DIRECT ESTIMATION OF POLICY IMPACTS

Hidehiko Ichimura
Christopher Taber

THE INSTITUTE FOR FISCAL STUDIES
WP 00/05

Direct Estimation of Policy Impacts

Hidehiko Ichimura[1]
Department of Economics
University College London
and
Institute for Fiscal Studies

Christopher Taber[2]
Department of Economics
and
Institute for Policy Research
Northwestern University

March, 1998
January, 2000

[1]The discussions both of us have had with Jim Heckman and Ichimura has had with Chris Sims and Ken Wolpin at different points in time have been useful in formulating the idea presented in this paper. We wish to acknowledge the indebtness here and thank them. This paper has been presented at the North American Econometric Society Winter Meeting in New York, 1998. We thank the comments we received from Josh Angrist, Guido Imbens, and Chuck Manski. We also thank useful comments received from seminar participants at Berkeley, Brown University, Northwestern University, Princeton University, Rice University, Texas A&M, University of Chicago, University of Bristol, University College London, and UCLA. In particular, we wish to thank Greg Chow for the Marschak reference and Andrew Chesher and Richard Blundell for clarifying comments on the role of conditioning variables.

Abstract

This paper specifies a general set of conditions under which the impacts of a policy can be identified using data generated under a different policy regime. We show that some of the policy impacts can be identified under relatively weak conditions on the data and structure of a model. Based on the identification result we develop estimators of policy impacts. We discuss a nonparametric method to implement the estimation but also discuss semiparametric methods in order to reduce the conditioning dimension. We then provide an empirical example of the impact of tuition subsidies using the ideas. While the framework used in this paper is fairly narrow, we believe this approach can be applied to a broad set of problems.

This paper specifies a general set of conditions under which the impacts of a policy can be identified using data generated under a different policy regime. We show that some of the policy impacts can be identified under relatively weak conditions on the data and structure of a model. Based on the identification result we develop estimators of policy impacts. We discuss a nonparametric method to implement the estimation but also discuss semiparametric methods in order to reduce the conditioning dimension. We then provide an empirical example of the impact of tuition subsidies using the ideas. While the framework used in this paper is fairly narrow, we believe this approach can be applied to a broad set of problems.

The standard formal econometric approach to estimation of a policy impact uses two stages. First a "structural" model is estimated, and second, these estimates are used to simulate the policy counter-factual. Sometimes the structural model takes the form of a regression model, and in other cases the model is specified from first principles of a behavioral model. In both cases parameters of a model are estimated first and then the estimate of a target parameter is constructed using these estimates.

Our approach here is to consider estimation of the policy impact directly rather than in two stages. When we have the limited objective of obtaining estimates of policy impacts we show that we can sometimes sidestep the problem of estimating the full behavioral models or even a regression model. We can still obtain consistent estimates of the policy impacts as captured by the parameters we specify.

When the conditions justifying our approach are applicable, it has three benefits over the standard structural approach. First, there are examples in which the full model is not identified, but the policy impact can be identified. In these cases the standard approach can not be carried out, but our approach may be applicable. For example, semiparametric identification of key parameters in the classic selection model is often achieved by "identification at infinity" making use of the subset of data where the probability of a particular event is close to 0 and 1.[1] If the support of the data is limited so that the probability is never close to the extremes, then the parameters of the model are not identified without strong (typically parametric) restrictions on the distribution of the unobservables.[2] The traditional two step policy analysis will not work without

[1] See Chamberlain (1986), Heckman (1990), Heckman and Honore (1990), or Angrist and Imbens (1991). Taber (2000) uses a similar strategy to show identification in discrete choice dynamic programming models.

[2] Heckman, Ichimura, Smith and Todd (1998) and Heckman Ichimura and Todd (1997, 1998) demonstrate that in practice these support conditions are very important.

a parametric distributional assumption in that case. We show that identification of the full joint distribution of the error terms is unnecessary for identifying the policy impact measure we define below and thus our approach avoids the "identification at infinity" problem. Related to this point is that even when all aspects of the first stage model are formally identified, using estimates of them may lead to inaccurate estimates of policy impact relative to ours if the estimation of those parameters can be done only inaccurately.

Second, as we do not require specification of the first stage model our approach is less prone to misspecification problems. In particular, often the two step approach relies on the specification of the additive error terms or parametric specification of the error distribution. Our approach does not rely on such specification.

Third, by the nature of two step procedures, the first stage estimation is carried out without regard to the second stage. Thus when the first stage is misspecified, the parameters that are tuned to approximate the first stage equation may not be adequate to approximate the policy impact measure in the second stage. The problem is essentially that the loss function used to estimate the parameters in the first stage is unrelated to the policy experiment for which the estimates will be used. By directly estimating the policy effect we avoid this problem by focusing on the variation of data that is directly linked to the policy impact.

In addition this approach shares with the structural approach an advantage over "instrumental variables" or "natural experiment" methods of being explicit about the policy and some aspects of the behavioral model underlying the estimation. The emphasis in the "natural experiment" literature is typically on finding exogenous variation. However, exactly how variation is linked to the policy under consideration is rarely made precise. We provide a framework to make this link. Making this link forces the empiricist to be explicit about which variation in the data corresponds to a policy equivalent variation.[3] Finding a policy equivalent variation typically requires stronger assumptions than in the instrumental variable case, but weaker assumptions than for full scale structural estimation. We view our approach as a hybrid between the two.

This relates to the debate over Instrumental Variable estimation of treatment effect models.[4] Imbens and Angrist (1994) show that the parameter being estimated by instrumental variables takes the form of a "Local Average Treatment Effect" (LATE). Heckman (1997) criticizes this pa-

[3] We formally define a policy equivalent variation in section 3.

[4] See e.g. Heckman (1997,1999a),Imbens and Angrist (1994), and Angrist and Imbens(1999).

rameter because it typically does not answer an economically interesting question. Our approach avoids this criticism by estimating a pre-specified policy counter-factual.

However, as we shall discuss, this type of reduced form approach is not always applicable. Clarification of the conditions under which we can and cannot identify a policy impact is the primary goal of this paper.

Our idea is an extension of the classic idea of making use of historical variation that corresponds to the policy under consideration. When there is such exogenous variation in the data, it may be used to identify the policy impact, but when there is no corresponding historical variation then there is difficulty using this approach. Marschak (1953) provides an example of a monopolistic firm trying to maximize profit. In his example, an output level correspond to a policy and outcome is profit. By randomly experimenting with different levels of output and tabulating the results, the firm would know the profit level that correspond to a particular output level without knowledge of any of the structural parameters. If we do not have data which correspond to certain level of outputs, then the profit function at those points won't be observed. The simple example makes it clear that when there is a variation in the data that correspond to a policy under consideration one would know the impact of a policy but that when we don't have the corresponding historical variation we do not.

Another limitation of the approach is that when there is a change in some parameters then the reduced form relationship examined will in general change and thus the approach requires variation under the new regime.[5] Marschak (1953) discusses this problem using a government contemplating imposing a tax on the demand for the monopolist's output. When the government changes the tax rate, the reduced form relationship of profit and output changes and thus the government would not be able to evaluate the impact of a change of the tax rate on the monopolist's profit by studying the reduced form relationship observed in the past. In this sense the reduced form analysis seems applicable for ex-post policy analysis but not for ex-ante policy analysis. Marschak (1953) points out that by making use of an economic model of the demand function and the specification of how it relates to tax and the profit function one can resolve this difficulty by estimating the demand function. In terms of the discussion above, his is a two step approach; the first step is the estimation of the demand function and the second step is to combine it with an economic model of demand that the demand only depends on tax through price to estimate the profit function

[5] See Hurwicz (1950), Marschak (1953), and Lucas (1976).

under new regime. The analysis provides an example of a possibility of substituting economic theory in place of lack of data.

It is important not to interpret the classic work on this subject as implying that estimation of all of the parameters of a structural model is necessary for predicting the effects of a new type of policy that has not been enacted in the past.[6] By making use of some aspects of a behavioral model one can exploit other types of variation in the data to mimic the effects of a policy change. Sims (1982) mentions that this may be possible with a change in the money stock. In our empirical work, we consider the example of a tuition subsidy. Even if a tuition subsidy has never been enacted before, if we are willing to assume that the tuition faced by individuals varies exogenously and that the tuition subsidy operates only by lowering the net tuition paid, then one can estimate its effect through reduced form nonparameteric regression. Knowing the effect of tuition on outcomes allows one to infer the effect of a tuition subsidy on outcomes without knowledge of the structural parameters of the model. However, we still need to impose some structure on the problem, namely that tuition subsidies operate only by lowering net tuition. Another example is taxes and labor supply. In a partial equilibrium setting workers will respond to changes in taxes in the same way they respond to changes in wages, so after invoking some structural assumptions one can use other types of variation in wages to estimate the effects of taxes on labor supply.

Marschak observes that one can substitute lack of historical variation in the data with economic modeling in the structural approach. We exploit this observation in a more reduced form approach and specify the conditions under which the effects of a particular policy can be identified directly. This paper focuses on the program evaluation model with two alternative choices, but the basic principle can be extended to more general contexts.

In section 2 we present our framework and section 3 establishes conditions under which policy effects can be identified. Section 4 describes the relationship between our approach and results available in the literature, and section 5 presents an empirical example. Section 6 concludes.

[6] See Heckman (1999b) for discussion of much of this previous literature.

1 Basic model and parameters of interest

There are three basic elements in our model: choice variables, outcome variables, and a policy under consideration. In this paper we consider a case in which the choice variable is binary and the outcome depends on the choice. We index the policy by π which we assume to lie in a space we call policy space Π. The choice variable under policy π is denoted by the random function $D(Z,\pi)$ which takes values 0 and 1, where Z is an observable random vector. $Y_1(Z)$ and $Y_0(Z)$ denote outcomes that correspond to choice $D(Z,\pi)=1$ and 0, respectively without reference to π.[7]

An important assumption we have made is that the outcome distribution of Y_0 and Y_1 is not altered by the introduction of the new policy.[8] We do not consider policies that change treatment intensity. Nor do we consider the general equilibrium effects of the policy change.[9] Heckman, Lochner, and Taber (1998) show that ignoring these effects may be disastrous for some national programs. However, for local programs there is no reason to believe this assumption will be particularly problematic.

In the context of program evaluation $D(Z,\pi)$ denotes program participation under policy π and the individual outcomes with and without enrolling in the program are denoted $Y_1(Z)$ and $Y_0(Z)$, respectively. Examples of π are subsidies or eligibility criteria of the program.[10]

Let

$$Y(Z,\pi) = D(Z,\pi)\,Y_1(Z) + \{1 - D(Z,\pi)\}\,Y_0(Z).$$

Since $Y_1(Z)$ is realized only if the person chooses $D(Z,\pi)=1$, and since $Y_0(Z)$ is realized only if the person chooses $D(Z,\pi)=0$, the econometrician can only observe $(D(Z,\pi), Y(Z,\pi), Z)$ for the policy π that is in place when the data is generated.

We use the following notational convention throughout the rest of this paper. If $H(Z)$ is

[7] Many of the results will make use of exclusion restrictions: elements that influence choice but not outcomes. However, to simplify the notation we write the outcome variables as a function of the whole vector Z. This notation includes cases where some elements of Z do not affect outcomes.

[8] We usually condition on Z but when we do not, we assume that the relevant Z distribution is the one under the old policy or it is not altered by a new policy.

[9] For these cases see Heckman and Smith (1997) and Heckman (1997). General equilibrium effects are considered by Heckman, Lochner, and Taber (1998) where changing enrollment in the program may change the value of the program through equilibrium effects.

[10] When possible we denote random variables or random vectors by upper case letters and the particular values of them by lower case letters.

a random function of Z, we take the expression $E\{H(z)\}$ to mean $E\{H(Z) \mid Z = z\}$.[11] Since virtually every expectation we consider conditions on Z, this simplifies the notation substantially.

The first two parameters we consider are

$$\Delta\left(z, \pi', \pi\right) = E\left[Y(z, \pi') - Y\left(z, \pi\right)\right],$$

the mean policy effect and

$$\Delta_c(z, \pi', \pi) = E\left[Y(z, \pi') - Y\left(z, \pi\right) \mid D(z, \pi') \neq D(z, \pi)\right],$$

the conditional policy effect. The first parameter $\Delta\left(z, \pi', \pi\right)$ captures the change in outcomes for the population with characteristic $Z = z$ when policy shifts to π' from π.[12] The second parameter $\Delta_c\left(z, \pi', \pi\right)$ captures the average gain for the population with characteristic $Z = z$ who would be affected by the policy shift.

In general,

$$\Delta(z, \pi', \pi) = \Delta_c(z, \pi', \pi) \Pr\left\{D(z, \pi') \neq D(z, \pi)\right\},$$

so that $|\Delta(z, \pi', \pi)| \leq |\Delta_c(z, \pi', \pi)|$. This observation highlights the distinction between the two parameters. The mean treatment policy embodies the notion of extensiveness of the impact measured by $\Pr\{D(z, \pi') \neq D(z, \pi)\}$, but the conditional policy effect is the measure that isolates the intensiveness of the impact once the choice is affected. Ideally we would want to identify both the extensive and intensive impacts. Identification of parameter $\Delta\left(z, \pi', \pi\right)$ can be achieved under weaker conditions than those for $\Delta_c(z, \pi', \pi)$.

The intensity measure $\Delta_c(z, \pi', \pi)$ described above is related to the local average treatment effect (LATE) of Imbens and Angrist (1994). They define LATE as the expected treatment effect for individuals who are influenced to change treatment status by a change in the value of a particular conditioning variable, which they refer to as an instrumental variable. Our parameter is the expected treatment effect for individuals who are influenced to change treatment status

[11] And similarly,

$$\Pr\{D(z, \pi) = 1\} = \Pr\{D(Z, \pi) = 1 \mid Z = z\}.$$

[12] It is a version of the policy effect in Heckman (1997) that conditions on Z.

by a change in a particular pre-specified policy. By separating the two explicitly, we provide a framework to discuss identification and measurement of policy impacts.[13]

The next two "marginal treatment effect" parameters we consider are normalized limits of the parameters $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$, and can be considered only for π that is defined on a policy space Π with a notion of "closeness". These parameters correspond to those discussed in the literature including studies by Bjorkland and Moffitt (1987), Heckman and Smith (1997), and Taber (1999). The nice aspect of these parameters is that with continuous data and policies they will be identified under weak support conditions.

To this point we have not put any structure on Π, so the value of π has no content unto itself except as an index to a policy option. In thinking about marginal treatment effects we specialize Π to be a finite dimensional vector of real valued functions. For example if the government considered a tuition subsidy, π could index the extent of the subsidy. In this case Π can be identified with the weakly positive real line. Suppose instead it considered a tuition subsidy with means testing, then a policy may be represented by two numbers $\pi = (\pi_1, \pi_2)$, where the amount of tuition subsidy is denoted by π_1 and the maximum eligible parental income by π_2. In this case Π can be identified with the two dimensional weakly positive real plane. If the amount of the subsidy depends on parental income then Π can be identified with a space of real valued functions.

We define a marginal treatment effect as the impact of an infinitesimal change in the extent of intervention starting at π and will consider two types. Let $\lambda > 0$ be a real number and let $\pi' = \pi + \lambda\tilde{\pi}$ for some element $\tilde{\pi}$ in Π. Letting '$\lambda \downarrow 0$' denote λ approaches 0 from above, one concept is the limit of the conditional mean impact on the switchers when policy shifts marginally in direction $\tilde{\pi}$,

$$\Delta_c^m(z, \tilde{\pi}, \pi) \equiv \lim_{\lambda \downarrow 0} \Delta_c(z, \pi', \pi).$$

The other concept is the normalized limit of the mean impact when policy shifts marginally in direction $\tilde{\pi}$,

$$\Delta^m(z, \tilde{\pi}, \pi) \equiv \lim_{\lambda \downarrow 0} \frac{\Delta(z, \pi', \pi)}{\lambda}.$$

[13] The distinction between an instrumental variable and a policy is important. The definition of the LATE parameter depends on the data so that by definition it will be identified. Since our parameter depends on a prespecified policy it is not necessarily identified. Angrist and Krueger (1992) provides a useful example for demonstrating the difference between policy parameters and instrumental variables. Compulsory schooling laws are a policy that one can consider. However, the source of variation used for identification comes from variation in quarter of birth rather than schooling laws. While they are related, they are not identical.

For example if $\tilde{\pi} = e_j$ where e_j has 0 in all but the jth element and 1 as the jth element then since $\Delta(z, \pi', \pi) = E[Y(z, \pi') - Y(z, \pi)]$,

$$\Delta^m(z, e_j, \pi) = \frac{\partial E[Y(z, \pi)]}{\partial \pi_j}.$$

We note that the concept of $\Delta_c^m(z, \tilde{\pi}, \pi)$ can be defined more generally than the concept of the directional limit utilized here but that the concept of $\Delta^m(z, \tilde{\pi}, \pi)$ depends crucially on the concept of the directional limit as we make use of the normalizing number λ in an essential way.

As we observed, in general,

$$\Delta(z, \pi', \pi) = \Delta_c(z, \pi', \pi) \Pr\left\{D(z, \pi') \neq D(z, \pi)\right\},$$

so that

$$\Delta^m(z, \tilde{\pi}, \pi) = \Delta_c^m(z, \tilde{\pi}, \pi) \lim_{\lambda \downarrow 0} \frac{\Pr\{D(z, \pi') \neq D(z, \pi)\}}{\lambda}.$$

That is, just as for parameters $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$, parameter $\Delta^m(z, \tilde{\pi}, \pi)$ is inclusive of the extensive impact whereas $\Delta_c^m(z, \tilde{\pi}, \pi)$ isolates the intensity of the marginal treatment effect.

Bjorkland and Moffitt (1987) examine a parameter analogous to $\Delta^m(z, \tilde{\pi}, \pi)$. They consider a case with

$$D(Z, \pi) = 1(Z'\gamma + \pi + U \geq 0),$$

where π is a real number representing costs of choosing 1 over 0, Z and U denote observable and unobservable random variables that affect the choice and study the parameter

$$\frac{\partial E\{Y(z, \pi)\}}{\partial \pi}.$$

Heckman and Vytlacil (1999) and Aakvik, Heckman and Vytlacil (1999) consider a related parameter they call local IV. In the same sense that $\Delta_c^m(z, \tilde{\pi}, \pi)$ is a limit form of $\Delta_c(z, \pi', \pi)$, their parameter is a limit form of LATE. They show that in a latent variable framework this parameter can be interpreted as the value of the treatment conditional on being indifferent between entering the program. Taber (1999) estimates a version of this parameter.

These parameters have two nice features. The first is that, as the definition makes clear, they can be approximated by $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$ where π' is defined by a small value of λ. The other nice feature is that, as we will show below, when the support $\mathcal{Z}$ is continuous, the parameters will be identified under weaker support conditions than those for $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$.

2 Identification

2.1 Identification of the Treatment Counter-factual

We first consider identification of the decision rule under the new policy function π', $D(Z,\pi')$. We then consider identification of the distribution of $Y_1(Z)$ given $D(Z,\pi')=1$ and $Z=z$ and that of the distribution of $Y_0(Z)$ given $D(Z,\pi')=0$ and $Z=z$. Clearly the first is relevant only when $\Pr\{D(Z,\pi')=1\}>0$ and the second is relevant only when $\Pr\{D(Z,\pi')=0\}>0$. Identification of the distribution of Y_1 given $D(Z,\pi')=0$ and that of the distribution of Y_0 given $D(Z,\pi')=1$ are not necessary for our purpose. Identification of the policy effects comes directly from these results.

Let $\mathcal{Z}$ be the support of Z. We want to identify the choice behavior under a new policy π', $D(Z,\pi')$, using the observations about choices made under old policy π, $D(Z,\pi)$. The following set plays the key role for this purpose:

$$\mathcal{D}(z,\pi',\pi)=\left\{z^*\in\mathcal{Z}:\Pr\{D(z,\pi')=D(z^*,\pi)\}=1\right\}.$$

For any point in this set, $z^*\in\mathcal{D}(z,\pi',\pi)$, the observed choice behavior $D(z^*,\pi)$ mimics the choice behavior under the new policy, $D(z,\pi')$. Thus if we could condition on elements of this set we could identify the choice behavior under the new policy.

Being able to condition on this set requires essentially two types of conditions. First, we need to be able to determine the values of z^* for which $D(z,\pi')=D(z^*,\pi)$. This will typically require some type of "structural" assumption. Second, these values of z^* must be contained within the support of Z.

In general, without any understanding of the relationship between z and π, the set $\mathcal{D}(z,\pi',\pi)$ is not known. However we show via examples below, that by exploiting some aspects of a behavioral model, in some cases we can identify the elements in this set.

The notation we use in the bulk of the paper hides an important aspect of a problem involved in the statement above. We use more complete notation just for a few paragraphs below to explain the assumption needed more explicitly. Implicit in the expression $\Pr\{D(z,\pi')=D(z^*,\pi)\}$ is the assumption that the concept of probability is well defined for two different points z and z^*. In particular, this requires that the stochastic element that drives the choice variable be independent from at least the part of Z that makes the equality holds. To express this more explicitly, let

$Z = \left(\tilde{Z}, Z_\pi\right)$ and corresponding fixed values $z = (\tilde{z}, z_\pi)$. Using this notation we write

$$\mathcal{D}(\tilde{z}, z_\pi, \pi', \pi) = \left\{(\tilde{z}, z_\pi^*) \in \mathcal{Z} | \Pr\{D(\tilde{z}, z_\pi, \pi'; \omega) = D(\tilde{z}, z_\pi^*, \pi; \omega)\} = 1\right\},$$

where ω expresses the stochastic element that drives the participation decision and z_π is the part of z that makes the equality holds. In order for this expression to make sense, Z_π and ω need to be independent given $\tilde{Z}$.

For convenience, we call Z_π "a policy-π equivalent variation given $\tilde{Z}$" or "a policy equivalent variation" when π and $\tilde{Z}$ are evident or not necessary to be made explicit in a discussion and ω "unobserved variation in the choice variable". We assume

Assumption 1 *A policy equivalent variation and the unobserved variation in choice variable are independent given some conditioning variables under two policies π and π'.*

The variables that correspond to the policy equivalent variation depends on the behavioral model assumed and Assumption 1 needs to be evaluated for each application.

For example in the empirical work we consider below, π is the current tuition subsidy level and π' is the contemplated tuition subsidy level. We assume a behavioral model in which college attendance depends only on net tuition so that individual choice depends only on $z_\pi - \pi$, where z_π is a level of tuition faced before the subsidy. It is conceivable that an individual's behavior could be different for different combination of z_π and π even if $z_\pi - \pi$ is the same but this is the "structural" assumption we are going to maintain and exploit. In addition to this assumption which defines the policy equivalent variation, we need to maintain Assumption 1. In our example, tuition is measured by the average tuition of 2 year colleges of the state in which the individual lived at age 17. We need to assume that this variable and the unobserved variation in choice variable are independent given some conditioning variables. Because we expect the Z_π variable to be correlated with some state characteristics which also can be correlated with the individual characteristics, we need to condition on certain variables such as race and parental education level.

We next consider identification of the distribution of $Y_1(z)$ given $D(z, \pi') = 1$ and identification of the distribution of $Y_0(z)$ given $D(z, \pi') = 0$. As we just discussed we simulate a decision under new policy, $D(z, \pi')$, by examining the choice made under the old policy by individuals with characteristic z^*, $D(z^*, \pi)$. Note that the corresponding outcome $Y_1(z^*)$ and $Y_0(z^*)$ need to

match $Y_1(z)$ and $Y_0(z)$, respectively. Thus the key assumption of the identification result is based on the following sets:

$$\begin{aligned}\mathcal{Z}_0(z,\pi) &= \{z^* \in \mathcal{Z} : \Pr\{Y_0(z) = Y_0(z^*) | D(z^*,\pi) = 0\} = 1\}, \\ \mathcal{Z}_1(z,\pi) &= \{z^* \in \mathcal{Z} : \Pr\{Y_1(z) = Y_1(z^*) | D(z^*,\pi) = 1\} = 1\}.\end{aligned}$$

Typically the assumption holds when some exclusion restrictions hold. In the tuition subsidy example, if tuition z_π does not enter directly into the outcome equation, the condition holds. More generally, while it may be possible to avoid them with some parametric specifications, for general nonparametric models exclusion restrictions will be required to satisfy these conditions.

Using the intersection of these sets with $\mathcal{D}(z,\pi',\pi)$, we can identify the distribution of $Y(z,\pi')$ in a manner similar to $\Pr(D(z,\pi')=1)$ above. In particular, for any $z \in \mathcal{Z}$, if we can find a value $z^* \in \mathcal{D}(z,\pi',\pi)$ that is contained in $\mathcal{Z}_0(z,\pi)$, and $\mathcal{Z}_1(z,\pi)$, then the distribution of $Y(z,\pi')$ is the same as the distribution of $Y(z^*,\pi)$. To see this note that for any $z^* \in \mathcal{D}(z,\pi',\pi) \cap \mathcal{Z}_0(z,\pi) \cap \mathcal{Z}_1(z,\pi)$,

$$\begin{aligned}\Pr(Y(z,\pi') < y) &= \Pr(Y_0(z) < y \mid D(z,\pi') = 0) \Pr(D(z,\pi') = 0) \\ &\quad + \Pr(Y_1(z) < y \mid D(z,\pi') = 1) \Pr(D(z,\pi') = 1) \\ &= \Pr(Y_0(z) < y \mid D(z^*,\pi) = 0) \Pr(D(z^*,\pi) = 0) \\ &\quad + \Pr(Y_1(z) < y \mid D(z^*,\pi) = 1) \Pr(D(z^*,\pi) = 1) \\ &= \Pr(Y_0(z^*) < y \mid D(z^*,\pi) = 0) \Pr(D(z^*,\pi) = 0) \\ &\quad + \Pr(Y_1(z^*) < y \mid D(z^*,\pi) = 1) \Pr(D(z^*,\pi) = 1) \\ &= \Pr(Y(z^*,\pi) < y).\end{aligned}$$

Thus conditioning on z^* allows us to identify the distribution of $Y(z,\pi')$. Once again this conditioning requires both that $\mathcal{D}(z,\pi',\pi) \cap \mathcal{Z}_0(z,\pi) \cap \mathcal{Z}_1(z,\pi)$ is known and that it is nonempty involving both "structural" and "support" conditions.

We now formalize this idea.

Assumption 2 *$\mathcal{Z}_0(z,\pi)$ and $\mathcal{Z}_1(z,\pi)$ are known and their intersection with $\mathcal{D}(z,\pi',\pi)$ is nonempty for $z \in \mathcal{Z}$.*

Lemma 1 *(i) Under Assumptions 1 and 2, if* $\Pr\{D(z,\pi')=0\}>0$ *the distribution of* Y_0 *given* $D(z,\pi')=0$ *is identified. (ii) Under Assumptions 1 and 2, if* $\Pr\{D(z,\pi')=1\}>0$ *the distribution of* Y_1 *given* $D(z,\pi')=1$ *is identified.*

(Proof In Appendix)

The lemma delivers identification of $\Delta(z,\pi',\pi)$.

Theorem 2 *If Assumptions 1 and 2 hold for the same* z *and if* $E\{D(z,\pi')Y_1(z)\}$ *and* $E[\{1-D(z,\pi')\}Y_0(z)]$ *are finite,* $\Delta(z,\pi',\pi)$ *is identified.*

(Proof In Appendix)

We next discuss identification conditions for the parameter $\Delta_c(z,\pi',\pi)$. Note that since our policy only influences outcomes through D and that $E\{Y(z,\pi')-Y(z,\pi)\mid D(z,\pi')=D(z,\pi)\}=0$, so that

$$\begin{aligned}\Delta_c(z,\pi',\pi) &= E\{Y(z,\pi')-Y(z,\pi)\mid D(z,\pi')\neq D(z,\pi)\}\\ &= \frac{E\{Y(z,\pi')-Y(z,\pi)\}}{\Pr\{D(z,\pi')\neq D(z,\pi)\}}\\ &= \frac{\Delta(z,\pi',\pi)}{\Pr\{D(z,\pi')\neq D(z,\pi)\}}.\end{aligned}$$

>From the theorem we know that $\Delta(z,\pi',\pi)$ is identified under Assumption 2. However these assumptions are not sufficient for identification of the denominator. All we can hope to identify about the joint distribution of $(D(z,\pi'),D(z,\pi))$ conditional on $Z=z$ is $\Pr\{D(z,\pi')=1\}$ and $\Pr\{D(z,\pi)=1\}$. Without further assumptions this will not be sufficient to identify $\Pr\{D(z,\pi')\neq D(z,\pi)\}$. To assure identification of $\Pr\{D(z,\pi')\neq D(z,\pi)\}$ we use a monotonicity assumption.

Assumption 3 *For any* $z\in\mathcal{Z}$, *either* $\Pr\{D(z,\pi')\geq D(z,\pi)\}=1$ *or* $\Pr\{D(z,\pi')\leq D(z,\pi)\}=1$.

Under this assumption

$$\Pr\{D(z,\pi')\neq D(z,\pi)\}=\left|\Pr\{D(z,\pi')=1\}-\Pr\{D(z,\pi)=1\}\right|$$

and thus $\Pr\{D(z,\pi')\neq D(z,\pi)\}$ is identified. Imbens and Angrist (1994) exploit this type of condition in the context of identification of treatment effects.

Corollary 3 *Under Assumptions 1,2 and 3,* $\Delta_c(z,\pi',\pi)$ *is identified.*

2.2 Identification of Marginal Treatment Effects

As we illustrate below, Assumption 2 is not likely to hold on all points in the support of Z. In this subsection we will establish conditions for identification of the marginal treatment effects defined above. We show that identification of the marginal treatment effects can be carried out under support conditions that are weaker. Recall that the marginal treatment effects are denoted $\Delta^m(z, \widetilde{\pi}, \pi)$ and $\Delta_c^m(z, \widetilde{\pi}, \pi)$, where the connection between π' and $\tilde{\pi}$ is that $\pi' = \pi + \lambda\tilde{\pi}$ and that we consider the limit of λ to zero from above. The key assumption is the following:

Assumption 4 *There exists $\lambda^*(z, \widetilde{\pi}, \pi) > 0$ such that Assumption 2 holds for all π' that correspond to λ such that $0 < \lambda < \lambda^*(z, \widetilde{\pi}, \pi)$.*

Under this assumption it is easy to show that the conditional marginal treatment effects are identified.

Corollary 4 *Under Assumptions 1 and 4, if $\Delta^m(z, \widetilde{\pi}, \pi)$ exists, then it is identified.*

Corollary 5 *Under Assumptions 1,3 and 4, if $\Delta_c^m(z, \widetilde{\pi}, \pi)$ exists, then it is identified.*

The argument here is in some sense the opposite of identification at infinity. In the typical identification at infinity we use the extremes of the distribution to produce the policy counterfactual. The corollaries above essentially use the part of the distribution that is infinitesimally close for identification. We will clarify this claim and discuss the extent to which these conditions are weaker in some of the examples below.

2.3 Examples

To demonstrate the ideas above and to examine some of the limitations of the approach we study four examples. These have been chosen to represent a wide variety of models and policies.

Example 1: Treatment on the Treated

One parameter that is often discussed in the program evaluation literature is the effect of the "treatment on the treated." It can be identified using an "identification at infinity" argument.

We will demonstrate that this result is a special case of Theorem 2 above. In our framework in which parameters are defined conditioned on Z, this parameter takes the form,

$$E\left(Y_1(z) - Y_0\left(z\right) | D(z,\pi) = 1\right).$$

It is interpreted as the effect of the program on those individuals who choose to enter it. It can be considered a special case of our conditional policy effect where the alternative policy π' corresponds to elimination of the program, so that for any $z \in \mathcal{Z}$, $D(z,\pi') = 0$. In that case,

$$\begin{aligned}\Delta_c(z,\pi',\pi) &= E(Y(\pi') - Y(\pi) \mid D(z,\pi') \neq D(z,\pi)) \\ &= -E\left(Y_1(z) - Y_0\left(z\right) | D(z,\pi) = 1\right).\end{aligned}$$

Suppose that we have exclusion restrictions as in the case discussed above so that $Z = (Z_1, Z_2)$ where Z_1 influences the decision to enter the program, but has no direct influence on outcome conditional on entry. Following the logic above, since $D(z,\pi') = 0$ with probability one, for each z_2 we need to find a value of z_1^* such that almost surely $D\left((z_1^*, z_2),\pi\right) = 0$. If we can find such a z_1^*, then with this type of exclusion restriction, Assumption 2 is satisfied. Thus for the treatment on treated parameter our identification conditions are met using "identification at infinity."

This example is extreme in two ways. On the one hand the conditions under which Assumption 2 was satisfied required very little structure on the model. An exclusion restriction was sufficient.[14] On the other hand the demands on the data are strong in the sense that for any $z \in \mathcal{Z}$, $\Delta(z,\pi',\pi)$ is only identified by values of z for which the probability of entering the program is zero. While for some small programs such as government job training programs it may be possible to find such a variable, but for larger "programs" such as college, finding such a variable may be infeasible.

We observe that the problem associated with "identification at infinity" is a special case of the problem of extrapolation in forecasting.

Example 2: Tuition Subsidy

In this example we consider the case of tuition subsidy which influences an individual's decision to attend college. In particular we consider a policy in which a student receives a tuition subsidy of level π' if they choose to attend college. We assume that there is no such policy in existence

[14] Alternatively one could use linear index assumptions.

.

today so $\pi = 0$,[15] and that we have data on tuition levels T faced by different individuals, and possibly other observables X. We also assume that tuition influences an individual's decision about whether to attend college, but does not influence earnings conditional on attending college. An important assumption is that it is only the net tuition and not tuition and subsidy separately that affects the college attendance decision. Thus $Z = (X, T)$ and

$$\begin{aligned} D(Z, \pi) &= D(X, T - \pi), \\ Y_0(Z) &= Y_0(X), \\ Y_1(Z) &= Y_1(X). \end{aligned}$$

The assumption that only the net tuition affects the college attendance decision can be justified in the model where individuals do not distinguish the sources of funding and that the net tuition is known enough in advance so that the attendance decision can be made with enough preparation time. The effect of a policy we measure under this assumption is that corresponds to the subsidy announced well in advance. More generally, the policy we can measure the effect of correspond to whatever the equivalent variation we use.

We assume that the support of T does not depend on X and bounded, $[T_\ell, T_u]$. In this case the set $\mathcal{D}(z, \pi', \pi) \cap \mathcal{Z}_0((x, t), \pi)$ satisfies the following:[16]

$$\begin{aligned} \mathcal{D}(z, \pi', \pi) \cap \mathcal{Z}_0((x, t), \pi) &= \left\{(x, t^*) \in \mathcal{Z} | \Pr\{D(x, t^*) = D(x, t - \pi')\} = 1\right\} \\ &\supseteq \left\{(x, t^*) \in \mathcal{Z} \mid t^* = t - \pi'\right\}. \end{aligned}$$

Clearly in this example any element of $\{(x, t^*) \in \mathcal{Z} \mid t^* = t - \pi'\}$ is also an element of $\mathcal{D}(z, \pi', \pi) \cap \mathcal{Z}_1((x, t), \pi)$. If $T_u > t > T_\ell + \pi'$ then $T_u > t^* = t - \pi' > T_\ell$, so (x, t^*) is in the support of (X, T) which means that this set is not empty. Thus Assumption 2 will hold and $\Delta(z, \pi', \pi)$ is identified.

However if $t < T_\ell + \pi'$, then $t^* = t - \pi < T_\ell$, so (x, t^*) is not in the support of (X, T). In this case Assumption 2 is likely to fail and we can not identify $\Delta(z, \pi', \pi)$. Thus for some values of z, we can identify the policy effect, but for others we can not. This means that we can only partially evaluate the policy, there will be a group of people for whom the effect of the policy is not identified.[17]

[15] Again we assume that this is a policy that only affects a small number of people so there are no general equilibrium effects.

[16] Implicitly we assumed monotonicity of the decision with respect to tuition. If it is not, then $\mathcal{Z}_0((x, t), \pi', \pi)$ includes the right hand side.

[17] Ichimura and Taber (1999) considers obtaining bounds for the impact in these cases.

The intuition here is straight forward. If we have an individual who faces tuition level \$1500 with other covariates x, to identify the effect of a \$1000 tuition subsidy, we need to find other individuals with the same covariates x, but who currently face a tuition level of \$500. If the minimal level of tuition in the data is \$0 then we can find such individuals, but if the minimal level is \$1000 then the effect of the policy change is not identified. In this case the policy effect is not identified for individuals who face a current tuition between \$1000 and \$2000.

In contrast, Assumption 4 does not fail for any interior points of $\mathcal{Z}$. For any interior point t if we choose $\widetilde{\pi} = t - T_\ell$, then when $\lambda < 1$,

$$t > T_\ell + \pi' = T_\ell + \lambda (t - T_\ell)$$

so Assumption 4 will hold. Thus the marginal treatment effects will be identified for all interior points of $\mathcal{Z}$. This is the sense in which the conditions for estimating the marginal treatment effects are weaker then the policy effects.

This case is somewhat special as tuition has two important roles. First, it is the central focus of the policy in that changing tuition levels has exactly the same effect on schooling attendance as changing the tuition subsidy. Second, it acts as an exclusion restriction in that it influences the decision to attend college, but does not influence earnings directly. The combination of these two characteristics allows us to put very little structure on the model but still be able to identify many of the policy effects. While this structure is special, it is not unique. Many programs have either subsidies or eligibility criteria that vary across individuals which may be of interest and these subsidies and criteria typically will not have a direct effect on outcomes.

Example 3: Linear Binary Choice Model

Our two earlier examples are special in that we needed to make only very weak assumptions about the form of $D(Z, \pi)$. Typically we will need to make stronger assumptions in order to verify Assumption 2. We often need an explicit structural model in which the parameters are policy invariant and need the model to predict how entrance to the program depends on the structural parameters. This third example has more structure than in the previous case, but less than in our fourth example.

We assume that program participation is determined by linear index binary choice model for $D(Z,\pi)$,

$$D(Z,\pi) = 1(Z'\beta(\pi) + U),$$

where $Z = (Z_1, Z_2)$ and Z_1 is an exclusion restriction that is independent of Y_1, but Z_2 need not be. We also assume that the relationship between $\beta(\pi)$ and $\beta(\pi')$ is well known in the sense that $\beta(\pi')$ is identified from $\beta(\pi)$.

In this case the key sets take the following form,

$$\begin{aligned}
\mathcal{D}(z,\pi',\pi) &= \left\{z^* \in \mathcal{Z} | \Pr\{D(z,\pi') = D(z^*,\pi)\} = 1\right\} \\
&= \left\{z^* \in \mathcal{Z} | z'\beta\left(\pi'\right) = z^{*\prime}\beta(\pi)\right\}.^{18} \\
\mathcal{Z}_0(z,\pi) \cap \mathcal{Z}_1(z,\pi) &= \left\{z^* \in \mathcal{Z} \mid z_2 = z_2^*\right\}.
\end{aligned}$$

This case turns out to be very similar to the tuition example above. Suppose that the support of Z_1 does not depend on Z_2 and that the support of $Z'\beta(\pi)$ is bounded, $[B_\ell, B_u]$. If $B_\ell < z'\beta(\pi') < B_u$ then if we choose z^* so that $z^{*\prime}\beta(\pi) = z'\beta(\pi')$ and $z_2^* = z_2$ then Assumption 2 will hold. However if $z'\beta(\pi')$ lies outside the support of $Z'\beta(\pi)$, then these assumptions will typically not hold. Thus in many cases Assumption 2 will hold for some of the values of $z \in \mathcal{Z}$ but not all. However, if as above $\pi' = \pi + \lambda\widetilde{\pi}$ for some $\widetilde{\pi}$, and $\lim_{\lambda \downarrow 0} \beta(\pi') = \beta(\pi)$ then for each $z \in \mathcal{Z}$, for some value of λ small enough, $B_\ell < z'\beta(\pi') < B_u$. Thus Assumption 4 will be satisfied under weaker support conditions.

Example 4: Search and Welfare

This example is loosely based on Wallace (1998). Devine and Kiefer (1991) provide an excellent survey of related empirical search models. Consider a women who currently participates in welfare. While on welfare she has the utility,

$$U_A(X, B(\pi)),$$

where $B(\pi)$ is the level of welfare benefits under the current welfare system π and X is observable factors.[19] This woman searches for a new job. The probability of a job arriving in some time

[19] We keep the model simple by abstracting from unobservable heterogeneity.

period is $\rho(X)$. When a job arrives, the wage is drawn from the distribution of wages $F(W;X)$. If the welfare mother chooses to accept the wage W she leaves welfare and receives utility,

$$U_L\left(X,W\right).$$

Under different types of assumptions one could derive the reservation wage $R(X,B\left(\pi\right))$ at which an individual is indifferent between working or not, where R increases with welfare benefits. If she receives at most one offer in a period, the probability that a woman who is on welfare at the beginning of the period works at the end of the period is,

$$\Pr(\text{Working} \mid X,u)=\rho(X)\left\{1-F(R(X,B\left(\pi\right));X)\right\}.$$

Even abstracting from issues about unobservable heterogeneity there is a fundamental identification problem that Flinn and Heckman (1982) point out. If R is bounded from below, it is impossible to distinguish ρ from $1-F$ below that point. In this case, ρ and thus the full structural model are fundamentally unidentified. However, in some cases, one can still evaluate the effects of policy changes.

In terms of our notation above, the observable variables are $Z=(X,B\left(\pi\right))$. The choice variable is welfare participation, the outcome is labor income, and the policy of interest is the welfare benefits. Thus, $D\left(Z,\pi\right)$ denotes welfare participation under policy π and,

$$\begin{aligned} Y_0(Z) &= W \\ Y_1(Z) &= 0. \end{aligned}$$

We assume that $X=(X_1,X_2)$ and that X_1 affects only the reservation wage and not the offered wage distribution or the job offer probability. Suppose we want to change welfare benefits to some new level under new policy π'. The key sets will have the following form,

$$\begin{aligned} \mathcal{D}(z,\pi',\pi) &= \left\{z^*\in\mathcal{Z}|\Pr\{D(z,\pi')=D(z^*,\pi)\}=1\right\} \\ &= \left\{z^*\in\mathcal{Z}|R(x_1,x_2,B\left(\pi'\right))=R(x_1^*,x_2,B\left(\pi\right))\right\}. \end{aligned}$$

We are worried about the problem that reservation wages may be bounded from below. If the counter-factual reservation wage $R(x,B\left(\pi'\right))$ falls below this bound, then the set $\mathcal{D}(z,\pi',\pi)$ will be empty and we will not be able to achieve identification. This case should depend on whether

the policy under consideration expands benefits or contracts them.[20] If it cuts them back then $R(X, B(\pi')) \leq R(X, B(\pi))$ so for some values of Z we are likely to have problems, however if the proposed policy increases current benefit level then $R(X, B(\pi')) > R(X, B(\pi))$ and we can still identify the impact of the policy even though we can not identify the full model.

3 Relationship with other approaches

The objective of this paper is to present a framework to consider direct estimation of policy impacts. However, since we carry this out in a binary choice framework our work can be linked directly to three strands in the literature of program evaluation. The first strand is the sample selection approach typified by Heckman and Robb (1985). The main criticism of this approach is that it requires strong assumptions to obtain consistent estimation of the parameters of the model (e.g. Lalonde, 1986). We note that the typical parameters studied in the program evaluation literature are not necessarily the parameters we study. As we observed in Example 1, one of the parameters we examine includes the average treatment on the treated parameter studied in the literature as a special case. The condition we place for its identification, in this case, coincides with the standard condition to identify the average treatment on the treated parameter. In this sense our framework can be seen as a generalization of the identification result to allow different types of policies.

A second strand is the instrumental variables or natural experiment approach typified by Imbens and Angrist (1994). The main criticism of this approach is that it either requires very strong assumptions or the coefficients do not converge to the policy relevant parameters of interest.[21] We draw on the natural experiment approach in two ways. First, we study the local effects which are similar in form to the "LATE" parameter defined by Imbens and Angrist (1994). Second, we share the idea of exploiting the variation in the data that are most relevant for the variation we wish to examine.

Our framework extends the natural experiment framework by formally considering a policy parameter separately from the conditioning variables. This allows us to explicitly define the policy impact parameters ex-ante and then to discuss conditions for identification and estimation of such parameters. There are special cases in which the parameters we examine and the LATE parameter

[20] It could be more complicated depending on the support of X.
[21] See Heckman (1997).

coincide which we discuss in the empirical section below.

Another benefit of making the policy parameter explicitly different from the conditioning variable is that we can meaningfully define what we mean by a policy equivalent variation and then make use of economic models to link variation in some variable with a policy under consideration.

A third strand is the matching method typified by Cochran and Rubin (1973), Rosenbaum and Rubin (1983), Heckman, Ichimura, Smith, and Todd (1998), and Heckman, Ichimura, and Todd (1997, 1998). The main criticism of this approach is that the identification condition is generally not testable within its framework and that it is consistent with a model that allows selection on unobservable only under special cases.[22] We draw on this approach by paying closer attention to the distribution of observables than much of the previous literature. In some sense our approach *is* matching except that we use some aspect of an economic model to justify the match rather than the distance of the regressors typically employed in the literature.

To see this consider the tuition subsidy example. In this context, what we want to estimate is, for example

$$\begin{aligned} & E\left(Y\left(t-\pi'\right)-Y\left(t\right)|T=t,X=x\right) \\ = \; & E\left(Y_1\left(x\right)D\left(t-\pi',x\right)+Y_0\left(x\right)\left(1-D\left(t-\pi',x\right)\right)|T=t,X=x\right) \\ & -E\left(Y\left(t\right)|T=t,X=x\right). \end{aligned}$$

The second term of the right-hand side can be identified directly in the data so that we concentrate on identifying the first term. Note that if T is independent with $\{D\left(t,X\right)\}_t$ given X and that T is excluded from outcome variables, then the following equalities holds:

$$\begin{aligned} & E\left(Y_1\left(x\right)D\left(t-\pi',x\right)+Y_0\left(x\right)\left(1-D\left(t-\pi',x\right)\right)|T=t,X=x\right) \\ = \; & E\left(Y_1\left(x\right)D\left(T,x\right)+Y_0\left(x\right)\left(1-D\left(T,x\right)\right)|T=t-\pi',X=x\right). \end{aligned}$$

As the right hand side is identified in the data, the left hand side is. In this sense, the approach can be viewed as matching.

When viewed in this manner our approach is also similar to Manski (1993). In our approach we show how the econometrician can study an individual who faces tuition $t-\pi'$, to learn about the behavior of an individual who faces t if the policy is enacted. Manski models how agents study other individuals to learn about their own outcomes under alternative choices.

[22] See Heckman and Robb (1985).

4 Estimation

4.1 Nonparametric method

We consider estimation of $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$ making use of the identification results discussed earlier. Estimation of the marginal parameters follow directly from estimators of $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$ and hence discussions of them are omitted.

Recall that

$$\Delta\left(z, \pi', \pi\right) = E\left[Y(z, \pi') - Y(z, \pi)\right].$$

As $E[Y(z, \pi)]$ can be estimated using the standard nonparametric regression method, we shall only discuss estimation of the counter factual parameter $E[Y(z, \pi')]$. Let

$$\mathcal{Z}^*\left(z, \pi', \pi\right) \equiv \mathcal{D}(z, \pi', \pi) \cap \mathcal{Z}_0(z, \pi) \cap \mathcal{Z}_1(z, \pi).$$

As our examples show, there are some cases in which set $\mathcal{Z}^*(z, \pi', \pi)$ is known ex-ante and others in which it needs to be estimated. Set $\mathcal{Z}^*(z, \pi', \pi)$ is known in Examples 1 and 2 without estimation of any parameters. However in Examples 3 and 4 we must first estimate some aspects of the model before we can construct $\mathcal{Z}^*(z, \pi', \pi)$. As estimation of set $\mathcal{Z}^*(z, \pi', \pi)$ is case specific, below we assume that the set is known or has been estimated.

Note that under our identification condition, for any z^* in $\mathcal{Z}^*(z, \pi', \pi)$,

$$E\left[Y(z, \pi')\right] = E\left[Y(z^*, \pi)\right]. \tag{1}$$

For each z^*, the right-hand side can be estimated using the standard nonparametric regression method. Thus when $\mathcal{Z}^*(z, \pi', \pi)$ is a singleton, the natural way to estimate $\Delta(z, \pi', \pi)$ is to contrast two nonparametric regression estimators, one centered at z^*, and the other centered at z.

When there are multiple elements in $\mathcal{Z}^*(z, \pi', \pi)$, however, we need to address the manner in which different z^* values would be combined. One possibility is to take a weighted average across the different points. Alternatively, one can exploit the following equality,

$$E\left[Y(z, \pi')\right] = E\left[Y(Z^*, \pi) | Z^* \in \mathcal{Z}^*\left(z, \pi', \pi\right)\right] \tag{2}$$

which is an implication of $E[Y(z, \pi')] = E[Y(z^*, \pi)]$. Note that it can happen that the dimension of vector z^* is higher than the dimension of the smallest linear space that includes $\mathcal{Z}^*(z, \pi', \pi)$.

The index model discussed above is an example. In that case, the weighting approach involves averaging of the higher dimensional nonparameteric estimation than that involved for an implementation based directly on this equation. Since the case with singleton $\mathcal{Z}^*(z, \pi', \pi)$ is a special case of the latter method, we define the estimator of $\Delta(z, \pi', \pi)$ as an estimator of,

$$E\left[Y(Z^*, \pi) | Z^* \in \mathcal{Z}^*\left(z, \pi', \pi\right)\right] - E\left[Y(z, \pi)\right].$$

Based on the relationship,

$$\Delta_c(z, \pi', \pi) = \frac{\Delta(z, \pi', \pi)}{|\Pr\{D(z, \pi') = 1\} - \Pr\{D(z, \pi) = 1\}|},$$

the problem of estimating $\Delta_c(z, \pi', \pi)$ is reduced to the problem of estimating $\Pr\{D(z, \pi') = 1\}$, which, by the same reasoning can be estimated using,

$$E\left[D(z, \pi')\right] = E\left[D(Z^*, \pi) | Z^* \in \mathcal{Z}^*\left(z, \pi', \pi\right)\right].$$

Thus we define the estimator of $\Delta_c(z, \pi', \pi)$ as an estimator of,

$$\frac{E\left[Y(Z^*, \pi) | Z^* \in \mathcal{Z}^*(z, \pi', \pi)\right] - E\left[Y(z, \pi)\right]}{|E\left[D(Z^*, \pi) | Z^* \in \mathcal{Z}^*(z, \pi', \pi)\right] - \Pr\{D(z, \pi) = 1\}|}.$$

4.2 Alternative Parameters

When the dimension of the linear space that includes $\mathcal{Z}^*(z, \pi', \pi)$ is high, we face the curse of dimensionality problem. There are two approaches established in the literature to deal with the curse of dimensionality. The first is to average the pointwise estimates as is done in this subsection. The second is to exploit parametric restrictions researchers are willing to impose which we discuss in the next subsection.

The averaging idea is to give up on estimating $\Delta(z, \pi', \pi)$ or $\Delta_c(z, \pi', \pi)$ and instead condition on a larger set of observables which can be estimated with smaller variance. For set $S \subset \mathcal{Z}$ we generalize the notation so that

$$\begin{aligned}
\Delta(S, \pi', \pi) &= E\left(Y(Z, \pi') - Y(Z, \pi) \mid Z \in S\right) \\
\Delta_c(S, \pi', \pi) &= E\left(Y(Z, \pi') - Y(Z, \pi) \mid Z \in S, d(Z, \pi') \neq d(Z, \pi)\right) \\
&= \frac{\Delta(S, \pi', \pi)}{\Pr\left(d(Z, \pi') \neq d(Z, \pi) \mid Z \in S\right)}
\end{aligned}$$

The choice of set S is dictated by two considerations. The first consideration is to define the group one is interested in studying. The second consideration is to define the subgroup of which

one can expect to estimate the impact. For example, in order to estimate $\Delta(z, \pi', \pi)$, we need to be able to estimate both $E[Y(Z^*, \pi)|Z^* \in \mathcal{Z}^*(z, \pi', \pi)]$ and $E[Y(z, \pi)]$ at the same time. This requires that the Lebesgue density of z and $\mathcal{Z}^*(z, \pi', \pi)$ be bounded away from 0 at all points in S.[23]

By the iterated expectation

$$E\left\{\Delta\left(Z, \pi', \pi\right)|Z \in S\right\} = E\left\{E\left[\Delta\left(z, \pi', \pi\right)|Z = z\right]|Z \in S\right\}.$$

Thus the averaged parameter can be obtained as the averages of the pointwise estimators.

Note that

$$\begin{aligned} & E\left\{\Delta\left(Z, \pi', \pi\right)|Z \in S\right\} \\ = \; & E\left\{E\left[Y(Z^*, \pi)|Z^* \in \mathcal{Z}^*\left(Z, \pi', \pi\right)\right]|Z \in S\right\} - E\left\{Y\left(Z, \pi\right)|Z \in S\right\} \end{aligned}$$

and that in some cases alternative methods for estimation can be considered because

$$E\left\{E\left[Y(Z^*, \pi)|Z^* \in \mathcal{Z}^*\left(Z, \pi', \pi\right)\right]|Z \in S\right\}$$

simplifies. For example consider a case, when $\mathcal{Z}^*(Z, \pi', \pi)$ is a singleton and for a given measurable function ζ, $Z^* = \zeta(Z, \pi', \pi)$. Then under Assumption 1,

$$\begin{aligned} & E\left\{E\left[Y(Z^*, \pi)|Z^* \in \mathcal{Z}^*\left(Z, \pi', \pi\right)\right]|Z \in S\right\} \\ & = E\left\{E\left[Y\left(\zeta\left(Z, \pi', \pi\right), \pi\right)|Z\right]|Z \in S\right\} \\ & = E\left\{Y\left(\zeta\left(Z, \pi', \pi\right), \pi\right)|Z \in S\right\}. \end{aligned}$$

There are a number of ways to estimate the model based on this equality. Two possibilities follow,

1. One could first estimate $Y(Z^*, \pi)$ using nonparametric regression. Once we have done that, for each Z we can construct $Y(\zeta(Z, \pi', \pi), \pi)$ and then average it over the data.

2. Letting g be the unconditional density of Z and g_S the distribution of Z conditional on $Z \in S$, we could use the density weighted average,

$$\begin{aligned} E\left\{Y\left(\zeta(Z, \pi', \pi), \pi\right) \mid Z \in S\right\} &= \int E(Y \mid Z^* = \zeta(Z, \pi', \pi)) g_c(Z) dZ \\ &= \int E(Y \mid Z^*) \frac{g_c(\zeta^{-1}(Z^*, \pi', \pi))}{g(Z^*)} \frac{\partial \zeta^{-1}(Z^*, \pi', \pi)}{\partial Z^*} g(Z^*) dZ^* \\ &= E\left(Y \frac{g(\zeta^{-1}(Z^*, \pi', \pi))}{g(Z^*)} \frac{\partial \zeta^{-1}(Z^*, \pi', \pi)}{\partial Z^*}\right). \end{aligned}$$

[23] The same issue is considered in Heckman, Ichimura, Smith, and Todd (1998).

Our empirical section below clarifies and implements this idea.

Note that in either case one needs to estimate either conditional mean of $Y(Z^*, \pi)$ or density $g(Z^*)$ both of which are high dimensional objects. However the averaging should yield higher convergence rates than for the parameters discussed in the previous section.

4.3 Parametric Restrictions

The second dimension reduction approach requires some form of parametric restrictions. Typically they are placed on the conditional mean function of the outcome and choice variables. For example, Let X denote the exogenous variables that affect both outcome and choice variables and $T(\pi)$ denote the policy related variables that only affect program participation. The potential outcome equations and choice variables may be specified as

$$\begin{aligned} Y_1 &= \alpha_1 + X'\beta_1 + V_1 \\ Y_0 &= \alpha_0 + X'\beta_0 + V_0 \\ D &= 1\left\{X'\theta_X + T(\pi)'\theta_Z + U \geq 0\right\} \end{aligned}$$

where $(X, T(\pi))$ and (V_0, V_1, U) are independent. With this restriction, $\Delta(z, \pi', \pi)$ and $\Delta_c(z, \pi', \pi)$ are both functions of three indices, $X'\beta_1$, $X_0'\beta_0$, and $X'\theta_X + T'(\pi)\theta_Z$. If we further restrict $\beta_0 = \beta_1$, then both are functions of two indices. The last formulation is the one we employ in the example below.

5 Empirical Example

5.1 Methodology

In this section we estimate the impact of a tuition subsidy following the example above. We assume that in the current state of the world there is a tuition level T_i in place for each individual i. A new tuition subsidy of the amount π' is proposed while there is currently no subsidy ($\pi = 0$). We take this to be a state level or narrowly targeted subsidy to rule out general equilibrium effects of the type discussed by Heckman, Lochner, and Taber (1998). Our goal is to estimate the impact that this subsidy will have on earnings.

We measure this impact using the conditional parameter,

$$\Delta_c\left(\mathcal{Z}\left(\pi'\right), \pi', \pi\right),$$

where $\mathcal{Z}(\pi')$ is a subset of the support of the observables. We assume that Z is composed of tuition, T, and other conditioning variables, X. The key assumption is that expected earnings conditional on (T, X) in the counter-factual world is equivalent to expected earnings conditioning on (T', X) in the current $(\pi = 0)$ world where $T' = T - \pi'$.

Consider estimation of the policy counter-factual $E(Y(T, X, \pi) | (T, X) \in \mathcal{Z}(\pi'))$. As in our example above we can write,

$$\begin{aligned} E\left(Y(T,X,\pi)\,|(T,X)\in\mathcal{Z}\left(\pi'\right)\right)&\Pr\left\{(T,X)\in\mathcal{Z}\left(\pi'\right)\right\} \\ &= \int_{\mathcal{Z}(\pi')} E\left(Y(t,x,\pi')\right) g(t,x)dtdx \\ &= \int_{\mathcal{Z}(\pi')} E(Y(t-\pi',x,0))g(t,x)dtdx \\ &= \int_{\mathcal{Z}(0)} E(Y(t^*,x,0))g(t^*+\pi',x)dt^*dx \\ &= \int_{\mathcal{Z}(0)} E(Y(t^*,x,0))\left(\frac{g(t^*+\pi',x)}{g(t^*,x)}\right) g(t^*,x)dt^*dx. \end{aligned}$$

where $t^* = t - \pi'$. Given an estimate $\hat{g}$ of the density g, we can create the sample analogue of this expression,

$$E\left(Y\left(T,X,\pi'\right)|(T,X)\in\mathcal{Z}\left(\pi'\right)\right) = \frac{\sum_{i=1}^{N} Y_i\left(\frac{\widehat{g}(T_i+\pi',X_i)}{\widehat{g}(T_i,X_i)}\right)1((T_i,X_i)\in\mathcal{Z}(0))}{\sum_{i=1}^{N}\left(\frac{\widehat{g}(T_i+\pi',X_i)}{\widehat{g}(T_i,X_i)}\right)1((T_i,X_i)\in\mathcal{Z}(0))}.$$

We can then use the same method to estimate the counter-factual college attendance,

$$E\left(D\left(T,X,\pi'\right)|(T,X)\in\mathcal{Z}\left(\pi'\right)\right),$$

and combine the estimates to form the parameter.

As we discussed, with a large number of covariates, the nonparametric strategy faces the curse of dimensionality problem. In particular, a high dimensional density function needs to be estimated for the case above. For this reason we consider a two dimensional index model as discussed above to obtain,

$$\begin{aligned} E\left\{Y(X,T,0)\,|X,T\right\} = &P\left(X_1'\gamma - T\right) g_1\left(X_1'\gamma - T, X_2'\beta\right) \\ &+ \left(1 - P\left(X_1'\gamma - T\right)\right) g_0\left(X_1'\gamma - T, X_2'\beta\right). \end{aligned}$$

where T represents tuition, and $X_1$and X_2 are composed variables in X that will typically have some elements in common. Note that we assume common β in functions g_0 and g_1.

This specification arises naturally in the standard selection model with additive error terms but that is not necessary to justify this specification. Under the standard additive error specification, one way of estimating the semiparametric model would be to first estimate the entire model including the full joint distribution of (u, ε_1) and the joint distribution of (u, ε_2), and then simulate the effect. While many semiparametric estimators do a good job estimating the slope parameters, they often perform poorly when estimating the joint distribution of the error terms.[24] Our approach is to estimate this parameter directly and avoid estimating the distribution of the error terms. In practice, often estimates of the distribution are represented by a low dimensional flexible form. This faces the challenge of approximating the full distribution by a small number of parameters. It seems reasonable that they may do a poor job in estimating the relatively small part of the joint distribution that is relevant for the policy simulation. We avoid this problem by essentially using only the part of the distribution that is relevant.

Under a maintained independence assumption there are a number of different methods one could use. One possibility would be to estimate the second stage model,

$$\begin{aligned} E(Y(X,T,0) \mid D,X,T) =& X_2'\beta + D(X,T,0)E\left(\varepsilon^1 \mid X,T,X_1'\gamma + T + u > 0\right) \\ &+ (1 - D(X,T,0))\, E\left(\varepsilon^0 \mid X,T,X_1'\gamma + T + u < 0\right) \\ \equiv& X_2'\beta + D(X,T,0)g_1\left(X_1'\gamma - T\right) + (1 - D(X,T,0))\, g_2\left(X_1'\gamma - T\right). \end{aligned}$$

One could then simulate the counter-factual since,

$$E\left(Y\left(Z,\pi'\right) - Y\left(Z,0\right) \mid D(Z,\pi') > D(Z,0), Z\right) =$$
$$\frac{p'g_1\left(X_1'\gamma - T + \pi'\right) + (1-p')\, g_2\left(X_1'\gamma - T + \pi'\right) - pg_1\left(X_1'\gamma - T\right) - (1-p')\, g_2\left(X_1'\gamma - T\right)}{p' - p}$$

where $p\prime = \Pr(X_1'\gamma - T + \pi' + u > 0)$ and $p = \Pr(X_1'\gamma - T + u > 0)$. This is essentially another reduced form approach to the problem. To show that this parameter is identified, you would need to show that $g_1\left(X_1'\gamma - T + \pi'\right)$ and $g_0\left(X_1'\gamma - T + \pi'\right)$ are identified which uses precisely the type of identification strategy we provide above. It typically does not require that the joint distribution of the error terms be globally identified. The primary advantage of our estimator versus this one is that we do not rely on the additive independent error terms. We use it only for convenience.

[24] See for example Heckman and Singer (1984) or Cameron and Taber (1996).

To simplify the exposition we define

$$\begin{aligned} T^* &\equiv Z^{'}\gamma + T \\ X^* &\equiv X'\beta \\ Y\left(T^*, X^*, \pi\right) &= Y\left((X, Z, T), \pi\right) \\ D\left(T^*, \pi\right) &= D\left((X, Z, T), \pi\right) \end{aligned}$$

Specifically to estimate the parameter we use the following procedure:

0. Dimension Reduction: Estimate γ and β from Semiparametric Least Squares,[25] call them $\hat{\gamma}$ and $\widehat{\beta}$ respectively.[26]

1. Estimate $g(T^*, X^*)$ with a kernel density estimator using $\left(Z\hat{\gamma}, X\hat{\beta}\right)$, call it $\hat{g}$.

2. For the trimming value α_t, find g_t such that $\hat{g}(t_i, x_i) > g_t$ with fraction $1 - \alpha_t$ in the data.

3. Define $\mathcal{Z}\left(\pi'\right) = \{(t, x) \in \Re^2 : \hat{g}(t - \pi', x) > g_t\}$.[27]

4. Construct estimates,

$$\begin{aligned} \hat{E}\left(Y\left(T^*, X^*, \pi\right) | (T^*, X^*) \in \mathcal{Z}\left(\pi'\right)\right) &= \frac{\sum_{i=1}^N Y_i \left(\frac{\widehat{g}(T_i^* + \pi', X_i^*)}{\widehat{g}(T_i^*, X_i^*)}\right) 1\left(\widehat{g}(T_i^*, X_i^*) > g_t\right)}{\sum_{i=1}^N \left(\frac{\widehat{g}(T_i^* + \pi', X_i^*)}{\widehat{g}(T_i^*, X_i^*)}\right) 1\left(\widehat{g}(T_i^*, X_i^*) > g_t\right)} \\ \hat{E}\left(Y\left(T^*, X^*, 0\right) | (T^*, X^*) \in \mathcal{Z}\left(\pi'\right)\right) &= \frac{\sum_{i=1}^N Y_i 1(\widehat{g}(T_i^* - \pi', X_i^*) > g_t)}{\sum_{i=1}^N 1(\widehat{g}(T_i^* - \pi', X_i^*) > g_t)} \\ \hat{E}\left(D\left(T^*, \pi\right) | (T^*, X^*) \in \mathcal{Z}\left(\pi'\right)\right) &= \frac{\sum_{i=1}^N D_i \left(\frac{\widehat{g}(T_i^* + \pi', X_i^*)}{\widehat{g}(T_i^*, X_i^*)}\right) 1\left(\widehat{g}(T_i^*, X_i^*) > g_t\right)}{\sum_{i=1}^N \left(\frac{\widehat{g}(T_i^* + \pi, X_i^*)}{\widehat{g}(T_i^*, X_i^*)}\right) 1\left(\widehat{g}(T_i^*, X_i^*) > g_t\right)} \\ \hat{E}\left(D\left(T^*, 0\right) | (T^*, X^*) \in \mathcal{Z}\left(\pi'\right)\right) &= \frac{\sum_{i=1}^N D_i 1(\widehat{g}(T_i^* - \pi', X_i^*) > g_t)}{\sum_{i=1}^N 1(\widehat{g}(T_i^* - \pi', X_i^*) > g_t)} \end{aligned}$$

[25] See Ichimura (1993).

[26] That is, we estimate β by finding the minimizer of

$$\sum_{i=1}^{N_1} (Y_{1i} - X_i'\beta - \hat{a}_1(Z_i'\gamma))^2 + \sum_{i=1}^{N_0} (Y_{0i} - X_i'\beta - \hat{a}_0(Z_i'\gamma))^2$$

where for $j = 0, 1$, $\hat{a}_j$ is obtained from a kernel regression of $(Y_j - X'\beta)$ on $Z'\gamma$.

[27] $\mathcal{Z}\left(\pi'\right)$ picks up both the fact that we have trimmed out some of the observations with low density and have eliminated the values of the observables for which the parameter is not observable. In practice the parameter is identified for more than 99 percent of the sample. This is driven in large part by the index model we are using. If we relied solely on variation from the tuition data, the identification problem would be much more severe.

5. Put the various terms together so that

$$\widehat{\Delta}_c\left(\mathcal{Z}\left(\pi'\right),\pi',\pi\right)=\frac{\hat{E}\left(Y\left(T^*,X^*,\pi\right)|(T^*,X^*)\in\mathcal{Z}\left(\pi'\right)\right)-\hat{E}\left(Y\left(T^*,X^*,0\right)|(T^*,X^*)\in\mathcal{Z}\left(\pi'\right)\right)}{\hat{E}\left(D\left(T^*,\pi\right)|(T^*,X^*)\in\mathcal{Z}\left(\pi'\right)\right)-\hat{E}\left(Y\left(T^*,0\right)|(T^*,X^*)\in\mathcal{Z}\left(\pi'\right)\right)}$$

This parameter will have the interpretation as a return to college. That is suppose we estimated a value $\hat{\Delta}_c\left(\pi'\right)=0.30$ where the dependent variable Y is log wages. We would interpret this parameter as implying that those people who are induced to attend college by a tuition subsidy of π' will see their wages grow by .30 log points in expectation.

5.2 Comparison with Instrumental Variables

It may be useful to compare this estimator to instrumental variables before estimating the model. For the sake of exposition we consider a case in which there are no conditioning variables so we focus only on the exclusion restriction T. The simplest case is in which there is no heterogeneity in the treatment effect so that for everyone in the population $Y^1-Y^0=\alpha$. In this case both estimators will yield consistent estimates of α.

In the case in which there is heterogeneity in the treatment effect, but tuition T, takes on only two values say t_1 and t_2, where $t_1<t_2$,the IV method using tuition as an instrumental variable will converge to LATE as Imbens and Angrist (1994) have shown. In this case LATE can be interpreted as the impact of the tuition subsidy policy of (t_2-t_1) which would be given to anyone facing t_2 originally. In this case LATE and $\Delta_c(\pi',\pi)$ would correspond exactly and the estimators would correspond exactly (once we have replaced the densities in the derivation above with probabilities).

The more realistic case in this example is where tuition takes on more than two values. Angrist, Graddy, and Imbens (1997) and Heckman and Vytlacil (1999) use alternative formulations to interpret the probability limit of the linear instrumental variables estimate in this type of case. For their application, Angrist, Graddy and Imbens show that it is a weighted average of derivatives of the demand functions. Heckman and Vytlacil show that it is a weighted average of Local IV parameters. In both cases the weights are very hard to interpret making the estimate very hard to interpret. In contrast to IV, when it is identified our estimator produces a parameter that is easy to interpret.

As an example consider the tuition result where in the current world tuition T measured

in thousands of dollars takes on J values $S = (t_1, ..., t_J)$ where $t_1 < \cdots < t_J$, with marginal probabilities $(P_1, ..., P_J)$. Assume for simplicity that these tuition values are equally spaced with Δ. In this case there are a number of different policy counter factuals we can identify. One is a Δ thousand dollars subsidy to anyone who would pay tuition in the current state of the world except for the minimum tuition t_1. Let $\tilde{S} = S \setminus \{t_1\}$. In this case our conditional parameter,

$$\begin{aligned}\Delta_c(\tilde{S}, \pi', \pi) &= E\left(Y\left(\pi'\right) - Y(\pi) \mid \tilde{S}, d(T, \pi') \neq d(T, \pi)\right) \\ &= \sum_{j=2}^{J} E\left(Y_1 - Y_0 \mid d(T, \pi') \neq d(T, \pi), T = t_j\right) \Pr\left(T = t_j \mid \tilde{S}, d(T, \pi') \neq d(T, \pi)\right).\end{aligned}$$

Note that

$$\Pr\left(T = t_j \mid \tilde{S}, d(T, \pi') \neq d(T, \pi)\right) = \frac{\Pr\left(d(T, \pi') \neq d(T, \pi) | T = t_j, \tilde{S}\right) \Pr\left(T = t_j | \tilde{S}\right)}{\Pr\left(d(T, \pi') \neq d(T, \pi) | \tilde{S}\right)}.$$

Thus our policy parameter is a weighted average of something related to the $LATE$ parameters and the weights are easily interpretable. For each tuition level the weight is the ratio of population affected at that particular tuition level to the population affected at some level of tuition. The key to the interpretability of the weights is the policy equivalent variation we introduced in this paper. The weight we would use is automatically adjusted appropriately when we define the policy of interest and its policy equivalent variation. Using this framework, at each level of tuition, we can hypothesize individuals choosing different schooling as they face two distinct policies. In the $LATE$ framework, in order for individuals to choose different schooling, they need to face two distinct tuitions. Thus the LATE parameter needs to be defined across two different tuition levels. When there are more than two potential outcomes, then, the weight takes a complicated form that is hard to make sense.

To see this we explicitly derive the weight for IV when there are more than 2 points in the support of the IV. We can show that

$$\begin{aligned}\beta_{IV} &= \frac{cov(T, Y)}{cov(T, d)} \\ &= \frac{\sum_{j=1}^{J} \sum_{k=1}^{J} \left[E\left(Y | T = t_j\right) - E\left(Y | T = t_k\right)\right] t_j \Pr\left(T = t_j\right) \Pr\left(T = t_k\right)}{\sum_{j=1}^{J} \sum_{k=1}^{J} \left[\Pr\left(d = 1 | T = t_j\right) - \Pr\left(d = 1 | T = t_k\right)\right] t_j \Pr\left(T = t_j\right) \Pr\left(T = t_k\right)} \\ &= \frac{\sum_{j=1}^{J} \sum_{k<j} \left[E\left(Y | T = t_j\right) - E\left(Y | T = t_k\right)\right] \left(t_j - t_k\right) \Pr\left(T = t_j\right) \Pr\left(T = t_k\right)}{\sum_{j=1}^{J} \sum_{k<j} \left[\Pr\left(d = 1 | T = t_j\right) - \Pr\left(d = 1 | T = t_k\right)\right] \left(t_j - t_k\right) \Pr\left(T = t_j\right) \Pr\left(T = t_k\right)}.\end{aligned}$$

As Imbens and Angrist (1994) showed, for any t and $t' \in S$

$$\begin{aligned} &E\left(Y|T=t\right)-E\left(Y|T=t'\right) \\ &\quad =E\left(Y_1-Y_0|d\left(t\right)>d\left(t'\right)\right)\Pr\left(d\left(t\right)>d\left(t'\right)\right)-E\left(Y_1-Y_0|d\left(t\right)<d\left(t'\right)\right)\Pr\left(d\left(t\right)<d\left(t'\right)\right) \end{aligned}$$

and that under monotonicity

$$\begin{aligned} &E\left(Y|T=t\right)-E\left(Y|T=t'\right) \\ &\qquad\qquad =E\left(Y_1-Y_0|d\left(t\right)\neq d\left(t'\right)\right)\left[\Pr\left(d=1|T=t\right)-\Pr\left(d=1|T=t'\right)\right]. \end{aligned}$$

Thus

$$\beta_{IV}=\sum_{j=1}^{J}\sum_{k<j}E\left(Y_1-Y_0|D\left(t_j\right)\neq D\left(t_k\right)\right)w\left(t_j,t_k\right)$$

where,

$$w\left(t_j,t_k\right)=\frac{\left[\Pr\left(d=1|T=t_j\right)-\Pr\left(d=1|T=t_k\right)\right]\left(t_j-t_k\right)\Pr\left(T=t_j\right)\Pr\left(T=t_k\right)}{\sum_{l=1}^{J}\sum_{m<l}\left[\Pr\left(d=1|T=t_l\right)-\Pr\left(d=1|T=t_m\right)\right]\left(t_l-t_m\right)\Pr\left(T=t_l\right)\Pr\left(T=t_m\right)}.$$

This weight can be interpreted as the contribution of the support points t_j and t_k in the overall covariance of d and T.[28] But why we should weight this way is much more difficult to justify.

5.3 Results

We now turn to the empirical exercise of estimating the effect of tuition subsidies on wages. We use data from the National Longitudinal Survey of Youth using a specification very similar to Cameron and Taber (1999).[29] Our experimentation indicates that tuition has a weak effect in the first stage, so we use tuition as well as the presence of a four year college in the county as exclusion restrictions.[30] In order to be consistent with our model above we choose schooling to be a binary variable indicating whether the individual attended college. Thus in terms of the notation above, D_i is an indicator of whether the student attended college, π indexes different levels of tuition subsidies, and Y_i is the log wage of individual i.

[28] We have derived this estimator as a weighted average of all of the local average treatment effects. Alternatively we could have derived it as a weighted average of the local-Late variabes, $E\left(Y(1)-Y(0)\mid D(t_j)\neq D(t_{j+1})\right)$ .This gives weights that are even harder to interpret (at least for us).

[29] Details about the data are provided there.

[30] Kane and Rouse (1993) also use tuition as an exclusion restriction in estimating the returns to schooling, and Card(1995) uses the presence of a college.

As we discussed, our tuition variable is measured by the average tuition of 2 year colleges in the state in which the individual lived at age 17. We need to assume that this variable and the unobservable in the choice equation are independent given the conditioning variables. Because we expect the tuition variable to be correlated with some state characteristics which also can be correlated with the individual characteristics, we condition on five background characteristics: race, parental education level, AFQT score, mean local income variables, and number of siblings. In addition, as we expect shorter experience for college graduates at the same age, we also condition on experience.

An issue that arises here as in other applications is the choice of bandwidth for the density g. We used the following procedure: After estimating $\hat{\gamma}$ and $\hat{\beta}$ we replace values for Y so that $Y_1 = 1$ and $Y_0 = 0$ for all individuals in the sample. We first choose a bandwidth for the first dimension. We then experiment with alternative values of the bandwidth of the second dimension so that the estimator of $\widehat{\Delta}_c\left(\mathcal{Z}\left(\pi'\right), \pi', \pi\right)$ on the simulated data is one. We have experimented with alternative values of the first and second dimension around those points and find that the results are not very sensitive to the bandwidth choices.[31]

The empirical results are presented in Tables 2 and 3. It should be kept in mind that these parameters represent the effect of attending college, not the return to a year of college. For comparison, in the first row of Table 2 we present the ordinary least square estimate of the returns to college and in the second we present the result from instrumental variables, instrumenting with tuition and with a dummy variable that indicates whether there is a college present in the county. We then estimate a selection model using a Heckman two step method and use that model to simulate the effect of several levels of tuition subsidies.We find that the selection results are lower than the OLS estimates, and that the IV estimates are higher. In results not reported, when we use tuition alone we find that the IV estimates are much higher than OLS, while using presence of a college yields estimates of approximately 0.17.

In Table 3 we present the estimates of the policy simulations using the methodology outlined above. As one can see, these estimates are fairly close to the IV results particularly for the larger subsidies. The \$100 yields a somewhat larger return of 0.410. These results suggest that students closer to the margin of whether to attend college have higher returns than others. There are a lot of caveats in interpreting these results. While most of these problems could be addressed, we view

[31] Changing a bandwidth by a factor of 2 typically yields a change in the estimated effect of approximately .02.

this exercise as an example of what one could do using these methods, rather than as an empirical exercise unto itself. Thus, for the sake of brevity we will refrain from a lengthy discussion of the many issues that arise.

6 Conclusions

When computational capacity is limited it is natural to construct and estimate a parsimonious model and then to use the result in many ways. The structural estimation approach shares this "estimate once, use many times" approach but takes advantage of the increased computational capacity by making the model more realistic in many dimensions in the way it was not possible before. In this paper we discuss an alternative way to take advantage of increased computational capacity. Our approach is to construct and estimate a different model tuned for each of a particular parameter we wish to estimate. We discuss this approach in the context of measuring policy impacts.

We present a framework to directly estimate the impact of a new policy using a reduced form approach. We provide precise conditions under which the policy counter-factual can be estimated directly. This requires essentially three types of conditions. First, it requires some structure to be placed on the problem. Second, it requires an exclusion restriction. Third, it requires support conditions on the data.

Our results are applicable to ex-ante as well as ex-post policy analysis. To make this point, we have considered estimation of a new policy effect using data generated under old policy regime.

We also presented an estimator that uses these ideas and applied it to the study of tuition policy. In this case the estimator takes the form of a simple density ratio weighted average of the outcome variable. The empirical work finds estimates of the payoff of tuition subsidies that are quite high and that smaller subsidies yield higher returns per individual.

When our goal is simply to estimate a policy impact, this approach improves over two stage methods that first estimate a full structural model and then simulate the policy effect for three reasons. First, there are cases in which the full model is not identified but the policy counter-factual can be identified. Second, we can often impose fewer assumptions and avoid spelling out preferences and the stochastic environment when they are not necessary for identification of the policy effect. Third, estimation is focused on the range of the data that is most informative for

estimating the policy counter-factual.

There are cases in which not all the policy impacts can be identified using the approach we have presented in this paper but some policies impacts are. In this case we need to resort to a more structural or parametric approach for the policy impacts our approach can not identify. Using the policy impact parameters both approaches identify we can examine the specification assumptions behind the more structural approach.

We see a number of extensions of this work. First, the estimator proposed can be formalized and extended to other contexts. Second, we believe the approach itself will prove useful in a wide range of empirical applications. For this purpose it will be useful to consider a decision framework where more than binary choice is involved.

References

Aakvik, A., Heckman, J., and Vytlacil, E. (1999), "Local Instrumental Variables and Latent Variable Models for Estimating Treatment Effects," unpublished manuscript, University of Chicago, 1999.

Angrist, J., and Imbens, G. (1991) "Sources of Identifying Information in Evaluation Models," NBER Technical Working Paper NO. 117.

Angrist, J., Graddy, K., and Imbens, G. (1997), "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," unpublished manuscript.

Angrist, J., and Imbens, G. (1999), "Comment on James J. Heckman, 'Instrumental Variables a Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, *The Journal of Human Resources,* 34, 823-827.

Angrist, J., and Krueger, A., (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106, 979-1014.

Bjorkland, A., and Moffitt, R. (1987) "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *The Review of Economics and Statistics*, 69, 42–49.

Cameron, S. and C. Taber, (1994), "Assessing Nonparametric Maximum Likelihood Models of Dynamic Discrete Choice", unpublished manuscript.

Cameron, S., and Taber, C., (1999) "Borrowing Constraints and the Returns to Schooling," unpublished manuscript.

Card, D., (1995) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," In L. Christofides, E. Grant, and R. Swidinsky eds., *Aspects of Labour Market Behavior: Essays in Honor of John Vanderkamp* (University of Toronto Press, Toronto): 201-222.

Chamberlain, G., (1986) "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Cochran, W., and Rubin, D., (1973), "Controlling Bias in Observational Studies: A Review," *Sankya* 35, 417-446.

Devine, T. and Kiefer, M. (1991), *Empirical Labor Economics.* Oxford University Press, New York.

Flinn, C., and Heckman, J. (1982) "New Methods for Analyzing Structural Models of Labor Force Dynamics," Journal of Econometrics 18, 115–168.

Heckman, J. (1990), "Varieties of Selection Bias ", *American Economic Review.*

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator ", *Journal of Human Resources,* 32, 1–40.

Heckman, J. (1999a), "Instrumental Variables, Response to Angrist and Imbens," *The Journal of Human Resources,* 34, 828-837.

Heckman, J. (1999b), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," NBER Working Paper No. 7333.

Heckman, J., and Honore, B. (1990) "The Empirical Content of the Roy Model," *Econometrica,* 58. 1121–1149.

Heckman, J., Ichimura, H., and Todd, P. (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies,* 64 605–654.

Heckman, J., Ichimura, H., and Todd, P. (1998) "Matching as an Econometric Evaluation Estimator," forthcoming, *Review of Economic Studies.*

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998) "Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA," forthcoming *Econometrica.*

Heckman, J., Lochner, L., and Taber, C. (1998) "General Equilibrium Treatment Effects," *American Economic Review Papers and Proceedings,* 88, 381-386.

Heckman, J., and Robb, R., "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer eds., *Longitudinal Analysis of Labor Market Data.* Cambridge, Cambridge University Press, 1985.

Heckman, J., and Singer, B., "A Method for Minimizing the Impact of Distributional Assumptions in Economic Models for Duration Data," *Econometrica*, 52(1984), 271- 320.

Heckman, J. and Smith, J. (1997), "Evaluating the Welfare State," in Strom, S. (ed.) *Frisch Centenary* (Cambridge: Cambridge University Press).

Heckman, J., and Vytlacil, E. (1999), "Local Instrumental Variables," unpublished manuscript.

Hurwicz, L. (1950), "Prediction and Least Squares," in *Statistical Inference in Dynamic Economic Models* edited by Tjalling Koopmans, John Wiley & Sons, 266-300.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics,* 58, 71-120.

Ichimura, H., and Taber, C. (1999), "Estimation of Policy Effects under Limited Support Conditions," unpublished manuscript.

Imbens, G., and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62.

Kane, T., and Rouse, C., (1993), "Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?", NBER Working Paper #4268.

Lalonde, R. (1986) "Evaluating the Econometric Evaluations of Training Programs Using Experimental Data," *American Economic Review,* 76, 602–620.

Lucas, R. E Jr. (1976) "Econometric policy evaluation: a critique," in *The Phillips curve and labor markets*, edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conference Series on Public Policy 1: pp. 19–46.

Manski, C., and Nagin, D. (1997), "Bounding Disagreements about Treatment Effects: A Case Study of Sentencing and Recidivism," Unpublished Manuscript.

Marschak, J. (1953) "Economic Measurements For Policy and Prediction," in *Studies in Econometric Method* edited by W. Hood and T. Koopmans, John Wiley 1–26.

Rosenbaum, P. and Rubin, D. (1983), "The Central Role of the Propensity Score in Obserational Studies for Causal Effects," *Biometrika,* **70,** 41-55.

Sims, C. (1982) "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, pp. 107–164.

Taber, C. (1999) "The College Premium in the Eighties: Return to College or Return to Ability," Unpublished Manuscript, Northwestern University

Taber, C. (2000) "Semiparametric Identification and Heterogeneity in Dynamic Programming Discrete Choice Models," forthcoming, *Journal of Econometrics*.

Wallace, G., (1998) "Searching for a Way off Welfare," unpublished manuscript, Northwestern University.

Appendix

Proof of Lemma 1: Note that for any $z^* \in \mathcal{Z}_0(z) \cap \mathcal{D}(z;\pi,\pi')$

$$\begin{aligned}
\Pr\left\{Y_0(z) \leq y | D(z,\pi') = 0\right\} &= \frac{E\left\{1\{Y_0(z) \leq y\}1\{D(z,\pi') = 0\}\right\}}{E\left\{1\{D(z,\pi') = 0\}\right\}} \\
&= \frac{E\left\{1\{Y_0(z) \leq y\}1\{D(z^*,\pi) = 0\}\right\}}{E\left\{1\{D(z^*,\pi) = 0\}\right\}} \\
&= \frac{E\left\{1\{Y_0(z^*) \leq y\}1\{D(z^*,\pi) = 0\}\right\}}{E\left\{1\{D(z^*,\pi) = 0\}\right\}} \\
&= \Pr\left\{Y_0(z^*) \leq y | D(z^*,\pi) = 0\right\}.
\end{aligned}$$

The first equality follows from the definition. The second equality follows from Assumption 2. The third equality follows from Assumption 2. The fourth equality follows from the definition. Since the last expression is uniquely determined in observable population, the first expression is also unique which implies the identification result (i). Result (ii) follows from an analogous argument as that for (i). ■

Proof of Theorem 2: We can write $\Delta(z,\pi',\pi)$ as the sum of three separate pieces,

$$\Delta(z,\pi',\pi) = E\left[D(z,\pi')Y_1(z)\right] + E\left[(1 - D(z,\pi'))Y_0(z)\right] - E[Y(z,\pi)].$$

Notice first that $E[Y(z,\pi)]$ is identified directly from the data. Also $E(D(z,\pi)Y_1(z))$ and $E[(1 - D(z,\pi))Y_0(z)]$ are identified using the results from Lemma 1 when the means are finite. ■

Proof of Corollary 3: By Theorem **2**, we know that $\Delta(z,\pi',\pi)$ is identified. Under Assumption 3 $\Pr\{D(z,\pi') \neq D(z,\pi)\}$ is identified, so by the observation in text, $\Delta_c(z,\pi',\pi)$ is identified. ■

Proof of Corollary 4: Under Assumption 4 and by Theorem 2 we know that $\Delta(z,\pi,\pi')$ is identified for all $\pi' = \pi + \lambda\widetilde{\pi}$ for which $\lambda < \lambda^*(z,\widetilde{\pi},\pi)$. Since,

$$\Delta^m(z,\widetilde{\pi},\pi) = \lim_{\lambda \downarrow 0} \frac{\Delta(z,\pi',\pi)}{\lambda}$$

then if $\Delta^m(z,\widetilde{\pi},\pi)$ exists, it is identified . ■

Proof of Corollary 5: Under Assumption 4 and by Corollary 3 we know that $\Delta_c(z,\pi,\pi')$ is identified for all $\pi' = \pi + \lambda\widetilde{\pi}$ for which $\lambda < \lambda^*(z,\widetilde{\pi},\pi)$. Since,

$$\Delta_c^m(z,\widetilde{\pi},\pi) = \lim_{\lambda \downarrow 0} \frac{\Delta_c(z,\pi',\pi)}{\lambda}$$

then if $\Delta_c^m(z,\widetilde{\pi},\pi)$ exits, it is identified . ■

Table 1

Summary Statistics,
Estimates of College Attendance,
and Estimates of Log Wage Equation
Males, National Longitudinal Survey of Youth

Variable	Mean	Standard Deviation	Stage 1‡ Coefficient	Stage 2* Coefficient
Tuition†	0.73	0.42	-6.04	
College in County	0.87	0.34	21.74	
Black	0.31	0.46	9.18	-0.10
Hispanic	0.19	0.39	1.00	-0.03
AFQT Test Score	0.205	22.14	11.79	0.005
Father's Highest Grade	10.46	4.08	-0.59	0.003
Mother's Highest Grade	10.76	3.19	1.45	0.002
Number of Siblings	3.75	2.63	0.19	0.007
Mean Local Income	13.57	2.74	-2.76	0.02
Experience	6.41	3.52		0.07
Experience Squared/100	0.53	0.49		-0.30
Sample Size§			2223	17068

† Tuition is the average tuition of 2 year colleges of the state in which the individual lived at age 17 measured in thousands of 1986 dollars.

‡ The first stage uses Semiparametric Least Squares to estimate the effects of these variables on college attendance.

* The second stage uses Semiparametric Least Squares to estimate a linear log wage equation model where the distribution of the error term is unspecified. The coefficients are restricted to be the same by college, but the conditional expectation of the error term differs.

§ The sample size differs because we have longitudinal data on wages.

Table 2

Estimates of College Return Using Standard Methods
Males, National Longitudinal Survey of Youth

Method	Level of Subsidy	Return to College	Change in College Attendance
Ordinary Least Squares:		0.217	
Instrumental Variables:	(LATE)	0.296	
Sample Selection with Mills's ratio:			
	(ATE)	0.116	1.0
	$ 1000	0.156	0.046
	$ 500	0.160	0.023
	$ 100	0.164	0.005

Table 3

Direct Estimation of Policy Effect
Males, National Longitudinal Survey of Youth

Level of Subsidy	Return to College	Change in College Attendance
$ 1000	0.346	0.055
$ 500	0.354	0.022
$ 100	0.410	0.005