*International Statistical Review* (2008), 76, 1, 1–27 doi:10.1111/j.1751-5823.2007.00024.x

# Econometric Causality

**James J. Heckman**[1,2,3]

[1] *University of Chicago, Chicago, Illinois 60637, USA*
[2] *American Bar Foundation, Chicago, Illinois, USA*
[3] *Geary Institute, University College Dublin, Ireland. E-mail: jjh@uchicago.edu*

## Summary

**This paper presents the econometric approach to causal modelling. It is motivated by policy problems. New causal parameters are defined and identified to address specific policy problems. Economists embrace a scientific approach to causality and model the preferences and choices of agents to infer subjective (agent) evaluations as well as objective outcomes. Anticipated and realized subjective and objective outcomes are distinguished. Models for simultaneous causality are developed. The paper contrasts the Neyman–Rubin model of causality with the econometric approach.**

*Key words*: Causality; econometrics; Roy model; Neyman–Rubin model; subjective and objective evaluations; anticipated vs. realized outcomes; counterfactuals; treatment effects.

## 1 Introduction

Economists and statisticians make causal inferences and draw, in part, on a common set of tools. Economists focus on causality from the perspective of policy evaluation. Causal parameters and causal inferences in economics are motivated by policy questions. Different policy questions require different parameters, so there is no universal causal parameter or set of parameters. This paper informs statisticians of developments in economics that are useful in addressing policy problems.

The econometric approach develops explicit models of outcomes where the causes of effects are investigated and the mechanisms governing the choice of treatment are analysed. The relationship between treatment outcomes and treatment choice mechanisms is studied. A careful accounting of the unobservables in outcome and treatment choice equations facilitates the design of estimators. Both objective and subjective evaluations are considered, where subjective valuations are those of the person receiving treatment as well as the persons assigning it. Differences between anticipated and realized objective and subjective outcomes are analysed. Models for simultaneous treatment effects are developed. A careful distinction is made between models for potential outcomes and empirical methods for identifying treatment effects.

The paper proceeds as follows. Section 2 distinguishes three distinct problems in analyzing causal models and defines the econometric approach. Section 3 discusses the variety of policy evaluation questions and causal parameters used by economists. Section 4 discusses counterfactuals, causality and structural econometric models and contrasts the econometric approach with the approach adopted in statistics. Section 5 presents a synthesis of the two approaches.

**Table 1**
*Three Distinct Tasks Arising in the Analysis of Causal Models*

| Task | Description | Requirements |
|---|---|---|
| **1** | Defining the set of hypotheticals or counterfactuals | A scientific theory |
| **2** | Identifying causal parameters from hypothetical population data | Mathematical analysis of point or set identification |
| **3** | Identifying parameters from real data | Estimation and testing theory |

## 2 The Econometric Approach

Counterfactuals are possible outcomes in different hypothetical states of the world. An example would be the health outcomes for a person associated with taking or not taking a drug. Causal comparisons entail contrasts between outcomes in possible states defined so that only the presence or absence of the drug varies across the states. The person receiving the drug is the same as the person who does not, except for treatment status and, possibly, the outcome associated with treatment status. The problem of causal inference is to assess whether manipulation of the treatment, holding all other factors constant, affects outcomes. The concept of causality developed in this paper and in the statistical treatment effect literature is based on the notion of controlled variation—variation in treatment holding other factors constant. It is distinct from other notions of causality based on prediction (e.g. Granger, 1969; Sims, 1972). Holland (1986) makes useful distinctions among commonly invoked definitions of causality. Cartwright (2004) discusses a variety of definitions of causality from a philosopher's perspective.

The econometric approach to causal inference carefully distinguishes three problems: (a) defining counterfactuals, (b) identifying causal models from idealized data of population distributions (infinite samples without any sampling variation), and (c) identifying causal models from actual data, where sampling variability is an issue. The contrast between (b) and (c) arises from the difference between empirical distributions based on sampled data and population distributions generating the data. Table 1 delineates the three distinct problems.

The first problem entails the application of science, logic and imagination. It is also partly a matter of convention. A model of counterfactuals is more widely accepted the more widely accepted are its ingredients, which are the rules used to derive a model, including whether or not the rules of logic and mathematics are followed, and its agreement with established theories. Models are descriptions of hypothetical worlds obtained by varying—hypothetically—the factors determining outcomes. Models are not empirical statements or descriptions of actual worlds. However, they are often used to make predictions about actual worlds and they are often abstract representations of empirical descriptions.

The second problem (b) is one of inference in very large samples. Can one recover counterfactuals (or means or distributions of counterfactuals) from data that are free of any sampling variation? This is the identification problem.

The third problem (c) is one of inference in practice. Can one recover a given model or a desired counterfactual from a given set of data? Solutions to this problem entail issues of inference and testing in real-world samples. This is the problem most familiar to statisticians and empirical social scientists. The boundary between problems (b) and (c) is permeable depending on how "the data" are defined.

Some of the controversy surrounding construction of counterfactuals and causal models is partly a consequence of analysts being unclear about these three distinct problems and

often confusing them. Particular methods of estimation (e.g. matching or instrumental variable estimation) have become associated with "causal inference" in some circles, and even the definition of certain "causal parameters", because issues of definition, identification and estimation have sometimes been conflated.

The econometric approach to policy evaluation separates these problems and emphasizes the provisional nature of causal knowledge. Some statisticians reject the notion of the provisional nature of causal knowledge and seek an assumption-free approach to causal inference (see, e.g., Tukey, 1986). However, human knowledge advances by developing theoretical models and testing them against data. The models used are inevitably provisional and depend on *a priori* assumptions. Even randomization, properly executed, cannot answer all of the relevant causal questions.

Many "causal models" in statistics are incomplete guides to interpreting data or for suggesting answers to particular policy questions. They are motivated by the experiment as an ideal. They do not clearly specify the mechanisms determining how hypothetical counterfactuals are realized or how hypothetical interventions are implemented except to compare "randomized" with "non-randomized" interventions. They focus only on outcomes, leaving the model for selecting outcomes only implicitly specified. The construction of counterfactual outcomes is based on appeals to intuition and not on formal models.

Because the mechanisms determining outcome selection are not modelled in the statistical approach, the metaphor of "random assignment" is often adopted. This emphasis on randomization or its surrogates, like matching or instrumental variables, rules out a variety of alternative channels of identification of counterfactuals from population or sample data. The focus on randomization has practical consequences because of the conflation of Task 1 with Tasks 2 and 3 in Table 1. Since randomization is used to define the parameters of interest, this practice sometimes leads to the confusion that randomization is the only way—or at least the best way—to identify causal parameters from real data. Extreme versions of this approach deny causal status to any intervention that cannot in principle be implemented by a practical, real-world experiment.

One reason why many statistical models are incomplete is that they do not specify the sources of randomness generating variability among agents, i.e. they do not specify why otherwise observationally identical people make different choices and have different outcomes given the same choice. They do not distinguish what is in the agent's information set from what is in the observing statistician's information set, although the distinction is fundamental in justifying the properties of any estimator for solving selection and evaluation problems. They do not distinguish uncertainty from the point of view of the agent whose behaviour is being analysed from variability as analysed by the observing analyst. They are also incomplete because they are recursive. They do not allow for simultaneity in choices of outcomes of treatment that are at the heart of game theory and models of social interactions and contagion (see, e.g., Brock & Durlauf, 2001; Tamer, 2003).

Economists since Haavelmo (1943, 1944) have recognized the value of precise models for constructing counterfactuals, for answering "causal" questions and addressing more general policy evaluation questions. The econometric framework is explicit about how models of counterfactuals are generated, the sources of the interventions (the rules of assigning "treatment"), and the sources of unobservables in treatment allocations and outcomes and their relationship. Rather than leaving the rules governing selection of treatment implicit, the econometric approach uses explicit relationships between the unobservables in outcome and selection mechanisms to identify causal models from data and to clarify the nature of identifying assumptions.

The goal of the econometric literature, like the goal of all science, is to understand the causes producing effects so that one can use empirical versions of the models to forecast the effects of interventions never previously experienced, to calculate a variety of policy counterfactuals and

to use scientific theory to guide the choices of estimators and the interpretation of the evidence. These activities require development of a more elaborate theory than is envisioned in the current literature on causal inference in statistics.

Many causal models in statistics are black box devices designed to investigate the impact of "treatments"—often complex packages of interventions—on observed outcomes in a given environment. Unbundling the components of complex treatments is rarely done. Explicit scientific models go into the black box to explore the mechanism(s) producing the effects. In the terminology of Holland (1986), the distinction is between understanding the "effects of causes" (the goal of the treatment effect literature as a large group of statisticians define it) or understanding the "causes of effects" (the goal of the econometric literature building explicit models).

By focusing on one narrow black box question, the treatment effect literature avoids many of the problems confronted in the econometrics literature that builds explicit models of counterfactuals and assignment mechanisms. This is its great virtue. At the same time, it produces parameters that are more limited in application. Without further assumptions, these parameters do not lend themselves to extrapolation out of sample or to accurate forecasts of impacts of other policies besides those being empirically investigated. By not being explicit about the contents of the black box (understanding the causes of effects), the treatment effect literature ties its hands in using information about basic behavioural parameters obtained from other studies as well as scientific intuition to supplement available information in the data in hand. It lacks the ability to provide explanations for estimated "effects" grounded in theory. When the components of treatments vary across studies, knowledge does not accumulate across treatment effect studies, whereas it does accumulate across studies estimating models generated from common parameters that are featured in the econometric approach.

## 3 Policy Evaluation Questions and Criteria of Interest

This section presents three central policy evaluation questions. Individual level treatment effects are defined and the evaluation problem is discussed in general terms.

### *3.1 Three Policy Evaluation Problems Considered in This Paper*

Three broad classes of policy evaluation problems are considered in economics. Policy evaluation problem one is:

**P1** *Evaluating the impact of historical interventions on outcomes including their impact in terms of the well-being of the treated and society at large*

By historical, I mean documented interventions. A variety of outcomes and criteria are used to form these evaluations depending on the question at hand. Economists distinguish objective or public outcomes that can, in principle, be measured by all external observers from "subjective" outcomes that are the evaluations of the agents experiencing treatment (e.g. patients) or the agents prescribing treatment (e.g. physicians). Objective outcomes are intrinsically *ex post* ("after the fact") in nature. The statistical literature on causal inference focuses exclusively on *ex post* objective outcomes. Subjective outcomes can be *ex ante* ("anticipated") or *ex post*. Thus the outcome of a medical trial produces both a cure rate and the pain and suffering of the patient. *Ex ante* anticipated pain and suffering may be different from *ex post* realized pain and suffering. Agents may also have *ex ante* evaluations of the objective outcomes that may differ from their *ex post* evaluations. By impact, I mean constructing either individual level or population-level counterfactuals and their valuations. By well-being, I mean the valuations of the outcomes

obtained from the intervention of the agents being analysed or some other party (e.g. the parents of the agent or "society" at large). They may be *ex ante* or *ex post*. P1 is the problem of *internal validity*. It is the problem of identifying a given treatment parameter or a set of treatment parameters in a given environment.

Most policy evaluation is designed with an eye toward the future and toward informing decisions about new policies and application of old policies to new environments. It is helpful to distinguish a second problem encountered in policy analysis.

**P2** *Forecasting the impacts (constructing counterfactual states) of interventions implemented in one environment in other environments, including their impacts in terms of well-being.*

Included in these interventions are policies described by generic characteristics (e.g. tax or benefit rates or therapy used, including intensity) that are applied to different groups of people or in different time periods from those studied in implementations of the policies on which data are available. This is the problem of *external validity*: taking a treatment parameter or a set of parameters estimated in one environment to another environment (see, e.g. Shadish & Cook, 2007). The environment includes the characteristics of individuals and of the treatments.

Finally, the most ambitious problem is forecasting the effect of a new policy, never previously experienced.

**P3** *Forecasting the impacts of interventions (constructing counterfactual states associated with interventions) never historically experienced to various environments, including their impacts in terms of well-being.*

This problem requires that one use past history to forecast the consequences of new policies. It is a fundamental problem in knowledge. P3 is a problem that economic policy analysts have to solve daily. I now present a framework within which analysts can address these problems in a systematic fashion. It is also a framework that can be used for causal inference.

### *3.2 Definition of Individual Level Treatment Effects*

To evaluate is to value and to compare values among possible outcomes. These are two distinct tasks. Define outcomes corresponding to state (policy, treatment) $s$ for an agent $\omega$ as $Y(s, \omega)$, $\omega \in \Omega$. The agent can be a household, a patient, a firm, or a country. One can think of $\Omega$ as a universe of agents. Assume that $\Omega = [0, 1]$. $Y(\cdot, \cdot)$ may be vector valued, but to simplify the exposition, I work with scalar outcomes. (See Heckman & Vytlacil, 2007a, for an analysis with vector outcomes.)

The $Y(s, \omega)$ are outcomes realized after treatments are chosen. In advance of treatment, agents may not know the $Y(s, \omega)$ but may make forecasts about them. These forecasts may influence their decisions to participate in the programme or may influence the agents who make decisions about whether or not an individual participates in the programme. Selection into the programme based on actual or anticipated components of outcomes gives rise to the selection problem in the evaluation literature.

Let $\mathcal{S}$ be the set of possible treatments with elements denoted by $s$. For simplicity of exposition, assume that this set is the same for all $\omega$. For each $\omega$, one obtains a collection of possible outcomes given by $\{Y(s, \omega)\}_{s \in \mathcal{S}}$. For simplicity, I assume that the set $\mathcal{S}$ is finite (Heckman & Vytlacil, 2007a, consider more general cases). For example, if $\mathcal{S} = \{0, 1\}$, there are two treatments, one of which may be a no-treatment state (e.g. $Y(0, \omega)$ is the outcome for an agent $\omega$ not getting a treatment like a drug, schooling or access to a new technology, while $Y(1, \omega)$ is the outcome in treatment state 1 for agent $\omega$ getting the drug, schooling or access). A two-treatment environment

receives the most attention in the theoretical literature, but the multiple treatment environment is the one most frequently encountered in practice.

Each "state" (treatment) may consist of a compound of subcomponent states. In this case, one can define $s$ itself as a vector (e.g. $s = (s_1, s_2, \ldots, s_K)$ for $K$ components) corresponding to the different components that comprise treatment. Thus a medical protocol typically consists of a package of treatments. One might be interested in the package of one (or more) of its components. Thus $s_1$ might be months of treatment with one drug, $s_2$ the quality of physicians, and so forth. No generality is lost by assuming that $s$ is a scalar, since each distinct treatment can be given a distinct label.

The outcomes may be time subscripted, with $Y_t(s, \omega)$ corresponding to outcomes of treatment measured at different times. The index set for $t$ may be the integers, corresponding to discrete time, or an interval, corresponding to continuous time. In principle, one could index $\mathcal{S}$ by $t$, which may be defined on the integers, corresponding to discrete time, or an interval corresponding to continuous time. The $Y_t(s, \omega)$ are realized or *ex post* (after treatment) outcomes. When choosing treatment, these values may not be known. Gill & Robins (2001), Abbring & Van den Berg (2003), Van der Laan & Robins (2003), Abbring & Heckman (2007a,b), and Heckman & Navarro (2007) develop models for dynamic counterfactuals, where time-subscripted and $\omega$-subscripted $\mathcal{S}$ arise as information accrues. Throughout this essay I keep the time subscript implicit.

The *individual treatment effect* for agent $\omega$ comparing objective outcomes of treatment $s$ with objective outcomes of treatment $s'$ is

$$Y(s, \omega) - Y(s', \omega), \qquad s \neq s', \tag{3.1}$$

for two elements $s, s' \in \mathcal{S}$. This is also called an *individual-level causal effect*. The causal effect is the Marshallian (1890) *ceteris paribus* change of outcomes for an agent across states $s$ and $s'$. Only $s$ and $s'$ are varied.

Other comparisons are of interest in assessing a programme. Economists are interested in the well-being of participants as well as the objective outcomes (see Heckman & Smith, 1998). Although statisticians often reason in terms of assignment mechanisms, economists recognize that agent preferences often govern actual choices. Comparisons across outcomes can be made in terms of utilities (personal, $R(Y(s, \omega), \omega)$, or in terms of planner preferences or physician preferences, $R_G$, or both types of comparisons might be made for the same outcome and their agreement or conflict evaluated). Utility functions produce subjective valuations of outcomes by the agents being treated or the planner.

To simplify the notation, and at the same time allow for more general possibilities for arguments of the valuation function, write $R(Y(s, \omega), \omega)$ as $R(s, \omega)$, suppressing the explicit dependence of $R$ on $Y(s, \omega)$. In this notation, one can ask if $R(s, \omega) > R(s', \omega)$, or not (is the agent better off as a result of treatment $s$ compared to treatment $s'$?). The difference in subjective outcomes is $R(s, \omega) - R(s', \omega)$, and is a type of treatment effect. Holding $\omega$ fixed holds all features of the agent fixed except the treatment assigned, $s$. Since the units of utility, $R(s, \omega)$, are arbitrary, one could, instead, record for each $s$ and $\omega$ an indicator if the outcome in $s$ is greater or less than the outcome in $s'$, i.e. $R(s, \omega) > R(s', \omega)$, or not. This is also a type of treatment effect. Agents making decisions about treatment may be only partially informed about realized payoffs at the time they make decisions. Modelling the distinction between anticipated and realized outcomes is an integral part of the econometric approach to causality and policy evaluation. A central feature of the econometric approach to programme evaluation is the evaluation of subjective valuations as perceived by decision makers and not just objective valuations.

The term "treatment" is used in multiple ways in various literatures. In its most common usage, a treatment assignment mechanism is a rule $\tau : \Omega \to \mathcal{S}$ which assigns treatment to each individual $\omega$. The consequences of the assignment are the outcomes $Y(s, \omega)$, $s \in \mathcal{S}$, $\omega \in \Omega$. The

collection of these possible assignment rules is $\mathcal{T}$ where $\tau \in \mathcal{T}$. There are two aspects of a policy under this definition. The policy selects who gets what. More precisely, it selects individuals $\omega$ and specifies the treatment $s \in \mathcal{S}$ received.

The econometric literature offers a more nuanced definition of treatment assignment that explicitly recognizes the element of choice by agent $\omega$ in producing the treatment assignment rule. Treatment can include participation in activities such as schooling, training, a medical therapy, adoption of a particular technology, and the like. Participation in treatment is often a choice made by agents. Modelling this choice process is a distinctive feature of the econometric approach. Under a more comprehensive definition of treatment, agents are assigned incentives like taxes, subsidies, endowments and eligibility that affect their choices, but the agent chooses the treatment selected. Agent preferences, programme delivery systems, market structures, and the like might all affect the choice of treatment. The treatment choice mechanism may involve multiple actors and multiple decisions that result in an assignment of $\omega$ to $s$. For example, $s$ can be schooling while $Y(s, \omega)$ is earnings given schooling for agent $\omega$. A policy may be a set of payments that encourage schooling, as in the PROGRESA programme in Mexico, and the treatment in that case is a choice of schooling with its consequences for earnings. The $s$ can also be a medical protocol that requires compliance by the patient (choice behaviour) to be effective.

The following description of treatment assignment recognizes individual choices and constraints and is more suitable to policy evaluation that recognizes the role of choice by agents. Specify assignment rules $a \in \mathcal{A}$ that map individuals $\omega$ into constraints (benefits) $b \in \mathcal{B}$ under different mechanisms. In this notation, a constraint assignment mechanism $a$ is a map $a : \Omega \to \mathcal{B}$ defined over the space of agents. The constraints may include endowments, eligibility, taxes, subsidies, and other incentives that affect agent choices of treatment. Elements of $b$ can be parameters of tax and benefit schedules that affect individual incentives. While a more general set-up is possible, where $\omega$-specific schedules are assigned to person $\omega$, the cost of such generality is more complicated notation. For simplicity, I confine attention to a fixed—but possibly very large—set of parameters defined for all agents. The map $a$ defines the rule used to assign $b \in \mathcal{B}$. It can include deterministic rules that give schedules mapping $\omega$ into $\mathcal{B}$, such as tax schedules or eligibility schedules. It can also include random assignment mechanisms that assign $\omega$ to an element of $\mathcal{B}$. Random assignment mechanisms add additional elements of randomness to the environment. Abusing notation, when randomization is used, redefine $\omega$ to include this new source of randomness.

Given $b \in \mathcal{B}$ allocated by constraint assignment mechanism $a \in \mathcal{A}$, agents pick treatments. Define treatment assignment mechanism $\tau : \Omega \times \mathcal{A} \times \mathcal{B} \to \mathcal{S}$ as a map taking agent $\omega \in \Omega$ facing constraints $b \in \mathcal{B}$ assigned by mechanism $a \in \mathcal{A}$ into a treatment $s \in \mathcal{S}$. (I use redundant notation to clarify concepts.) In settings with choice, $\tau$ is the choice rule used by agents where $\tau \in \mathcal{T}$, a set of possible choice rules. It is conventional to assume a unique $\tau \in \mathcal{T}$ is selected by the relevant decision makers, although that is not required in this definition. A policy regime $p \in \mathcal{P}$ is a pair $(a, \tau) \in \mathcal{A} \times \mathcal{T}$ that maps agents denoted by $\omega$ into elements of $s$. In this notation, $\mathcal{P} = \mathcal{A} \times \mathcal{T}$.

Incorporating choice into the analysis of treatment effects is an essential and distinctive ingredient of the econometric approach to the evaluation of social programmes. The traditional treatment–control analysis in statistics equates mechanisms $a$ and $\tau$. An assignment in that literature is an assignment to treatment, not an assignment of incentives and eligibility for treatment with the agent making treatment choices. In this notation, the traditional approach has only one assignment mechanism and treats non-compliance with it as a problem rather than as a source of information on agent preferences, which is a central feature of the econometric approach (Heckman & Smith, 1998). Thus, under full compliance, $a : \Omega \to \mathcal{S}$ and $a = \tau$, where $\mathcal{B} = \mathcal{S}$.

Policy invariance is a key assumption for any study of policy evaluation. It allows analysts to characterize outcomes without specifying how those outcomes are obtained. Policy invariance has two aspects. The first aspect is that, for a given $b \in \mathcal{B}$ (incentive schedule), the mechanism $a \in \mathcal{A}$ by which $\omega$ is assigned a $b$ (e.g. random assignment, coercion at the point of a gun, etc.) and the incentive $b \in \mathcal{B}$ are assumed to be irrelevant for the values of realized outcomes for each $s$ that is selected. Second, for a given $s$ for agent $\omega$, the mechanism $\tau$ by which $s$ is assigned to the agent under assignment mechanism $a \in \mathcal{A}$ is irrelevant for the values assumed by realized outcomes. Both assumptions define what economists mean by policy invariance. Policy invariance was first defined and formalized by Marschak (1953) and Hurwicz (1962).

Policy invariance allows one to describe outcomes by $Y(s, \omega)$ and ignore features of the policy and choice environment in defining outcomes. If one has to account for the effects of incentives and assignment mechanisms on outcomes, one must work with $Y(s, \omega, a, b, \tau)$ instead of $Y(s, \omega)$. The more complex description is the outcome associated with treatment state $s$ for person $\omega$, assigned incentive package $b$ by mechanism $a$ which are arguments of assignment rule $\tau$. See Heckman & Vytlacil (2007a) for precise definitions of invariance.

The invariance assumptions state that for the same treatment $s$ and agent $\omega$, different constraint assignment mechanisms $a$ and $a'$ and associated constraint state assignments $b$ and $b'$ produce the same outcome. For example, they rule out the possibility that the act of randomization or the act of pointing a gun at an agent to secure cooperation with planner intentions has an effect on outcomes, given that the agent ends up in $s$. This is a strong assumption.

The second invariance assumption invoked in the literature is that for a fixed $a$ and $b$, the outcomes are the same, independent of the treatment assignment mechanism. This assumption states that the actual mechanism used to assign treatment does not affect the outcomes. It rules out, among other things, social interactions, contagion and general equilibrium effects. Heckman (1992), Heckman & Smith (1998), Heckman *et al.* (1999) and Heckman & Vytlacil (2007b) discuss evidence against this assumption, and Heckman *et al.* (1998a,b,c) show how to relax it.

If treatment effects based on subjective evaluations are also considered, as is distinctive of the econometric approach, it is necessary to broaden invariance assumptions to produce invariance in rewards for certain policies and assignment mechanisms (see Heckman & Vytlacil, 2007a). The required invariance assumptions state, for example, that utilities are not affected by randomization or the mechanism of assignment of constraints. Heckman (1992), Heckman *et al.* (1999) and Heckman & Vytlacil (2007b) present evidence against this assumption. Another invariance assumption rules out social interactions in both subjective and objective outcomes. It is useful to distinguish invariance of objective outcomes from invariance of subjective outcomes. Randomization may affect subjective evaluations through its effect of adding uncertainty into the decision process but it may not affect objective valuations. The econometric approach models how assignment mechanisms and social interactions affect choice and outcome equations rather than postulating *a priori* that invariance postulates for outcomes are always satisfied for outcomes.

### *3.2.1 More general criteria*

There are many comparisons the analyst might make (see, e.g. Heckman *et al.* 1997). One might compare outcomes in different sets that are ordered. Define $\underset{s \in \mathcal{S}}{\text{argmax}} \{Y(s, \omega)\}$ as the value of $s$ that produces the maximal $Y(s, \omega)$ for $s \in \mathcal{S}$. Thus if $Y(s, \omega)$ is scalar income and one compares outcomes for $s \in \mathcal{S}_A$ with outcomes for $s' \in \mathcal{S}_B$, where $\mathcal{S}_A \cap \mathcal{S}_B = \emptyset$, then one might compare $Y_{s_A}$ to $Y_{s_B}$, where

$$s_A = \underset{s \in \mathcal{S}_A}{\text{argmax}} \{Y(s, \omega)\} \quad \text{and} \quad s_B = \underset{s \in \mathcal{S}_B}{\text{argmax}} \{Y(s, \omega)\},$$

where I suppress the dependence of $s_A$ and $s_B$ on $\omega$. This compares the best in one choice set with the best in the other. Another contrast compares the best choice with the next best choice. To do so, define $s' = \underset{s \in \mathcal{S}}{\text{argmax}} \{Y(s, \omega)\}$ and $\mathcal{S}_B = \mathcal{S} \backslash \{s'\}$ and define the treatment effect as $Y_{s'} - Y_{s_B}$. This is the comparison of the highest outcome over $\mathcal{S}$ with the next best outcome. Many other individual-level comparisons might be constructed, and they may be computed using personal preferences, $R(\omega)$, using the preferences of the planner, $R_G$, or using the preferences of the planner over the preferences of agents. Heckman (2005) and Heckman & Vytlacil (2007a,b) present a comprehensive discussion of alternative criteria.

### 3.3 The Evaluation Problem

In the absence of a theory, there are no well-defined rules for constructing counterfactual or hypothetical states or constructing the rules for assignment to treatment. Lewis (1974) defines admissible counterfactual states without an articulated theory as "closest possible worlds". His definition founders on the lack of any meaningful metric or topology to measure "closeness" among possible worlds. Articulated scientific theories provide algorithms for generating the universe of internally consistent, theory-consistent counterfactual states. These hypothetical states are possible worlds. They are products of a purely mental activity. Different theories produce different $Y(s, \omega)$ and different assignment mechanisms.

The *evaluation problem* is that the analyst observes each agent in one of $\bar{S}$ possible states. One does not know the outcome of the agent in other states that are not realized, and hence cannot directly form individual level treatment effects. The *selection problem* arises because one only observes certain agents in any state. Thus one observes $Y(s, \omega)$ only for agents who choose (or are chosen) to be in that state. In general, the outcomes of agents found in $S = s$ are not representative of what the outcomes of agents would be if they were randomly assigned to $s$.

The evaluation problem is an identification problem that arises in constructing the counterfactual states and treatment assignment rules produced by these abstract models using data. This is the second problem presented in Table 1. This problem is not precisely stated until the data available to the analyst are precisely defined. Different areas of knowledge assume access to different types of data.

For each policy regime, at any point in time one observes agent $\omega$ in some state but not in any of the other states. Thus one does not observe $Y(s', \omega)$ for agent $\omega$ if one observes $Y(s, \omega)$, $s \neq s'$. Let $D(s, \omega) = 1$ if one observes agent $\omega$ in state $s$ under policy regime $p$, where I keep the policy regime $p$ implicit to simplify the notation. In this notation, $D(s, \omega) = 1$ implies that $D(s', \omega) = 0$ for $s \neq s'$.

$Y(s, \omega)$ is observed if $D(s, \omega) = 1$ but not $Y(s', \omega)$, for $s \neq s'$. One can define observed $Y(\omega)$ as

$$Y(\omega) = \sum_{s \in \mathcal{S}} D(s, \omega) Y(s, \omega). \tag{3.2}$$

Without further assumptions, constructing an empirical counterpart to the individual-level causal effect (3.1) is impossible from the data on $(Y(\omega), D(\omega))$, $\omega \in \Omega$, where $D(\omega) = \{(D(s, \omega)\}_{s \in \mathcal{S}}$. This formulation of the evaluation problem is known as Quandt's switching regression model (Quandt, 1958) and is attributed in statistics to Neyman (1923), Cox (1958) and Rubin (1978). A version of it is formulated in a linear equations context for a continuum of treatments by Haavelmo (1943). The Roy model (Roy (1951)) is another version of this framework with two possible treatment outcomes ($\mathcal{S} = \{0, 1\}$) and a scalar outcome measure and a particular assignment mechanism $\tau$ which is that $D(1, \omega) = \mathbf{1}[Y(1, \omega) \geq Y(0, \omega)]$, where $\mathbf{1}[\cdot]$ means $\mathbf{1}[\cdot] = 1$ if the argument "$\cdot$" is true and $= 0$ otherwise. Thus $\tau(\omega) = 1$ for $\omega$ satisfying $Y(1, \omega) \geq Y(0, \omega)$

and $\tau(\omega) = 0$ for $\omega$ satisfying $Y(1, \omega) < Y(0, \omega)$. The mechanism of selection depends on the potential outcomes. Agents choose the sector with the highest income so that the actual selection mechanism is not a randomization. Versions of this model with more general self-selection mechanisms are surveyed in Heckman (1990), Heckman & Smith (1998), Heckman & Vytlacil (2007a,b), and Abbring & Heckman (2007b).

The evaluation literature in macroeconomics analyzes policies with universal coverage at a point in time (e.g. a tax policy or social security) so that $D(s, \omega) = 1$ for some $s$ and all $\omega$. It uses time series data to evaluate the impacts of policies in different periods and typically uses mean outcomes (or mean utilities) to evaluate policies.

The problem of self-selection is an essential aspect of the evaluation problem when data are generated by the choices of agents. The agents making choices may be different from the agents receiving treatment (e.g. parents making choices for children). Such choices can include compliance with the protocols of a social experiment as well as ordinary choices about outcomes that people make in everyday life. As a consequence of self-selection, the distribution of the $Y(s, \omega)$ observed are not the population distribution of randomly sampled $Y(s, \omega)$.

In the prototypical Roy model, the choice of treatment (including the decisions not to attrite from the programme) is informative on the relative evaluation of $Y$ $(s, \omega)$. This point is more general and receives considerable emphasis in the econometrics literature (e.g. Heckman & Smith, 1998; Heckman & Vytlacil, 2007a). Choices by agents provide information on subjective evaluations which are of independent interest.

The evaluation problem arises from the absence of information on outcomes for agent $\omega$ other than the outcome that is observed. Even a perfectly implemented social experiment does not solve this problem (Heckman, 1992). Randomization identifies only one component of $\{Y(s, \omega)\}_{s \in \mathcal{S}}$ for any agent. In addition, even with large samples and a valid randomization, some of the $s \in \mathcal{S}$ may not be observed if one is seeking to evaluate new policies never experienced.

There are two main avenues of escape from this problem. The first avenue, featured in explicitly formulated econometric models, often called "structural econometric analysis", is to model $Y(s, \omega)$ explicitly in terms of its determinants as specified by theory. This entails describing the random variables characterizing $\omega$ and carefully distinguishing what agents know and what the analyst knows. This approach also models $D(s, \omega)$ and the dependence between $Y(s, \omega)$ and $D(s, \omega)$ produced from variables common to $Y(s, \omega)$ and $D(s, \omega)$. The Roy model explicitly models this dependence. See Heckman & Honoré (1990) and Heckman (2001) for a discussion of this model. Heckman (1990), Heckman & Smith (1998), Carneiro *et al.* (2003) and Cunha *et al.* (2005) extend the Roy model. This approach stresses understanding of the factors underlying the outcomes and the choice of outcome equations and their dependence. Empirical models explicitly based on scientific theory pursue this avenue of investigation. Some statisticians call this the "scientific approach" and are surprisingly hostile to it. See Holland (1986).

A second avenue of escape, and the one pursued in the recent treatment effect literature, redirects attention away from estimating the determinants of $Y(s, \omega)$ towards estimating some population version of (3.1), most often a mean, without modelling what factors give rise to the outcome or the relationship between the outcomes and the mechanism selecting outcomes. Agent valuations of outcomes are ignored. The statistical treatment effect literature focuses exclusively on policy problem P1 for the subset of outcomes that is observed. It ignores the problem of forecasting a new policy in a new environment (problem P2), or a policy never previously experienced (problem P3). Forecasting the effects of new policies is a central task of science.

### 3.4 New Population Level Treatment Parameters Introduced in Economics

Economists and statisticians often draw on the same set of population level treatment parameters such as the average treatment effect (ATE), $(E(Y(s) - Y(s')))$, treatment on the treated, (TT) $E((Y(s) - Y(s')) \mid D(s) = 1)$, and treatment on the untreated (TUT) $E((Y(s) - Y(s')) \mid D(s') = 1)$. In this subsection, we keep the "$\omega$" implicit to simplify the notation. In some discussions in statistics, ATE is elevated to primacy as *the* causal parameter. Economists use different causal parameters for different policy problems. The distinction between the marginal and average return is a central concept in economics. It is also of interest in medicine where the effect of a treatment on the marginal patient is an important question. It is often of interest to evaluate the impact of marginal extensions (or contractions) of a programme or treatment regime. Incremental cost–benefit analysis is conducted in terms of marginal gains and benefits.

The *effect of treatment for people at the margin of indifference* (EOTM) between $j$ and $k$, given that these are the best two choices available, is defined with respect to personal preferences and with respect to choice-specific costs $C(j)$. Formally, making the dependence of the reward on $Y(s)$, $C(s)$ explicit, i.e., writing $R(Y(s), C(s))$ as the reward in state $s$,

$$\mathrm{EOTM}^R(j,k) = E\left( Y(j) - Y(k) \;\middle|\; \begin{array}{l} R(Y(j), C(j)) = R(Y(k), C(k)); \\ \left.\begin{array}{l} R(Y(j), C(j)) \\ R(Y(k), C(k)) \end{array}\right\} \geq R(Y(l), C(l)) \\ \qquad\qquad\qquad\qquad\qquad\qquad l \neq j, k \end{array} \right). \tag{3.3}$$

This is the mean gain to agents indifferent between treatments $j$ and $k$, given that these are the best two options available. In a parallel fashion, one can define $\mathrm{EOTM}^{R_G}(Y(j) - Y(k))$ using the preferences of another agent (e.g. the parent of a child, a paternalistic bureaucrat, etc.). This could be a subjective evaluation made by a physician, for example. Analogous parameters can be defined for mean setwise comparisons (see Heckman & Vytlacil, 2005, 2007a,b). A generalization of this parameter called the *marginal treatment effect* (MTE), introduced into the evaluation literature by Björklund & Moffitt (1987) and further developed in Heckman & Vytlacil (1999, 2005, 2007b) and Heckman *et al.* (2006), plays a central role in organizing and interpreting a wide variety of econometric and statistical estimators as weighted averages on marginal treatment effects.

Many other mean treatment parameters can be defined depending on the choice of the conditioning set. Analogous definitions can be given for median and other quantile versions of these parameters (see Heckman *et al.*, 1997; Abadie *et al.*, 2002). Although means are conventional, distributions of treatment parameters are also of considerable interest. I discuss distributional parameters in the next subsection.

Of special interest in policy analysis is the *policy-relevant treatment effect* (PRTE). It is the effect on aggregate outcomes of one policy regime $p \in \mathcal{P}$ compared to the effect of another policy regime. Under invariance assumptions,

$$\text{PRTE:} \qquad E_p(Y(s)) - E_{p'}(Y(s)), \qquad \text{where } p, p' \in \mathcal{P},$$

where the expectations are taken over different spaces of policy assignment rules. Heckman & Vytlacil (2007b) show how to identify this parameter.

Mean treatment effects play a special role in the statistical approach to causality. They are the centerpiece of the Holland (1986)–Rubin (1978) model and in many other studies in statistics and epidemiology. Social experiments with full compliance and no disruption can identify these means because of a special mathematical property of means. If one can identify the mean of $Y(j)$ and the mean of $Y(k)$ from an experiment where $j$ is the treatment and $k$ is the baseline,

one can form the average treatment effect for $j$ compared to $k$. These can be formed over two different groups of agents. By a similar argument, TT or TUT can be formed by randomizing over particular subsets of the population (those who would select treatment and those who would not select treatment, respectively), assuming full compliance and no bias arising from the randomization. See Heckman (1992), Heckman & Vytlacil (2007b) and Abbring & Heckman (2007b).

The case for randomization is weaker if the analyst is interested in other summary measures of the distribution or the distribution itself. In general, randomization is not an effective procedure for identifying median gains, or the distribution of gains or many other key parameters. The elevation of population means as the primary "causal" parameters promotes randomization as an ideal estimation method.

### *3.5 Distributions of Counterfactuals*

Although means are traditional, the answers to many interesting evaluation questions require knowledge of features of the distribution of programme gains other than some mean. Thus it is of interest to know if some fraction of the population benefits from a treatment even if on average there is zero benefit or a negative mean outcome. Let $s_p$ be the shorthand notation for assignment of $\omega$ to outcome $s$ under policy $p$ and the associated set of treatment assignment mechanisms. For any two regimes $p$ and $p'$, the proportion that benefits is $\Pr(Y(s_p(\omega), \omega) \geq Y(s_{p'}(\omega), \omega))$. This is called the *voting criterion* (Heckman *et al.*, 1997). It requires knowledge of the joint distribution of the two arguments in the inequality. Experiments, without further assumptions, can only identify marginal distributions and not the joint distributions required to identify the voting criterion (Heckman 1992).

For particular treatments within a policy regime $p$, it is also of interest to determine the proportion who benefit from $j$ compared to $k$ as $\Pr(Y(j, \omega) \geq Y(k, \omega))$. One might be interested in the quantiles of $Y(s_p(\omega), \omega) - Y(s_{p'}(\omega), \omega)$ or of $Y(j, \omega) - Y(k, \omega)$ for $s_p(\omega) = j$ and $s_p(\omega) = k$ or the percentage who gain from participating in $j$ (compared to $k$) under policy $p$. More comprehensive analyses would include costs and benefits. Distributional criteria are especially salient if programme benefits are not transferrable or if restrictions on feasible social redistributions prevent distributional objectives from being attained. Abbring & Heckman (2007b) present a comprehensive survey of approaches to identifying joint distributions of counterfactual outcomes.

### *3.6 Accounting for Uncertainty*

Systematically accounting for uncertainty introduces additional considerations that are central to economic analysis but that are largely ignored in the statistical treatment effect literature as currently formulated. Persons do not know the outcomes associated with possible states not yet experienced. If some potential outcomes are not known at the time treatment decisions are made, the best that agents can do is to forecast them with some rule. Even if, *ex post*, agents know their outcome in a benchmark state, they may not know it *ex ante*, and they may always be uncertain about what they would have experienced in an alternative state. This creates a further distinction: that between *ex post* and *ex ante* evaluations of both subjective and objective outcomes. The economically motivated literature on policy evaluation makes this distinction. The statistical treatment effect literature does not.

Because agents typically do not possess perfect information, a simple voting criterion that assumes perfect foresight over policy outcomes may not accurately predict choices and requires modification. Let $\mathcal{I}_\omega$ denote the information set available to agent $\omega$. He or she evaluates policy

$j$ against $k$ using that information. Under an expected utility criterion, agent $\omega$ prefers policy $j$ over policy $k$ if

$$E(R(Y(j,\omega),\omega) \mid \mathcal{I}_\omega) \geq E(R(Y(k,\omega),\omega) \mid \mathcal{I}_\omega).$$

The proportion of people who prefer $j$ is

$$PB(j \mid j,k) = \int \mathbf{1}\left(E[R\left(Y(j,\omega),\omega\right) \mid \mathcal{I}_\omega) \geq E(R\left(Y(k,\omega),\omega\right) \mid \mathcal{I}_\omega)\, d\mu\left(\mathcal{I}_\omega\right), \quad (3.4)$$

where $\mu(\omega)$ is the distribution of $\omega$ in the population whose preferences over outcomes are being studied. The voting criterion presented in the previous section is the special case where the information set $\mathcal{I}_\omega$ contains $(Y(j,\omega), Y(k,\omega))$, so that there is no uncertainty about $Y(j)$ and $Y(k)$. Cunha *et al.* (2005, 2006) and Abbring & Heckman (2007b) offer examples of the application of this criterion. See Cunha *et al.* (2005, 2006) for computations regarding both types of joint distributions.

Accounting for uncertainty in the analysis makes it essential to distinguish between *ex ante* and *ex post* evaluations. *Ex post*, part of the uncertainty about policy outcomes is resolved although agents do not, in general, have complete information about what their potential outcomes would have been in policy regimes they have not experienced and may have only incomplete information about the policy they have experienced (e.g. the policy may have long-term consequences extending after the point of evaluation).

In advance of choosing an activity, agents may be uncertain about the outcomes that will actually occur. They may also be uncertain about the full costs they will bear. In general the agent's information is not the same as the analyst's, and they may not be nested. The agent may know things in advance that the analyst may never discover. On the other hand, the analyst, benefitting from hindsight, may know some information that the agent does not know when he is making his choices.

Let $\mathcal{I}_a$ be the information set confronting the agent at the time choices are made and before outcomes are realized. Agents may only imperfectly estimate consequences of their choices. One can write the evaluation of $s$ by an agent, using somewhat non-standard notation, as

$$R\left(s,\mathcal{I}_a\right) = \mu_R\left(s,\mathcal{I}_a\right) + \upsilon\left(s,\mathcal{I}_a\right),$$

reflecting that *ex ante* valuations are made on the basis of *ex ante* information where $\mu_R(s,\mathcal{I}_a)$ is determined by variables that are known to the econometrician and $\upsilon\left(s,\mathcal{I}_a\right)$ are components known to the agent but not the econometrician. *Ex post* evaluations can also be made using a different information set $\mathcal{I}_{ep}$ reflecting the arrival of information after the choice is realized. It is possible that

$$\underset{s\in\mathcal{S}}{\operatorname{argmax}}\{R(s,\mathcal{I}_a)\} \neq \underset{s\in\mathcal{S}}{\operatorname{argmax}}\{R(s,\mathcal{I}_{ep})\},$$

in which case there may be *ex post* regret or elation about the choice made.

The *ex ante* vs. *ex post* distinction is essential for understanding behaviour. In environments of uncertainty, agent choices are made in terms of *ex ante* calculations. Yet the treatment effect literature largely reports *ex post* returns.

The econometrician may possess yet a different information set, $\mathcal{I}_e$. Choice probabilities computed against one information set are not generally the same as those computed against another information set. Operating with hindsight, the econometrician may be privy to some information not available to agents when they make their choices.

*Ex post* assessments of a programme through surveys administered to agents who have completed it may disagree with *ex ante* assessments of the programme. Both may reflect honest valuations of the programme (Katz *et al.*, 1975; Hensher *et al.*, 1999). They are reported when

agents have different information about it or have their preferences altered by participating in the programme. Before participating in a programme, agents may be uncertain about the consequences of participation. An agent who has completed programme $j$ may know $Y(j, \omega)$ but can only guess at the alternative outcome $Y(k, \omega)$ which they have not experienced. In this case, *ex post* "satisfaction" with $j$ relative to $k$ for agent $\omega$ who only participates in $k$ is synonymous with the following inequality,

$$R(Y(j, \omega), \omega) \geq E(R(Y(k, \omega), \omega) \mid \mathcal{I}_\omega), \tag{3.5}$$

where the information is post-treatment. Survey questionnaires about "client" satisfaction with a programme capture subjective elements of programme experience not captured by "objective" measures of outcomes that usually exclude psychic costs and benefits. Heckman *et al.* (1997), Heckman & Smith (1998), and Heckman *et al.* (1999) present evidence on this question. Carneiro *et al.* (2001, 2003); Cunha *et al.* (2005, 2006) and Heckman & Navarro (2007) develop econometric methods for distinguishing *ex ante* from *ex post* evaluations of social programs. See Abbring & Heckman (2007b) for an extensive survey of this literature. Heckman & Vytlacil (2007a) discuss the data needed to identify these criteria, and present examples of Roy models and their extensions that allow for more general decision rules and imperfect information by agents. They show how to use economic models to form treatment parameters.

### *3.7 A Specific Model*

To crystallize the discussion in this section, it is helpful to present a prototypical econometric model for policy evaluation. A patient can be given two courses of treatment "1" and "0" with outcomes $Y_1(\omega)$ and $Y_0(\omega)$. I drop the "$\omega$" notation to simplify the notation.

$Y_1$ is an index of well-being of the patient if treated; $Y_0$ if untreated. At any point in time, a person can be either treated or untreated. The decision to treat may be made on the basis of the expected outcomes $E(Y_1 \mid \mathcal{I})$ and $E(Y_0 \mid \mathcal{I})$ and costs $E(C \mid \mathcal{I})$ where the expectations are those of the relevant decision maker—the patient, the doctor or possibly the parent if the patient is a child. The costs might be the pain and suffering of the patient and/or the direct medical costs of the patient. For any problem, the costs $C$ and expectations $\mathcal{I}$ are for the relevant decision maker who decides who gets treatment.

From the point of view of the patient the expected utility or value of treatment is $E(Y_1 \mid \mathcal{I}) - E(C \mid \mathcal{I})$. The value of no treatment is $E(Y_0 \mid \mathcal{I})$. The expected net value is

$$E(Y_1 \mid \mathcal{I}) - E(C \mid \mathcal{I}) - E(Y_0 \mid \mathcal{I}). \tag{3.6}$$

Then for patients who pick a treatment based on maximum gain

$$D = \begin{cases} 1, & \text{if } [E(Y_1 \mid \mathcal{I}) - E(C \mid \mathcal{I}) - E(Y_0 \mid \mathcal{I})] \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.7}$$

or, more succinctly, $D = \mathbf{1}[(E(Y_1 \mid \mathcal{I}) - E(C \mid \mathcal{I}) - E(Y_0 \mid \mathcal{I})) \geq 0]$. This is the generalized Roy model developed in Cunha *et al.* (2005). See Heckman & Vytlacil (2007a) for a survey of such models.

If the doctor makes the decision to treat, then the relevant $C$ and $\mathcal{I}$ are those of the doctor. Instead of a drug, the treatment can be schooling, migration, installation of a technology and the potential outcomes are the counterfactuals with or without treatment. The *ex post* treatment effect is $Y_1 - Y_0$. The *ex ante* effect is $E(Y_1 \mid \mathcal{I}) - E(Y_0 \mid \mathcal{I})$.

Behavioural or scientific theory motivates the construction of $(Y_0, Y_1)$ and the decision to assign treatment. The most basic model in economics is the Roy model previously mentioned.

The decision maker's information is perfect. There are no direct costs of treatment ($C = 0$) and the decision rule is

$$D = \mathbf{1}(Y_1 \geq Y_0). \tag{3.8}$$

Those who get treatment are those who benefit from it. Thus the treated are a non-random sample of the general population, and there is a selection bias in using the treated sample to infer what the average person would experience if selected at random.

The econometric approach models the dependence between observed $Y = DY_1 + (1 - D)Y_0$ and $D$ to suggest alternative estimators to identify causal parameters. Recent work identifies various mean treatment effects, distributions of treatment effects and the cost of treatment including the pain and suffering of the patients.

Commonly used specifications are

$$\begin{aligned} Y_1 &= X\beta_1 + U_1 \\ Y_0 &= X\beta_0 + U_0 \\ C &= Z\gamma + U_C, \end{aligned} \tag{3.9}$$

where $(X, Z)$ are observed by the analyst and $(U_1, U_0, U_C)$ are unobserved. The patient may know more or less than the analyst. Econometric models allow for the patient to know more (observe more) than the analyst and analyse patient selection into treatment accounting for the asymmetry in knowledge between the patient and the analyst. (Matching assumes that, conditional on $X$ and $Z$, $D$ is independent of $Y_0$, $Y_1$ and so assumes a lot of information is available to the analyst.) The Roy model sets $\gamma = 0$, $U_C = 0$ and assumes normality for $(U_0, U_1)$. These distributional and parametric assumptions are relaxed in the recent econometric literature (see Heckman & Vytlacil, 2007a, for a review).

The statistical approach does not model the treatment assignment rule or its relationship to potential outcomes. The econometric approach makes the treatment assignment equation the centrepiece of its focus and considers both objective and subjective valuations as well as *ex ante* ($E(Y_1 \mid \mathcal{I})$, $E(Y_0 \mid \mathcal{I})$, $E(C \mid \mathcal{I})$) and *ex post* outcomes ($Y_1$, $Y_0$, $C$). For this model, EOTM is $E(Y_1 - Y_0 \mid E(Y_1 \mid \mathcal{I}) - E(Y_0 \mid \mathcal{I}) - E(C \mid \mathcal{I}) = 0)$, i.e. the gain to people just indifferent between treatment and no treatment.

## 4 Counterfactuals, Causality and Structural Econometric Models

The literature on policy evaluation in statistics sometimes compares econometric "structural" approaches with "treatment effect" or "causal" models (see, e.g. Angrist & Imbens, 1995; Angrist *et al.*, 1996). The comparison is not clear because the terms are not precisely defined. Heckman (2005) and Heckman & Vytlacil (2007a) formally define "structural" models and use them as devices for generating counterfactuals. They consider both outcome and treatment choice equations. This section presents a brief introduction to the econometric approach and compares it with models for causal inference in statistics.

### *4.1 Generating Counterfactuals*

The traditional model of econometrics is the "all causes" model. It writes outcomes as a deterministic mapping of inputs to outputs:

$$y(s) = g_s(x, u_s), \tag{4.1}$$

where $x$ and $u_s$ are fixed variables specified by the relevant economic theory. The notation anticipates the distinction between observable ($x$) and unobservable ($u_s$) that is important in

empirical implementation. The role of the two types of variables in (4.1) is symmetric. This notation allows for different unobservables $u_s$ to affect different outcomes. $\mathcal{D}$ is the domain of the mapping $g_s : \mathcal{D} \longrightarrow \mathcal{R}^y$, where $\mathcal{R}^y$ is the range of $y$. There may be multiple outcome variables. All outcomes are explained in a functional sense by the arguments of $g_s$ in (4.1). If one models the *ex post* realizations of outcomes, it is entirely reasonable to invoke an all causes model since the realizations are known (*ex post*) and all uncertainty has been resolved. Implicit in the definition of a function is the requirement that $g_s$ be "stable" or "invariant" to changes in $x$ and $u_s$. The $g_s$ function remains stable as its arguments are varied. Invariance is a key property of a causal model.

Equation (4.1) is a production function relating inputs (factors) to outputs. $g_s$ maps $(x, u_s)$ into the range of $y$ or image of $\mathcal{D}$ under $g_s$, where the domain of definition $\mathcal{D}$ may differ from the empirical support. Thus, equation (4.1) maps admissible inputs into possible *ex post* outcomes. This notation allows for different unobservables from a common list $u$ to appear in different outcome equations.

A "deep structural" version of (4.1) models the variation of the $g_s$ in terms of $s$ as a map constructed from generating characteristics $q_s, x$ and $u_s$ into outcomes:

$$y(s) = g\,(q_s, x, u_s)\,, \tag{4.2}$$

where now the domain of $g$, $\mathcal{D}$, is defined for $q_s, x, u_s$ so that $g : \mathcal{D} \longrightarrow \mathcal{R}^y$. The components $q_s$ provide the basis for generating the counterfactuals across treatments from a base set of characteristics. $g$ maps $(q_s, s, u_s)$ into the range of $y$, $g : (q_s, x, u_s) \longrightarrow \mathcal{R}^y$, where the domain of definition $\mathcal{D}$ of $g$ may differ from the empirical support. In this specification, different treatments $s$ are characterized by different bundles of a set of characteristics common across all treatments. This framework provides the basis for solving policy problem P3 since new policies (treatments) are generated from common characteristics, and all policies are put on a common basis. If a new policy is characterized by known transformations of $(q_s, x, u_s)$ that lie in the domain of definition of $g$, policy forecasting problem P3 can be solved. The argument of the maps $g_s$ and $g$ are part of the *a priori* specification of a causal model. Analysts may disagree about appropriate arguments to include in these maps.

One benefit of a treatment effect approach that focuses on problem P1 is that it works solely with outcomes rather than inputs. However, it is silent on how to solve problems P2 and P3 and provides no basis for interpreting the population-level treatment effects.

Consider alternative models of schooling outcomes of pupils where $s$ indexes the schooling type (e.g. regular public, charter public, private secular and private parochial). The $q_s$ are the observed characteristics of schools of type $s$. The $x$ are the observed characteristics of the pupil. The $u_s$ are the unobserved characteristics of both the schools and the pupil. If one can characterize a proposed new type of school as a new package of different levels of the same ingredients $x$, $q_s$, and $u_s$ and one can identify (4.2) over the domain of the function defined by the new package, one can solve problem P3. If the same schooling input (same $q_s$) is applied to different students (those with different $x$) and one can identify (4.1) or (4.2) over the new domain of definition, one can solve problem P2. By digging deeper into the "causes of the effects" one can do more than just compare the effects of treatments in place with each other. In addition, modeling the $u_s$ and its relationship with the corresponding unobservables in the treatment choice equation is highly informative on the choice of appropriate identification strategies.

Equations (4.1) and (4.2) are sometimes called Marshallian causal functions. Assuming that the components of $(x, u_s)$ or $(q_s, x, u_s)$ are variation-free, a feature that may or may not be produced by the relevant theory, one may vary each argument of these functions to get a *ceteris paribus* causal effect of the argument on the outcome. (See Heckman & Vytlacil, 2007a, for a precise definition of variation-free.) Some components may be variation-free while others are

not. These thought experiments are conducted for hypothetical variations. Recall that the *a priori* theory specifies the arguments in the causal functions and the list of things held fixed when a variable is manipulated.

Changing one coordinate while fixing the others produces a Marshallian *ceteris paribus* causal effect of a change in that coordinate on the outcome variables. Varying $q_s$ fixes different treatment levels. Variations in $u_s$ among agents explain why people with the same $x$ characteristics respond differently to the same treatment $s$. The *ceteris paribus* variation need not be for a single variable of the function. A treatment generally consists of a package of characteristics and if one varies the package from $q_s$ to $q_{s'}$ one gets different treatment effects.

I use the convention that lower-case values are used to define fixed values and upper-case notation denotes random variables. In defining (4.1) and (4.2), I have explicitly worked with fixed variables that are manipulated in a hypothetical way as in the algebra of elementary physics. In a purely deterministic world, agents respond to these non-stochastic variables. Even if the world is uncertain, *ex post*, after the realization of uncertainty, the outcomes of uncertain inputs are deterministic. Some components of $u_s$ may be random shocks realized after decisions about treatment are made.

Thus if uncertainty is a feature of the environment, (4.1) and (4.2) can be interpreted as *ex post* realizations of the counterfactual as uncertainty is resolved. *Ex ante* versions may be different. From the point of view of agent $\omega$ with information set $\mathcal{I}_\omega$, the *ex ante* expected value of $Y(s, \omega)$ is

$$E\left(Y\left(s,\omega\right) \mid \mathcal{I}_\omega\right) = E\left(g\left(Q(s,\omega), X(\omega), U(s,\omega)\right) \mid \mathcal{I}_\omega\right), \tag{4.3}$$

where $Q\left(s, \omega\right), X\left(\omega\right), U(s, \omega)$ are random variables generated from a distribution that depends on the agent's information set indexed by $\mathcal{I}_\omega$.

The expectation might be computed using the information set of the relevant decision maker (e.g. the parents in the case of the outcomes of the child) who might not be the agent whose outcomes are measured. These random variables are drawn from agent $\omega$'s subjective distribution. This distribution may differ from the distribution produced by "reality" or nature if agent expectations are different from objective reality. In the presence of intrinsic uncertainty, the relevant decision maker acts on (4.3) but the *ex post* counterfactual is

$$Y\left(s,\omega\right) = E(Y\left(s,\omega\right) \mid \mathcal{I}_\omega) + \nu\left(s,\omega\right), \tag{4.4}$$

where $\nu\left(s, \omega\right)$ satisfies $E(\nu(s, \omega) \mid \mathcal{I}_\omega) = 0$. In this interpretation, the information set of agent $\omega$ is part of the model specification but the realizations come from a probability distribution, and the information set includes the technology $g$. This representation clarifies the distinction between deterministic *ex post* outcomes and intrinsically random *ex ante* outcomes. Abbring & Heckman (2007b) survey econometric evaluation models accounting for uncertainty.

This restatement of the basic deterministic model reconciles the "all causes" model (4.1) and (4.2) with the intrinsic uncertainty model favoured by some statisticians (see, e.g. Dawid, 2000, and the discussion following his paper). *Ex ante*, there is uncertainty at the agent ($\omega$) level but *ex post* there is not. The realizations of $\nu\left(s, \omega\right)$ are ingredients of the *ex post* "all causes" model, but not part of the subjective *ex ante* "all causes" model. The probability law used by the agent to compute the expectations of $g(Q(s, \omega), X(\omega), U_s(\omega))$ may differ from the objective distribution that generates the observed data. In the *ex ante* all causes model, manipulations of $\mathcal{I}_\omega$ define the *ex ante* causal effects.

Thus, from the point of view of the agent, one can vary elements in $\mathcal{I}_\omega$ to produce Marshallian *ex ante* causal response functions. The *ex ante* treatment effect from the point of view of the

agent for treatment $s$ and $s'$ is

$$E(Y(s, \omega) \mid \mathcal{I}_\omega) - E(Y(s', \omega) \mid \mathcal{I}_\omega). \tag{4.5}$$

The data used to determine these functions may be limited in their support. In that case analysts cannot fully identify the theoretical relationships over hypothetical domains of definition. In addition, in the support, the components of $X$, $U(s)$ and $\mathcal{I}_\omega$ may not be variation-free even if they are variation-free in the hypothetical domain of definition of the function. If the $X$ in a sample are functionally dependent, it is not possible to identify the Marshallian causal function with respect to variations in $x$ over the available support even if one can imagine hypothetically varying the components of $x$ over the domains of definition of the functions (4.1) or (4.2).

I next turn to an important distinction between fixing and conditioning on factors that gets to the heart of the distinction between causal models and correlational relationships. This point is independent of any problem with the supports of the samples compared to the domains of definition of the functions.

### *4.2 Fixing vs. Conditioning*

The distinction between *fixing* and *conditioning* on inputs is central to distinguishing true causal effects from spurious causal effects. In an important paper, Haavelmo (1943) made this distinction in linear equation models. Haavelmo's distinction is the basis for Pearl's (2000) book on causality that generalizes Haavelmo's analysis to non-linear settings. Pearl defines an operator "do" to represent the mental act of fixing a variable to distinguish it from the action of conditioning, which is a statistical operation. If the conditioning set is sufficiently rich, fixing and conditioning are the same in an *ex post* all causes model. Pearl suggests a particular physical mechanism for fixing variables and operationalizing causality, but it is not central to his or any other definition of causality.

The distinction between fixing and conditioning is most easily illustrated in the linear regression model analysed by Haavelmo (1943). Let $y = x\beta + u$. While $y$ and $u$ are scalars, $x$ may be a vector. The linear equation maps every pair $(x, u)$ into a scalar $y \in \mathbb{R}$. Suppose that the support of random variable $(X, U)$ in the data is the same as the domain of $(x, u)$ that are fixed in the hypothetical thought experiment and that the $(x, u)$ are variation-free (i.e. can be independently varied coordinate by coordinate). Abstract from the problem of limited support that is discussed in the preceding section. Dropping the "$\omega$" notation for random variables, write

$$Y = X\beta + U.$$

$U$ is assumed to have a finite mean. Here "nature" or the "real world" picks $(X, U)$ to determine $Y$. $X$ is observed by the analyst and $U$ is not observed, and $(X, U)$ are random variables. This is an "all causes" model in which $(X, U)$ determine $Y$. The variation generated by the hypothetical model varies one coordinate of $(X, U)$, fixing all other coordinates to produce the effect of the variation on the outcome $Y$. Nature (as opposed to the model) may not permit such variation.

Formally, one can write this model defined at the population level as a conditional expectation,

$$E(Y \mid X = x, U = u) = x\beta + u.$$

Since conditioning is on both $X$ and $U$, there is no further source of variation in $Y$. This is a deterministic model that coincides with the "all causes" model. Thus on the support, which is also assumed to be the domain of definition of the function, this model is the same model as the deterministic, hypothetical model, $y = x\beta + u$. Fixing $X$ at different values corresponds to doing different thought experiments with the $X$. Fixing and conditioning are the same in this case.

If, however, one only conditions on $X$, one obtains

$$E(Y \mid X = x) = x\beta + E(U \mid X = x). \tag{4.6}$$

This relationship does not generate $U$-constant $(Y, X)$ relationships. It generates only an $X$-constant relationship. Unless one conditions on all of the "causes" (the right-hand side variables), the empirical relationship (4.6) does not identify causal effect of $X$ on $Y$. The variation in $X$ also moves the conditional mean of $U$ given $X$. This analysis can be generalized to a non-linear model $y = g(q, x, u)$ (see Pearl, 2000). It can be generalized to account for the temporal resolution of uncertainty if one includes $\nu(s, \omega)$ as an argument in the *ex post* causal model. The outcomes can include both objective outcomes $Y(s, \omega)$ and subjective outcomes $R(Y(s, \omega), \omega)$.

Parallel to causal models for outcomes are causal models for the choice of treatment (see Heckman & Vytlacil, 2007a). Accounting for uncertainty and subjective valuations of outcomes (e.g. pain and suffering for a medical treatment) is a major contribution of the econometric approach (see, e.g. Carneiro *et al.*, 2003; Chan & Hamilton, 2006; Cunha *et al.*, 2005, 2006; Cunha & Heckman, 2007; Heckman & Navarro, 2007). The factors that lead an agent to participate in treatment $s$ may be dependent on the factors affecting outcomes. Modelling this dependence is a major source of information used in the econometric approach to construct counterfactuals from real data. A parallel analysis can be made if the decision maker is not the same as the agent whose objective outcomes are being evaluated.

### *4.3 The Econometric Model vs. the Neyman–Rubin Model*

Many statisticians and social scientists use a model of counterfactuals and causality attributed to Donald Rubin by Paul Holland (1986). The framework was developed in statistics by Neyman (1923), Cox (1958) and others. Parallel frameworks were independently developed in psychometrics (Thurstone, 1927) and economics (Haavelmo, 1943; Quandt, 1958, 1972; Roy, 1951). The statistical treatment effect literature originates in the statistical literature on the design of experiments. It draws on hypothetical experiments to define causality and thereby creates the impression in the minds of many of its users that random assignment is the most convincing way to identify causal models. Some would say it is the only way to identify causal models.

Neyman and Rubin postulate counterfactuals $\{Y(s, \omega)\}_{s \in \mathcal{S}}$ without modelling the factors determining the $Y(s, \omega)$ as is done in the econometric approach (see equations (4.1)–(4.4)). Rubin and Neyman offer no model of the choice of which outcome is selected. No lower case, "all causes" model explicitly specified in this approach nor is there any discussion of the social science or theory producing the outcomes studied.

In this notation, Rubin (1986) invokes versions of traditional econometric invariance assumptions called "SUTVA" for Stable Unit Treatment Value Assumption. Since he does not develop choice equations or subjective evaluations, he does not consider the more general invariance conditions for both objective and subjective evaluations discussed in Section 3.2. Invariance assumptions were developed in Cowles Commission econometrics and formalized in Hurwicz (1962). They are surveyed in Heckman & Vytlacil (2007a).

The Rubin model assumes

(R-1) $\{Y(s, \omega)\}_{s \in \mathcal{S}}$, *a set of counterfactuals defined for* ex post *outcomes. It does not analyze agent valuations of outcomes nor does it explicitly specify treatment selection rules, except for contrasting randomization with non-randomization;*

(R-2) *Invariance of counterfactuals for objective outcomes to the mechanism of assignment within a policy regime;*

(R-3) *No social interactions or general equilibrium effects for objective outcomes;*
and
(R-4) *There is no simultaneity in causal effects, i.e. outcomes cannot cause each other reciprocally.*

Two further implicit assumptions in the application of the model are that P1 is the only evaluation problem of interest and that mean causal effects are the only objects of interest.

The econometric approach considers a wider array of policy problems than the statistical treatment effect approach. Its signature features are:

1. Development of an explicit framework for outcomes $Y(s, \omega)$, $s \in \mathcal{S}$, measurements and the choice of outcomes where the role of unobservables ("missing variables") in creating selection problems and justifying estimators is explicitly developed.
2. The analysis of subjective evaluations of outcomes $R(s, \omega)$, $s \in \mathcal{S}$, and the use of choice and compliance data to infer them.
3. The analysis of *ex ante* and *ex post* realizations and evaluations of treatments. This analysis enables analysts to model and identify regret and anticipation by agents. Points 2 and 3 introduce agent decision making into the treatment effect literature.
4. Development of models for identifying and evaluating entire distributions of treatment effects (*ex ante* and *ex post*) rather than just the traditional mean parameters. These distributions enable analysts to determine the proportion of people who benefit from treatment, a causal parameter not considered in the statistical literature on treatment effects.
5. Models for simultaneous causality.
6. Definitions of parameters made without appeals to hypothetical experimental manipulations.
7. Clarification of the need for invariance of parameters with respect to different classes of manipulations to answer different classes of questions.

I now amplify these points.

Selection models defined for potential outcomes with explicit treatment assignment mechanisms were developed by Gronau (1974) and Heckman (1974, 1976, 1978, 1979) in the economics literature before the Neyman–Rubin model was popularized in statistics. The econometric discrete choice literature (McFadden, 1974, 1981) uses counterfactual utilities or subjective evaluations as did its parent literature in mathematical psychology (Thurstone, 1927, 1959). The model sketched in Section 3.7 considers both choices and outcomes of choices. Unlike the Neyman–Rubin model, these models do not start with the experiment as an ideal but they start with well-posed, clearly articulated models for outcomes and treatment choice where the unobservables that underlie the selection and evaluation problem are made explicit. The hypothetical manipulations define the causal parameters of the model. Randomization is a metaphor and not an ideal or "gold standard".

In contrast to the econometric model, the Holland (1986)–Rubin (1978) definition of causal effects is based on randomization. The analysis in Rubin's 1976 and 1978 papers is a dichotomy between randomization ('ignorability') and non-randomization, and not an explicit treatment of particular selection mechanisms in the non-randomized case as developed in the econometrics literature. There is no explicit discussion of treatment selection rules like (3.8) and their relationship with realized outcomes. Even under ideal conditions, randomization cannot answer some very basic questions, such as what proportion of a population benefits from a programme (Heckman, 1992). See Carneiro *et al.* (2001, 2003), where this proportion is identified using choice data and/or supplementary proxy measures. See also Cunha *et al.* (2005, 2006) and Cunha & Heckman (2007). Abbring & Heckman (2007b) discuss this work. In practice, contamination and cross-over effects make randomization a far from sure-fire solution even for constructing $E(Y_1 - Y_0)$. See the evidence on disruption bias and contamination bias

arising in randomized trials that is presented in Heckman *et al.* (1999) and Heckman *et al.* (2000).

Many leading causal analysts conflate the three points of Table 1. The analysis of Holland (1986) illustrates this point and the central role of the randomized trial to the Holland–Rubin analysis. After explicating the "Rubin model", Holland gives a very revealing illustration that conflates the first two tasks of Table 1. He claims that there can be no causal effect of gender on earnings because analysts cannot randomly assign gender. This statement confuses the act of defining a causal effect (a purely mental act) with empirical difficulties in estimating it. These are tasks 1 and 2 in Table 1.

As another example of the same point, Rubin (1978, p. 39) denies that it is possible to define a causal effect of sex on intelligence because a randomization cannot *in principle* be performed. "Without treatment definitions that specify actions to be performed on experimental units, I cannot unambiguously discuss causal effects of treatments" Rubin (1978, p. 39). In this and many other passages in the statistics literature, a causal effect is defined by a randomization. Issues of definition and identification are confused. This confusion continues to flourish in the literature in applied statistics. For example, Berk *et al.* (2005) echo Rubin and Holland in insisting that if an experiment cannot "in principle" be performed, a causal effect cannot be defined. The local average treatment effect (LATE) parameter of Imbens & Angrist (1994) is defined by an instrument and conflates tasks 1 and 2 (definition and identification). Imbens & Angrist (1994) use instrumental variables as surrogates for randomization. Heckman & Vytlacil (1999, 2005) and Heckman *et al.* (2006) define the LATE parameter abstractly and separate issues of definition of parameters from issues of identification.

The act of definition is logically distinct from the acts of identification and inference. A purely mental act can define a causal effect of gender. That is a separate task from identifying the causal effect. The claim that causality can only be determined by randomization reifies randomization as the "gold standard" of causal inference.

The econometric approach to causal inference is more comprehensive than the Neyman–Rubin model of counterfactuals. It analyzes models of the choice of counterfactuals $\{D(s, \omega)\}_{s \in \mathcal{S}}$ and the relationship between choice equations and the counterfactuals. The $D(s, \omega)$ are explicitly modelled as generated by the collection of random variables $(Q(s, \omega), C(s, \omega), Y(s, \omega) \mid \mathcal{I}_\omega), s \in \mathcal{S}$, where $Q(s, \omega)$ is the characteristic of treatment $s$ for agent $\omega$, $C(s, \omega)$ are costs and $\{Y(s, \omega)\}_{s \in \mathcal{S}}$ are the outcomes and the "|" denotes that these variables are defined conditional on $\mathcal{I}_\omega$ (the agent's information set). (Recall the discussion in section 3.6.) If other agents make treatment assignment decisions, then the determinants of $D(s, \omega)$ are modified according to what is in their information set. The variables determining choices are analysed.

Modeling $Y(s, \omega)$ in terms of the characteristics of treatment, and of the treated, facilitates comparisons of counterfactuals and derived causal effects across studies where the composition of programmes and treatment group members may vary. It also facilitates the construction of counterfactuals on new populations and the construction of counterfactuals for new policies. The Neyman–Rubin framework focuses exclusively on population-level mean "causal effects" or treatment effects for policies actually experienced and provides no framework for extrapolation of findings to new environments or for forecasting new policies (problems P2 and P3).

### *4.4 Simultaneous Causality*

One major limitation of the Neyman–Rubin model is that it is recursive. It does not model causal effects of outcomes that occur simultaneously. Since Haavelmo (1943, 1944), econometricians have used simultaneous equations theory to define causality in nonrecursive models where causes

are simultaneous and interdependent. Heckman (2005) and Heckman & Vytlacil (2007a) present extensive discussions of simultaneous causality.

Consider the following non-linear simultaneous equations system where identification of causal effects can be defined by exclusion of variables. Let $(Y_1, Y_2)$ be a pair of jointly determined (internal) variables with externally specified variables $X_1$, $X_2$, $U_1$, $U_2$. Externally specified variables are variables that are specified independently of the system being analysed. $(Y_1, Y_2)$ are internal variables determined by the system. We may represent the system determining the internal variables as

$$Y_1 = g_1(Y_2, X_1, X_2, U_1) \tag{4.7}$$

$$Y_2 = g_2(Y_1, X_1, X_2, U_2). \tag{4.8}$$

$\frac{\partial g_1}{\partial Y_2}\big|_{Y_2=y_2,\, X_1=x_1,\, X_2=x_2,\, U_1=u_1}$ is the causal effect of $Y_2$ on $Y_1$, holding $X_1, X_2$, and $U_1$ fixed. This is Haavelmo's definition of the causal effect applied to a simultaneous equations system. Assuming the existence of local solutions, one can solve these equations to obtain the internal variables in terms of the external variables $Y_1 = \varphi_1(X_1, X_2, U_1, U_2)$ and $Y_2 = \varphi_2(X_1, X_2, U_1, U_2)$. These functions can be determined from the data under standard exogeneity conditions for $X$ (see, e.g. Amemiya, 1985). By the chain rule, one can define the causal effect of $Y_2$ on $Y_1$, using exclusion ($\frac{\partial g_1}{\partial X_1} = 0$ for all $(Y_2, X_1, X_2, U_1)$) and identify the causal effect of $Y_2$ on $Y_1$, by

$$\frac{\partial g_1}{\partial Y_2} = \frac{\partial Y_1}{\partial X_1} \Big/ \frac{\partial Y_2}{\partial X_1} = \frac{\partial \varphi_1}{\partial X_1} \Big/ \frac{\partial \varphi_2}{\partial X_1}.$$

One may define causal effects of $Y_1$ on $Y_2$ using partials with respect to $X_2$ if there is exclusion with respect to $X_2$ in equation (4.8).

A simple example is the simultaneous equations model of Haavelmo (1944),

$$Y_1 = \gamma_{12} Y_2 + b_{11} X_1 + b_{12} X_2 + U_1 \tag{4.9}$$

$$Y_2 = \gamma_{21} Y_1 + b_{21} X_1 + b_{22} X_2 + U_2. \tag{4.10}$$

Assume $(U_1, U_2)$ is independent of $(X_1, X_2)$. Under the condition that

$$\begin{bmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{bmatrix}$$

is full rank, one can solve out for the internal variables $(Y_1, Y_2)$ as a function of external variables $(X_1, X_2, V_1, V_2)$,

$$Y_1 = \pi_{11} X_1 + \pi_{12} X_2 + V_1 \tag{4.11}$$

$$Y_2 = \pi_{21} X_1 + \pi_{22} X_2 + V_2, \tag{4.12}$$

where $(X_1, X_2)$ is assumed to be statistically independent of $(V_1, V_2)$. We can estimate the $\pi_{ij}$, $i, j = 1, 2$, by ordinary least squares. If $b_{11} = 0$, then $X_1$ affects $Y_1$ only through its effect on $Y_2$ (see equation (4.9)). From (4.11) and (4.12),

$$\frac{\partial Y_1}{\partial X_1} = \pi_{11}, \qquad \frac{\partial Y_2}{\partial X_1} = \pi_{21}.$$

From the definition of $\pi_{11}, \pi_{21}$ as coefficients of the solution of the external variables in terms of the external variables, it follows that

$$\frac{\pi_{11}}{\pi_{21}} = \gamma_{12}.$$

Thus one can define and identify the causal effect of $Y_2$ on $Y_1$. This is the method developed by Tinbergen (1930) and when applied as an estimation method is called indirect least squares.

The intuition is simple. Because $X_1$ is excluded from (4.9) (or (4.7)), it can be used to shift $Y_2$ in that equation keeping $X_2$ and $U_1$ fixed. This enables the analyst to determine the causal effect of $Y_2$ on $Y_1$ holding the other determinants of $Y_1$ fixed.

This definition of causal effects in an interdependent system generalizes the recursive definitions of causality featured in the statistical treatment effect literature (Pearl, 2000). The key to the more general econometric definition is manipulation of external inputs and exclusion, not a particular method such as randomization, matching, or instrumental variables. One can use the population simultaneous equations model to define the class of admissible variations and address problems of definitions (task 1 of Table 1). If, for a given model, the functions (4.7) or (4.8) shift when external variables are manipulated, or if external variables cannot be independently manipulated, causal effects of one internal variable on another cannot be defined within that model.

## 5 Marschak's Maxim and the Relationship Between the Econometric Literature and the Statistical Treatment Effect Literature: A Synthesis

The absence of explicit models of outcomes and choice is a prominent feature of the statistical treatment effect literature. Scientifically well-posed models make explicit the assumptions used by analysts regarding preferences, technology, the information available to agents, the constraints under which they operate, and the rules of interaction among agents in market and social settings and the sources of variability among agents. These explicit features make econometric models useful vehicles (a) for interpreting empirical evidence using theory; (b) for collating and synthesizing evidence across studies using economic theory; (c) for measuring the various effects of policies; and (d) for forecasting the welfare and direct effects of previously implemented policies in new environments and the effects of new policies.

These features are absent from the modern treatment effect literature. At the same time, that literature makes fewer statistical assumptions in terms of independence, functional form, exclusion and distributional assumptions than the standard structural estimation literature in econometrics. These are the attractive features of this approach.

However, the econometric literature has advanced greatly in recent years in terms of producing a robust version of its product. Major advances summarized in Powell (1994), Heckman & Vytlacil (2007b), and Matzkin (2007) have relaxed the strong parametric assumptions that characterized the early econometric literature.

In reconciling these two literatures, I reach back to a neglected but important paper by Marschak (1953). Marschak noted that for many specific questions of policy analysis, it is not necessary to identify fully specified models that are invariant to classes of policy modifications. All that may be required for any policy analysis are combinations of subsets of the structural parameters, corresponding to the parameters required to forecast particular policy modifications, which are often much easier to identify (i.e. require fewer and weaker assumptions).

I call this principle *Marschak's Maxim* in honour of this insight. The modern statistical treatment effect literature implements Marschak's Maxim where the policies analysed are the treatments available under a particular policy regime and the goal of policy analysis is restricted to evaluating policies in place (task 1 in Table 1) and not in forecasting the effects of new policies or the effects of old policies on new environments. What is often missing from the literature on treatment effects is a clear discussion of the policy question being addressed by the particular treatment effect being identified and why it is interesting.

Population mean treatment parameters are often identified under weaker conditions than are traditionally assumed in structural econometric analysis. Thus, to identify the average treatment effect given $X$ for $s$ and $s'$ one only requires $E(Y(s, \omega) \mid X = x) - E(Y(s', \omega) \mid X = x)$. Under invariance conditions about outcome equations, this parameter answers the policy question of determining the average effect on outcomes of moving an agent from $s'$ to $s$ when there are no social interaction or contagion effects. The parameter is not designed to evaluate a whole host of other policies. One does not have to know the functional form of the generating $g_s$ functions nor does $X$ have to be exogenous. One does not have to invoke strong conditions about invariance of the choice equations. However, if one seeks to identify $E(Y(s, \omega) \mid X = x, D(s, \omega) = 1) - E(Y(s', \omega) \mid X = x, D(s, \omega) = 1)$, one needs to invoke invariance of choice equations recognizing the conditioning on a choice variable. No conditioning on a choice is required in defining average treatment effects.

Treatment effects are causal effects for particular policies that move agents from $s \in \mathcal{S}$ to $s' \in \mathcal{S}, s' \neq s$, keeping all other features of the agent and environment the same. These effects are designed to address policy problem P1. Treatment effects and causal models can be generated from explicit economic models and are more easily interpreted. Invariant, explicitly formulated, economic models are useful for addressing policy problems P2 and P3: extrapolation and predicting the effects of new policies, respectively.

If the goal of an analysis is to predict outcomes, and the environment is stable, then accurate predictions can be made without causal or structural parameters. Consider Haavelmo's analysis of fixing vs. conditioning discussed in Section 4.2. He analysed the linear regression model $Y = X\beta + U$ and defined the causal effect of $X$ on $Y$ as the $U$-constant effect of variations in $X$. If the goal of the analysis is to predict the effect of $X$ on $Y$, and if the environment is stable so that the historical data have the same distribution as the data in the forecast sample, least squares projections are optimal predictors under mean square error criteria (see, e.g. Goldberger, 1964). One does not need to separate out the causal effect of $X$ on $Y$, $\beta$, from the effect of $X$ on the unobservables operating through $E(U \mid X)$.

Marschak's Maxim urges analysts to formulate the problem being addressed clearly and to use the minimal ingredients required to solve it. The treatment effect literature addresses the problem of comparing treatments under a particular policy regime for a particular environment. The original econometric pioneers considered treatments under different policy regimes and with different environments. As analysts ask more difficult questions, it is necessary to specify more features of the models being used to address the questions.

Marschak's Maxim is an application of Occam's Razor to policy evaluation. For certain classes of policy interventions, designed to answer problem P1, the treatment effect approach may be very powerful and more convincing than explicitly formulated models because it entails fewer assumptions. However, as previously noted, considerable progress has been made in relaxing the parametric structure assumed in the early economic models. As the treatment effect literature is extended to address the more general set of policy forecasting problems entertained in the econometric literature, including the evaluation of subjective outcomes, and the parametric assumptions of the original econometric approach are relaxed, the two literatures will merge in a creative synthesis.

## Acknowledgements

This paper was presented at the ISI Conference in Seoul, Korea, on August 27, 2001. It has benefitted from discussions with Jaap Abbring, Pedro Carneiro, Steve Durlauf, Seong Moon, Salvador Navarro, T.N. Srinivasan, John Trujillo, Ed Vytlacil, Jordan Weil and Yu

Xie; and comments by the editor, Eugene Seneta, and two anonymous referees. This research was supported by NIH R01-HD043411, NSF SES-0241858 and the Geary Institute, University College Dublin, Ireland. The views expressed in this paper are those of the author and not necessarily those of the funders listed here.

## References

Abadie, A., Angrist, J.D. & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, **70**, 91–117.

Abbring, J.H. & Heckman, J.J. (2007a). Dynamic policy analysis. *The Econometrics of Panel Data*, Eds. L. Matyas & P. Sevestre. Dordrecht: Kluwer Academic Publishers, 3rd edn., forthcoming.

Abbring, J.H. & Heckman, J.J. (2007b). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In *Handbook of Econometrics*, vol. 6B, Eds. J. Heckman & E. Leamer, pp. 5145–5303. Amsterdam: Elsevier.

Abbring, J.H. & Van den Berg, G.J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, **71**, 1491–1517.

Amemiya, T. (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Angrist, J.D. & Imbens, G.W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Stati. Assoc.*, **90**, 431–442.

Angrist, J.D., Imbens, G.W. & Rubin, D. (1996). Identification of causal effects using instrumental variables. *J. Amer. Stat. Assoc.*, **91**, 444–455.

Berk, R., Li, A. & Hickman, L.J. (2005). Statistical difficulties in determining the role of race in capital cases: A re-analysis of data from the state of Maryland. *J. Quant. Criminol.*, **21**, 365–390.

Björklund, A. & Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection. *Rev. Econ. Stat.*, **69**, 42–49.

Brock, W.A. & Durlauf, S.N. (2001). Interactions-based models. *Handbook of Econometrics*, vol. 5. Eds. J. J. Heckman & E. Leamer, pp. 3463–3568. New York: North-Holland.

Carneiro, P., Hansen, K. & Heckman, J.J. (2001). Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Econ. Pol. Rev.*, **8**, 273–301.

Carneiro, P., Hansen, K. & Heckman, J.J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *Int. Econ. Rev.*, **44**, 361–422.

Carneiro, P., Heckman, J.J. & Vytlacil, E.J. (2006). Estimating marginal and average returns to education. *Amer. Econ. Rev.*, under review.

Cartwright, N. (2004). Causation: One word many things. *Philos. Sci.*, **71**, 805–819.

Chan, T.Y. & Hamilton, B.H. (2006). Learning, private information and the economic evaluation of randomized experiments. *J. Polit. Econ.*, **114**, 997–1040.

Cox, D.R. (1958). *Planning of Experiments*. New York: Wiley.

Cunha, F. & Heckman, J.J. (2007). The evolution of inequality, heterogeneity and uncertainty in labor earnings in the U.S. economy. *J. Polit. Economy*, forthcoming.

Cunha, F., Heckman, J.J. & Navarro, S. (2005). Separating uncertainty from heterogeneity in life cycle earnings, The 2004 Hicks Lecture. *Oxford Econ. Pap.*, **57**, 191–261.

Cunha, F., Heckman, J.J. & Navarro, S. (2006). Counterfactual analysis of inequality and social mobility. *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, Eds. S.L. Morgan, D.B. Grusky & G.S. Fields, chap. 4, pp. 290–348. Stanford, CA: Stanford University Press.

Dawid, A. (2000). Causal inference without counterfactuals. *J. Amer. Stat. Assoc.*, **95**, 407–424.

Gill, R.D. & Robins, J.M. (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Stat.*, **29**, 1785–1811.

Goldberger, A.S. (1964). *Econometric Theory*. New York: Wiley.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *J. Polit. Econ.*, **82**, 1119–1143.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, **11**, 1–12.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, **12**, iii–vi and 1–115.

Heckman, J.J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, **42**, 679–694.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Social Measure.*, **5**, 475–492.

Heckman, J.J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*. **46**, 931–959.
Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–162.
Heckman, J.J. (1990). Varieties of selection bias. *Amer. Econ. Rev.*, **80**, 313–318.
Heckman, J.J. (1992). Randomization and social policy evaluation. *Evaluating Welfare and Training Programs*, Eds. C. Manski & I. Garfinkel, pp. 201–230. Cambridge, MA: Harvard University Press.
Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *J. Poli. Econo.*, **109**, 673–748.
Heckman, J.J. (2005). The scientific model of causality. *Sociol. Methodol.*, **35**, 1–97.
Heckman, J.J., Hohmann, N. Smith, J. & Khoo, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quart. J. Econ.*, **115**, 651–694.
Heckman, J.J. & Honoré, B.E. (1990). The empirical content of the Roy model. *Econometrica*, **58**, 1121–1149.
Heckman, J.J., LaLonde, R.J. & Smith, J.A. (1999). The economics and econometrics of active labor market programs. *Handbook of Labor Economics*, vol. 3A, Eds. O. Ashenfelter & D. Card, chap. 31, pp. 1865–2097. New York: North-Holland.
Heckman, J.J., Lochner, L.J. & Taber, C. (1998a). Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Rev. Econ. Dynam.*, **1**, 1–58.
Heckman, J.J., Lochner, L.J. & Taber, C. (1998b). General-equilibrium treatment effects: A study of tuition policy. *Amer. Econ. Rev.*, **88**, 381–386.
Heckman, J.J., Lochner, L.J. & Taber, C. (1998c). Tax policy and human-capital formation. *Amer. Econ. Rev.*, **88**, 293–297.
Heckman, J.J. & Navarro, S. (2007). Dynamic discrete choice and dynamic treatment effects. *J. Econ.*, **136**, 341–396.
Heckman, J.J. & Smith, J.A. (1998). Evaluating the welfare state. *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, Ed. S. Strom, pp. 241–318. New York: Cambridge University Press.
Heckman, J.J., Smith, J.A. & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.*, **64**, 487–536.
Heckman, J.J., Urzua, S. & Vytlacil, E.J. (2006). Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Stat.*, **88**, 389–432.
Heckman, J.J. & Vytlacil, E.J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *P Natl. Acad. Sci. USA.*, **96**, 4730–4734.
Heckman, J.J. & Vytlacil, E.J. (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, **73**, 669–738.
Heckman, J.J. & Vytlacil, E.J. (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics*, vol. 6B, Eds. J. Heckman & E. Leamer, pp. 4779–4874. Amsterdam: Elsevier.
Heckman, J.J. & Vytlacil, E.J. (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. *Handbook of Econometrics*, vol. 6B, Eds. J. Heckman & E. Leamer, pp. 4875–5144. Amsterdam: Elsevier.
Hensher, D., Louviere, J. & Swait, J. (1999). Combining sources of preference data. *J. Econ.*, **89**, 197–221.
Holland, P.W. (1986). Statistics and causal inference. *J. Amer. Stat. Assoc.*, **81**, 945–960.
Hurwicz, L. (1962). On the structural form of interdependent systems. *Logic, Methodology and Philosophy of Science*, Eds. E. Nagel, P. Suppes & A. Tarski, pp. 232–239. Stanford University Press.
Imbens, G.W. & Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.
Katz, D., Gutek, A., Kahn, R. & Barton, E. (1975). *Bureaucratic Encounters: A Pilot Study in the Evaluation of Government Services*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.
Lewis, H.G. (1974). Comments on selectivity biases in wage comparisons. *J. Polit. Econ.*, **82**, 1145–1155.
Marschak, J. (1953). Economic measurements for policy and prediction. *Studies in Econometric Method*, Eds. W. Hood & T. Koopmans, pp. 1–26. New York: Wiley.
Marshall, A. (1890). *Principles of Economics*. New York: Macmillan and Company.
Matzkin, R.L. (2007). Nonparametric identification. *Handbook of Econometrics*, vol. 6B, Eds. J. Heckman & E. Leamer. Amsterdam: Elsevier.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, Eds. P. Zarembka. New York: Academic Press.
McFadden, D. (1981). Econometric models of probabilistic choice. *Structural Analysis of Discrete Data with Econometric Applications*, Eds. C. Manski & D. McFadden. Cambridge, MA: MIT Press.
Neyman, J. (1923). Statistical problems in agricultural experiments. *J. Roy. Stat. Soc.*, II (Supplement), 107–180.

Pearl, J. (2000). *Causality*, Cambridge, UK: Cambridge University Press.
Powell, J.L. (1994). Estimation of semiparametric models. *Handbook of Econometrics*, *Volume 4*, Eds. R. Engle & D. McFadden, pp. 2443–2521. Amsterdam: Elsevier.
Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Stat. Assoc.*, **53**, 873–880.
Quandt, R.E. (1972). A new approach to estimating switching regressions. *J. Amer. Stat. Assoc.*, **67**, 306–310.
Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Econ. Pap.*, **3**, 135–146.
Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Stat.*, **6**, 34–58.
Rubin, D.B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *J. Amer. Stat. Assoc.*, **81**, 961–962.
Shadish, W.R. & Cook, T.D. (2007). *Experimental and Quasi-Experimental Designs for Field Research*. Hillsdale, NJ: Lawrence Erlbaum, forthcoming.
Sims, C.A. (1972). Money, income, and causality. *Ameri. Econ. Rev.*, **62**, 540–552.
Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Rev. Econ. Stud.*, **70**, 147–165.
Thurstone, L.L. (1927). A law of comparative judgement. *Psychol. Rev.*, **34**, 273–286.
Thurstone, L.L. (1959). *The Measurement of Values*. Chicago: University of Chicago Press.
Tinbergen, J. (1930). Bestimmung und deutung von angebotskurven. *Zeitschrift für Nationalökonomie*, **1**, 669–679.
Tukey, J.W. (1986). Comments on alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inferences from Self-Selected Samples*, Ed. H. Wainer, pp. 108–110. New York: Springer-Verlag (Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum).
Van der Laan, M.J. & Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer-Verlag.

[*Received March 2005, accepted August 2007*]