Using Machine Learning Methods to Address Selection on Unobservables

Rui Sun[*]

January 2020

Abstract

In this paper I present a new estimation strategy when strong prior information is unavailable regarding to the exogeneity of treatment variable. First of all, I use machine learning method to determine the right set of observed control variables without assuming the exact identities of these variables. Secondly, I estimate casual effects based on the idea that the amount of selection on observed explanatory variables in a model provides a guide to the amount of selection on unobservables. Lastly, I provide an empirical example that examines the effect of abortion on crime rates using the method proposed in this paper.

JEL classification: C52, C55

Keywords: causal inference, selection on unobservables, machine learning, model selection

[*]Department of Economics, University of Connecticut, 341 Mansfield Road, Unit 1063, Storrs, CT 06269-1063; email: rui.sun@uconn.edu

1 Introduction

The goal of many empirical analysis is to understand the causal effect of a treatment, such as participation in a training program or a government policy, on economic and other outcomes. Such analyses are often complicated by the fact that treatments or policies are rarely randomly assigned. The lack of true random assignment has led to the adoption of a variety of quasi-experimental approaches to estimating treatment effects that are based on observational data. Such approaches include instrumental variable methods in cases where treatment is not randomly assigned but there is some other external variable, such as eligibility for receipt of a government program or service, that is either randomly assigned or researcher is willing to take as exogenous conditional on a set of control variables. Another common approach is to assume treatment variable itself may be taken as exogenous after conditioning on a right set of controls which leads to regression for estimating treatment effects. Let's consider a model,

$$y_i = \alpha d_i + x_i'\beta + u_i \tag{1}$$

where d_i is the treatment variable and α is the parameter of interest. Some researchers are willing to take the endogenous treatment variable d_i as exogenous conditioning on a right set of control variables x_i. Alternatively, one may find a instrumental variable z_i which is correlated with endogenous variable d_i but uncorrelated with the unobserved error term u_i. In most of the cases, doubt remains about the validity of these identifying assumptions and, therefore, the inferences based on them. This is mainly because researchers can hardly verify these identifying assumptions in practice which becomes even worse when strong prior knowledge is not available. Another practical issue faced by empirical researchers is that how to determine the "right" set of control variables that should be included in the model. Typically, what most researchers do in practice is to use economic intuitions suggest a set of variables that might be important in determining the outcome, however, they will not identify which exact variables are important or the functional forms of these variables.

In this paper, instead of point identification, I estimate treatment effects under some weaker assumptions when doubt remains about the exogeneity of the treatment variable. In particular, I follow the idea proposed by Altonji, Elder and Taber (2005*b*) (hence, AET) that "the degree of selection on observables is like the degree of selection on unobservables" to address the issue of selection on unobservables. This method is first developed to address the issue of selection on unobserved variables in estimating the effect of attending catholic schools(see Altonji, Elder and Taber (2005*a*) and Altonji, Elder and Taber (2005*b*)). The main idea of this approach is that when one is able to control for a large set of observed explanatory variables, one can regard the set of observed variables as a random subset of all potential underlying variables including both observed and unobserved variables. With this randomness selection of observed control variables, it is relatively comfortable for us to draw the conclusion that the amount of selection on observed variables is somehow close to the amount of selection on unobserved variables. Therefore, one can use the amount of selection on observed explanatory variables to suggest a bound on the treatment effect of interest. I have to point out that this method is not a strategy that can substitute for point identification, but rather a generalization of results in bounds. Intuitively, if there is a lot of selection on the observables then one can expect the bounds to be wide, but the desirable case is that there is very little selection on observables and explanatory power of the observables is high, the bounds will be tight. If there is no selection on unobservables, meaning all the selection is through observed dimensions, the treatment variable can be taken as exogenous conditional on the right set of observed control variables. Whether or not there is selection on unobserved variables, determining the right set of control variables is one of the crucial steps one needs to consider in order to have a good internal validity. This is also part of the reason why we need to introduce machine learning method in this paper to help us determine the right set of control variables.

The machine learning method discussed in this paper refers to the modern model selection methods, in particular, such as Lasso. These methods are characterized by having many potential predictors or control variables of which only a few are actually important in predicting the outcome variable of interest. The goal of these model selection methods is to obtain a good out-of-sample forecast of the outcome variable without assuming the exact identities of these predictors or control variables. In this paper, I use model selection method to choose a right set of observed control variables that should enter the model and also take into account the fact that the goal of doing model selection in this paper has been shifted from prediction to causal inference. By doing model selection, the amount of selection extracted from observables can be more accurate and precise than without selection procedures which turns out helping us achieve a more reliable estimate than AET's original method. In particular, I adopt the double selection approach proposed by Belloni, Chernozhukov and Hansen (2014*a*) to perform the model selection procedure. The main attractive feature of this method is that it allows for imperfect model selection of control variables and provides confidence intervals that are valid uniformly across a large class of models. Double selection is named by the fact that the selection procedures are performed in two stages: in the first stage, selection is performed in the reduced form model where treatment variable is excluded from the selection equation and only potential control variables are being selected for the purpose of predicting the outcome variable of interest; in the second stage, selection is performed in the first stage equation where the same set of potential control variables is being selected for the purpose of predicting the treatment variable. The union of these two sets of selected control variables is the final “right” set of control variables.

This paper relates to two categories of literatures in general. First of all, this paper adds new perspectives to the partial identification and bound identification literatures. My paper mainly adopts the method that has been used in a series of paper by Altonji, Elder, and Taber including Altonji, Elder and Taber (2002), Altonji, Elder and Taber (2005*a*), andAltonji, Elder and Taber (2005*b*) in which they study the effect of attending catholic schools by varying identification strategies. In Altonji et al. (2013), they further provide the theoretical foundation for using selection on observables to address the issue of selection on observables. In addition to that, they also propose another estimator that using factor loading model to address the same issue of selection on unobservables. To address the issue of omitted variable and selection bias, Rosenbaum and Rubin (1983) and Rosenbaum (1995) propose an estimation strategy by assessing the sensitivity of treatment effect to varying the amount of selection on unobserved variables, which is closely related to AET's method. There are a lot of other studies formally argue the exogeneity of an explanatory variable by examining the sensitivity of point estimates by including additional control variables or by assessing the relationship between the variable of interest and a set of observed characteristics (see, e.g., Grogger, Bronars and Grogger (1994), Currie and Thomas (1998), Engen, Gale and Scholz (1996), Udry (1996), Angrist and Evans (1998), Angrist and B. Krueger (1999), Jacobsen, Pearce and Rosenbloom (1999), or (2004)). The second area my paper can be related to is the application of model selection in social science study when goal is causal inference. There is a rapidly rising emphasis in using machine learning method to address the high dimensional problem. Among all the literatures, I relate my paper to Belloni et al. (2012), Belloni and Chernozhukov (2013), Belloni, Chernozhukov and Hansen (2014*a*), Belloni, Chernozhukov and Hansen (2014*b*), and Belloni, Chernozhukov and Hansen (2014*c*) where they combine machine learning method with causal inference and use Lasso estimator to perform model selections.

The remainder of this paper is organized as follows. Section 2 defines the econometric model for selection on observables and unobservables. Section 3 formally discusses the double selection procedures. Section 4 provides the estimation strategy for estimating the treatment effect as a bound. Section 5 applies this proposed method to the an empirical application provided by Donohue III and Levitt (2001). Section 6 concludes.

2 Link Between the Observed and Unobserved Determinants of the Treatment and Outcome

In this section, I formally build the connection between selection on observables and selection on unobservables. This is the first step in developing theoretical foundation for using the relationship between endogenous variable and observed variables to make inferences about the relationship between the endogenous variable and unobserved variables. I first discuss what does it mean by "selection on observables is like the selection on unobservables" and explain under what conditions this statement is valid. Then, I provide the main theoretical results for this estimation strategy. This method is proposed by Altonji, Elder and Taber (2005*a*) and my contribution is mainly adding some refinements to this estimation strategy. Particularly, I bring model selection techniques into the framework. More details about the model selection techniques will be discussed in Section 3.

2.1 Selection on Observables is like Selection on Unobservables

To better understand the main idea of "the selection on observables is like the selection on unobservables", let us first consider the model in which some outcome variable y_i is determined by

$$y_i = \alpha d_i + z_i' \beta_z + x_i^{*'} \beta^* + \xi_i \tag{2}$$

where d_i is the treatment variable and α is the causal effect of d_i on y_i. z_i is a set of variables that has relatively important roles in determining the outcome variable y_i. They are assumed to be available in the dataset. For example, one can consider z_i as gender or schooling in a wage regression. x_i^* is a vector of additional characteristics that are relevant in determining outcome of interest which may or may not be observed in the dataset. Since the main concern of this paper is the observed and unobserved components in x_i^* and z_i plays no important role in the analysis, for simplicity, I residualize all the variables by regressing each variable on z_i and taking residuals to remove z_i from the analysis of this paper. Hereafter, all variables in this article are defined as residuals from regression of that variable on z_i. ξ_i represents the idiosyncratic shock that is unrelated to all the components in the model. After residualization, one can rewrite the model as[1]

$$y_i = \alpha d_i + x_i^{*'} \beta^* + \xi_i \tag{3}$$

where all the variables are residuals from partialling out z_i. As we have discussed above, x_i^* may contain observed and unobserved components. Let's further decompose $x_i^{*'} \beta^*$ into two parts, i.e. observed and unobserved parts, indexed by X and X_u, respectively. Specifically, $X = x_i' \beta$ and $X_u = x_i^{u'} \beta^u$ where x_i and x_i^u represent observed variables and unobserved variables, respectively. To fix the idea that "the selection on observables is like the selection on unobservables", one can consider a linear projection of the treatment variable d_i onto the observed and unobserved determinants of outcome variable under the assumption that idiosyncratic shock $\xi_i = 0$

$$Proj(d_i | X, X_u) = \phi_0 + \phi X + \phi_u X_u. \tag{4}$$

When the idiosyncratic shock $\xi_i = 0$, there are two possible scenarios that associated with selection on unobserved variables:

Scenario 1.

$$\phi_u = 0 \tag{5}$$

Scenario 1 is the simplest case where there is no selection on unobservables, in other words, all the selection is through the observed dimensions. This case is actually what most researchers assume in the empirical studies when lack of randomized experiment. After controlling for a set of observed variables, they are willing to take the treatment variable as exogenous. However, in many cases this assumption is unlikely to be valid. More importantly, one can barely verify this assumption in practice which often raises doubt about the identifying assumptions. With this doubt in mind, we can consider a second scenario, which takes into account the role of unobserved variables,

Scenario 2.

$$\phi = \phi_u \tag{6}$$

The idea "selection on observables is like the selection on unobservables" implies the partial correlations between the treatment variable and the observed and unobserved components of outcome are the same, in other words, $\phi = \phi_u$. This equality indicates that the

[1] For simplicity, we do not distinguish the notations before and after residualizations.

selection on observables is exactly the same as the selection on unobservables. Notice that I haven't presented any conditions under which this argument is true. I just present a possible implication of equal selection on observables and unobservables. Therefore, this scenario is not true in general. Remember this result is obtained under the restriction that there is no disturbance term, which is essentially not the case in practice. Therefore, a more realistic scenario is that the idiosyncratic shock is nonzero and the composite term $(X_u+\xi_i)$ is all that one can approximate for the unobserved components. Considering about the compositeness of the unobserved term, the analog of equation (4) can be written as[2]

$$Proj(d_i|X,(X_u+\xi_i)) = \phi_0 + \phi X + \phi_u(X_u+\xi_i) \tag{7}$$

With (7) defined above, the idea of same amount of selection on observed and unobserved components becomes a slightly different from 4. Same amount of selection suggests a equal partial correlation of d_i with X and X_u, however, in (7) a equal partial correlation implies an inequality of ϕ and ϕ_u. This is due to the attenuation bias in the latter coefficient and therefore leads to the third scenario,

Scenario 3.

$$|\phi| \geq |\phi_u| \tag{8}$$

So far, I have formalized the idea that "the selection on observables is like the selection on unobservables". Yet, I haven't provided any justification or conditions under which this statement is valid. Next, I will discuss under what conditions one can apply this idea to the analysis in practice and the implications in the context of estimation strategy.

2.2 Random Selection of Observables

Whether or not one can consider the selection on observables is like the selection on unobservables is largely determined by how the observed variables are selected from the full set of underlying variables that determine the outcome of interest. To begin with, I provide some discussions about under what situations one can consider the selection on observables is like the selection on unobservables. Generally speaking, in order to consider the amount of selection is close to the amount of selection on unobservables, one needs to have the number of observed variables to be relatively large and, in addition to that, to maintain high explanatory powers. Most of the datasets are created to address multiple issues rather than a specific research question. Dataset contents are compromises among the interests of multiple researches, budget, burden on respondents and so on. Some factors that affect a large set of outcomes are more likely to be collected and of course some variables are left out. In most of the cases, when the number of observed variables is limited, one may not be able to consider the selection on observed variables is like the selection on unobserved variables. This is because, which I will discuss in details in the following sections, this idea requires researcher to consider the set of observed variables as a random subset of the full underlying

[2] For a slightly abuse of the notation we continue to use the same notions as in (4) to denote for the coefficients.

set of variables that determines the outcome of interest. Obviously, when the number of observed variables is limited, it is more likely that the set of observables is suffering from extreme selection. One can infer very little information from the selection relationships between observables and treatment variable. In other words, if one is able to recover the full set of observed variables that determines the outcome of interest, the amount of explained variation by observables would suggest a reasonable amount of variations could have been explained by the unobservables. Intuitively, this assumption requires the set of observed variables to be as large as possible but also remain relatively high explanatory power, which means the best scenario is to recover all the observe variables that are truly relevant to the outcome variable. This is also the main reason why I artificially create a relatively large set of variables, where the number of variables is possibly larger than the number of observations, using the existing observed variables and then use machine learning method to select the high explanatory power variables. I discuss the procedure of the model selection in detail in Section 2. To formalize the idea of random selection, I consider the model[3]

$$y_i = \alpha d_i + x_i^{*'} \beta^* + \xi_i \tag{9}$$

where x_i^* contains all the possible explanatory variables, i.e. both observed variables x_i and unobserved variables x_i^u. Now, one can rewrite the model as

$$y_i = \alpha d_i + x_i' \beta + x_i^{u'} \beta^u + \xi_i \tag{10}$$

In order to model the process of random selection of observed variables, I introduce a random variable b_j which indicates whether covariate j is observed in the dataset. More specifically, if covariate j is observed in the data set, $b_j = 1$, and if is not, $b_j = 0$. I assume b_j as an *iid* binary random variable which equals to one with probability p_b for all covariates in x_i^*. As I model b_j as an *iid* random variable, the process of selecting observed variables actually can be considered as a random process. With this random selection assumption, one can further treat the set of observed covariates as a random subset of the full set of underlying variables. This assumption is critical to the main analysis of this paper, in the sense that without this treatment one is not able to draw the key conclusion that "the selection on observables is like the selection on unobservables". In other words, an implication of random selection of observables is that one can treat observed variables and unobserved variables as symmetric in the model, therefore the amount of selection on observables is like the amount selection on unobservables.

Following the modeling strategy in Joseph G. Altonji, Timothy Conley, Todd E. Elder and Christopher R. Taber (2013), I define outcome variable is determined by a sequence of models indexed by K^*, where K^* is the number of elements in x_i^*. To understand the logic behind this way of modeling, let's first note that the assumption I make, i.e. "selection on observables is like the selection on unobservables", requires us to have a relatively large set of observables so that the set of observables could be treated as somehow a random subset of the full set of underlying variables. Additionally, the asymptotic analysis also requires the number of variables, both observed and unobserved, goes to infinity. Secondly, the number of

[3] All variables are residuals from regression on z_i.

covariates varies across models can be explained by the idea that as one adds more and more covariates into the model, the importance of each variable would decline. Before I discuss the data generating process for this sequence of models framework, let me first introduce the information set $\mathcal{G}^{K^*}$ as the realization of the indicator b_j, true coefficients β_j, and the joint distribution of x_{ij} conditional on $j = 1, \cdots, K^*$. With this information set $\mathcal{G}^{K^*}$ well defined, the data generating process can be expressed as a two stage process,

1. In the first stage, for a given number of covariates K^*, the model of interest is drawn, which means the joint distribution of x_{ij}, d_i, ξ_i, and b_j are drawn.

2. In the second stage, individual data is then drawn form these underlying distributions.

This two stage data generating process can be summarized as the following assumption,

Assumption 1

$$y_i = \alpha d_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} x_{ij}\beta_j + \xi_i \tag{11}$$

where (x_{ij}, β_j) is unconditionally stationary.

In this set up, x_{ij} and β_j are different among different models, in other words, their values depend on the number of the potential variables K^*. Different number of K^* stands for different models. Putting a scaling factor $\frac{1}{\sqrt{K^*}}$ in the model is aim to capture the idea that a large number of factors are important in determining outcomes in most of the social science studies and as the number of factors increases the importance of each variable will decline. Next, I borrow the assumptions from Altonji et al. (2013), which are necessary to get the main asymptotic results. Let $\sigma_{j,l}^{K^*} = E(x_{ij}x_{il}|\mathcal{G}^{K^*})$. To guarantee that $var(y_i)$ is bounded as K^* becomes large, we assume that

Assumption 2 $0 < \lim\limits_{K^* \to \infty} \frac{1}{K^*} \sum_{j=1}^{K^*} \sum_{l=1}^{K^*} E(\sigma_{j,l}^{K^*} \beta_j \beta_l) < \infty;$
$\lim\limits_{K^* \to \infty} Var\left(\frac{1}{K^*} \sum_{j=1}^{K^*} \sum_{l=1}^{K^*} \sigma_{j,l}^{K^*} \beta_j \beta_l\right) \to 0$

Next assumption guarantees that $cov(d_i, y_i)$ is well behaved as K^* grows,

Assumption 3 *For any $j = 1, ..., K^*$, define $\mu_j^{K^*}$ so that $E\left(d_i x_{ij}|\mathcal{G}^K\right) = \frac{\mu_j^{K^*}}{\sqrt{K^*}}$. We assume that $E(\mu_j^{K^*}\beta_j) < \infty$; $\lim\limits_{K^* \to \infty} Var\left(\frac{1}{K^*} \sum_{j=1}^{K^*} \mu_j^{K^*} \beta_j\right) \to 0$.*

I also provide the assumption about the random process under which observables are chosen,

Assumption 4 *For $j = 1, ..., K^*, b_j$ is independent and identically distributed with $0 < \Pr(b_j = 1) \equiv p_b \leq 1$. b_j is also independent of all other random variable in the model.*

Assumption 5 *ξ_i is mean zero and uncorrelated with d_i and x_{ij}.*

As noted in Altonji et al. (2013), the asymptotic analysis is non standard in the sense that it allows the number of underlying covariates K^* increase and the randomness of x_{ij} is different from that of β_j and b_j. For any single covariate indexed by j, the observations of β_j and b_j are the same across all the individuals in the population. However, each individual i has their own values of x_{ij}. Now, with the model defined in Assumption (1) one can define ϕ and ϕ_u such that

$$\begin{aligned} Proj & \left(d_i | \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} b_j x_{ij} \beta_j, \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - b_j) x_{ij} \beta_j + \xi_i ; \mathcal{G}^K \right) \\ & = \phi \left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} b_j x_{ij} \beta_j \right) + \phi_u \left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - b_j) x_{ij} \beta_j + \xi_i \right) \end{aligned}$$

and leads us to the Theorem 1,

Theorem 1 [4]*Under Assumption (1)–(5), if the probability limit of ϕ is nonzero, then*

$$\frac{\phi_u}{\phi} \xrightarrow[K^* \to \infty]{p} \frac{(1 - p_b) A}{(1 - p_b) A + \sigma_\xi^2}$$

where $A \equiv \lim_{K^* \to \infty} E\left(\frac{1}{K^*} \sum_{j=1}^{K^*} \sigma_{j,j}^{K^*} \beta_j^2 \right)$ If the probability limit of ϕ is zero, then the probability limit of ϕ_u is also zero.

This is the main result of AET's approach, but a few things need to be noted. First of all, this theorem requires us to observe the true values of β_j and b_j. Secondly, the assumptions for this theorem also require a large set of observables to account for the random selection of observables. As I mentioned in the introduction, I introduce machine learning techniques into this framework to make the assumptions more plausible.

3 Model Selection

So far I have built a framework within which one can use the idea of "the selection on observables is like the selection on unobservables" to have a better understanding of the treatment effect. A natural question that one might ask is why we bother to bring machine learning techniques into this framework? In this section, I explain the importance of bringing model selection procedures into AET's framework. Then, I also discuss how to apply machine learning techniques in social science studies when the ultimate goal is causal inference.

3.1 Why We Need Model Selection?

In order to properly apply AET's idea of "the selection on observables is like the selection on unobservables", an important assumption one needs to make is that the set of observed

[4] The proof of Theorem 1 can be found in Appendix A of Altonji et al. (2013)

variables is relatively large so that the process of selecting variables through the observed dimensions is somehow random. As a consequence of this random selection of observed variables, one can further conclude that the amount of selection on unobserved variables can be approximated by the amount of selection on observed variables. In practice, there are several issues associated with this argument. First of all, for most of the dataset, it does not contain a relative large number of variables. [5] Secondly, even though when the set of observed variables is large, one actually has no clue either which exact variables should be included in our model or the functional forms of the these variables. In practice, most of the researchers use economic intuitions or reply on the previous literatures to justify a set of variables that should be controlled for. Then, they are willing to take the treatment variable as exogenous conditional on this set of control variables. The problem of this approach is that economic intuition is not always reliable and no one can really verify whether or not the economic intuition suggested variables are important in determining the outcome variable.

Fortunately, machine learning method can help us address these two issues at the same time. When there are not enough observed variables, one can follow the strategy developed by Belloni, Chernozhukov and Hansen (2014*a*) to construct potentially large number of technical regressors through taking transformations of existing observed variables. Note that we do not require every artificially constructed variable makes intuitive sense. They only serve as the purpose of technical controls. Through this approach, one can recover the potential distribution of observed variables to the largest extend. However, it also creates another issue to this method, that is some regressors constructed by doing transformations might have relatively low correlations with the outcome variable. We then have to rely on some other approaches to help us determine which variables are actually important in determining the outcome variable. This is place where machine learning method comes in to play an role. In the next section, I first discuss the procedure of using machine learning method to perform model selection and then I use an empirical example to illustrate a possible application of this strategy in practice.

3.2 Lasso Estimator and Naive Approach

Before we turn into the so called "double selection" procedure, let me first introduce the Lasso estimator and discuss some drawbacks of the simple one stage model selection procedure. The machine learning method I have been referring to in this paper is actually the Lasso estimator, i.e. Absolute Shrinkage and Selection Operator Estimator. To begin with the discussion of model selection procedures, let me introduce some features of Lasso estimator. The well known Lasso estimator was designed for estimating the parameters of sparse high dimensional linear model, introduced by Frank and Friedman (1993) and Tibshirani (1996).

[5] Of course, all the variables containing in the data set are indeed observed variables. The unobserved variables I refer to in this paper are either not available to econometricians or unable to be measured by researchers.

The estimator is given by

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} b_j \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \tag{12}$$

where the coefficients are chosen to minimize the sum of squared residuals plus a penalty term that penalizes the size of the model through the sum of absolute values of the coefficients. λ is the penalty level that controls the degree of penalization. To avoid the issue of overfitting, the practical choice of λ has been discussed in Belloni et al. (2012). It is also common to choose λ by using cross-validation when the goal is prediction, whereas if the end goal of doing Lasso is not prediction, cross-validation may not be the best way of choosing the penalty level. The intuition behind the model selection feature of Lasso is that the penalty function in Lasso is special in the sense that it has a kink at 0. If one takes a look at the Lasso estimator as in (12), what Lasso does is to choose the values of coefficients to minimize the whole objective function consisting with two parts: the first part is the sum of squared residuals and the second part is the sum of absolute values of coefficients. Because of the property of the penalty function, which is always non-negative, Lasso estimator reports us with a sparse estimator with many coefficients set exactly to zero. Therefore, Lasso estimator can be used as a tool for model selection. A large part of the appealing feature of Lasso estimator when compared with other model selection methods is that Lasso problem is indeed a convex optimization problem and there exists highly computational efficient algorithms to solve it. Not only exists highly efficient algorithms to Lasso estimator, but also it has been shown that Lasso estimator can allow for approximation errors, heteroskedasticity, clustering, fixed effects, and non-normality (see Bickel, Ritov and Tsybakov (2009), Belloni et al. (2012)).

Lastly, it is important to be aware that model selection techniques are originally developed for the purpose of predictions. Those techniques can estimate which variables have strong associations with an outcome variable in a sparse model framework. However, naively using Lasso estimators to draw casual inference could be problematic. Part of the reason is that the results of Lasso estimator tend to be substantially biased toward zero. One possible solution to this problem is to employ the idea of post Lasso estimator. The post Lasso estimator works in two stages: in the first stage, one can apply Lasso to determine which variables should be excluded from the standpoint of prediction; in the second stage, one can estimate the coefficients of remaining variables by running OLS. This post Lasso estimator has been shown works better than Lasso estimator in terms of convergence rate and bias. However, the potential issue of post Lasso based estimator is that it relies on perfect model selection which rarely holds true in practice. If one can be sure about the model selection is perfect in the sense that Lasso always chooses variables with exactly nonzero coefficients, and then any conventional estimation and inference procedures can be applied to those selected variables. As I mentioned, perfect model selection is highly unrealistic, and a more realistic scenario is that model selection tends to drop some variables that are indeed correlated with the outcome variable but might have very small nonzero coefficients. Once one allows for model selection mistakes, the estimations and inference based on the mistakenly selected variables would also be incorrect. In other words, the validity of post Lasso estimators is highly associated with the perfectness of model selection and because Lasso is originally designed to target on prediction, not learn about the values of specific parameters, perfect model selection in the context of causal inference is not guaranteed.

Noting these features of Lasso, the problem of post Lasso estimator in the context of treatment effect estimation can be easily understood: any variable that is highly correlated with the treatment variable is very likely to be dropped out of the model since including such variable will not add too much predictive power of the model given the treatment variable is already in the model. However, missing such variables will lead to the model suffer from substantial omitted variable bias. One natural extension of this post Lasso estimator in the context of treatment effect estimation is that one can apply model selection procedures by forcing treatment variable remain in the model but dropping the coefficient for treatment effect, i.e. α, from the penalty function. Specifically, if one decomposes the objective function of Lasso estimator (12) into two parts: sum of squared residuals and a penalty function, it can be easily noted that, in this extension, we force the treatment variable remain in the first part of the objective function which is the sum of squared residuals part and exclude the treatment effect which is the coefficient of the treatment variable from the penalty function. In such a way, one is able to ensure that the model selection procedure is only being performed among the potential control variables, and avoid the case when some candidates of control variable that are highly correlated with treatment variable but are important in determining outcome variable are dropped out by the model selection procedure. After the model selection stage one can estimate and do inference about the treatment effect by any conventional approach[6]. As it has been shown in Belloni, Chernozhukov and Hansen (2014*c*), this approach still does not perform well in terms of the robustness to model selection mistakes. Therefore, we need a more sophisticated and robust model selection based estimator to address the issue especially when the ultimate goal is causal inference.

3.3 Model Selection When the Goal is Causal Inference

How to effectively combine the modern machine learning techniques with causal inference has become a challenge to the researchers in social science studies. Notice that simply using Lasso as a procedure to estimate the coefficients can be useful when the ultimate goal is obtaining forecasting rules and estimating which variables have strong association with an outcome variable in the context of a sparse model. But, for most of the social science studies, the ultimate goals are normally to know the values of some parameters and then draw inferences based on those parameters. As I have mentioned, model selection mistakes seem inevitable, therefore naively using Lasso procedure to draw causal inference could be problematic as well. One of the major but straightforward reasons is that Lasso is designed for prediction and the estimations of parameters might suffer from omitted variable bias. Even though one can apply the idea of post Lasso estimator as I mentioned in Section (3.2) to do inference, the probability of having omitted variable bias is still relatively high(see Belloni, Chernozhukov and Hansen (2014*c*) for details). Observing the fact that Lasso tends to focus on prediction rather than obtaining inference, a more desirable procedure could be

[6] In this paper, I estimate bounds for the treatment effect.

obtained if one focuses on model selection over the predictive part of an economic model rather than uses model selection techniques in structural model directly. Normally, in the context of treatment effect estimation, the predictive parts of the model are the reduced form and first-stage. In the spirit of model selection, Lasso can be applied to these two parts to select the relevant determinants of treatment variable and outcome variable. This idea was first proposed up by Alexandre Belloni, Victor Chernozhukov and Christian Hansen (2014*b*) and they later provide the theoretical results in Alexandre Belloni, Victor Chernozhukov and Christian Hansen (2014*c*). The usual post model selection methods rely on a single selection step, whereas this approach uses two different model selection steps followed by a final estimation. The procedure can be summarized into three steps as follows:

1. In the first step, a set of control variables that are useful for predicting the treatment d_i is selected. This step helps to ensure validity of post model selection inference by finding control variables that are strongly related to the treatment and thus potentially important confounding factors.

2. In the second step, additional control variables that are important in predicting outcome y_i are selected. This step helps to ensure that the model selection procedure has captured important elements in the equation of interest, providing an additional chance to find important confounds.

3. In the final step, the treatment effect α of interest is obtained by the linear regression of y_i on the treatment d_i and the union of the set of variables selected in the two model selection steps.

The main attractive feature of this method is that it allows for imperfect selection. Alexandre Belloni, Victor Chernozhukov and Christian Hansen (2014*c*) shows that this method provides confidence intervals that are valid uniformly across a large class of models. In contrast, standard post model selection estimators fail to provide uniform inference even in simple cases with a relatively small and fixed number of controls. Now, I will begin to review this method and discuss how it can be applied to AET's framework in practice. In a partially linear model framework

$$y_i = \alpha d_i + g(w_i) + g^u(w_i^u) + \xi_i \tag{13}$$

where d_i is the treatment variable of interest, α is the treatment effect, w_i and w_i^u are sets of observed and unobserved control variables respectively, and ξ_i is the idiosyncratic shock that satisfies $E[\xi_i|w_i, w_i^u, d_i] = 0$[7]. In order to illustrate the model selection procedure, I assume that there is no effect of unobserved variables on outcome, which means the treatment variable can be taken as exogenous after controlling for a right set of observed variables. This leads us to the model selection procedure whose main purpose is to determine what exactly should be the right set of control variables. With this simplification, the original model can be rewritten as follows

$$y_i = \alpha d_i + g(w_i) + \xi_i \tag{14}$$

[7] This set up is analogous to the set up in Section2 where I specify the functional forms as linear combinations of control variables.

The ultimate goal of the econometric analysis is to conduct inference on the treatment effect α. Particularly, what model selection approach does is to select a set of variables from p potential regressors $x_i = P(w_i)$, which may consist of the original observables w_i and transformations of w_i, to properly approximate the unknown functional $g(w_i)$. At the same time we allow the dimension of potential regressors p to be greater than the number of observations n. When the number of potential regressors is greater than the number of observations, one apparently cannot conduct inference of treatment effect α in this model by using any conventional econometrics techniques. We need further impose some structural restrictions on this model. One crucial assumption we make here is that the model selection framework is approximate sparse. Simply speaking, with this sparsity assumption one can regard the treatment variable d_i as exogenous once controls for a relatively small number $s < n$ of variables in x_i whose identities might be unknown. This is the most important assumption in model selection frameworks, which also implies a linear combination of these s regressors provide approximation to $g(w_i)$. The details of this sparsity assumption are discussed in Section 3.3.2. The key idea of double selection is to perform another model selection for first stage equation, therefore one needs to specify another selection procedure for the first stage.

3.3.1 Framework

Again, we consider a partially linear model and assume there are no unobservables

$$y_i = \alpha d_i + g(w_i) + \xi_i, \quad E[\xi_i|w_i, d_i] = 0, \tag{15}$$

$$d_i = m(w_i) + \nu_i, \qquad E[\nu_i|w_i] = 0, \tag{16}$$

where all the variables are defined as before. The confounding factors w_i affect treatment variable d_i through function $m(\cdot)$ and affect outcome variable through function $g(\cdot)$. Both $g(\cdot)$ and $m(\cdot)$ are unknowns functions and potentially complicated. I use linear combinations of control terms $x_i = P(w_i)$ to approximate $g(w_i)$ and $m(w_i)$, rewriting equation 15 and 16 as

$$y_i = \alpha d_i + \underbrace{x_i'\beta_g + r_{gi}}_{g(w_i)} + \xi_i, \tag{17}$$

$$d_i = \underbrace{x_i'\beta_m + r_{mi}}_{m(w_i)} + \nu_i, \tag{18}$$

where $x_i'\beta_g$ and $x_i'\beta_m$ are the approximations to $g(w_i)$ and $m(w_i)$, and r_{gi} and r_{mi} are the corresponding approximation errors. Note that p, the dimension of potential control variables x_i, could be larger than n, the number of observations. The identity of x_i is also potentially complicated, which could be w_i itself or some technical transformations of elementary regressors w_i such as B-splines, dummies, polynomials, and various interactions(see, e.g., Newey (1997), Chen (2007), Chen and Pouzo (2009), or Chen and Pouzo (2012)). Having too many controls creates challenge to econometricians to make estimation and inference on treatment effect. It requires further restrictions on the structure of the model especially on control variables in order to make reliable estimation and inference. Such restriction is known as sparsity conditions.

3.3.2 Sparsity Conditions

Intuitively speaking, sparsity conditions ensure the existence of $x_i'\beta_g$ and $x_i'\beta_m$ in (17) and (18) such that approximation errors r_{gi} and r_{mi} are relatively small compared to estimation errors. Note that $x_i'\beta_g$ and $x_i\beta_m$ are approximations to $g(w_i)$ and $m(w_i)$, sparsity conditions require that there exist a small number of nonzero coefficients to render the approximation errors r_{gi} and r_{mi} to be small relative to estimation errors. More formally, sparsity conditions require

Condition 1 [8]

$$\|\beta_m\|_0 \leq s \textit{ and } \|\beta_g\|_0 \leq s$$

where $s \ll n$.

Condition 1 tells us there exist β_m and β_g such that at most $s \ll n$ elements of β_m and β_g are nonzeros.

Condition 2 [9]

$$\{\bar{E}[r_{gi}^2]\}^{1/2} \lesssim \sqrt{s/n} \textit{ and } \{\bar{E}[r_{mi}^2]\}^{1/2} \lesssim \sqrt{s/n}$$

Condition 2 requires the size of the resulting approximation errors to be small relative to the conjectured size of the estimator error. These two conditions are critical to the high dimensional sparse models in the sense that they allow us to extend the standard treatment effect framework, which assumes the identities of the control variables are known and the number of control variables is much smaller than the number of observations, to a high dimensional model selection framework. Within this new framework, we assume, instead of the identities of control variables, there are potentially large number of control variables available to us and we do not know the exact identities of these relevant control variables however at most s controls variables suffice to achieve a desirable approximations to the unknown functions $g(\cdot)$ and $m(\cdot)$ specified in equation 15 and equation 16. Replying on these sparsity assumptions, we use selection method, i.e. Lasso, to help determine the approximately right set of control variables.

[8] $\|\cdot\|_0$ denotes the l_0-norm: the number of non-zero elements of a vector.

[9] $\bar{E}[f]$ denotes the average expectation operator: $\bar{E}[f] = \sum_{i=1}^{n} E[f(\omega_i)]/n$.

3.3.3 Least Squares After Double Selection

To fix the idea of double selection, I first write the corresponding reduced forms of (15) and (16)

$$y_i = x_i'\bar{\beta} + \bar{r}_i + \bar{\xi}_i, \tag{19}$$

$$d_i = x_i'\beta_m + r_{mi} + \nu_i, \tag{20}$$

where $\bar{\beta} = \alpha\beta_m + \beta_g$, $\bar{r}_i = \alpha r_{mi} + r_{gi}$, $\bar{\xi}_i = \alpha\nu_i + \xi_i$.[10] One can apply model selection methods to these two equations to select a right set of control variables for each of them and take union of these two selected sets afterwards. This final set is therefore considered as the selected control variables for structural model. Given the set of selected variables, one can estimate the treatment effect by a least squares regression of y_i on d_i and the selected set of control variables. Inference can be performed by conventional inference methods for parameters estimated by least squares. One advantage of this post double selection estimator is that it does not reply on the highly unrealistic perfect model selection assumption, which is often the necessary assumption for performing inference after estimation in a model selection framework. Intuitively speaking, this approach is more desirable in the sense that it increases the probability of recovering all relevant control variables by selection variables in two equations rather than selecting only through one of (19) and (20). Belloni, Chernozhukov and Hansen (2014*c*) shows that in finite sample experiments, single selection methods essentially fail, providing poor inference compared to the double selection methods outlined above. This argument is also theoretically supported by the fact that double selection method requires weaker regularity conditions for its validity and for attaining the semi-parametric efficiency bound than single selection method. The regularity conditions and theory and the main result of post double selection estimator are omitted in this paper(see Belloni, Chernozhukov and Hansen (2014*c*) for details).

The idea of post selection estimators is using the selected variables to perform estimations. In fact, the estimator for AET's method is also considered as another post selection estimator, which is presented in Section 4. Here I present another post selection estimator, i.e. post double selection estimator(Belloni, Chernozhukov and Hansen (2014*c*)), in the context of treatment effect estimation. Let $\hat{I}_1$ and $\hat{I}_2$ denote the controls selected using data $(\widetilde{y}_i, \widetilde{x}_i) = (d_i, x_i)$ and $(\widetilde{y}_i, \widetilde{x}_i) = (y_i, x_i)$, $i = 1, ..., n$, respectively. The post double selection estimator $\check{\alpha}$ of α is defined as the least squares estimator obtained by regression y_i on d_i and the selected control x_{ij} with $j \in \hat{I} \supseteq \hat{I}_1 \cup \hat{I}_2$,

$$(\check{\alpha}, \check{\beta}) = \underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \left\{ \mathbb{E}_n[(y_i - \alpha d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \hat{I} \right\}. \tag{21}$$

Note that if one has strong belief on the conditional exogeneity of treatment variable, meaning if one can justify that treatment variable can be taken as exogenous once control for a set of observed variables, this post double selection estimator can be directly applied. However, when doubt remains about the exogeneity of treatment variable, the estimator proposed in Section 4 of this paper is preferred.

[10] To get equation (19), plug (18) into (17).

3.3.4 Feasible Lasso Estimator

Now I present the Feasible Lasso Estimator that is geared for heteroscedastic and non-Gaussian cases. Unlike the post double selection estimator (21), any Lasso estimator cannot be used to do inference directly because it only serves as a model selection procedure. Note that each of the equation (19) and (20) is of the form

$$\widetilde{y}_i = \underbrace{\widetilde{x}_i'\beta + r_i}_{f(\widetilde{w}_i)} + \epsilon_i \tag{22}$$

where $f(\widetilde{w}_i)$ is the regression function, $\widetilde{x}_i'\beta$ is the approximation based on the dictionary $\widetilde{x}_i = P(\widetilde{w}_i)$, r_i is the approximation error, and ϵ_i is the error. Given this set up, a Feasible Lasso Estimator solves

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(\widetilde{y}_i - \widetilde{x}_i'\beta)^2] + \frac{\lambda}{n}\|\hat{\Psi}\beta\|_1 \tag{23}$$

where $\hat{\Psi}$ =diag$(\hat{l}_1, ..., \hat{l}_p)$ is the diagonal matrix of penalty loadings and $\|\hat{\Psi}\beta\|_1 = \sum_{j=1}^p |\hat{l}_j\beta_j|$. The penalty level λ and loadings $\hat{l}_j$'s are set as

$$\lambda = 2 \cdot c\sqrt{n}\Phi^{-1}\frac{1-\gamma}{2p} \text{ and } \hat{l}_j = l_j + op(1), l_j = \sqrt{\mathbb{E}_n[\widetilde{x}_{ij}^2\epsilon_{ij}^2]}, \text{uniformly in } j = 1, ..., p, \tag{24}$$

where $c > 1$, $1 - \gamma$ is a confidence level, and γ is set such that[11]

$$\gamma = o(1) \text{ and } \log(1/\gamma) \lesssim \log(\max\{p, n\}).$$

The term Feasible Lasso Estimator refers to a Lasso estimator that solves for (23) and (24). To sum up, in this paper the "right set" of control variables in treatment effect estimation framework is obtained by applying Feasible Lasso Estimator to (19) and (20) separately. Then, apply any post Lasso estimator using the union of these two sets of selected variables as controls.

4 Estimator of Treatment Effect

In Section 3, I describe the model selection technique when goal is causal inference. I assume there are no unobservable variables so that the treatment variable can be taken as exogenous once we control for the right set of observed variables. However, this high level assumption is unlikely to hold true in many situations. Therefore, we need to bring back the impact of unobservables back to the framework and re-apply AET's idea of "the selection on unobservables is like the selection on observables." Let's go back to the basic model (10) in Section 2, where there is also a set of unobserved variables that affect the outcome of interest ,and rewrite (10) as

$$\begin{aligned} y_i &= \alpha d_i + x_i'\beta + x_i^{u'}\beta^u + \xi_i \\ &= \alpha d_i + x_i'\beta + u_i, \end{aligned} \tag{25}$$

[11] The practical recommendations of c and γ are given in Belloni, Chernozhukov and Hansen (2014c).

Even though we select the right set of observables x_i via double selection procedure for the model we specified, one can not directly apply the post double selection estimators when the effect of unobserved variables is non-zero[12]. I bring in the idea of AET's method, using the degree of selection on observables to guide us the degree of selection on unobservables, to the model selection framework. The contribution of bringing model selection into AET's method can be understood intuitively by noticing the fact that through recovering the observed variables as many as possible, the degree of selection on observables is much more informative than simply using a relatively small and limited set of observables that are available in the original dataset as observed controls.

Here I go back to AET's framework and replace the moment condition $E(d_i u_i) = 0$ in (25) with the restriction as in (8), i.e. $|\phi| \geq |\phi_u|$. This restriction is the implication of random selection of observed variables under the condition that the number and explanatory powers of observables are relatively high. However, another problem for AET's method is that mean independence of observed variables x_i and unobserved factors u_i is unlikely to hold. Mean independence is important and maintained in virtually all observational studies of selection problem because without it the treatment effect is unidentifiable even if one has valid exclusion restriction. In the context of using selection on observed variable to infer the selection on unobservable, the mean independence is even harder to justify. Intuitively, if the observed variables are likely to be correlated with each other, which is true for most of the applications, then it is easy to draw the conclusion that the observed and unobserved determinants of outcome variable are also likely to be correlated to some extend. Remember the unobserved factor u_i includes both unobserved control variables and error term, so if one believes in this argument, any estimation of treatment effect would be biased and, as a consequence, any inference based on this estimator would also be incorrect whether one uses the condition $E(d_i u_i) = 0$ or $|\phi_u| \leq |\phi|$.

Before turning into the estimator, similar to Assumption 1, I continue to provide an explicit model for the treatment variable d_i which leads to the following assumption,

Assumption 6

$$d_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} x_{ij} \delta_j + \nu_i$$

In this assumption, ν_i is independent of all observed and unobserved variables. For convenience, we rewrite the model for d_i as

$$d_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{s} x_{ij} \delta_j + \psi_i \tag{26}$$

where $\psi_i = \frac{1}{\sqrt{K^*}} \sum_{j=s+1}^{K^*} x_{ij} \delta_j + \nu_i$ and s is the number of observed variables among K^* control variables. In order to address the issue of mean independence between observed and

[12] If there is no effect of unobservables, we will have $u_i = \xi_i$ and post double selection estimator (21) can be directly applied.

unobserved factors, instead of imposing mean independence of u_i and x_i, I assume $E(u_i|x_i)$ is of the linear form. More specifically,

$$E(y_i - \alpha d_i|x_i) = x_i'\beta + E(u_i|x_i) \equiv x_i'\gamma \tag{27}$$

and

$$y_i - E(y_i - \alpha d_i|x_i) \equiv \varepsilon_i \tag{28}$$

Note that if $E(u_i|x_i) = 0$, which means the assumption of mean independence is satisfied, then $\beta = \gamma$ and one can directly apply Theorem 1.

Let $\phi_{x_i'\gamma}$ and ϕ_ε be the coefficients of the projection of d_i on $x_i'\gamma$ and ε_i as defined in (27) and (28), that is

$$\begin{aligned} &Proj\left(d_i \left| \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} b_j x_{ij}\gamma_j, \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}(1-b_j)x_{ij}\beta_j + \xi_i \,; \mathcal{G}^K \right.\right) \\ &= \phi_{x_i'\gamma}\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} b_j x_{ij}\beta_j\right) + \phi_\varepsilon\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}(1-s_j)x_{ij}\beta_j + \xi_i\right) \end{aligned}$$

Theorem 2 [13] *Under Assumptions (1)-(5) and Assumption (6), if the probability limit of* $\phi_{x_i'\gamma}$ *is nonzero, then*

$$\frac{\phi_\varepsilon}{\phi_{x'\gamma}} \xrightarrow[K^*\to\infty]{p} \frac{\sum_{l=-\infty}^{\infty} E(x_{ij}x_{ij-l}\beta_j\beta_{j-l})}{\sum_{l=-\infty}^{\infty} E(x_{ij}x_{ij-l}\beta_j\beta_{j-l}) + \sigma_\xi^2}$$

As a result, the estimator of AET's method should be the solutions to a system of equations (29) and (30) subject to the restriction (31),

$$y_i = \alpha d_i + \frac{1}{\sqrt{K^*}} x_i'\gamma + \varepsilon_i \tag{29}$$

$$d_i = \frac{1}{\sqrt{K^*}} x_i'\delta + \psi_i \tag{30}$$

$$0 \le \left|\frac{cov(\psi_i, \varepsilon_i)}{var(\varepsilon_i)}\right| \le \left|\frac{Cov(x_i'\delta, x_i'\gamma)}{Var(x_i'\gamma)}\right| \tag{31}$$

Note that one improvement to the original estimator of AET's method is that I use the selected variables x_i as controls for the estimation.

[13] The proof of Theorem 4 can be found in Appendix A of Altonji et al. (2013).

5 Empirical Example

In the preceding sections, I provide some theoretical results demonstrating how model selection methods can be used to help us address the issue of selection on unobservables. I further apply this proposed method to an empirical example in the following section by re-examining a research conducted by Donohue III and Levitt (2001). In their original paper, they study of the impact of legalized abortion on crime rates. I briefly review the background of the original work first, and then present some estimates obtained using the methods proposed in this paper.

5.1 Effects of Legalized Abortion on Crime

Donohue III and Levitt (2001) estimates the effect of abortion on crime rates in which they argue two possible casual channels relating abortion to crime: the first channel is that the more abortion among a cohort will result in an otherwise smaller cohort size and so lower the crimes 15-20 years later given the smaller cohort size. The second is that abortion gives women more flexibilities over the timing of their fertility allowing them to more easily assure that childbirth occurs at a time when a more favorable environment is available during a childs life. More specifically, having access to abortion can help women ensure a child is born at a time when the family environment is stable, the mother is more well educated, or household income is stable. Consequently, given the unchanged fertility rates, the crime rates would be lower given a better family environment.

A major problem associated with this study is that crime rates are not randomly assigned among the states. It is very likely that some factors affect the abortion rate would affect the crime rate as well. Those factors could be due to the existence of persistent state-to-state differences in policies, attitudes, and demographics. It is also important to take the aggregate trends into consideration. For example, it could be the case that national crime rates were falling over some period while national abortion rates were rising but these trends were driven by completely different factors. In addition to these overall differences across states and times, it is also reasonable to control for some other time varying factors such as state level income and policy that could also relate current crime rate to past abortion. To address this issue Donohue III and Levitt (2001) estimates a model for state level crime rates running from 1985 to 1997 in which they condition on a number of observed factors. Their basic specification is

$$y_{cit} = \alpha_c a_{cit} + w_{it}\beta_c + \delta_{ci} + \gamma_{ct} + \varepsilon_{cit} \tag{32}$$

The dependent variable y_{cit} denotes the crime rate for crime type c, which is categorized between violent, property, and murder, in state i and year t. δ_{ci} are state specific effects that control for any time invariant state specific characteristics, γ_{ct} are time specific effects that control flexibly for any aggregate trends, w_{it} is a set of control variables to control for time varying confounding state level factors, a_{cit} is a measure of the abortion rate relevant for type of crime c[14].

[14] This variable is constructed as weighted average of abortion rates where weights are determined by the

5.2 Selecting Observables Through Double Selection

For the analysis of this section, I follow the original specification of Donohue III and Levitt (2001) but relax the assumption that abortion rates can be taken as exogenous relative to crime rates conditional upon a set of factors, which they originally include the log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, the generosity of the Aid to Families with Dependent Children (AFDC) at time $t15$, a dummy for concealed weapons law, and beer consumption per capita for w_{it}. My analysis differs from the original study of Donohue III and Levitt (2001) in two aspects: first, I relax the conditional exogeneity assumption of treatment variable, i.e. abortion, which means I estimate the treatment effect as a bound and take the effect of unobserved factors into account. Secondly, I do not assume the identities of those observed control variables and create additional control variables by allowing for smooth, flexible trends to account for factors that may influence both abortion and crime but smoothly trend over time. Once I have those artificially created technical regressors, I use model selection method, i.e. double selection, to help me select the relative important control variables.

Given the obvious importance of controlling for state and time fixed effects, I account for these fixed effects by adding a full set of state and time dummies. I consider the following two model selection equations

$$y_{it} = \alpha a_{it} + w_{it}'\beta_y + \delta_{y,i} + \gamma_{y,t} + g(z_{it}, t) + \varsigma_{it} \tag{33}$$

$$a_{it} = w_{it}'\beta_a + \delta_{a,i} + \gamma_{a,t} + m(z_{it}, t) + \nu_{it} \tag{34}$$

where $g(z,t)$ and $m(z,t)$ are smooth functions of observed variables z_{it}, which includes w_{it}, time-invariant characteristics of $\{y_{it}, a_{it}, w_{it}\}_{t=1}^{T}$ such as initial conditions or state level averages and time. I approximate $g(z_{it}, t)$ and $m(z_{it}, t)$ by a large number of observed controls. Specifically, I form 27 factors to include in z_{it},

$$z_{it} = \left(a_{i0}, \frac{1}{T}\sum_t a_{it}, y_{i0}, w_{i0}', \frac{1}{T}\sum_t w_{it}', w_{it}'\right)' \tag{35}$$

and 9 smoothed function of time,

$$f_t = \left(t, t^2, t^3, \sin(\pi\frac{t}{T}), \sin(2\pi\frac{t}{T}), \sin(3\pi\frac{t}{T}), \cos(\pi\frac{t}{T}), \cos(2\pi\frac{t}{T}), \cos(3\pi\frac{t}{T})\right)' \tag{36}$$

and then assume that

$$g(z_{it}, t) \approx \sum_{r=1}^{27}\sum_{s=1}^{9} \beta_{g,r,s} z_{it,r} f_{t,s} = h_{it}'\beta_g \tag{37}$$

$$m(z_{it}, t) \approx \sum_{r=1}^{27}\sum_{s=1}^{9} \beta_{m,r,s} z_{it,r} f_{t,s} = h_{it}'\beta_m \tag{38}$$

fraction of the type of crime committed by various age groups.

where h_{it} is a vector containing all the interactions, and β_g and β_m are vectors of coefficients for each equation. As a result, I artificially construct 243 additional control variables to the original model (32) and use double selection procedure discussed in this paper to search for the right set of observed control variables among these additional 243 variables.[15]To allow state and time fixed effects, and w_{it} to enter each equation without shrinkage, I use the double selection method based on $\tilde{y}_{it}$, $\tilde{a}_{it}$, and $\tilde{h}_{it}$ where $\tilde{y}_{it}$ is the residual from the regression of y_{it} on w_{it} and a full set of state and time dummies and $\tilde{a}_{it}$ and $\tilde{h}_{it}$ are defined similarly.

Having the g and m functions defined above, I then turn to the double selection procedures for the following two models

$$\tilde{y}_{it} = \tilde{h}_{it}'\beta_g + \tilde{\varsigma}_{it} \tag{39}$$

$$\tilde{a}_{it} = \tilde{h}_{it}'\beta_m + \tilde{\nu}_{it} \tag{40}$$

Controlling for a large set of observed variable describing above is desirable from the point that it might fully recover the underlying distribution of observed variables, which will turn out to help us better understand the selection on the unobserved determinants. However, the downside of including such large set of control variables is that it lessens the ability to identify the effect of interest and therefore makes the results far less precise. In order to illustrate the relative importance of model selection, I first estimate the abortion effect on crime without model selection by simply including all the 68 original control variables as in Donohue III and Levitt (2001) plus 243 additional controls I constructed via (39) and (40). The first row of Table 1 is from Donohue III and Levitt (2001) Table III and second row is estimated using all 311 controls. As we can see, adding too many controls makes the results very imprecise and one is essentially unable to learn about the causal effect of abortion.

Table 1: Estimated Effects of Abortion on Crime Rates

	Violent Crime		Property Crime		Murder	
	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
Donohue III and Levitt (2001)	-0.129	0.024	-0.091	0.018	-0.121	0.047
Fixed Effects	-0.131	0.045	-0.091	0.016	-0.131	0.058
All Controls	0.183	0.447	0.013	0.067	0.855	0.974

Note: This table displays the estimated coefficient on the abortion rate and its estimated standard error. Numbers from the first row are taken from Donohue III and Levitt (2001) Table IV, column (2), (4), and (6). Numbers from the second row are estimated by OLS of the crime rate on the abortion rate along with a full set of state dummies and a full set of time dummies. The third row is estimated by OLS of the crime rate on the abortion rate and all 311 controls described in the text above, including the original 68 controls from Donohue III and Levitt (2001) along with 243 variables constructed, and use standard errors clustered at the state-level.

Having a large set of control variables creates a challenge for researchers, but it is a trade-off between the precision of the estimates and the plausibility of using observables address the issue of selection on unobservables. By including the additional 243 variables makes AET's method more plausible. But at the same time, it potentially reduces the precision of the estimates. This is the reason why we need model selection to help us to achieve a balance.

[15] I do not include the original set of controls into the selection procedure and treat them as important factors that determining both abortion and crime.

With this relatively large set of observed control variables, I use double selection approach discussed in Section 3 to help me determine the right set of relevant observed variables that should be controlled for. The selection procedures are performed in two stages, the first stage is conducted upon the abortion equation (39) and second stage is conducted in the crime equation (40). The selected variables are described in Table 2. As we can see, a relatively small number of variables are selected out of the 243 potential control variables. For violent crime, 8 in abortion equation and 7 in crime equation; for property crime, 10 in abortion equation and 7 in crime equation with one in common; for murder, 8 in abortion equation and 2 in crime equation.

Table 2: Selected Variables for Abortion and Crime

	Violent Crime	Property Crime	Murder
Abortion Equation	8	10	8
Crime Equation	7	7	2
Total Selected	15	16	10

Note:

[1] For violent crime: in the abortion equation, the selected variables are average abortion×t, average abortion×$cos(\pi\frac{t}{T})$, initial crime×t^2, initial crime×$cos(2\pi\frac{t}{T})$, average income×t^3, average income×$sin(\pi\frac{t}{T})$, average income×$cos(2\pi\frac{t}{T})$, initial poverty×$cos(2\pi\frac{t}{T})$; in the crime equation, the selected variables are average abortion×t^3, initial abortion×t^3, initial abortion×$sin(\pi\frac{t}{T})$, initial poverty×$sin(2\pi\frac{t}{T})$, initial poverty×$cos(\pi\frac{t}{T})$, police$_{it}$ × t^3, and beer$_{it}$ × $sin(3\pi\frac{t}{T})$.

[2] For property crime: in the abortion equation, the selected variables are average abortion×$cos(\pi\frac{t}{T})$, initial abortion×$sin(3\pi\frac{t}{T})$, initial crime×$cos(\pi\frac{t}{T})$, average income×t, average income×$cos(\pi\frac{t}{T})$, initial poverty×$cos(2\pi\frac{t}{T})$, initial beer×$cos(2\pi\frac{t}{T})$, prison$_{it}$ × $cos(\pi\frac{t}{T})$, income$_{it}$ × $cos(\pi\frac{t}{T})$, and AFDC$_{it}$ × $cos(2\pi\frac{t}{T})$; in the crime equation, the selected variables are average abortion×t^3, initial crime×$sin(2\pi\frac{t}{T})$, initial crime×$cos(\pi\frac{t}{T})$, average police×$cos(2\pi\frac{t}{T})$, average AFDC×t, initial AFDC×t, and initial AFDC×t^2).

[3] For murder: in the abortion equation, the selected variables are average abortion×t^2, average abortion×$cos(\pi\frac{t}{T})$, initial crime×t^3, initial crime×$cos(2\pi\frac{t}{T})$, average income×t^3, average income×$sin(\pi\frac{t}{T})$, average income×$cos(2\pi\frac{t}{T})$, and average income×$cos(3\pi\frac{t}{T})$; in the crime equation, the selected variables are average abortion×$sin(\pi\frac{t}{T})$ and initial abortion×$sin(\pi\frac{t}{T})$.

5.3 Estimating Bounds for Treatment Effects of Abortion

With the selected observed control variables, I proceed to AET's method by considering a system of equations,

$$\tilde{y}_{it} = \alpha\tilde{a}_{it} + \tilde{h}'_{it}\beta_g + \tilde{\varsigma}_{it} \tag{41}$$

$$\tilde{a}_{it} = \tilde{h}'_{it}\beta_m + \tilde{\nu}_{it} \tag{42}$$

and estimating the effect of abortion on crime that satisfies the following condition,

$$0 \leq \left|\frac{cov(\tilde{\nu}_{it}, \tilde{\varsigma}_{it})}{var(\tilde{\varsigma}_{it})}\right| \leq \left|\frac{Cov(\tilde{h}'_{it}\beta_m, \tilde{h}'_{it}\beta_g)}{Var(\tilde{h}'_{it}\beta_g)}\right| \tag{43}$$

Table 3 shows the estimated bounds for the effect of abortion on crime in the context of model selection. Each of the estimates is estimated under the restriction (43) along with the observed control variables selected by double selection method. As we can see, for all these three outcome variables, with the new estimation strategy, the standard errors become larger than the original ones from Donohue III and Levitt (2001)[16], especially for the outcome variable violent crime and murder. This observation indicates that as one increases the number of control variables, the results become less precise. However, if one looks at the estimates themselves, they are quite different from the original estimates from Donohue III and Levitt (2001). For violent crime, as we can see from Table 1, the original estimate of abortion on violent crime is -0.129. With the new estimation strategy, the lower bound is pretty close to the original estimate as in Donohue III and Levitt (2001), which is -0.134, but the upper bound jumps to 0.133. The magnitude of increasing is almost as double as the lower bound. In addition, I found that lower bound in this case coincides with the case where there is no selection on unobservables. Jumping from a negative lower bound to a positive upper bound and not only the sign has changed but also the magnitude of change is so wide, one can infer that there is relative strong selection through the unobserved dimensions. If merely rely on the conditional exogeneity assumption of the treatment variable might result in a problematic estimate. For the outcome variables property crime and murder, the sign of the bounds do not change and the magnitude of change is also not as wide as violent crime. The percentage of increasing from lower bound to upper bound is 45% and 82% for property crime and murder, respectively. Therefore, the degree of selections on unobserved factors for property crime and murder are not as severe as violent crime.

Table 3: Estimated Bounds of for Treatment Effects of Abortion

	Violent Crime		Property Crime		Murder	
	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
Lower Bound	-0.134	0.330	-0.097	0.048	-0.692	0.438
Upper Bound	0.133	0.303	-0.053	0.044	-0.122	0.420

Note: This table displays estimated bounds for treatment effect of abortion rate and its estimated standard error clustered at the state-level. Lower bound and upper bound are estimated under the restriction 43 with the selected observed control variables using double selections.For violent crime, there are 83 control variables in total. For property crime, there are 84 control variables in total. For murder, there are 78 control variables in total.

It is interesting to note that one could draw quite qualitatively different conclusions from the estimates using bounds estimation along with model selection procedures than from the estimates obtained using only a small set of intuitively selected control variables. One cannot precisely determine the effect of the abortion rate on crime rates once control for a relatively large set of variables through model selection. This is a trade-off between precisions of estimates and strong exogeneity of the treatment variable. If one has strong belief regarding to the exogeneity of the treatment variable, we definitely recommend to go with the conventional estimation strategy. On the other hand, if the strong prior knowledge is not available to the researchers or one has doubts about the identifying assumptions, our strategy provides another way to assess the causality between the treatment variable and outcome variable. I also would like to point out I do not regard this strategy as a substitution for point estimation, rather as a way to assess the possible selection amount on unobservables.

[16] The original estimates from Donohue III and Levitt (2001) are displayed in the first row of Table 1.

6 Conclusion

In this paper, I use machine learning method to revisit AET's approach. My approach is different from the original approach of AET in two ways: first of all, I use the observed variables in the original data set to create a large set of "observed variables" by doing technical transformations. The reason of doing variable transformations is trying to recover the full distribution of all underlying observed variables so that I can use as large number of observables as possible to be better comply with AET's assumptions. Secondly, in order to avoid the case where the number of variables is larger than the number of observations and identify which exact variables should enter the model, I use modern model selection techniques to help me search for a smaller but relative important set of observables. This procedure helps me reduce the dimension of the covariates while maintain a relatively high explanatory power for these selected variables. I consider my approach as an improvement of AET's original method. Notice that the assumptions of Theorem 1 indicate that it is problematic, according to Altonji et al. (2013), "to infer too much about selection on unobservables from selection on the observables if the observables are small in number and explanatory power, or if they are unlikely to be representative of the full range of factors that determine an outcome." Therefore, model selection technique is important in AET's framework in the sense that it can help us achieve a balance between having large number of observables and maintaining high explanatory powers. In the future study, I still need to show some evidence that by creating technical controls and selecting them through machine learning method is indeed a better approach than AET's original method. Or I can directly model the relationship between observables and unobservables. My ultimate goal of studying selection on unobservables is to extend this internal selection framework to external selection framework where I can understand a broader picture of external validity issues.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 91(5): 1369–1401.

Altonji, Joseph G., Timothy Conley, Todd E. Elder, and Christopher R. Taber. 2013. "Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables." *Working paper.*

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2002. "The Effectiveness of Catholic School." Northwestern University.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005*a*. "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling." *Journal of Human Resources*, 40(4): 791–821.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005*b*. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–184.

Angrist, Joshua, and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." *Handbook of Labor Economics*, 3: 1277–1366.

Angrist, Joshua D., and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review*, 88(3): 450–477.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica*, 80(6): 2369–2429.

Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli*, 19(2): 521–547.

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

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

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

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2016. "Inference in High-Dimensional Panel Models With an Application to Gun Control." *Journal of Business & Economic Statistics*, 34(4): 590–605.

Belloni, A., V. Chernozhukov, and C. Hansen. 2010. "LASSO Methods for Gaussian Instrumental Variables Models." *ArXiv e-prints*.

Belloni, A., V. Chernozhukov, and C. Hansen. 2012. "Inference for High-Dimensional Sparse Econometric Models." *ArXiv e-prints*.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen. 2017. "Program Evaluation and Causal Inference With High-Dimensional Data." *Econometrica*, 85(1): 233–298.

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

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics*. Vol. 6 of *Handbook of Econometrics*, , ed. J.J. Heckman and E.E. Leamer. Elsevier.

Chen, Xiaohong, and Demian Pouzo. 2009. "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals." *Journal of Econometrics*, 152(1): 46–60.

Chen, Xiaohong, and Demian Pouzo. 2012. "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals." *Econometrica*, 80(1): 277–321.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach." *Annual Review of Economics*, 7(1): 649–688.

Currie, Janet, and Duncan Thomas. 1998. "Does Head Start Make a Difference?" *American Economic Review*, 85.

Donohue III, John J., and Steven D. Levitt. 2001. "The Impact of Legalized Abortion on Crime." *The Quarterly Journal of Economics*, 116(2): 379–420.

Donohue, III, John J., and Steven D. Levitt. 2008. "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz." *Quarterly Journal of Economics*, 123(1): 425–440.

Engen, Eric M., William G. Gale, and John Karl Scholz. 1996. "The Illusory Effects of Saving Incentives on Saving." *Journal of Economic Perspectives*, 10(4): 113–138.

Estimation of Educational Borrowing Constraints Using Returns to Schooling. 2004. “Estimation of Educational Borrowing Constraints Using Returns to Schooling.” 112(1): 132–182.

Foote, Christopher L., and Christopher F. Goetz. 2008. “The Impact of Legalized Abortion on Crime: Comment.” *Quarterly Journal of Economics*, 123(1): 407–423.

Frank, Ildiko E., and Jerome H. Friedman. 1993. “A Statistical View of Some Chemometrics Regression Tools.” *Technometrics*, 35(2): 109–135.

Grogger, Jeffrey, Stephen Bronars, and Jeff Grogger. 1994. “The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment.” *American Economic Review*, 84: 1141–56.

Hansen, Christian, and Damian Kozbur. 2014. “Instrumental variables estimation with many weak instruments using regularized JIVE.” *Journal of Econometrics*, 182(2): 290 – 308.

Jacobsen, Joyce, James Wishart Pearce, and Joshua Rosenbloom. 1999. “The Effects of Childbearing on Married Women’s Labor Supply and Earnings: Using Twin Births as a Natural Experiment.” *Journal of Human Resources*, 34(3): 449–474.

Newey, Whitney K. 1997. “Convergence rates and asymptotic normality for series estimators.” *Journal of Econometrics*, 79(1): 147–168.

Rosenbaum, Paul R. 1995. *Observational Studies.* Springer-Verlag New York.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. “The central role of the propensity score in observational studies for causal effects.” *Biometrika*, 70(1): 41–55.

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

Udry, Christopher. 1996. “Gender, Agricultural Production, and the Theory of the Household.” *Journal of Political Economy*, 104(5): 1010–1046.