J. (2017), 'Machine learning tests for effects on multiple outcomes', *arXiv preprint arXiv:1707.01473* .

Luo, Y. & Spindler, M. (2016), 'High-dimensional l_2 boosting: rate of convergence', *arXiv preprint arXiv:1602.08927* .

Luo, Y. & Spindler, M. (2017), 'l_2-boosting for economic applications', *American Economic Review* **107**(5), 270–73.

Mitchell, T. M. (1997), 'Machine learning (mcgraw-hill international editions computer science series)'.

Mullainathan, S. & Spiess, J. (2017), 'Machine learning: an applied econometric approach', *Journal of Economic Perspectives* **31**(2), 87–106.

R Core Team (2017), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Sawa, T. (1969), 'The exact sampling distribution of ordinary least squares and two-stage least squares estimators', *Journal of the American Statistical association* **64**(327), 923–937.

Staiger, D. O. & Stock, J. H. (1994), 'Instrumental variables regression with weak instruments'.

Stock, J. & Yogo, M. (2005), *Asymptotic distributions of instrumental variables statistics with many instruments*, Vol. 6, Chapter.

Theil, H. (1953), 'Repeated least squares applied to complete equation systems', *The Hague: central planning bureau* .

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Varian, H. R. (2014), 'Big data: new tricks for econometrics', *Journal of Economic Perspectives* **28**(2), 3–28.

Wooldridge, J. M. (2009), 'Introductory econometrics a modern approach. usa: South-western cengage learning'.

Wright, P. G. (1928), *Tariff on animal and vegetable oils*, Macmillan Company, New York.

Appendix

In Tables 8 and 9, `x` signifies that the instrument has been selected by the variable selection method, and thus included in the IV regression. `QTR.1` is a dummy variable for being born in the first quarter (of any year), `QOBYOB.10` is a dummy variable for being born the first quarter of 1930 and `QOBState.1_01` is a dummy variable for being born in the first quarter and in state number 01 (Alabama). The other variables follow the same logic.

Table 8: For age excluded specification

	Instruments	λ_{1se}	λ_{min}	m_{cv}	$m_{.25}$	$m_{.5}$	$m_{.75}$
1	QTR.1	X	X	X	X	X	X
2	QTR.2	.	X	X	X	X	X
3	QTR.3	.	.	.	.	.	.
4	QOBYOB.10	X	X	X	X	X	X
5	QOBYOB.11	.	X	.	.	.	.
6	QOBYOB.12	.	X	X	X	X	X
7	QOBYOB.13	.	.	X	X	.	.
8	QOBYOB.14	.	X	X	X	.	.
9	QOBYOB.15	.	X	X	X	X	X
10	QOBYOB.16	.	X	.	.	.	.
11	QOBYOB.17	.	X	X	X	.	.
12	QOBYOB.18	.	.	.	.	.	.
13	QOBYOB.20	.	X	X	X	X	X
14	QOBYOB.21	X	X	X	X	X	.
15	QOBYOB.22	.	X	X	.	.	.
16	QOBYOB.23	.	X	X	.	.	.
17	QOBYOB.24	.	.	.	.	.	.
18	QOBYOB.25	.	.	.	.	.	.
19	QOBYOB.26	.	X	.	.	.	.
20	QOBYOB.27	.	X	X	X	.	.
21	QOBYOB.28	.	X	.	.	.	.
22	QOBYOB.30	.	X	X	X	X	X
23	QOBYOB.31	.	X	X	.	.	.
24	QOBYOB.32	.	X	X	X	X	X
25	QOBYOB.33	.	X	X	.	.	.
26	QOBYOB.34	.	X	X	X	X	X
27	QOBYOB.35	.	X	.	.	.	.
28	QOBYOB.36	.	.	.	.	.	.
29	QOBYOB.37	.	.	.	.	.	.
30	QOBYOB.38	.	X	X	X	X	X
31	QOBState.1_01	.	X	X	X	.	.
32	QOBState.1_02	.	X	X	.	.	.
33	QOBState.1_04	.	X	X	.	.	.
34	QOBState.1_05	.	X	X	.	.	.
35	QOBState.1_06	.	.	X	X	.	.
36	QOBState.1_08	.	X	X	X	X	X
37	QOBState.1_09	.	.	.	.	.	.
38	QOBState.1_10	.	X	X	.	.	.
39	QOBState.1_11	.	X	X	X	X	.
40	QOBState.1_12	.	X	X	.	.	.
41	QOBState.1_13	.	.	.	.	.	.
42	QOBState.1_15	.	X	.	.	.	.
43	QOBState.1_16	.	.	.	.	.	.
44	QOBState.1_17	.	X	X	X	X	X
45	QOBState.1_18	.	.	.	.	.	.
46	QOBState.1_19	.	X	X	.	.	.
47	QOBState.1_20	.	X	X	X	.	.
48	QOBState.1_21	X	X	X	X	X	X
49	QOBState.1_22	.	X	X	.	.	.
50	QOBState.1_23	.	X	.	.	.	.
51	QOBState.1_24	X	X	X	X	X	X
52	QOBState.1_25	.	X	X	.	.	.
53	QOBState.1_26	.	.	.	.	.	.
54	QOBState.1_27	.	X	X	X	X	X
55	QOBState.1_28	.	X	X	.	.	.
56	QOBState.1_29	.	.	X	.	.	.
57	QOBState.1_30	.	.	.	.	.	.
58	QOBState.1_31	.	X	X	X	.	.
59	QOBState.1_32	.	.	.	.	.	.
60	QOBState.1_33	.	.	.	.	.	.
61	QOBState.1_34	.	.	.	.	.	.
62	QOBState.1_35	.	X	X	.	.	.
63	QOBState.1_36	.	X	X	X	.	.
64	QOBState.1_37	.	X	X	X	X	.
65	QOBState.1_38	.	X	X	X	X	X
66	QOBState.1_39	.	X	X	.	.	.
67	QOBState.1_40	.	.	.	.	.	.
68	QOBState.1_41	.	.	.	.	.	.
69	QOBState.1_42	.	X	X	.	.	.
70	QOBState.1_44	.	.	.	.	.	.
71	QOBState.1_45	.	X	.	.	.	.
72	QOBState.1_46	.	X	X	X	.	.
73	QOBState.1_47	.	X	X	X	X	X
74	QOBState.1_48	.	X	X	.	.	.
75	QOBState.1_49	.	X	X	.	.	.
76	QOBState.1_50	.	X	X	.	.	.
77	QOBState.1_51	.	X	X	X	X	X
78	QOBState.1_53	.	.	.	.	.	.
79	QOBState.1_54	.	X	.	.	.	.
80	QOBState.1_55	.	.	X	X	.	.
81	QOBState.2_01	.	X	X	X	.	.
82	QOBState.2_02	.	X	X	X	X	X
83	QOBState.2_04	.	X	X	X	X	X
84	QOBState.2_05	X	X	X	X	X	X
85	QOBState.2_06	.	X	X	X	X	X
86	QOBState.2_08	.	.	.	.	.	.
87	QOBState.2_09	.	X	X	X	X	.
88	QOBState.2_10	.	.	X	.	.	.
89	QOBState.2_11	.	.	.	.	.	.
90	QOBState.2_12	.	X	X	X	.	.
91	QOBState.2_13	X	X	X	X	X	X
92	QOBState.2_15	.	.	X	X	.	.
93	QOBState.2_16	.	.	X	.	.	.
94	QOBState.2_17	.	.	X	X	X	X
95	QOBState.2_18	.	.	.	.	.	.
96	QOBState.2_19	.	.	.	.	.	.
97	QOBState.2_20	.	.	X	X	.	.
98	QOBState.2_21	X	X	X	X	X	X
99	QOBState.2_22	.	.	X	X	.	.
100	QOBState.2_23	.	.	X	.	.	.
101	QOBState.2_24	.	.	.	X	.	.
102	QOBState.2_25	.	.	.	.	.	.
103	QOBState.2_26	.	.	X	X	X	.
104	QOBState.2_27	.	.	.	.	.	.
105	QOBState.2_28	.	.	X	X	.	.
106	QOBState.2_29	.	.	X	X	.	.
107	QOBState.2_30	.	.	.	X	.	.
108	QOBState.2_31	.	.	.	.	.	.
109	QOBState.2_32	.	.	X	X	.	.
110	QOBState.2_33	.	.	X	X	.	.
111	QOBState.2_34	.	.	.	.	.	.
112	QOBState.2_35	.	.	X	.	.	.
113	QOBState.2_36	.	.	X	X	X	X
114	QOBState.2_37	X	X	X	X	X	X
115	QOBState.2_38	.	.	X	X	X	.
116	QOBState.2_39	.	.	X	.	.	.
117	QOBState.2_40	.	.	X	X	X	.
118	QOBState.2_41	.	.	X	.	.	.
119	QOBState.2_42	.	.	.	.	.	.
120	QOBState.2_44	.	.	X	X	X	.
121	QOBState.2_45	.	.	X	X	X	.
122	QOBState.2_46	.	.	.	.	.	.
123	QOBState.2_47	X	X	X	X	X	X
124	QOBState.2_48	X	X	X	X	X	X
125	QOBState.2_49	.	.	X	X	.	.
126	QOBState.2_50	.	.	X	.	.	.
127	QOBState.2_51	.	.	X	X	X	.
128	QOBState.2_53	.	.	X	X	.	.
129	QOBState.2_54	.	.	X	X	.	.
130	QOBState.2_55	.	.	X	X	X	X
131	QOBState.3_01	.	.	X	X	X	X
132	QOBState.3_02	.	.	X	X	X	.
133	QOBState.3_04	.	.	.	.	.	.
134	QOBState.3_05	.	.	X	X	.	.
135	QOBState.3_06	.	.	X	.	.	.
136	QOBState.3_08	.	.	X	X	.	.
137	QOBState.3_09	.	.	.	.	.	.
138	QOBState.3_10	.	.	X	X	.	.
139	QOBState.3_11	.	.	.	.	.	.
140	QOBState.3_12	.	.	X	.	.	.
141	QOBState.3_13	.	.	.	.	.	.
142	QOBState.3_15	.	.	X	X	X	X
143	QOBState.3_16	.	.	.	X	.	.
144	QOBState.3_17	.	.	X	X	X	.
145	QOBState.3_18	.	.	.	.	.	.
146	QOBState.3_19	.	.	.	X	.	.
147	QOBState.3_20	.	.	X	X	.	.
148	QOBState.3_21	.	.	X	X	.	.
149	QOBState.3_22	.	.	.	.	.	.
150	QOBState.3_23	.	.	X	X	.	.
151	QOBState.3_24	.	.	.	X	.	.
152	QOBState.3_25	.	.	.	.	.	.
153	QOBState.3_26	.	.	.	.	.	.
154	QOBState.3_27	.	.	.	X	.	.
155	QOBState.3_28	.	.	X	.	.	.
156	QOBState.3_29	.	.	.	X	.	.
157	QOBState.3_30	.	.	.	.	.	.
158	QOBState.3_31	.	.	.	.	.	.
159	QOBState.3_32	.	.	.	X	.	.
160	QOBState.3_33	.	.	.	.	.	.
161	QOBState.3_34	.	.	X	X	.	.
162	QOBState.3_35	.	.	X	X	.	.
163	QOBState.3_36	.	.	X	X	.	.
164	QOBState.3_37	.	.	X	X	X	.
165	QOBState.3_38	.	.	.	.	.	.
166	QOBState.3_39	.	.	.	.	.	.
167	QOBState.3_40	.	.	.	.	.	.
168	QOBState.3_41	.	.	.	.	.	.
169	QOBState.3_42	.	.	X	X	X	.
170	QOBState.3_44	.	.	X	.	.	.
171	QOBState.3_45	.	.	X	X	X	.
172	QOBState.3_46	.	.	.	X	.	.
173	QOBState.3_47	.	.	X	X	X	.
174	QOBState.3_48	.	.	X	.	.	.
175	QOBState.3_49	.	.	.	.	.	.
176	QOBState.3_50	.	.	.	X	.	.
177	QOBState.3_51	.	.	.	.	.	.
178	QOBState.3_53	.	.	X	.	.	.
179	QOBState.3_54	.	.	X	.	.	.
180	QOBState.3_55	.	.	X	.	.	.
181	Count	11	119	111	61	40	32

'QTR.q' indicates being born in quarter q. 'QOBYOB.qy' means being born in quarter q year y. 'QOBSTATE.q_ss' means being born quarter q in state ss. Marked 'x' means the instrument was chosen by the algorithm in the column. m_{cv} indicates m^* selected using cross validation while $m_{.25}$, $m_{.5}$, and $m_{.75}$ indicate m^* for different values of C denoted in the subscript. λ_{1se} indicates λ chosen by the one standard error stopping rule, and λ_{min} indicates the cross validated λ.

Table 9: For age included specification

	Instruments	λ_{1se}	λ_{min}	m_{cv}	$m_{.25}$	$m_{.5}$	$m_{.75}$
1	QTR.1	X	X	X	X	X	X
2	QTR.2	.	.	.	.	.	.
3	QTR.3	.	X	X	.	.	.
4	QOBYOB.10	X	X	X	X	X	X
5	QOBYOB.11	.	.	X	.	.	.
6	QOBYOB.12	.	X	X	X	.	.
7	QOBYOB.13	.	X	X	X	X	X
8	QOBYOB.14	.	X	X	.	.	.
9	QOBYOB.15	.	X	X	X	X	.
10	QOBYOB.16	.	X	X	.	.	.
11	QOBYOB.17	.	X	X	X	.	.
12	QOBYOB.18	.	X	X	.	.	.
13	QOBYOB.20	.	X	X	.	.	.
14	QOBYOB.21	.	X	X	.	.	.
15	QOBYOB.22	.	.	.	.	.	.
16	QOBYOB.23	.	X	X	.	.	.
17	QOBYOB.24	.	.	.	.	.	.
18	QOBYOB.25	.	X	X	.	.	.
19	QOBYOB.26	.	X	X	.	.	.
20	QOBYOB.27	.	X	X	X	X	.
21	QOBYOB.28	.	X	X	.	.	.
22	QOBYOB.30	.	X	X	X	X	.
23	QOBYOB.31	.	X	X	X	X	X
24	QOBYOB.32	.	X	X	.	.	.
25	QOBYOB.33	.	X	X	X	.	.
26	QOBYOB.34	.	X	X	.	.	.
27	QOBYOB.35	.	X	X	X	.	.
28	QOBYOB.36	.	.	.	.	.	.
29	QOBYOB.37	.	X	X	.	.	.
30	QOBState.1_01	.	X	X	X	.	.
31	QOBState.1_02	.	X	X	.	.	.
32	QOBState.1_04	.	X	X	X	.	.
33	QOBState.1_05	.	.	X	.	.	.
34	QOBState.1_06	.	X	X	X	.	.
35	QOBState.1_08	.	.	X	X	X	X
36	QOBState.1_09	.	X	.	.	.	.
37	QOBState.1_10	.	X	X	.	.	.
38	QOBState.1_11	.	X	X	X	X	.
39	QOBState.1_12	.	.	X	.	.	.
40	QOBState.1_13	.	X	.	.	.	.
41	QOBState.1_15	.	.	X	.	.	.
42	QOBState.1_16	.	X	.	.	.	.
43	QOBState.1_17	.	.	X	X	X	X
44	QOBState.1_18	.	X	X	.	.	.
45	QOBState.1_19	.	.	.	.	.	.
46	QOBState.1_20	.	X	X	X	.	.
47	QOBState.1_21	X	X	X	X	X	X
48	QOBState.1_22	.	X	X	.	.	.
49	QOBState.1_23	.	X	.	.	.	.
50	QOBState.1_24	X	X	X	X	X	X
51	QOBState.1_25	.	.	X	.	.	.
52	QOBState.1_26	.	X	.	.	.	.
53	QOBState.1_27	.	X	X	X	X	X
54	QOBState.1_28	.	.	X	.	.	.
55	QOBState.1_29	.	.	X	.	.	.
56	QOBState.1_30	.	X	.	.	.	.
57	QOBState.1_31	.	.	X	X	.	.
58	QOBState.1_32	.	.	X	.	.	.
59	QOBState.1_33	.	.	.	.	.	.
60	QOBState.1_34	.	X	.	.	.	.
61	QOBState.1_35	.	X	X	.	.	.
62	QOBState.1_36	.	X	X	X	.	.
63	QOBState.1_37	.	X	X	X	X	.
64	QOBState.1_38	.	X	X	X	X	X
65	QOBState.1_39	.	.	X	X	.	.
66	QOBState.1_40	.	.	X	.	.	.
67	QOBState.1_41	.	X	.	.	.	.
68	QOBState.1_42	.	.	X	.	.	.
69	QOBState.1_44	.	X	.	.	.	.
70	QOBState.1_45	.	X	X	.	.	.
71	QOBState.1_46	.	X	X	X	.	.
72	QOBState.1_47	.	X	X	X	X	X
73	QOBState.1_48	.	X	X	X	.	.
74	QOBState.1_49	.	X	X	X	.	.
75	QOBState.1_50	.	X	X	.	.	.
76	QOBState.1_51	.	.	X	X	X	X
77	QOBState.1_53	.	X	.	.	.	.
78	QOBState.1_54	.	.	.	.	.	.
79	QOBState.1_55	.	X	X	X	.	.
80	QOBState.2_01	.	X	X	X	X	.
81	QOBState.2_02	.	X	X	X	X	X
82	QOBState.2_04	.	X	X	X	X	X
83	QOBState.2_05	X	X	X	X	X	X
84	QOBState.2_06	.	.	X	X	X	.
85	QOBState.2_08	.	X	.	.	.	.
86	QOBState.2_09	.	.	X	X	.	.
87	QOBState.2_10	.	.	X	.	.	.
88	QOBState.2_11	.	.	.	.	.	.
89	QOBState.2_12	.	X	X	.	.	.
90	QOBState.2_13	X	X	X	X	X	X
91	QOBState.2_15	.	X	X	.	.	.
92	QOBState.2_16	.	X	.	.	.	.
93	QOBState.2_17	.	.	X	X	X	X
94	QOBState.2_18	.	.	.	.	.	.
95	QOBState.2_19	.	X	X	X	.	.
96	QOBState.2_20	.	X	X	.	.	.
97	QOBState.2_21	X	X	X	X	X	X
98	QOBState.2_22	.	X	X	.	.	.
99	QOBState.2_23	.	.	X	.	.	.
100	QOBState.2_24	.	.	.	.	.	.
101	QOBState.2_25	.	X	.	.	.	.
102	QOBState.2_26	.	.	X	.	.	.
103	QOBState.2_27	.	X	X	.	.	.
104	QOBState.2_28	.	X	X	.	.	.
105	QOBState.2_29	.	.	X	X	.	.
106	QOBState.2_30	.	.	X	.	.	.
107	QOBState.2_31	.	.	.	.	.	.
108	QOBState.2_32	.	X	X	.	.	.
109	QOBState.2_33	.	.	X	.	.	.
110	QOBState.2_34	.	.	.	.	.	.
111	QOBState.2_35	.	X	.	.	.	.
112	QOBState.2_36	.	X	X	X	.	.
113	QOBState.2_37	.	X	X	X	X	X
114	QOBState.2_38	.	X	X	X	.	.
115	QOBState.2_39	.	X	X	X	.	.
116	QOBState.2_40	.	X	X	X	.	.
117	QOBState.2_41	.	.	.	.	.	.
118	QOBState.2_42	.	X	.	.	.	.
119	QOBState.2_44	.	X	X	X	X	.
120	QOBState.2_45	.	.	X	X	X	X
121	QOBState.2_46	.	X	.	.	.	.
122	QOBState.2_47	X	X	X	X	X	X
123	QOBState.2_48	X	X	X	X	X	X
124	QOBState.2_49	.	X	X	.	.	.
125	QOBState.2_50	.	X	.	.	.	.
126	QOBState.2_51	.	X	X	X	X	X
127	QOBState.2_53	.	X	.	.	.	.
128	QOBState.2_54	.	X	X	X	.	.
129	QOBState.2_55	.	.	X	X	X	X
130	QOBState.3_01	.	X	X	X	X	.
131	QOBState.3_02	.	.	X	X	.	.
132	QOBState.3_04	.	X	.	.	.	.
133	QOBState.3_05	.	X	X	.	.	.
134	QOBState.3_06	.	X	.	.	.	.
135	QOBState.3_08	.	.	.	.	.	.
136	QOBState.3_09	.	X	.	.	.	.
137	QOBState.3_10	.	.	X	.	.	.
138	QOBState.3_11	.	X	.	.	.	.
139	QOBState.3_12	.	.	.	.	.	.
140	QOBState.3_13	.	X	X	.	.	.
141	QOBState.3_15	.	.	X	X	X	X
142	QOBState.3_16	.	X	X	.	.	.
143	QOBState.3_17	.	.	X	X	.	.
144	QOBState.3_18	.	.	.	.	.	.
145	QOBState.3_19	.	X	X	X	.	.
146	QOBState.3_20	.	X	X	.	.	.
147	QOBState.3_21	.	.	X	X	.	.
148	QOBState.3_22	.	X	.	.	.	.
149	QOBState.3_23	.	.	X	.	.	.
150	QOBState.3_24	.	.	.	.	.	.
151	QOBState.3_25	.	.	.	.	.	.
152	QOBState.3_26	.	.	.	.	.	.
153	QOBState.3_27	.	X	X	.	.	.
154	QOBState.3_28	.	.	.	.	.	.
155	QOBState.3_29	.	.	X	.	.	.
156	QOBState.3_30	.	.	.	.	.	.
157	QOBState.3_31	.	.	.	.	.	.
158	QOBState.3_32	.	.	X	.	.	.
159	QOBState.3_33	.	.	.	.	.	.
160	QOBState.3_34	.	X	X	.	.	.
161	QOBState.3_35	.	X	X	.	.	.
162	QOBState.3_36	.	X	X	.	.	.
163	QOBState.3_37	.	.	X	X	.	.
164	QOBState.3_38	.	.	.	.	.	.
165	QOBState.3_39	.	.	.	.	.	.
166	QOBState.3_40	.	.	.	.	.	.
167	QOBState.3_41	.	X	.	.	.	.
168	QOBState.3_42	.	X	X	X	.	.
169	QOBState.3_44	.	X	.	.	.	.
170	QOBState.3_45	.	.	X	X	.	.
171	QOBState.3_46	.	X	.	.	.	.
172	QOBState.3_47	.	X	X	X	X	.
173	QOBState.3_48	.	.	.	.	.	.
174	QOBState.3_49	.	.	.	.	.	.
175	QOBState.3_50	.	.	X	.	.	.
176	QOBState.3_51	.	X	.	.	.	.
177	QOBState.3_53	.	X	.	.	.	.
178	QOBState.3_54	.	.	X	.	.	.
179	QOBYOB.38	.	X	.	.	.	.
180	Count	9	113	122	65	35	25

'QTR.q' indicates being born in quarter q. 'QOBYOB.qy' means being born in quarter q year y. 'QOBSTATE.q_ss' means being born quarter q in state ss. Marked 'x' means the instrument was chosen by the algorithm in the column. m_{cv} indicates m^* selected using cross validation while $m_{.25}$, $m_{.5}$, and $m_{.75}$ indicate m^* for different values of C denoted in the subscript. λ_{1se} indicates λ chosen by the one standard error stopping rule, and λ_{min} indicates the cross validated λ.