Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluations of Social Programs

James J. Heckman*

University of Chicago and American Bar Foundation

Edward Vytlacil

Stanford University

Revised June 2001

*This research was supported by NSF 97-09-873 and NCHD:40-4043-000-85-261 and a grant from the Donner Foundation. Heckman also received support from the American Bar Foundation.

1 Introduction

The evaluation of public policies is a central task of economics. Two distinct approaches are currently pursued in evaluating microeconomic programs: (a) the structural approach and (b) the treatment effect approach. This essay examines the benefits and limitations of each approach and the role of social experiments in recovering the parameters required to answer different questions using each approach. We consider the policy evaluation problem, the problem of causal inference, and the policy forecasting problem. We establish how treatment effects approximate well-defined causal effects. We present a *ceteris paribus* model-based definition of causal effects that economists have used since the time of Marshall (1961 version of 1890 edition). We compare this definition with a recent model-free definition of causality presented in the statistics literature as the Rubin (1978) model. We argue that the Rubin model is incomplete.

The *policy evaluation problem* is the problem of comparing outcomes of a policy in place with outcomes under alternative policies. *The problem of causal inference* consists of determining which causes affect outcomes and measuring their quantitative importance. The policy evaluation problem is a special case of the problem of causal inference which entails comparisons between a hypothetical state and the observed state where the "causes" are different policies. In the policy evaluation problem, one (or more) of the many possible policy regimes is observed. In the problem of causal inference, all potential states can be hypothetical with none observed. The *policy forecasting problem* is also related to the problem of causal inference. In this problem, the goal is to forecast and evaluate the effects of a new policy never previously experienced. There are two distinct problems subsumed under this general topic: (a) forecasting the effects of a policy experienced and evaluated in one environment that is proposed for application in a different social and economic environment and (b) forecasting the effects of a new policy never previously experienced - one with different characteristics than any policy previously experienced or, in a more tractable version, one with different bundles of the same characteristics than those that characterize in previous policies. These forecasting problems are special cases of the problem of causal inference in which extrapolation from knowledge of currently experienced states is required to forecast and evaluate states not previously experienced.

The recent literature on program evaluation contrasts structural equation methods with treatment effect methods. (See, e.g. Angrist, 2000). The two competing approaches have different objectives. The structural equation approach, which originated in the work of the Cowles Commission, seeks a low dimensional representation of causal functions as defined in Section 2. The goal of this approach is to produce a model that can do many things at once: serve as a framework for causal inference, evaluate the effects of various policies currently in place, and forecast and evaluate the effects of current policies applied to different environments and the effects of new policies never previously experienced. The whole edifice of Cowles Commission econometrics is designed to produce a structural econometric model to address all of these issues using a common set of parameters. Various definitions of exogeneity and their refinements such as weak exogeneity, strong exogeneity, non-Granger causality and super-exogeneity (Engle, Hendry and Richard, 1983; Ericsson, 1995) are required to identify structural parameters and make counterfactual forecasts and policy simulations.

The goals of the treatment effect literature are typically much more limited. It is primarily aimed at the policy evaluation problem. The goal is to evaluate whether an existing policy "works," and not to forecast the effects of the same policy in a new environment or to forecast the effects of a new policy. For this more modest objective, a variety of treatment effects can be defined that do not require for their identification the assumptions of exogeneity or super-exogeneity that are essential for multiple-purpose structural econometrics.[1] However, when analysts seek to move out of sample and consider the policy forecasting problem to estimate the impact of treatments on new environments or to estimate the effects of new treatments never previously experienced, the apparatus of Cowles

[1] As noted below, a cross-sectional analog of a Granger noncausality (or no feedback) requirement for the conditioning variables is needed in order to identify the full effect of an intervention being evaluated.

econometrics comes into its own. Thus much of the contrast between the structural equation approach and the treatment effect approach arises because of the different objectives of the two literatures.

To focus our discussion and make it more concrete, we consider both treatment effects and structural equations in the context of a standard model of labor supply. We consider how different types of randomization and different assumptions about the structure of the labor supply equation identify different parameters which answer different policy questions.

The plan of this paper is as follows. In Section 2, we present a framework for causal inference that is rooted in economic theory. We define various causal effects within the context of a well-defined economic model and introduce the notion of Marshallian causal functions and structural equations. In Section 3, we consider the problem of causal inference at the level of the individual. In Section 4, we reformulate the evaluation problem at the population level. Section 5 establishes relationships among population treatment effects, causal parameters and structural equation models. Section 6 considers the value of structural equations in making policy forecasts. Section 7 discusses the Rubin (1978) - Holland (1986) model and compares it to the econometric model of counterfactuals developed in this paper. Section 8 develops a specific labor supply example and considers the case where treatments are continuous. We discuss the two cases for social experiments in the context of a labor supply example: (a) randomization to identify structural parameters and (b) randomization to identify treatment effects. Section 9 concludes.

2 A Framework For Counterfactuals And Causal Inference

We construct a counterfactual world where we enumerate all possible outcomes. For simplicity, we assume:

(A-1) Two potential states, "untreated" and "treated," with corresponding potential outcomes given by the variables (y_0, y_1), y_0, $y_1 \in \Re$.[2] Extension to more than two potential outcomes is straightforward. Our labor supply example in Section 8 considers a continuum of outcomes.

(A-2) No market or social interactions among agents in the hypothetical world.[3]

(A-3) A static model.[4]

We start off by defining a causal function for each state,

[2] For convenience, we are defining y_0 and y_1 to be elements of $\Re^1$, though the analysis immediately generalizes to the case where they are elements of $\Re^K$ or a more general space.

[3] Heckman, Lochner and Taber (1998, 2000) demonstrate the empirical importance of accounting for these interactions in large scale programs.

[4] Implicit in our analysis is a process evolving in real time but for simplicity we only formally consider a static environment.

$$y_0 = g_0(x)$$

$$y_1 = g_1(x).$$

Letting $\mathcal{D}_i$ be the domain of function g_i, $g_0 : \mathcal{D}_0 \mapsto \Re$, $g_1 : \mathcal{D}_1 \mapsto \Re$ and $\mathcal{D}_i \subseteq \Re^J$.[5] The variables x are causal variables that completely determine the outcomes y_0, y_1. The function g_j describes how each possible vector of causal variables, $x \in \mathcal{D}_j$, is mapped into a resulting outcome. Tracing g_j over different x values reveals how the outcome varies as a function of the causal variables. We do not necessarily impose any cross-equation restriction between g_0 and g_1.[6] We do not impose any restriction on the range of g_0 and g_1 other than that the range of each function is a subset of $\Re^1$.

Causality is in the mind. The arguments in, and functional forms of, g_0 and g_1 are specified by economic theory. In defining the potential outcomes, the elements of x are assumed to be externally specified by a well defined hypothetical variation. External manipulation of causes is one essential idea in the definition of a causal relation. The other essential idea is the stability of the relationship mapping causes into effects. Economic theory produces causal functions in which the inputs (or externally specified x variables) affect outputs. Different conceptual experiments define different causal relations.

[5] For simplicity, we are taking the domain to be a subset of $\Re^J$. The analysis immediately generalizes to allow the domain to be a more general space.

[6] For example, the g_0 and g_1 may be nontrivial functions of different elements of x.

Thus, following Samuelson (1947), consider a microeconomic demand curve where y is the quantity of a good demanded and x is a vector of price, income and preference parameters. In this example, we may think of different x evaluation points set by a scientist through some mechanism which may be randomization or which may be some other deliberate manipulation of inputs. There is a deterministic relationship connecting x and y. In the conceptual experiment where the agent is a price taker, and preferences and incomes are externally specified, $y = g(x)$ is the Marshallian demand curve, and the x are the causal variables. In a different conceptual experiment, the roles of these variables may be reversed. Thus in a choice experiment examining "willingness-to-accept" functions, quantities y may be specified externally and the minimum price one would be willing accept to give up a unit of y (a component of x in the Marshallian demand) is the causal outcome of interest. In this case, variations in y *cause* a component of x to vary, say x_1. Depending on the exact question, the answer to the second problem may, or may not, be derived by inverting the $g(x)$ function specified in the first problem interchanging the roles of y and the component of x, x_1[7]

[7] Thus if income effects are small, $g(x)$ is the utility constant demand function. Varying quantities to produce associated marginal willingness to accept values would entail inverting g to obtain $x_1 = \varphi(y, \tilde{x})$ where $\tilde{x} = (x_2, ..., x_J)$, assuming that a local implicit function theorem is satisfied so $\dfrac{\partial \varphi}{\partial y} = \left(\dfrac{\partial g}{\partial x_1}\right)^{-1}$ where $\dfrac{\partial g}{\partial x_1} \neq 0$. However, if income effects are nonzero, the causal function required to answer the willingness to accept question cannot be obtained simply by inverting g. One would have to derive the Hicksian demand from the Marshallian demand and derive φ from the Hicksian demand.

Causal functions are thus derived from conceptual experiments where externally specified causes are varied. There are as many causal functions as there are conceptual experiments. The specification of these hypothetical external variations is a crucial part of model specification, and lies at the heart of any rigorous definition of causality. The "*ceteris paribus*" variations of Marshall (1961) or the comparative statics variations studied by Samuelson (1947) are examples of these hypothetical variations. An external mechanism is assumed to operate to vary the causes independently of each other. The goal of structural econometric estimation is to determine the g_0 and g_1 functions from sample data but estimation and identification are different problems distinct from the definition of external variables and causal effects which is the task at hand.[8]

In a production function example output $y = F(x)$. If inputs x are externally specified, F is a causal function. To determine the amount of x_1 required to produce at output y holding $(x_2, ..., x_J)$ at prespecified values, one would invert F to obtain $x_1 = M(y, x_2, ..., x_J)$, assuming $\dfrac{\partial F}{\partial x_1} \neq 0$. M is a causal function associated with the conceptual thought experiment of varying $y, x_2, ..., x_J$ which are externally specified while x_1 is determined.

[8] A more abstract approach to the definition of a causal relation that does not require formulation of explicit economic models builds on the work of Simon (1952) and Sims (1977) and specifies properties of the input space (X) and the output space (Y) and their relationship. The crucial idea is that inputs can be manipulated in ways that do not affect the structure of the causal relation but that affect the realized outputs.

Thus consider an abstract space S of possible features of models, both inputs and outputs. Consider two sets of restrictions: $A \subset S$ restricts inputs and $B \subset S$ restricts outputs. Suppose that S is mapped into two spaces: $P_X : A \rightarrow X$; $P_Y : B \rightarrow Y$. Then (A, B) defines a *causal ordering* from X to Y if A restricts X (if at all) but not Y and B restricts Y (if at all) without further restricting X. More formally (A, B), restrictions on S, determine a causal ordering from X to Y iff $P_Y(A) = Y$ and $P_X(A \cap B) = P_X(A)$. Geweke (1984) and Sims (1977) provide examples. The leading example is $S = \{(x, y) \in R^2\}$, $X = a$ (corresponds to A), $x + bx = c$ (corresponds to B). (A, B) is a causal ordering from X to Y because A determines x without affecting y. B, along with A, determines y without further restricting x. There may be many pairs of restrictions on S that produce the same causal ordering. A version of this example with uncorrelated error terms across the two equations produces the Strotz-Wold (1960) causal chain model.

To link our purely deterministic model to the literature on causal models in statistics, we formally define a probability model. Let $(\Omega, \mathcal{A}, P)$ denote a probability space, and let ω denote an element of Ω. We will think of Ω as a set of individuals in a hypothetical superpopulation. Each person has a pair random variables for the two potential outcomes, only one of which is realized at any time in the hypothetical world. The random variables for the potential outcomes are $(Y_0(\omega), Y_1(\omega))$ in the two possible states "0"and "1". We do not impose any restrictions on the support of these variables.[9] Associated with each potential outcome is a random vector $X(\omega)$ of J dimensional explanatory or "causal" variables that are *causes* of $Y_0(\omega)$ and $Y_1(\omega)$. The $X(\omega)$ are the particular causes associated with person ω. The following consistency property ensures that evaluating the causal functions at the random vector $X(\omega)$ produces the random variable $Y(\omega)$,[10]

The Simon-Sims definition of a causal order is for a given pair of restrictions (A, B). The notion of causality is intimately involved with the idea of a stable relationship *i.e.* that if A is changed, the outcome will still be $A \cap B$ with B (the input-output relation) unchanged. Otherwise, when A is changed a different causal ordering may result. To guarantee that this does not occur we require the following condition: for any $A \subset S$ which constrains only X (*i.e.* $P_X^{-1}(P_X(A)) = A$), (A, B) determines a causal ordering from X to Y. (This is sometimes called "$B \subseteq S$" accepts X as input"). Thus a full specification of a causal model entails a description of admissible input processes and the notion that B is unchanged when A is manipulated (and hence the X is changed). The requirement "for any A" can be replaced by "for some subsets A" and the stable causal relationship can be defined with respect to those subsets. For the model to be "correct," the set $B \subseteq S$ must be such that if B accepts X as an input, and when any set $C \subseteq X$ is implemented (A is manipulated), then $P_Y(P_X^{-1}(C) \cap B)$ is "true" *i.e.* in some sense depicts reality.

[9] For example, these random variables may be discrete, continuous, or mixed discrete/continuous.

[10] An alternative definition of a causal model would enumerate all possible realizations of all possible X, Y variables that can be achieved for each person ω. For this definition, we could work with the realized values from all possible worlds to define counterfactuals at the individual level for each possible realization. This approach is notationally more complex so we use the simpler (and more traditional) definition of causal models using variations of X within the support of $X(\omega)$.

(1a) $\quad Y_0(\omega) = g_0(X(\omega))$

(1b) $\quad Y_1(\omega) = g_1(X(\omega))$

where we now impose that g_0 and g_1 are measurable functions.[11] Included among the $X(\omega)$ may be *ex post* "random shocks".

Note that the g_0, g_1 functions are playing two roles in our analysis. They are describing not only how the random vector $X(\omega)$ is functionally related to the random variables $Y_0(\omega)$, $Y_1(\omega)$, but also are specifying what values the outcomes would have taken had the causes $X(\omega)$ taken alternative values. Saying that one random variable is a deterministic function of another is not enough to define a causal function.[12]

By assumption, the support of $X(\omega)$ lies within the domain of the functions, and the support of $Y(\omega)$ lies within the range of the functions. However, the support of $X(\omega)$ need not equal the domain of the functions. The mismatch between the domain of definition of the function and the population support gives rise to the central problem of identification of causal effects, a topic we do not consider in this paper.

Our definition of causal relationships within a well defined economic model differs radically from a recent influential definition in statistics associated with Rubin (1978) that

[11] Assumptions (1a) and (1b) with g_0 and g_1 measurable are equivalent to assuming that Y_0 and Y_1 are measurable $\sigma(X(\omega))$, where $\sigma(X(\omega))$ is the sigma-field generated by $X(\omega)$. See, e.g., Billingsley (1995, Theorem 20.1, p. 255).

[12] Saying that $Y(\omega)$ is measurable with respect to $X(\omega)$ is not enough to define a causal function. If $Y = X + Z$, then $X = Y - Z$. Y is measurable with respect to X and $Z; X$ is measurable with respect to Y and Z.

seeks to define the “effects of causes” rather than the “causes of effects” without specifying a complete theory (Holland, 1986). In our view, a causal relationship is only well defined if a theory relating causes to outcomes is articulated and a mechanism generating variation in the causes is clearly specified. Both steps are pre-statistical and require thought experiments involving counterfactuals in imaginary worlds. Operationalizing the theory with data and identifying causal mechanisms comes later and is only fruitfully conducted after a theory is constructed. We discuss the Rubin model in greater detail in Section 7 after we present our own definition of a causal model.

We now discuss in more detail the use of the causal functions to study causal effects. The causal effect of x_j on $y_i (i = 0, 1)$ is obtained from varying x_j. Thus consider two values of $x_j : x_j'$ and x_j''. For this variation, we define *the causal effect of* x_j *on* y_i, fixing the remaining coordinates of x, as

$$(2) \qquad [\Delta y_i \mid \Delta x_j = x_j' - x_j''] = g_i(x_1, ..., x_j', ..., x_J) - g_i(x_1, ..., x_j'', ..., x_J).$$

This is a definition of a causal effect corresponding to a scientist manipulating causes (the x) to produce effects. Note that, in general, the causal effect of x_j on y_i will depend on the values of x_j as well as values assumed by the other arguments.

For definition (2) to be meaningful requires that the x_j can be independently varied when the other variables are fixed so there are no functional restrictions connecting the arguments. In order to define a causal effect for all values of x_j in the hypothetical domain of the functions g_0, g_1, it is thus required that these variables be variation-free: $(x_1, ..., x_J) \in \mathcal{X}_1 \times ... \times \mathcal{X}_J$.[13]

The inability to vary x_j independently of the other causal variables is the key idea in Holland's (1986) distinction between causal variables and associated variables. Causal variables are variation-free in the sense we use that term; associated variables are not.

If the variables are variation-free, we may define the *person-specific* causal effect of x_j for person ω by evaluating equation (2) at the characteristics of person ω except for the variable being manipulated:

$$g_i(X_1(\omega), ..., X_{j-1}(\omega), x_j', X_{j+1}(\omega), ..., X_J(\omega)) - g_i(X_1(\omega), ..., X_{j-1}(\omega), x_j, X_{j+1}(\omega), ..., X_J(\omega)).$$

If the g_i are differentiable with respect to x, then the *Limit Causal Effect* of x_j on y_i is

$$(3) \qquad \textit{Limit Causal Effect of } x_j \textit{ on } y_i = \frac{\partial y_i}{\partial x_j} = \frac{\partial g_i(x)}{\partial x_j}.$$

These notions of causal effects are exactly the *ceteris paribus* definitions of a causal effect that economists have used since Marshall (1961).[14] These are also the comparative statics

[13] It also possible to define causal parameters for subsets of the domain so that this variation-free condition can be weakened considerably. One could define variation in subsets and could define an entire class of restricted variations, something we do not pursue in this paper.

[14] Thus Marshall writes, "*It is sometimes said that the laws of economics are 'hypothetical'. Of course, like every other science, it undertakes to study the effects which will be produced by certain causes, not absolutely, but subject to the condition that other things are equal and that the causes are able to work out their effects undisturbed. Almost every scientific doctrine, when carefully and formally stated, will be found to contain some proviso to the effect that other things are equal; the action of the causes in question is supposed to be isolated; certain effects are attributed to them, but only on the hypothesis that no cause is permitted to enter except those distinctly allowed for*" (A. Marshall, 1961, p. 36).

causal effects analyzed by Samuelson (1947) and Hicks (1939).[15] For this reason, we define the causal relationships (1a) and (1b) as Marshallian causal functions. Again, if the variables are variation free, we may define on effect x_j for person ω by evaluating equation (3) at the characteristics of person ω.

Treatment effect causal effects are versions of Marshallian causal effects. Define $D(\omega)$ as an indicator denoting whether in a hypothetical population person ω is in regime "1" or not. Thus $D(\omega) = 1$ if $Y_1(\omega)$ is observed and $D(\omega) = 0$ if $Y_0(\omega)$ is observed. Irrespective of whether $D\left(\omega\right) = 1$, the person has an associated vector of potential outcomes $\left(Y_0\left(\omega\right), Y_1\left(\omega\right)\right)$. The "treatment" is like any other cause. For any cause, a complete definition of a causal model requires that a model of manipulation of the cause be specified. We have been implicit about how the X are manipulated. We are now explicit about how $D(\omega)$ can be manipulated. A fully symmetric analysis would present models of manipulation for both X and D.

$D(\omega)$ is assumed to be a random variable determined by the random vector $Z(\omega)$.[16] $Z(\omega)$ may contain some or all of the components of $X(\omega)$, and may contain other variables as well. Let $\mathcal{Z}$ denote the support of $Z(\omega)$. The assumption that $D(\omega)$ is measurable $\sigma(Z(\omega))$ is equivalent to the following representation:

$$(4)\quad D(\omega) = 1 \text{ if } Z(\omega) \in \bar{\mathcal{Z}}, \qquad D(\omega) \ = 0 \text{ if } Z(\omega) \in \mathcal{Z} \setminus \bar{\mathcal{Z}}$$

[15] See Heckman (2000) for a discussion of this definition of a causal effect and its intellectual history.

[16] Formally, $D(\omega)$ is measurable $\sigma(Z(\omega))$, where $\sigma(Z(\omega))$ is the sigma-field generated by $Z(\omega)$.

where $\bar{\mathcal{Z}}$ is a measurable subset of the support of $Z(\omega)$.

By analogy with the previous analysis of Marshallian causal functions, we replace D by d and write the deterministic relationship over a more general domain than just the support of $Z(\omega)$,

$$d = 1 \text{ if } z \in \tilde{\bar{\mathcal{Z}}},\ d = 0 \text{ if } z \in \mathcal{D}_d \setminus \tilde{\bar{\mathcal{Z}}}$$

where $\mathcal{D}_d$ is the domain on which the function is defined and $\bar{\mathcal{Z}} = \tilde{\bar{\mathcal{Z}}} \bigcap \mathcal{Z}$. Notice that this function satisfies the consistency property that the function evaluated at $Z(\omega)$ equals $D(\omega)$.

In this framework, we define the measured outcome, random variable $Y(\omega)$ in a hypothetical population, as:

(5) $\quad Y(\omega) = D(\omega)Y_1(\omega) + (1 - D(\omega))Y_0(\omega).$[17]

In this notation, the causal effect of D on Y for a given evaluation point x due to a manipulation in z is

(6) *Treatment Effect Causal Effect:*

$$\Delta(x) = y_1(x) - y_0(x) = g_1(x) - g_0(x).$$

This expression evaluated at the characteristics of individual ω is

[17] This representation first appeared in the switching regressions literature in econometrics. See Quandt (1972, 1988) or Goldfeld and Quandt (1976). Statisticians sometimes call this the Rubin (1978) model (see, *e.g.* Holland, 1986).

$$\Delta(X(\omega)) = Y_1(\omega) - Y_0(\omega) = g_1(X(\omega)) - g_0(X(\omega)).$$

No new idea is entailed in this definition of a causal parameter beyond the *ceteris paribus* definition of a causal effect economists have used since the time of Marshall. Like the other causal effects previously discussed, these are just the *ceteris paribus* effects of changing one variable by external manipulation while keeping the others constant. Define the causal effect of changing D while holding X constant at $X = x$, through the following equations:

$$h(d(z), y_1(x), y_0(x)) = d(z)y_1(x) + (1 - d(z))y_0(x),$$

and

$$\tilde{h}(z, x) = d(z)y_1(x) + (1 - d(z))y_0(x).$$

Then provided that there is some mechanism for changing D while holding X fixed it is meaningful in the hypothetical world to define the treatment effect causal effect as

$$\Delta(x) = h(1, y_1(x), y_0(x)) - h(0, y_1(x)), y_0(x))$$

or in terms of $\tilde{h}$ as

$$\Delta(x) = \tilde{h}(z, x) - \tilde{h}(z', x)$$

for any $z \in \bar{\mathcal{Z}}$ and $z' \in \bar{\mathcal{Z}}^c$. There is no essential distinction between the causal effects previously defined and the treatment effect causal effect. Both define the causal effects by hypothetical external manipulations of the causal variables in hypothetical worlds. Notice that if there is a functional relationship connecting Z and X then this expression cannot be evaluated for all x given z. X and Z have to be variation free to generate a treatment effect causal effect.

Any definition of a causal parameter is intrinsically incomplete unless a mechanism is defined specifying how the change in the causal variable is implemented. This mechanism is a hypothetical mechanism about a hypothetical world. But a full and rigorous specification of a causal parameter requires specification of both the outcome equations and the mechanism of intervention that makes causes vary.

Suppose that X and Z are not variation free. Then if the only possible intervention inducing a change in D is a change in X, no causal effect of D on Y can be defined. For example, consider a model in which Z and X are the same. This cases arises in a Roy model with all X observed. In the Roy model $D = 1(Y_1 \geq Y_0)$. Since X determines D and (Y_0, Y_1), for some X values $D = 1$ while for others $D = 0$. (This is equation (4) where the $Z = X$). If the only way to vary D is to vary X, then no definition of a *ceteris paribus* causal effect of D on Y is possible. If, on the other hand, we postulate a variable in Z not in X such that Z is variation free, (support of Z given X is unrestricted), we can define the causal parameter at least over the region of Z given X that causes persons to switch from $D = 0$ to $D = 1$.

Statisticians such as Holland (1986) and Rubin (1978) implicitly assume that a randomization assigning persons to D status is available that does not affect outcomes[18]. More generally, some exclusion restriction or instrument inducing people to change treatment states given X must be explicitly postulated as part of any complete definition of a causal parameter in a hypothetical population.

Discussions concerning whether or not a variable is causal are discussions about models of intervention in hypothetical worlds.[19] Whether or not the hypothetical interventions have empirical, operational, counterparts is an entirely separate matter concerning the identification of a causal parameter. This analysis applies symmetrically to all of the variables that are candidate causal variables including the components of X.[20]

As a consequence of (1a), (1b), and (4), the random variables $Y_0(\omega), Y_1(\omega)$ and $D(\omega)$ in a hypothetical population are degenerate given the causes, $X(\omega)$ and $Z(\omega)$, respectively. That is what is meant by a full accounting of causes. Most statisticians implicitly assume an unexplained source of randomness without justifying its origin. A complete causal model justifies the source of randomness in any statistical model. Among the $X(\omega)$ and $Z(\omega)$ may

[18] This is the absence of "randomization bias" as defined by Heckman (1992). The assumed randomization bias plays the role of Z if there is no randomization bias.

[19] See Heckman and Honoré, 1990, for a formal presentation of the Roy model.

[20] Pearl (2000) defines a specific intervention mechanism using a recursive econometric model to define causal parameters. He fixes values of variables by "shutting equations down," *i.e.* fixing one equation to a constant in a chain choice model. In a more general simultaneous equations model there are many possible ways to fix a variable. Accordingly, there are many possible causal effects corresponding to the different ways the variables are fixed at pre-specified values.

be variables that cannot, in principle, be observed. Random variables that are degenerate conditional on the causes arise in many economic models, and are widely used in scientific models. However, such causal models are often dismissed by many applied statisticians even though they do not explicitly specify the sources of randomness in their models.[21] Random variables used to characterize actual samples are rarely degenerate conditional on observed covariates because some causes (components of $X(\omega)$) are usually not observed.[22]

To account for random variables that are nondegenerate conditional on the observed covariates, a common convention in econometrics breaks $X(\omega)$ and $Z(\omega)$ into observed ($X_o(\omega)$, $Z_o(\omega)$) and unobserved components $(X_u(\omega), Z_u(\omega))$. This dichotomy is made with respect to the information available to the econometrician. A different dichotomy would in general be applicable to the agents being analyzed. Many commonly-used evaluation estimators implicitly assume that econometricians have the same relevant information as the agents they analyze. (See Heckman and Vytlacil, 2000b).

[21] The entire controversy over the determinacy or indeterminacy of quantum mechanics centered on the existence of a sufficient array $X(w)$ variables that perfectly predict physical outcomes. See, *e.g.* Bell (1964), Einstein, Podolsky and Rosen (1935) and the essays in Suppes and Zanotti (1996). Holland (1986, 1988) is a good example of the view in the statistics literature that rejects the possibility of representations like (1a), (1b) and (5). Fraser (1979) is an exception to this tradition in statistics. His discussion of structural models and measurement models is exactly what we mean by equation (1a), (1b) and (5) being deterministic functions.

[22] For example, Hansen and Sargent (1990) break the perfect predictability or “stochastic singularity” implied by the economic model they analyze by assuming that observing economists have less information than the agents they model.

This partial observability of causes makes the potential outcomes in a hypothetical population conditional on the observed values nondegenerate random variables. The approach followed is then to work with mean values of potential outcomes given observables. This practice is usually accompanied by an additive separability assumption for the g_0 and g_1 functions:

(7a) $\quad Y_0(\omega) = g_{0o}(X_o(\omega)) + g_{0u}(X_u(\omega))$

(7b) $\quad Y_1(\omega) = g_{1o}(X_o(\omega)) + g_{1u}(X_u(\omega)).$

The error terms for these models are defined as

$$U_0(\omega) = g_{0u}(X_u(\omega)) \text{ and } U_1(\omega) = g_{1u}(X_u(\omega))$$

for measurable functions g_{0u}, g_{1u}. With the further assumptions that $(U_0(\omega),\, U_1(\omega))$ are (mean) independent of $X_o(\omega)$ with zero mean,[23]

(8a) $\quad E(U_0(\omega) \mid X_o(\omega)) = 0$

(8b) $\quad E(U_1(\omega) \mid X_o(\omega)) = 0,$

we thus obtain the mean outcome equations

(9a) $\quad E(Y_0(\omega) \mid X_o(\omega)) = g_{0o}(X_o(\omega))$

(9b) $\quad E(Y_1(\omega) \mid X_o(\omega)) = g_{1o}(X_o(\omega)).$

These assumptions imply that $g_{0u}(X_o(\omega))$ and $g_{1u}(X_o(\omega))$ are conditional expectations of $Y_0(\omega)$ and $Y_1(\omega)$ given $X_o(\omega)$ that can be used to construct the population average versions

[23] Setting this expectation to zero is an arbitrary but convenient normalization.

of the causal effects previously defined.

Consider two points of evaluation of $X_o(\omega), x_o = (x_{o,1}, ..., x_{o,J})$ and $x'_o = (x'_{o,1}, ..., x'_{o,J})$. Define the causal effect of a change in x_o for person ω as

$$\Delta Y_i(\omega) = g_i(x_o, X_u(\omega)) - g_i(x'_o, X_u(\omega)).$$

Then as a consequence of assumptions (7) and (8),

$$E(Y_i(\omega) \mid X_o(\omega) = x_o) - E(Y_i(\omega) \mid X_o(\omega) = x'_o) = \Delta Y_i(\omega) = g_{io}(x_o) - g_{io}(x'_o).$$

Failure of (8a) and (8b) gives rise to *simultaneous equations bias:* $E(Y_i(\omega) \mid X_o(\omega)) \neq g_{io}(X_o(\omega))$. Even if $X_o(\omega)$ and $X_u(\omega)$ are independent but conditions (7a) and (7b) do not hold, as a consequence of the mean value theorem for integrals, the parameters defined by conditional expectations are defined at unknown points of evaluation of the unobserved variables under additional regularity conditions. Assuming that the support of $X(\omega)$, $\mathcal{X}$, is open and connected, and the conditional means of Y_i given $X(\omega)$ are continuous and bounded functions of $X(\omega)$, then for each value of $X_o(\omega)$ there is a point in the support of $X(\omega)$ say $(X_o(\omega),\, X_u^*(\omega))$ where, by the mean value function for integrals,

$$(10) \int_{\mathcal{X}_u} E(Y_i(\omega) \mid X_o(\omega), X_u(\omega)) g(X_u(\omega) \mid X_o(\omega)) dX_u(\omega) = E(Y_i(\omega) \mid X_o(\omega), X_u^*(X_o(\omega))$$

where $\mathcal{X}_u$ is the support of $X_u(\omega)$ and where $g\left(X_u\left(\omega\right)|X_o\left(\omega\right)\right)$ is the density of $X_u\left(\omega\right)$ given $X_o\left(\omega\right)$.[24] Observe that $X_u^* = X_u^*\left(X_o\left(\omega\right)\right)$. The right hand side expression of (10) is

[24] See Buck, (1965, p. 106).

not a Marshallian causal function because variations in $X_o(\omega)$ shift the implicit point of evaluation of the unobservables $X_u^*(X_o(\omega))$. The causal effects are defined on unobserved values of $X_u(\omega)$ that depend on $X_o(\omega)$. Variations in $X_o(\omega)$ implicitly change the point of evaluation $X_u^*(X_o(\omega))$. In general $E(Y_i(\omega) \mid X_o(\omega), X_u(\omega)) \neq E(Y_i(\omega) \mid X_o(\omega))$. Both additive separability and condition (8a) and (8b) are required to make conditional expectations average Marshallian causal functions.

3 The Problem of Causal Inference

The problem of causal inference and the related problems of policy evaluation and policy forecasting arise in attempting to find empirical counterparts to the hypothetical counterfactuals defined in the previous section, and in particular from missing values in available samples for the arguments of the Marshallian causal functions and their extensions to cover treatment effects. Some of the comparisons one would like to make in defining causal parameters cannot be made because the outcome functions cannot be determined on the relevant empirical supports of the random variables. Some components of $(X(\omega), Z(\omega))$ in the available data may not be observed. Alternatively, limited support of $(X(\omega), Z(\omega))$ may prevent the analyst from determining the g_0 and g_1 functions over the domain of interest. In either case, the causal parameters cannot be identified over full domain of the original hypothetical population causal functions.

It is often not possible to observe the same person (*i.e.*, the same ω) simultaneously in both the $D(\omega) = 1$ and the $D(\omega) = 0$ state, so that it is often not possible to simultaneously observe $Y_0(\omega)$ and $Y_1(\omega)$ for the same ω. Thus, it is typically not possible to form *Treatment Effect Causal Effect* (6) for any person.[25] This observational problem has been called the "Fundamental Problem of Causal Inference," by Holland (1986). However, from equations (1a) and (1b), it is not necessary to observe the same person in both the treated and untreated states to form the *Treatment Effect Causal Effect* (6) for that person. It is sufficient to observe two individuals with the same values of X but different values of D. In other words, if ω, ω' are such that $X(\omega) = X(\omega')$, and $D(\omega) = 1$ but $D(\omega') = 0$, then $Y_0(\omega)$ can be determined by $Y_0(\omega')$ and $Y_1(\omega')$ can be determined by $Y_1(\omega)$. Such (ω, ω') pairs are possible if there are pairs such that $X(\omega) = X(\omega')$ but $Z(\omega) \in \bar{\mathcal{Z}}$, $Z(\omega') \in \bar{\mathcal{Z}}^c$. From our discussion in Section 2, a sufficient condition for such pairs to exist is that there are some variables in $Z(\omega)$ distinct from $X(\omega)$.[26] The ability to find variables that shift persons across treatment states but do not affect potential outcomes underlies many of the econometric methods used in the evaluation literature.[27]

[25] However, certain "fixed effect" and repeated cohort panel data solutions to the evaluation problem, discussed in Heckman and Smith (1998), Heckman, LaLonde and Smith (1999), Heckman and Vytlacil (2000b), assume that a person who occupies both states but at different times is the counterpart twin that can solve the problem of causal inference and generate (6). Genetic twins are sometimes used in the same way in the study of the effects of schooling on earnings. See, *e.g.* the studies in Taubman (1977).

[26] Formally, $\sigma\left(Z\left(\omega\right)\right) \nsubseteq \sigma\left(X\left(\omega\right)\right)$.

[27] An approach based on controlled variation is in direct opposition to the approach advocated by Hol-

Only if it were possible to compare persons with different $D(\omega)$ but the same $X(\omega)$ is it possible to produce causal parameters for individuals. If values of g_0 or g_1 cannot be obtained for certain values of the $X(\omega)$ variables, it is not possible to define the causal parameters for those values. One response to this problem of causal inference, pursued in the economics or statistics literatures, and advocated by Holland (1986, 1988a), is to reformulate the problem of estimating causal effects from the individual level to the group level. Causal parameters are defined at the population level, but this entails a whole set of implicit assumptions which we now make explicit. The population level causal parameters so generated generally do not correspond to the Marshallian causal parameters, except under special conditions, and give rise to the treatment parameters that occupy center stage in the recent literature.

land (1986, 1988). Holland refers to the assumption that $(Y_1(\omega), Y_0(\omega)) = (Y_1(\omega'), Y_0(\omega'))$ as a "unit homogeneity" assumption. In our model, any x defines a set $X^{-1}(x) = \{\omega : X(\omega) = x\}$, where unit homogeneity holds among elements of the set. Holland describes this method of paired comparisons as the "scientific approach," and argues against its use. Instead, Holland advocates what he terms the "statistical approach," which consists of imposing population level independence assumptions without reference to any causal model or unit level homogeneity assumptions to estimate certain population average parameters defined in the next section. As discussed in Section 2, a more complete definition of a causal effect requires a model of interventions.

4 Reformulating the Evaluation Problem to the Population Level

The assumption justifying population level definition of treatment parameters is that $0 < Pr(D(\omega) = 1 \mid X_o(\omega)) < 1$ so that in large samples both treated and untreated persons are observed with the same $X_o(\omega)$ characteristics. Thus it is possible to construct the distribution of $Y_1(\omega)$ given $D(\omega) = 1$ and the distribution of $Y_0(\omega)$ given $D(\omega) = 0$ conditional on $X_o(\omega)$ for any point in the support of $X_o(\omega)$. Implicit in this assumption of the existence of some observed or unobserved factors that cause persons of the same $X_o(\omega)$ to switch treatment states.

This assumption is so important to the entire treatment effect literature that we give it special attention:

(A-4) $\quad 0 < Pr(D(\omega) = 1 \mid X_o(\omega)) < 1.$

Although certain mean outcomes receive the most attention, recent research has also considered identification of population distributions of treatment effects (see Heckman, Smith and Clements, 1997). The transition from individual level treatment parameters to parameters defined at the population level is fundamental, and creates a major disconnect between the traditional structural equations approach in econometrics and the treatment effect or causal parameters approach in statistics. In particular, distinctions between "exogenous" and "endogenous" variables, so central to structural econometrics, become unimportant in the context of the treatment effect literature formulated at the population level provided that the goal is solely to evaluate various programs in place. Before examining this disconnect in depth, it is useful to consider the parameters most often examined in the treatment effect literature and its recent extensions.

A variety of population counterpart treatment parameters have been defined as substitutes for (2) and (4). Here we list a few of the parameters that are widely used. The *Average Treatment Effect (ATE)* is prominent in this literature:

$$ATE(X_o(\omega) = x) = E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x). \tag{11}$$

This is the average treatment effect for persons selected "at random" or by an external mechanism independent of $X_u(\omega)$ from the population for whom $X_o(\omega) = x$. When pressed, most empirical analysts claim that they are estimating this parameter.

In fact, the most commonly estimated parameter, and the one produced from many recent social experiments, is the effect of *Treatment on the Treated (TT)*:

$$TT(X_o(\omega) = x) = E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x, D(\omega) = 1). \tag{12}$$

This is the average of (4) over subpopulations for which $D(\omega) = 1$ and $X_o(\omega) = x$. *Treatment on the Untreated* can be defined in a similar fashion:

$$TUT(X_o(\omega) = x) = E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x, D(\omega) = 0). \tag{13}$$

These definitions do not require that $X_o(\omega) \perp\!\!\!\perp X_u(\omega)$. However, in order to identify the full effect of treatment, the $X_o(\omega)$ must satisfy a no-feedback condition. Thus the $X_o(\omega)$ need not be exogenous in the ordinary usage of this term. (Engle, Hendry and Richard, 1983).

A sufficient no-feedback condition works with a counterfactual $X_{o,d}(\omega)$ process, defined as the realization of the $X_o(\omega)$ process when D is fixed externally at d ($D = d$, $d \in \{0, 1\}$). The no-feedback condition is that

$$\text{(A-5)} \qquad X_{o,1}(\omega) = X_{o,0}(\omega) \qquad a.e.$$

i.e. that the realized values of the $X_o(\omega)$ process are essentially the same irrespective of the values assumed by D.[28] $X_{o,1}(\omega)$ and $X_{o,0}(\omega)$ need not be stochastically independent of $(Y_0(\omega), Y_1(\omega))$: $X_{o,d} \not\perp\!\!\!\perp (Y_0(\omega), Y_1(\omega)$. If (A-5) is violated, conditioning on $X_o(\omega)$ masks the full effect of $D(\omega)$ on outcomes. Then the estimated effect of treatment is marginal of its effect as it operates through the conditioning variables.

Assumption (A-5) is a cross-section no-feedback assumption analogous to, but not identical with, the no-feedback assumptions of Granger noncausality in the time series literature.[29] As discussed below, if we seek to project estimated treatment parameters to other environments, *i.e.* if we seek to use these parameters structurally, then it is necessary to

[28] We can weaken (A-5) to be $\Pr(X_{o,1} \leq t \mid D = d) = \Pr(X_{o,0} \leq t \mid D = d)$. In a Roy model, $(D(\omega) = 1(Y_1(\omega) > Y_0(\omega))$. In this context, the weakened condition is implied by the formal condition $X_0(\omega) \perp\!\!\!\perp D(\omega) \parallel Y_0(\omega), Y_1(\omega)$, the form used in Heckman, Ichimura, Smith and Todd (1998) and Heckman, LaLonde and Smith (1999).

[29] See Florens and Mouchart (1985a, 1985b). However, (A-5) is not explicitly stated in the time series causality literature although it is the precise form of the no-feedback condition required there.

invoke structural assumptions and exogeneity is useful in that context.

A variety of other parameters can also be constructed in a similar fashion for various subpopulations (*e.g.* nontreatment on the treated, etc.). The mean finite change in $Y(w) = D(Z(\omega))Y_1(X(\omega)) + (1 - D(Z(\omega)))Y_0(X(\omega))$ with respect to the finite change in the jth coordinate of $Z_o(\omega)$, $Z_j(\omega)$ is

$$(14) \qquad E\left[\left.\frac{\Delta Y(\omega)}{\Delta Z_j(\omega)}\right| X_o\left(\omega\right) = x, Z_o(\omega) = z, \Delta Z_j(\omega) \neq 0\right].$$

This parameter can be defined conditional on all of the $X_o\left(\omega\right)$, $Z_o(\omega)$ components or only on subsets of these components. However, in constructing a rigorous definition, it is necessary to specify the relationship between the variables on which conditioning is made and those not conditioned on in deciding whether the relationship is causal. Only if the $\Delta Z_j(\omega)$ are externally specified would the effect be causal.[30] This parameter is closely related to the local average treatment effect *(LATE)* of Imbens and Angrist (1994) and is their parameter under additional assumptions given in Heckman and Vytlacil (2000a,b). These conditions clearly state the relationship between the variables being changed and those not accounted for. Under additional assumptions, the limit of expectation (14), as $\Delta Z_j(\omega) \rightarrow 0$, if it exists, is the *Marginal Treatment Effect (MTE)* parameter of Heckman and Vytlacil (1999, 2000a,b,c).[31] It is formally defined as

[30] Thus a condition analogous to (A-5) needs to hold *i.e.* defining $Z_{0,d}$ analogous to $X_{0,d}$, $Z_{o,0}(\omega) = Z_{o,1}(\omega)$ *a.e.* In addition, we require that Z is nondegenerate conditional on X.

[31] They also refer to this parameter as LIV for local instrumental variable, the estimator analog of *MTE.*

$$(15) \qquad MTE\left(X_o(\omega) = x, Z_o(\omega) = z\right) = \lim_{\Delta Z_j(\omega)\to 0} E\left[\frac{\Delta Y(\omega)}{\Delta Z_j(\omega)}\middle| X_o(\omega) = x, Z_o(\omega) = z\right].$$

Again, one can condition on other components in $X_o(\omega), Z_o(\omega)$ or integrate out all other components except $Z_j(\omega)$ but in doing so the causal status of the parameter may be affected unless a condition like (A-5) applies to both the X and the Z.

One can also replace these definitions with median, mode or general quantiles of the gain distributions. Various distributions of the treatment parameter (4) are also of interest, especially in evaluating programs with a broad social impact. (See, Heckman, Smith and Clements, 1997, Heckman and Smith, 1998 or Heckman and Vytlacil, 2000b). Despite the interest in the questions they answer, the distribution parameters require more information to identify and are rarely estimated and we do not discuss them further in this paper. See Heckman, Smith and Clements, (1997), for a discussion and estimates of such parameters and the analyses of Aakvik, Heckman and Vytlacil (1999, 2000).

5 Relationships Among Population Treatment Effects Causal Parameters, Marshallian Causal Functions and Structural Equation Models

To investigate relationships among treatment parameters, Marshallian causal parameters and parameters of structural equations, it is fruitful first to distinguish between structural equations and Marshallian causal functions. *Structural equations* are low dimensional parameterizations of the Marshallian causal functions. Thus a structural representation of (1a) and (1b) writes

(16a) $\quad Y_0 = g_0\left(X\left(\omega\right)\right) = f_0\left(X_o\left(\omega\right), X_u\left(\omega\right); \theta_0\right)$

(16b) $\quad Y_1 = g_1\left(X\left(\omega\right)\right) = f_1\left(X_o\left(\omega\right), X_u\left(\omega\right); \theta_1\right)$

where θ_0 and θ_1 are parameters that generate the g_0 and g_1 functions. The θ_0 and θ_1 are "deep structural" parameters. The derivatives or finite changes of these functions are the Marshallian causal parameters previously defined. For low dimensional θ_0 and θ_1, structural functions have two distinct advantages: (a) they reduce the computational burden of determining g_0 and g_1 and (b) they can be used to extrapolate functions fitted on the support of $X(\omega)$, where (θ_0, θ_1) can be identified, to other domains of definition to make policy forecasts and to construct policy counterfactuals for new policies that entail different combinations of $X(\omega)$ characteristics. This extrapolation feature is essential to the evaluation of new policies and to extrapolating the effects of old policies on new domains.

In the standard linear structural equations case

$$f_0 = X_o\left(\omega\right)' \theta_o + \nu_0\left(\omega\right) \tag{17a}$$

$$f_1 = X_o\left(\omega\right)' \theta_1 + \nu_1\left(\omega\right) \tag{17b}$$

where the $\nu_i\left(\omega\right)$ are scalar-valued functions of $X_u\left(\omega\right)$. The structural parameters in this instance are the causal parameters in the sense of Marshall. But in general, the θ_0 and θ_1 in (16a) and (16b) are *not* causal parameters, but rather are structural parameters that generate the Marshallian causal functions, which when differentiated (or differenced) with respect to external variables, produce Marshallian causal relationships.

The population level treatment effect parameters discussed in the previous subsection condition on $X_o(\omega)$, leaving the dependence between $X_o(\omega)$ and $X_u(\omega)$ unspecified. Changes in $X_o(\omega)$ implicitly change the point of evaluation of X_u^* as in (10). Thus variations in the ATE and TT treatment parameters induced by $X_o(\omega)$ are not *ceteris paribus* changes analogous to those produced from the Marshallian causal functions that control for the unobservables (*i.e.* that hold them fixed at a prespecified value). They represent the average effect of changing treatment status for different $X_o(\omega)$ values, given the associated $X_u(\omega)$ values. This means that, in general, the treatment effect parameters that are determined in one population do not apply to another population unless the relationship between the observables and unobservables is the same in both populations, a point we illustrate in Section 8 in the context of our discussion of a labor supply example.

Consider *ATE:*

$E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x) =$

$\int [g_1(X_o(\omega) = x, X_u(\omega))$ - $g_0(X_o(\omega) = x, X_u(\omega)] dF(X_u(\omega) \mid X_o(\omega) = x)$

where $F(X_u(\omega) \mid X_o(\omega))$ is the conditional distribution of $X_u(\omega)$ given $X_o(\omega)$. For two different populations (different F), defined on the same support, ATE will in general be different. For two different distributions, F and $F,^*$ ATE generalizes to different populations as long as

$$E_F(U_0(\omega) \mid X_0(\omega) = x) = E_{F^*}(U_0(\omega) \mid X_0(\omega) = x).$$

If the supports do not overlap, then ATE determined in one population will apply only to the subset of the support in the extrapolated population.[32] Similar remarks apply to the other treatment effect parameters.[33]

The important point is that ATE and the other treatment parameters are defined conditional on $X_o(\omega) = x$. Variations in ATE across values of $X_o(\omega)$ do not, in general, hold constant values of $X_u(\omega)$ unless representation (7) is invoked and $X_o(\omega) \perp\!\!\!\perp X_u(\omega)$

[32] In the separable case, the requirement for extrapolation is that $E_F(g_u(X_u(\omega)) = E_{F^*}(g_u(X_u(\omega))$ where F and F^* are the conditional distributions of $X_u(\omega)$ given $X_o(\omega)$ for the two populations and that the new distribution F^* does not change the underlying structural relationship between $Y(\omega)$ and $(X_o(\omega); X_u(\omega))$.

[33] Thus consider treatment on the treated: $E(Y_1(\omega) - Y_o(\omega) \mid X_o(\omega), D(\omega) = 1)$. Extrapolation of this parameter to other populations requires analogous conditions on $F(X_u(\omega), Z_u(\omega) \mid X_o(\omega), Z_o(\omega), D(\omega) = 1)$, where $Z(\omega)$ is partitioned analogous to $X(\omega)$ into observed and unobserved components.

or least condition 8(a) and 8(b) holds. These variations are not the Marshallian causal variations as defined in Section 2.

From knowledge of f_0 and f_1, and their arguments, structural equations (16a) and (16b) can be used to generate causal parameters. However, estimation of the population level treatment effect parameters does not necessarily entail estimation of the structural parameters.

At the population level, the treatment effect parameters do *not* control for the effects of unobservables except in special cases. Thus *ATE*, *TT*, and *MTE* are not in general population versions of Marshallian causal parameters. They cannot be used to extrapolate to other populations except under the special conditions previously noted. They cannot in general be used to make *ceteris paribus* counterfactual statements that change values of the conditioning variables holding constant values of unobservables. Offsetting these disadvantages, estimation of these parameters does not entail the full set of identifying assumptions required to estimate structural models that can be applied to general populations provided support conditions are met. The parameters θ_0 and θ_1 need not be identified to identify the treatment effect parameters.[34] For that reason, many of the standard econometric distinctions about exogenous and super-exogenous variables used to identify structural pa-

[34] In an early contribution, Marschak (1953) defined classes of decision problems where knowledge of the full structural parameters was not required to answer well-posed economic questions. See Heckman (2000) for a discussion of Marschak's contribution to this problem. In the treatment effect literature, no structural parameter need be determined to define the treatment effects.

rameters and justify forecasts are of limited value in the analysis of the population causal parameters used in the treatment effect literature used solely to solve the program evaluation problem. By focusing attention on population treatment effects, many of the deep problems of identifying and estimating structural equations that generate causal laws (1a) and (1b) are bypassed. At the same time, the advantages of structural parameters and Marshallian causal parameters in defining economically interpretable parameters that can be used to forecast the effects of new policies and to construct the criteria of traditional welfare economics are lost, except under special conditions. In the next subsection, we consider how structural equations can be used to make valid policy forecasts. These properties are to be contrasted with what can be obtained from conventional treatment effect parameters, unless restrictions comparable to those made in the structural equations literature are invoked.

6 The Value of Structural Equations in Making Policy Forecasts

Structural equations are useful for three different purposes. First, their derivatives or finite changes generate the comparative statics *ceteris paribus* variations produced by economic theory. Tests of economic theory and measurements of economic parameters (price elasticities, measurements of consumer surplus, etc.) are often based on these causal functions or their structural equation counterparts. This is a feature shared with nonparametric Marshallian causal functions.

Second, structural equations and Marshallian causal functions can be used to forecast the effects of policies evaluated in one population in other populations, provided that the parameters are invariant across populations. A purely nonparametric Marshallian causal function cannot be extrapolated to other populations with different supports. Third, as emphasized by Marschak (1953), Marshallian causal functions and structural equations are one ingredient required to forecast the effect of a new policy, never previously implemented.

The problem of forecasting the effects of a policy evaluated on one population but applied to another population can be formulated in the following way. Let $Y = \varphi(X_o(\omega), X_u(\omega))$, where $\varphi : \mathcal{D} \mapsto \mathcal{Y}$, $\mathcal{D} \subseteq \Re^J$, $\mathcal{Y} \subseteq \Re$. φ is a Marshallian causal function determining outcome Y, and we assume that it is known only over $\text{Supp}(X_o(\omega), X_u(\omega)) = \mathcal{X}_o \times \mathcal{X}_u$. X_o and X_u are the random variables in the source population. The mean outcome conditional on $X_o(\omega) = x$ is

$$E_S(Y(\omega) \mid X_o(\omega) = x) = \int_{\mathcal{X}_u} \varphi(X_o(\omega) = x, X_u(\omega)) dF_S(X_u(\omega) \mid X_o(\omega) = x)$$

where $F_S(X_u(\omega) \mid X_o(\omega))$ is the distribution in the source (S) population. We seek to forecast the outcome in a target population which may have a different support. The average outcome in the target population (T) is

$$E_T(Y(\omega) \mid X_o(\omega) = x) = \int_{\chi_u^T} \varphi(X_0(\omega) = x, X_u(\omega)) dF_T(X_u(\omega) \mid X_o(\omega) = x)$$

where χ_u^T is the support of U in the target population. Provided the support of $(X_o(w), X_u(w))$ is the same in the source and the target populations, from knowledge of F_T it is possible to produce a correct value of $E_T(Y(\omega) \mid X_o(\omega) = x)$ on the target population. Otherwise, it is possible to evaluate this expectation only over the intersection set $Supp_T(X(\omega)) \cap Supp_S(X(\omega))$ where $Supp_A(X(\omega))$ is the support of $X(\omega)$ in the A population. In order to extrapolate over the whole set $Supp_T(X(\omega))$ it is necessary to adopt a structural representation of the φ function (*e.g.* equations 17a and 17b or equations 16a and 16b). Additive separability in φ simplifies the extrapolation problem. If φ is additively separable

$$\varphi = \varphi_o(X_o(\omega)) + \varphi_u(X_u(\omega)),$$

$\varphi_o(X_o(\omega))$ applies to all populations for which we can condition on $X_o(\omega)$. However, some structure may have to be imposed to extrapolate from $Supp_S(X_o(\omega))$ to $Supp_T(X_o(\omega))$ if $\varphi_o(X_o(\omega))$ on T is not determined nonparametrically from S.

The problem of forecasting the effect of a new policy, never previously experienced, is similar in character to the policy forecasting problem just discussed. It shares many elements in common with the problem of forecasting the demand for a new good, never previously consumed.[35] Without imposing some structure on this problem, it is impossible

[35] Quandt and Baumol (1966), Lancaster (1971), Gorman (1980), McFadden (1974) and Domencich and

to solve. The literature in structural econometrics associated with the work of the Cowles Commission adopts the following five step approach to this problem.

(1) Structural or Marshallian causal functions are determined (*e.g.* $\varphi(X(\omega))$ in the previous discussion).

(2) The new policy is characterized by an invertible mapping from observed random variables to the characteristics associated with the policy: $c(\omega) = q(X(\omega))$, where $c(\omega)$ is the set of characteristics associated with the policy and q, $q : R^J \rightarrow R^J$, is a *known* invertible mapping.

(3) $X(\omega) = q^{-1}(c(\omega))$ is solved to associate characteristics that in principle can be observed with the policy. This places the characteristics of the new policy on the same footing as those of the old.

(4) It is assumed that $Supp(q^{-1}(c(\omega)) \subseteq Supp(X(\omega))$. This ensures that the support of the new characteristics mapped into $X(\omega)$ space is contained in the support of $X(\omega)$. If this condition is not met, structural versions of the nonparametric Marshallian causal functions must be used to forecast the effects of the new policy, to extend it beyond the support of the source population.

McFadden (1975) consider the problem of forecasting the demand for a new good. Marschak (1953) is the classic reference for evaluating the effect of a new policy. Ichimura and Taber (2000) discuss this problem in a nonparametric context with limited support conditions.

(5) The forecast effect of the policy on Y is $Y_c(\omega) = \varphi(q^{-1}(c(\omega))$.

The leading example of this approach is Lancaster's method for estimating the demand for a new good (Lancaster, 1971). New goods are viewed as bundles of old characteristics. McFadden's conditional logit scheme (1974) is based on a similar idea.[36] Marschak's analysis of the effect of a new commodity tax is another example. Let $P(\omega)$ be the random variable denoting the price facing consumer ω. The tax changes the product price from $P(\omega)$ to $P(\omega)(1+t)$, where t is the tax. With sufficient price variation so that the assumption in Step 4 is satisfied (so the support of the price after tax $Supp_{\text{post tax}}(P(\omega)(1+t)) \subseteq Supp_{\text{pretax}}(P(\omega))$), it is possible to use reduced form demand functions fit on a pretax sample to forecast the effect of a tax never previously put in place. Marschak uses a linear structural equation to solve the problem of limited support. From linearity, determination of the structural equations over a small region determines it everywhere. We develop this point further in the labor supply example presented in Section 8 below.

Marshallian or structural causal functions are an essential ingredient in constructing such forecasts because they explicitly model the relationship between $X_u(\omega)$ and $X_o(\omega)$.

[36] McFadden's stochastic specification is different from Lancaster's specification. See Heckman and Snyder (1997) for a comparison of these two approaches. Lancaster assumes that the $X_u(\omega)$ are the same for each consumer in all choice settings. (They are preference parameters in his setting). McFadden allows for $X_u(\omega)$ to be different for the same consumer across different choice settings but assumes that the $X_u(\omega)$ in each choice setting are draws from a common distribution that can be determined from the demand for old goods.

The treatment effect approach does not explicitly model this relationship so that treatment parameters cannot be extrapolated in this fashion, unless the dependence of potential outcomes on $X_u(\omega)$ and $X_o(\omega)$ is specified, and the required support conditions are satisfied. The influential Rubin (1978) - Holland (1986) model does not specify the required relationships.

7 Comparing the Rubin Model with the Model of this Paper

The Rubin (1978) model as exposited by Holland (1986, 1988a) is widely regarded in statistics as a proper paradigm of defining causal effects and guiding causal analysis. Like the Marshallian causal parameters, the causal parameters in the Rubin model are *ceteris paribus* in nature without an explicit statement of what is held constant or how treatment is assigned. We argue that the Rubin model is seriously incomplete because it is not a model of scientific explanation. Indeed, Holland (1986, 1988b) regards it as a virtue of the Rubin approach that it "models the effects of causes and not the causes of the effects," i.e. that it does not model either the determinants of potential outcomes (equations (1a) and (1b)) or the mechanisms by which potential outcomes are observed. It is a model of (Y_0, Y_1, D) without establishing how these variables are related to X and

Z or how they are interrelated. We regard these features as serious limitations to the use of the framework when considering the counterfactual questions that are central to causal analysis and econometric policy evaluation.

The lack of an explicit model of assignment to treatment gives rise to the logical difficulties in defining a causal effect holding X fixed that were discussed in Section 2. If a change in X is required to induce a change in D, it is not possible to even define an X-constant treatment effect if there is no other way to change D except through a change in X. In the Rubin model, some implicitly specified Z is assumed to affect D where Z is assumed to be variation free of X. The literature on the Rubin model is not clear on this question. In some cases treatment status is assumed to be determined by an experiment that does not affect the outcome equations. For other cases randomization is assumed to be impossible.[37] No explicit discussion of treatment assignment mechanisms is provided and the relationship between potential outcomes and assignment to treatment is never articulated although this relationship plays a central role in the econometric approach to evaluating programs (see Heckman and Vytlacil, 2000b).

The absence of any explicit model for potential outcomes limits the usefulness of the

[37] Thus Rubin (1970) and Holland (1986) claim that gender, race and many other statuses cannot be causal because there is no way to randomize them. This ignores the possibility of defining causal effects by controlling on other characteristics (the “scientific approach” eschewed by Holland (1986)) although fixed effect and paired comparison methods have proven to be scientifically fruitful in economics and other fields. The argument also confuses what is perceived to be empirically possible with what is in principle possible.

Rubin model for generalizing findings from any study; for forecasting the effects of new policies (modeled as different bundles of X distinct from what has been experienced in the past) and for forecasting the effects of existing policies in new environments or comparing treatments across environments. The implicit theorizing that underlies this framework precludes the application of economic theory as a guide to interpreting evidence on treatment effects or for justifying the identification assumptions required to recover the hypothetical causal parameters from data. Indeed, the vaguely specified "treatments" considered by the Rubin model discourage the application of precisely formulated economic theory as embodied in (1a) and (1b) to empirical evaluation problems.

8 What Treatment Effect and Structural Parameters Are Identified From Social Experiments?

In order to make our discussion of treatment effects and structural parameters specific, it is helpful to consider how social experiments identify different parameters of labor supply equations that are useful for answering distinctly different policy questions. There are two distinct cases for the application of the method of randomized trials to identify economic evaluation parameters corresponding to the distinction between treatment parameters and structural parameters.[38] The first, and historically older, case in economics advocates randomization to identify structural parameters. The second and more recent case seeks to use randomization to identify treatment effects.[39] The assumptions required to identify structural parameters are more stringent than the assumptions required to identify treatment effect parameters. At the same time, the treatment parameters identified by social experiments are usually not structural parameters and cannot be used, without additional assumptions, to evaluate programs that differ from the one evaluated by the social experiment, or even to apply the results of the same program to different populations, or to compare results across different programs.[40]

A. Treatment Effects vs. Structural Parameters: A Labor Supply Example

As we have previously noted, a central distinction in the evaluation literature is the one between Marshallian causal parameters (or structural parameters) and treatment effect parameters which we have previously discussed. We revisit this problem in the context of a structural econometric model of labor supply. This distinction accounts for the two different cases for social experimentation that are made in the literature. It also accounts for the centrality of exogeneity assumptions in structural econometrics and their irrelevance in the treatment effect literature that focuses solely on evaluating existing programs.

[38] See Orcutt and Orcutt (1968) or Hausman and Wise (1985).

[39] The two cases for randomization were first discussed in Heckman (1992) and Heckman and Smith (1993). See also the discussion in Heckman, LaLonde and Smith (1999).

[40] These points were first made in Heckman (1992) and Heckman and Smith (1993, 1995).

As previously noted, the goals of the literature on structural equation estimation and on the estimation of treatment parameters are different. As first formulated by Marschak (1953), the goal of structural estimation is to solve a variety of decision problems.[41] Those decision problems entail such distinct tasks as (a) evaluating the effectiveness of an existing policy, (b) projecting the effectiveness of a policy to different environments from the one where it was experienced, or (c) forecasting the effects of a new policy, never previously experienced. In addition, structural parameters can be linked to economic theory. Estimates of these parameters can be used to test propositions of the theory and to make quantitative estimates of the relative importance of different causes within a theory in a way that is ruled out in the Rubin model. Although Marschak was concerned about policy evaluation, similar distinctions arise in estimating the demand for goods. Structural methods can be used to estimate the parameters of demand equations in a given economic environment, in forecasting the demand for goods in a different environment, and in forecasting the demand for a new good never previously consumed. Knowledge of the parameters of demand functions is crucial in testing alternative theories of consumer demand and measuring the strength of complementaries and substitution among goods.

Modern structural econometrics takes as its credo the estimation of "policy invariant structural parameters." By this is usually meant the entire set of structural parameters

[41] Recall the opening sentence of his seminal article: "Knowledge is useful if it helps us make the best decisions". (Marschak, 1953, p.1).

generating the models. Yet Marschak (1953) already realized that for certain decision problems, knowledge of individual structural parameters, or any structural parameter, may be unnecessary. A second, and neglected, contribution of his paper was the notion of decision-specific parameters. In his linear equation examples, the decision-specific parameters involve ratios of structural parameters. One need only determine these ratios in order to answer certain policy questions (*i.e.* solve certain decision problems).

Marschak's insight presages the modern treatment effect literature which takes as its main goal the estimation of one or another treatment effect parameters - not the full range of parameters pursued in structural econometrics - to answer specific policy evaluation questions. Accordingly, the literature is more focused than structural econometrics and at the same time is more limited. The information required to solve a particular decision problem does not necessarily require knowledge of structural parameters. As we have seen, distinctions between endogenous and exogenous variables developed in the literature on structural parameter estimation are typically irrelevant in identifying treatment effects used to evaluate programs in place. One goal of this paper is to clarify when these concepts are relevant in the treatment effect literature

By focusing on one particular decision problem, the treatment effect literature achieves its objectives under weaker assumptions than are invoked in the structural econometrics literature. At the same time, the parameters so generated are less readily transported to different environments to estimate the effects of the same policy in a different setting or the effects of a new policy. The treatment effect literature has to be extended to make such projections and unsurprisingly, the extensions are nonparametric versions of the assumptions used by structural econometricians.

To illustrate these points, and broaden the analysis of the previous sections to consider the case of continuous treatments, consider the prototypical problem of determining the impact of taxes on labor supply. This problem motivated the early literature on social experiments in economics (Cain and Watts, 1973) and remains an important policy problem to this day.

Write an interior solution labor supply equation of hours of work h in terms of wages, W, and other variables including assets, demographic structure and the like. Denote these other variables by X. Let ε denote an unobservable. In the most general form for h,

$$h = h(W, X, \varepsilon). \tag{18}$$

Assume for simplicity that h is differentiable in all of its arguments. Equation (18) is a Marshallian causal function. Its derivatives produce the *ceteris paribus* effect of a change in the argument being varied in h. Suppose that we wish to evaluate the effect of a change in a proportional wage tax on labor supply. Proportional wage taxes at rate t make the after tax wage $W(1-t)$. We assume that agents correctly perceive the tax and we ignore any general equilibrium effects of the tax. Equation (18) is also a model of potential outcomes.

For each value of (W, X, ε) we obtain a value of h. Thus $h : (W, X, \varepsilon) \longrightarrow h$ produces counterfactual states for each value of (W, X) holding ε fixed. We assume no feedback between t and (W, X, ε).

An additively separable version of the Marshallian causal function (18) is

(19) $\qquad h = h(W, X) + \varepsilon.$

This version of the labor supply function enables the analyst to define the *ceteris paribus* effects of W and X on h without having to know the level of the unknown (to the econometrician) unobservable ε. A structural version of (18) is

(18′) $\qquad h = h(W, X, \varepsilon; \theta)$

where θ is a low dimensional parameter that generates h. A structural version of (19) is

(19′) $\qquad h = h(W, X; \theta) + \varepsilon.$

The structural parameters reduce the dimensionality of the identification problem from that of identifying an infinite-dimensional function to that of identifying a finite set of parameters. They play a crucial role in forecasting the effects of an old policy on different populations and forecasting the effects of a new policy. Finally, for the sake of familiarity, we write a linear-in-parameters Cowles Commission type representation of h :

(20) $\qquad h = \alpha' X + \beta \ell n W + \varepsilon$

where we adopt a semi-log specification to represent models widely used in the literature on labor supply. (See Killingsworth, 1983).

Following Marschak, we distinguish three cases. (1) The case where tax t has been implemented in the past and we wish to forecast the effects of the tax in the future in a population with the same distribution of (W, X, ε) variables as prevailed when measurements of tax variation were made. (2) A second case where tax t has been implemented in the past but we wish to project the effects of the same tax to a different population of (W, X, ε) variables. (3) A case where the tax has never been implemented and we wish to forecast the effect of a tax either on an initial population used to estimate (18) or on a different population.

Suppose that the goal of the analysis is to determine the effect of taxes on average labor supply on a relevant population with distribution $G(W, X, \varepsilon)$. In case 1, we have data from the same population as the one on which we wish to construct a forecast. Assume data from a randomized trial, in which persons face tax rate t_j in regime j, $j = 1, ..., J$ assigned so that $\Pr(T = t_j \mid X, W, \varepsilon) = \Pr(T = t_j \mid X, W)$. In the sample from each regime we can identify

$$(21) \qquad E(h \mid W, X, T = t_j) = \int h(W(1-t_j), X, \varepsilon) dG(\varepsilon \mid X, W, T = t_j).$$

For the entire population this function is

$$(21') \qquad E(h \mid T = t_j) = \int h(W(1-t_j), X, \varepsilon) dG(\varepsilon, X, W \mid T = t_j).$$

This function is assumed to apply to the target population. Knowledge of (21) or (21′) determined from historical data can be projected into all future periods provided the joint distributions of data are temporally invariant. If one regime has been experienced in the past, lessons from it apply to the future, provided that the same $h(\cdot)$ and $G(\cdot)$ prevail. No knowledge of any Marshallian causal function or structural parameter is required to do policy analysis for case one. It is not necessary to break apart (21) or (21′) to isolate h or G.

Case two resembles case one except for one crucial difference. Because we are now projecting the same policy onto a different population, it is necessary to break (21) or (21′) into its components and determine $h(W(1-T), X, \varepsilon)$ separately from $G(\varepsilon, X, W, T)$. The assumptions required to project the effects of the old policy in a new regime are as follows.

(1) Knowledge of $h(\cdot)$ is needed on the new population. This may entail determination of h on a different support from that used to determine h in an initial sample if the target population has a different support than the original source population. At this stage, structural estimation comes into its own. It sometimes enables us to extrapolate h from the source population to the target population. A completely nonparametric solution to this problem is impossible even if we adopt structural additive separability assumption (19′) unless the supports of target and source populations coincide.

Some structure must be placed on h even if (19′) characterizes the labor supply model. Parametric structure (20) is a traditional one in the labor supply literature and versions of a linear in parameters model dominate applied econometric research.[42]

Knowledge of $G(\cdot)$ for the target population is also required. In this context, exogeneity enters as a crucial facilitating assumption. If we define exogeneity by

(A-6) $\quad (X, W, T) \perp\!\!\!\perp \varepsilon$

then

$$G(\varepsilon \mid X, W, T) = G(\varepsilon).^{43}$$

In this case, if we assume that the distribution of unobservables is the same in the sample as in the forecast or target regime, $G(\varepsilon) = G'(\varepsilon)$, where $G'(\varepsilon)$ is the distribution of unobservables in the target population, we can project to a new population using the relationship

$$E(h \mid W, X, T = t_j) = \int h(W(1-t_j), X, \varepsilon) dG(\varepsilon)$$

provided we can determine $h(\cdot)$ over the new support of (W, X, ε). If, however, $G' \neq G$, G' must somehow be determined. This entails some structural assumptions to determine the relationship between G and G'. This case does not require that the effect of W on h be

[42] The assumption that $h(W, X)$ is real analytic so that it can be extended to other domains is another structural assumption. This assumption is exploited in Heckman and Singer (1984) to solve a censoring problem in duration analysis.

[43] There are many definitions of this term. See Engle, Hendry and Richard (1983). Assumption (A-6) is often supplemented by the additional assumption that the distribution of X does not depend on the parameters of the model (*e.g.* θ in $(19')$ or $(18')$).

determined. In principle, knowing the effects of t on h over the target support is all that is required.

In the third case, where no tax has previously been introduced, knowledge of the target population is required. Taxes operate through the term $W(1-t)$. If there is no wage variation in samples extracted from the past, there is no way to identify the effect of taxes on labor supply since by assumption $t=0$, and it is not possible to determine the effect of the first argument on labor supply because of the assumed absence of wage variation. Even though wage variation substitutes for tax variation, there is no way to identify the effect of taxes on labor supply in the target population, since there is no wage variation.

If wages vary in the presample period, it may not be necessary to decompose (21) into h and G, or to do structural estimation, in order to estimate the effect of taxes on labor supply in a regime that introduces taxes for the first time. If the support of $W(1-t)$ $\overset{\text{def}}{=} W^*$ in the target regime is contained in the support of W in the historical regime, and the conditional distributions of W^* and W given X, ε are the same, and the supports of (X, ε) are the same in both regimes, then knowledge of (21) over the support of W in the historical regime is enough to determine the effect of taxes in the target regime. More precisely, letting "historical" denote the past data, and "target" denote the target population for projection, we may write these assumptions as:

(a) Support $(W^*)_{\text{target}} \subseteq$ Support $(W)_{\text{historical}}$

(b) $G(W^* \mid X, \varepsilon)_{\text{target}} = G(W^* \mid X, \varepsilon)_{\text{historical}}$

(c) Support $(X, \varepsilon)_{\text{target}} =$ Support $(X,\varepsilon)_{\text{historical}}$

where $W^* = W(1-t)$ for random variables W defined in the new regime and $(W^*)_{\text{target}} = W_{\text{historical}}$.

In this case, no structural estimation is required to forecast the effect of taxes on labor supply in the target population. A fully nonparametric policy evaluation is possible estimating (21) or (21′) nonparametrically (and not decomposing $E(h \mid W, X)$ into the $h(\cdot)$ and $G(\cdot)$ components). Under assumption (a), we may find a counterpart value of $W(1-t) = W^*$ in the target population to insert in the nonparametric version of (21) (or (21′)). If these conditions are not met, it is necessary to build up the G and the h functions over the new supports using the appropriate distributions. We enter the realm where structural estimation is required, either to extend the support of the $h(\cdot)$ functions or to determine $G(W \mid X, \varepsilon)$ or both. It is necessary to determine the relationship among (W, X, ε) in the target population.

B. Two Different Cases for Social Experiments

Using the labor supply example of Subsection A, we now demonstrate the contrasting nature of the two commonly made cases for social experiments. We also present a form of experimentation that identifies the marginal effect of policy changes (or other variables) on outcomes.

The first, and historically older, case in economics seeks to use randomization to identify Marshallian causal functions and structural parameters. (See Orcutt and Orcutt, 1968). The goal of this type of analysis is to form counterfactuals for policies never tried or to project the effects of policies experienced in one environment to new environments. The second, and more recent, case seeks to use social experiments to evaluate the effectiveness of various "treatments" in place for various treatment parameters defined in the literature. Most often, treatment on the treated is the parameter of interest in these evaluations. Throughout this subsection we abstract from attrition, compliance and randomization bias problems discussed in Heckman, LaLonde and Smith (1999).

In the following example, the "treatment" is a tax policy - say a proportional tax on wages.[44] Thus the goal of experiments under the second case for social experiments is to determine how labor supply responds to taxes t in an experimentally determined population. No explicit attention is given to forecasting the effects of the tax on different populations or in different economic environments for the same population. The labor supply equation is $h = h(t, W, X, \varepsilon)$, where X may include unearned income and asset income so that labor supply depends on both wage and unearned income. Assume that

[44] Historically, randomization was first used in economics to vary wage and income parameters facing individuals in order to estimate wage and income effects in labor supply to examine the implications of negative income taxes on labor supply. Part of the goal of randomization was to produce variation in wages and incomes to determine estimates of income and substitution effects. See Cain and Watts (1973). Ashenfelter (1983) shows how estimates of income and substitution effects can be used to estimate the impact of negative income taxes on labor supply.

taxes T are assigned to persons so that

(A-7) $\quad (T \perp\!\!\!\perp \varepsilon) \parallel (W, X)$

so $\Pr(T = t \mid W, X, \varepsilon) = \Pr(T = t \mid W, X)$. Assuming full compliance with the assignment, we may compute the labor supply given t ("treatment" or taxes) as,

$$E(h \mid T = t, W, X) = \int h(t, W, X, \varepsilon) dG(\varepsilon \mid T = t, W, X) = \int h(t, W, X, \varepsilon) dG(\varepsilon \mid W, X)$$

where the second equality follows from (A-7). Similarly, for the same fixed population, but for tax rate t',

$$E(h \mid T = t', W, X) = \int h(t', W, X, \varepsilon) dG(\varepsilon \mid W, X).$$

The treatment effect of taxation t relative to a base state t' on the same fixed population is

$$E(h \mid T = t, W, X) - E(h \mid T = t', W, X) = \int [h(t, W, X, \varepsilon) - h(t', W, X, \varepsilon)] dG(\varepsilon \mid W, X).$$

We may remove the conditioning on (W, X) by integrating out (W, X) using a common distribution. Let $F_c(W, X)$ denote the common distribution. It could be any benchmark, or one selected to match the features of a particular target population.[45] Then the population average treatment effect for taxes (t, t') is

[45] There may be different distributions of W, X given t by virtue of the assignment rule producing (A-7). If assignment to treatment is the same for all (W, X)(so $T \perp\!\!\!\perp (W, X, \varepsilon)$), then a common distribution of (W, X) is produced by randomization in all treatment categories.

$$E_{F_c}(h \mid t) - E_{F_c}(h \mid t') = \int [E(h \mid T = t, W, X) - E(h \mid T = t', W, X)] dF_c(W, X),$$

for all, t, t', $t \neq t'$. These treatment effects combine structural (and causal parameters) in an economically uninterpretable form. Yet at the same time, they answer the specific question of how labor supply responds to taxes t and t' in the populations over which randomization is conducted.

Applying the results of the experiment to a new population, or forecasting the effects of tax rates not previously experienced, requires additional assumptions. It is still necessary to decompose $E(h \mid t, W, X)$ into $h(\cdot)$ and $G(\cdot)$ components and to determine these functions over the target supports for the distributions for the target population and the target tax rates. Structural assumptions must be invoked to extrapolate to the target population.

Additive separability of h in ε facilitates this task. Thus if

$$h = h(W, T, X) + \varepsilon,$$

under assumption (A-7)

(22) $$E(h \mid W, X, T = t) - E(h \mid W, X, T = t') = h(W, X, t) - h(W, X, t').$$

Then the treatment effect is the difference between two Marshallian causal functions. With additional structure imposed, it is possible to move from treatment effects to combinations of explicit structural parameters that are determined by interactions between T and (X, W)

and the main effects of T. Thus suppose that we further specialize the Marshallian causal functions so that there are only main effects in t:

$$h = \alpha_0 + \alpha_1 \ell n(W(1-t)) + \alpha_2' X + \varepsilon = \alpha_0 + \alpha_1 \ell n W + \alpha_1 \ell n(1-t) + \alpha_2' X + \varepsilon.$$

Recall that in this case, the derivatives of the Marshallian parameters are the slopes and the structural parameters. Under (A-7), equation (22) specializes to

$$E(h \mid W, t, X) - E(h \mid W, t', X) = \alpha_1[\ell n(1-t) - \ell n(1-t')]$$

and α_1 is identifiable from the treatment effects, [just divide both sides by the expression in brackets if $t \neq t'$]. More generally, under additive separability and (A-7) we can identify the combinations of structural parameters represented in (22). Randomization governed by (A-7) does not identify α_2. In general, ε and (W, X) are stochastically dependent, and the variation induced in T by virtue of a randomization that implements (A-7) does not make W or X exogenous (independent of ε). The only reason why the coefficient on the wage term is identified in this example is that it is the same as the coefficient on taxes.

This is a general point about the data produced from social experiments. Social experiments only identify treatment terms and terms that interact with treatment. Main effects for (W, X) are not identified. Thus consider the additively separable case $h(W, X, t, \varepsilon) = h(W, X, t) + \varepsilon$. Under assumption (A-7) for randomization, and assuming full compliance, we can recover $h(W, X, t) - h(W, X, t')$ for various treatment (tax) combinations,

$t \neq t'$. However, decomposing $h(W, X, t)$ into a main effect term $\varphi(W, X)$ and an interaction term plus main effect for treatment term $\eta(W, X, t)$ we may write $h(W, X, t) = \varphi(W, X) + \eta(W, X, t)$. $\varphi(W, X)$ differences out in all contrasts. Only differences in $\eta(W, X, t)$ can be identified.

Randomization identifies the treatment effect not by creating exogeneity between the "right hand" variables and the error term and identifying the Marshallian causal parameters, but rather by *balancing the bias.* Thus, as a consequence of (A-7)

$$E(\varepsilon \mid T = t', W, A) = E(\varepsilon \mid T = t, W, A).$$

These "control functions" (or conditional bias terms) are equated across treatment groups as a consequence of randomization and can be differenced out across treatments. This is a feature of randomization shared by matching, nonparametric instrumental variables and an entire class of "control function" methods introduced in Heckman and Robb (1985,1986) and discussed in Heckman and Vytlacil (2000b). W and X are not exogenous and randomization of t does not make them exogenous. Exogeneity of the conditioning variables is not required to construct the treatment effect that compares the mean outcomes under the two treatments. However exogeneity becomes an important issue if we seek to apply the results from one experiment to another environment, or if we seek to predict the effects of tax rates not previously experienced on labor supply.

Thus if we seek to project the findings from one experiment to a new population with the same distribution of ε but different distributions of (W, X), the task is greatly simplified by assuming

$$\varepsilon \perp\!\!\!\perp (W, X).$$

Then it is no longer necessary to determine the distribution of ε given W, X $(G(\varepsilon \mid W, X))$ in the target population. It is still necessary to determine $h(W, X, t, \varepsilon)$ over the support of the target population which remains a formidable task.

9 Summary and Conclusions

This paper defines causal parameters and structural parameters and relates the parameters of the treatment effect literature to the parameters of structural econometrics and formal causal models in economics. We present precise definitions of causal effects within an economic model that are inclusive of the specification of a mechanism (a formal model) by which causal variables are externally manipulated. We argue that model-free definitions of causality advocated in statistics under the rubric of the Rubin model are incomplete because they do not specify the mechanisms of external variation that are central to the definition of causality. In addition, by not determining the causes of effects, or modeling the relationship between potential outcomes and assignment to treatment, the Rubin model cannot be used to answer numerous counterfactual questions required for economic policy analysis, and does not exploit relationships among potential outcomes, assignment to treatment and the variables causing potential outcomes, that can be used to devise econometric evaluation estimators.

Treatment effects are typically proposed to answer a more limited set of questions than are addressed by structural equation models and it is not surprising that they can do so under weaker conditions than are required to identify structural equations. At the same time, if treatment effects are used structurally - *i.e.* to forecast the effect of a program on new populations or to forecast the effects of new programs, it is not surprising that stronger assumptions are required of the sort used in standard structural econometrics. We present the conditions required to validly extrapolate estimated treatment effects to new populations, and evaluate the impact of a program never previously experienced.

We make our discussion specific by analyzing a labor supply example. We consider what treatment and structural parameters of labor supply equations are identified by randomization and the conditions required to apply the results of the social experiments to different populations, and to evaluate the effects of a tax-subsidy program never previously experienced.

References

[1] Aakvik, A., J. Heckman and E. Vytlacil, (2000), "Treatment Effects for Discrete Outcomes when Responses to Treatment Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs," unpublished manuscript, University of Chicago.

[2] _____, (1999), "Semiparametric Program Evaluation Lessons from an Evaluation of a Norwegian Training Program," unpublished manuscript, University of Chicago.

[3] Angrist, J., (2000), "Estimation of Limited-Dependent Variable Models with Binary Endogenous Regressors: Simple Strategies for Empirical Practice," forthcoming in the *Journal of Business* and *Economics and Statistics*.

[4] Ashenfelter, O., (1983), "Determining Participation in Income-Tested Social Programs," *Journal of the American Statistical Association*, **78**, 517-25.

[5] Bell, J.S. (1964), "On the Einstein-Podolsky-Rosen Paradox," *Physics*, **1**, 195-200.

[6] Buck, R., (1965), *Advanced Calculus*, Second edition. (New York: McGraw Hill).

[7] Cain, G. and H. Watts, (1973), "Introduction," in G. Cain and H. Watts, *Income Maintenance and Labor Supply:Econometric Studies*, (Chicago: Rand McNally).

[8] Domencich, T. and D. McFadden, (1975), *Urban travel demand : a behavioral analysis.* Oxford : North-Holland.

[9] Einstein, A., B. Podolsky and N. Rosen, (1935), *Physical Review*, **47**, 777-80.

[10] Engle, R.F., D.F. Hendry and J.F. Richard, (1983), "Exogeneity," *Econometrica*, **51**, 277-304.

[11] Ericsson, N., (1995), "Testing Exogeneity: An Introduction," in N.R. Ericsson and J.S. Irons, eds., *Testing Exogeneity*, (Oxford: Oxford University Press).

[12] Florens, J.P. and M. Mouchart, (1985a), "Conditioning in Dynamic Models," *Journal of Time Series Analysis*, **6**, 15-34.

[13] Florens, J.P. and M. Mouchart, (1985b), "A Linear Theory for Noncausality," *Econometrica*, **53**, 157-175.

[14] Fraser, D.A.S. (1979), *Inference and Linear Models*, (Wiley: New York).

[15] Goldfeld, S. and R. Quandt, (1976), "Techniques for Estimating Switching Regressions," in S. Goldfeld and R. Quandt, (eds.), *Studies in Nonlinear Estimation*, (Cambridge, MA: Ballinger).

[16] Gorman, W. (1980), "A Possible Procedure for Analysing Quality Differentials in the Egg Market," *Review of Economic Studies*, **47**, 843-856.

[17] Hansen, L. and T. Sargent (1980), "Formulating and Estimating Dynamic Linear Rational Expectations Models," *Journal of Economic Dynamics and Control*, **2**, 7-46.

[18] Hausman, J. and D. Wise (1985), "Technical Problems In Social Experimentation: Costs Versus Ease of Analysis," in J. Hausman and D. Wise, (eds), *Social Experimentation*, (Chicago: University of Chicago Press).

[19] Heckman, J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics*, **115**, 45-97.

[20] Heckman, J. (1992), "Randomization and Social Policy Evaluation," in C. Manski and I. Garfinkel, (eds.), *Evaluating Welfare and Training Programs*, (Cambridge, MA: Harvard University Press), 201-230.

[21] Heckman, J. and Honoré, Bo (1990), "The Empirical Content of the Roy Model," *Econometrica*, **58**, 1121-1149.

[22] Heckman, J., H. Ichimura, J. Smith and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica,* **66**, 1017-1098.

[23] Heckman, J., R. LaLonde and J. Smith, (1999), "The Economics and Econometrics of Active Labor Market Programs," in O. Ashenfelter and D. Card, (eds.), *Handbook of Labor Economics*, (North Holland: Amsterdam), Vol. III, Chapter 31, 1865-2097.

[24] Heckman, J., L. Lochner and C. Taber, (1998), "General Equilibrium Treatment Effects: A Study of Tuition Policy," *American Economic Review*, **88**, 381-386.

[25] _____, (2000), "General Equilibrium Cost Benefit Analysis of Education and Tax Policies," in G. Ranis and L. Raut, (eds.), *Trade, Growth and Development: Essays in Honor of T.N. Srinivasan*, (Amsterdam: Elsevier).

[26] Heckman, J. and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, (eds.), *Longitudinal Analysis of Labor Market Data* (New York: Cambridge University Press for Econometric Society Monograph Series), 156-246.

[27] Heckman, J. and R. Robb, (1986), "Alternative Methods For Solving The Problem of Selection Bias in Evaluating The Impact of Treatments on Outcomes," in H. Wainer, (ed.), *Drawing Inferences From Self-Selected Samples*, (Berlin: Springer Verlag), 63-107.

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

[29] Heckman, J. and J. Smith, (1993), "Assessing the Case for Randomized Evaluation of Social Programs," in *Measuring Labour Market Measures: Evaluating the Effects*

of Active Labour Market Policy Initiatives. Proceedings from the Danish Presidency Conference "Effects and Measuring of Effects of Labour Market Policy Initiatives," (Copenhagen: Denmark Ministry of Labour), 35-95.

[30] Heckman, J. and J. Smith, (1995), "Assessing the Case for Randomized Social Experiments," *Journal of Economic Perspectives*, **9**, 85-110.

[31] Heckman, J. and J. Smith, (1998), "Evaluating The Welfare State," in S. Strom, (ed.), *Econometrics and Economics in the 20th Century: The Ragnar Frisch Centenary* (New York: Cambridge University Press for Econometric Society Monograph Series), Chapter 8, 241-318.

[32] Heckman, J., J. Smith and N. Clements (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting For Heterogeneity In Programme Impacts," *Review of Economic Studies,* 64(4):487-535.

[33] Heckman, J. and J. Snyder, (1997), "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators," *RAND Journal of Economics*, **28**, S142-S189.

[34] Heckman, J. and E. Vytlacil, (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, **96**, 4730-4734.

[35] Heckman, J. and E. Vytlacil, (2000a), “Local Instrumental Variables,” in C. Hsiao, K. Morimune, and J. Powell (eds.), *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, (Cambridge University Press: Cambridge).

[36] Heckman, J. and E. Vytlacil, (2000b), “Econometric Evaluation of Social Programs,” in J. Heckman and E. Leamer, (eds.), *Handbook of Econometrics*, Vol. 5, (Amsterdam: Elsevier).

[37] Heckman, J. and E. Vytlacil, (2000c), “The Relationship Between Treatment Parameters within a Latent Variable Framework,” *Economics Letters*, **66**, 33-39.

[38] Hicks, J. R. (1939), *Value and Capital,* (Oxford, England: Oxford University Press).

[39] Holland, P., (1986), “Statistics and Causal Inference,” *Journal of the American Statistical Association*, **81**, 945-960.

[40] _____, (1988a), “Causal Inference, Path Analysis and Recursive Structural Equation Models,” in C. Clogg, (ed.), *Sociology Methodology*, (Washington, DC: American Sociological Association).

[41] _____, (1988b), “Comment: Causal Mechanism or Causal Effect: Which is Best for Statistical Science?,” *Statistical Science*, **3**, 186-188.

[42] Ichimura, H. and C. Taber, (2000), "Direct Estimation of Policy Impacts," NBER Technical Working Paper 254, June.

[43] Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica,* **62**, 467-76.

[44] Killingsworth, M. (1983), *Labor Supply*, (Cambridge: Cambridge University Press).

[45] Lancaster, K. (1971), *Consumer Demand: A New Approach,* (New York: Columbia University).

[46] Marschak, J. (1953), "Economic Measurements For Policy and Prediction," in W. Hood and T. Koopmans, (eds.), *Studies in Econometric Method*, (Wiley: New York), 1-26.

[47] Marshall (1961), *Principles of Economics,* 9th edition, (New York: Macmillan).

[48] McFadden, D., (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, (ed.), *Frontiers in Econometrics*, (New York: Academic Press).

[49] Orcutt, A. and G. Orcutt (1968), "Experiments for Income Maintenance Policies," *American Economic Review*, **58**, 754-772.

[50] Pearl, J., (2000), *Causality*, (Cambridge: Cambridge University Press).

[51] Quandt, R., (1972), "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association,* **67**, 306-310.

[52] Quandt, R., (1988), *The Econometrics of Disequilibrium,* (New York: Blackwell).

[53] Quandt, R. and W. Baumol, (1966), "The Demand for Abstract Transport Modes: Theory and Measurement," *Journal of Regional Science*, **6**, 13-26.

[54] Rubin, D., (1978), "Bayesian Inference For Causal Effects: The Role of Randomization," *Annals of Statistics*, **7**, 34-58.

[55] Samuelson, P., (1947), *Foundations of Economic Analysis*, (Cambridge, MA: Harvard University Press).

[56] Suppes, P. and M. Zanotti, (1996), *Foundations of Probability with Applications: Selected Papers 1974-1995*, (Cambridge: Cambridge University Press).

[57] Taubman, P., (1977), *Kinometrics: Determinants of Socioeconomic Success Within and Between Families*, (Amsterdam: North Holland).