

**An Econometric Framework for Analyzing Health Policy with Nonexperimental Data**

by

Joseph V. Terza  
Center for Health Economic and Policy Studies  
Medical University of South Carolina  
Charleston, SC 29425  
[terza@musc.edu](mailto:terza@musc.edu)  
843-792-2202

(June 2004)

This research was supported by Grant # 45500 from the Substance Abuse Policy Research Program of the Robert Wood Johnson Foundation.

## Abstract

If asked, most researchers in applied health economics would say that the goal of their empirical research is to provide solid scientific evidence that will serve to inform current and future health policy. Unfortunately, many research studies, though adherent to the scientific method (hypothesize-observe-test), conclude without reporting their findings in a form that is useful to health policy makers who are typically interested in causal effects – i.e. the effect of a policy variable ( $x_p$ ) on the outcome of interest ( $y$ ). The present paper offers a rigorous and, therefore, useful definition of the *policy effect of  $x_p$  on  $y$* . Moreover, we offer a generic and unified framework for the estimation and statistical testing of policy effects using nonexperimental (survey) data. Specific versions of the generic framework are detailed for many of the parametric regression formulations that are most commonly applied in empirical health economics. The approach is illustrated in an analysis of the effect of substance abuse on employment using the 1992 National Longitudinal Alcohol Epidemiologic Survey.

## 1. Introduction

Much research in applied health economics is conducted with the goal of providing solid scientific evidence that will serve to inform current and future health policy. Nearly every such policy analytic study in health economics is based on nonexperimental (survey) data and has as its specific objective the estimation of the *effect* of a change in a variable of interest that is at present, or will in the future be, under the control of a health policy decision-making entity. Put more succinctly, health policy analysts typically seek to answer the following question: What is the effect of the policy variable ( $x_p$ ) on the outcome of interest ( $y$ )?

In many cases, however, the definition of the *effect of  $x_p$  on  $y$*  is not given, or is so vague as to be of little or no use to health policy makers. Because of this lack of completeness and rigor, results are often left as uninterpreted parameter estimates that fail to inform health policy beyond basic inference with regard to the sign and statistical significance of the effect in question. Other undesirable consequences of this lack of rigor and definitional specificity include inaccurate measurement of causal effects, and misinterpreted empirical findings. The present paper offers a well-defined generic framework for analyzing health policy using nonexperimental data. We begin by offering a rigorous and policy relevant definition of the *effect of  $x_p$  on  $y$*  – the *policy effect*, and discuss the problems encountered in estimating policy effects with nonexperimental data. In section 3, we propose a generic framework for health policy effect estimation that is amenable to all types of econometric models: linear, parametric nonlinear, semi-parametric, and non-parametric. In section 4, we show how the model can be used to analyze policy effects for subgroups of the population, we discuss a partial derivative version of the policy effect framework, and derive the asymptotic standard error of the policy effect estimator. An illustrative example is detailed in Section 5 – estimation of the effect of substance abuse on employment status using the 1992

National Longitudinal Alcohol Epidemiologic Survey (NLAES). Additional examples are briefly discussed in Section 6. The final section summarizes and concludes.

## 2. Strict Definition of the term *Policy Effect* and the Problem with Nonexperimental Data

Loosely speaking, the term *policy effect* is defined as the amount by which the outcome of interest ( $y$ ) would differ between two counterfactual scenarios – one in which all individuals in the relevant population are mandated a given value of the policy variable ( $x_p$ ); another in which a different value of  $x_p$  is imposed. These scenarios are described as counterfactual because, in a nonexperimental (survey) context, they do not coincide with the observable data. Formally, the policy effect that we seek to estimate is

$$PE = E[y_{x_{p2}}] - E[y_{x_{p1}}] \quad (1)$$

where  $x_{p1}$  and  $x_{p2}$  are the respective mandated values of  $x_p$ , and  $y_{x_p^*}$  is the random variable denoting the value of  $y$  that would prevail under the counterfactual scenario in which the value of  $x_p$  is imposed at  $x_p^*$ .<sup>1</sup> Because of the potential endogeneity of  $x_p$ , attempts to estimate PE by applying conventional econometric methods to experimental data will often be biased. To understand the problem, consider the following example. Suppose  $y$  is a binary variable indicating whether ( $y=1$ ) or not ( $y=0$ ) the individual is employed, and the policy variable of interest is also a binary variable that represents the individual's substance abuse (SA) status ( $x_p = 1$  if abuse, 0 otherwise). The health policy analyst is interested in how exogenous reductions in SA will effect improvements in employability. Such exogenous reductions in SA might be attained through effective prevention and

---

<sup>1</sup>In general, superscript “\*” is used to denote counterfactual (mandated) values.

treatment policies. Consider the population of 10 individuals shown in Table 1, and define  $x_p^* = 1$  ( $x_p^* = 0$ ) if everyone (no one) in the population is a substance abuser.

**Table 1**

| Person        | $y_0 (x_p^* = 0)$ | $y_1(x_p^* = 1)$ |
|---------------|-------------------|------------------|
| ( $x_p = 0$ ) | A                 | C                |
| 1             | 1                 | 1                |
| 2             | 1                 | 1                |
| 3             | 1                 | 1                |
| 4             | 1                 | 0                |
| 5             | 0                 | 0                |
| ( $x_p = 1$ ) | B                 | D                |
| 6             | 1                 | 1                |
| 7             | 1                 | 0                |
| 8             | 0                 | 0                |
| 9             | 0                 | 0                |
| 10            | 0                 | 0                |

The second column in Table 1 contains the employment outcomes for the counterfactual “no SA” ( $x_p^* = 0$ ) scenario. The third column is similarly defined for the “SA” ( $x_p^* = 1$ ) scenario. It is clear that the value of the desired policy effect (1) in this case is the difference of the full column averages, i.e.

$$PE = E[y_1] - E[y_0] = .4 - .6 = -.2 \quad (2)$$

indicating that a fully effective SA treatment policy would, on average, result in a .2 increase in the probability of employment. The main problem in estimating the policy effect is that the population outcomes on  $y$  as displayed in the full second and third columns of Table 1 are not completely observable via survey (nonexperimental) sampling. In survey sampling, data on  $y_1$  is only observable for those individuals who happened to be substance abusers. Likewise, observations on  $y_0$  are only

available for those who are observed to be non-abusers. Let  $x_p$  be the random variable indicating observed SA status ( $x_p = 1$  if the individual is an abuser,  $x_p = 0$  otherwise). Let us suppose that the first five persons listed in the table are observed not to be substance abusers ( $x_p = 0$ ), while the remaining five are observed to be substance abusers ( $x_p = 1$ ). The shaded cells A and D contain the observable (factual) data on  $y$  – i.e. the values of  $y$  in cell D correspond to people who are observed to be abusers; the values of  $y$  in A correspond to people who are observed not to be abusers. Unshaded cells B and C contain the unobservable (counterfactual) data on  $y$  – i.e. the values of  $y$  in B are those that *would have been observed* for individuals observed to be abusers ( $x_p = 1$ ) *had they not been abusers* ( $x_p^* = 0$ ); the values of  $y$  in C are those of observed non-abusers ( $x_p = 0$ ) that *would have been observed if they were abusers* ( $x_p = 1$ ). In summary, the shaded cells contain factual (observable) data; the unshaded cells contain counterfactual (unobservable) data. Through survey sampling, we are unable to obtain unrestricted access to the relevant data for the estimation of the true (but counterfactual) policy effect as given in (1) [expression (2) in the context of our example]. In our example, we would like to freely sample from the full population, including the unshaded blocks of Table 1 (B and C). Unfortunately, we are restricted to sampling from the population data in the shaded blocks (A and D). Because of this sampling restriction, conventional unbiased methods applied to such data will be biased for the true policy effect. The reason for this is simple. When we sample from blocks A and D we are actually sampling from the *conditional distributions of  $y$  given the observed outcomes of the policy variable* – i.e.,  $(y | x_p = 0)$  and  $(y | x_p = 1)$ , respectively. So, for instance, the conventional difference-of-means (DOM) estimator will be unbiased for  $E[y | A = 1] - E[y | A = 0]$ , but it may not be unbiased for (1). For the example in Table 1 we have

$$E[y | A = 1] - E[y | A = 0] = .2 - .8 = -.6. \quad (3)$$

Clearly, in this example, the conventional DOM estimator will be biased for (2). As this discussion makes clear, if conventional unbiased estimators applied to the observable data are to produce unbiased estimates of PE as defined in (1), the following condition must be true

$$E[y_{x_p^*}] = E[y | x_p = x_p^*]. \quad (4)$$

Why would (4) fail to hold? It fails because there exist confounders – variables that affect  $y$  and are correlated with  $x_p$ . In such situations  $x_p$  is described as *endogenous*. So what can we do if (4) fails? We must find a way to control for the confounders.

### 3. Estimation of the Health Policy Effect While Taking Account of Confounders

As we have seen, it is the fact that condition (4) does not typically hold that is the main drawback in using survey data to estimate the policy effect defined as in (1). In such cases, it is clear that we must find a way to establish equality between the counterfactual expectation of interest on the left-hand side of (4), and the right-hand-side factual conditional expectation that characterizes sampling in nonexperimental contexts. We do this through the explicit inclusion of both observable ( $x_o$ ) and unobservable ( $x_u$ ) confounders.<sup>2</sup> We begin by rewriting the counterfactual value of the random variable representing the outcome ( $y$ ) at a fixed value of the policy variable in the following way

---

<sup>2</sup>For the purpose of exposition, we henceforth place the discussion in the context of parametric nonlinear regression models. The concepts involved here are, however, easily extended to semi- and nonparametric models.

$$y_{x_p^*} = H(x_p^*, x_o^*, x_u^*, \epsilon, \tau) \quad (5)$$

where  $H(\cdot)$  is a nonlinear function,  $\epsilon$  is the random error term (possibly a vector), and  $\tau$  is the vector of parameters. The random error  $\epsilon$  is tautologically independent of  $x_p^*$ ,  $x_o^*$  and  $x_u^*$ , given that the  $x$ 's are counterfactually defined.

A few comments are in order here. First note that  $H(\cdot)$  is the fundamental characterization of the policy relationship between the outcome  $y$  and the policy variable  $x_p$ . As we will see, it is the first principle from which we derive the fundamental tool for policy analysis. Secondly, for the purpose of policy analysis,  $x_p$  will be fixed at two different values by the policy analyst, but  $x_o^*$  and  $x_u^*$  may or may not be fixed values. It is, for example, rare that the policy analyst will know the relevant fixed values of  $x_o^*$  and  $x_u^*$ . Instead, for the purpose of evaluating the policy effect (1), one might take the expectation over the distribution of observable values of  $x_o$  and  $x_u$ . In any case, for the purpose of health policy analysis,  $x_o^*$  and  $x_u^*$  are viewed as counterfactual in the sense that their values are not affected by the counterfactually chosen values of  $x_p$ . It is in this sense that the policy effect (1) can be viewed as measuring the *direct effect* of  $x_p$  on  $y$ . In other words, by definition  $H(\cdot)$  does not allow for indirect effects of  $x_p$  through other variables.

As an example of (5), consider the case in which the outcome variable is binary. We might write (5) as

$$y_{x_p^*} = I(x_p^* \gamma + x_o^* \beta + x_u^* \pi + \epsilon > 0) \quad (6)$$

where  $I(C)$  denotes the index function which takes the value 1 if condition  $C$  is true (0 otherwise),

$\tau = [\gamma \ \beta' \ \pi']'$ , and  $\epsilon$  is standard normally distributed. This is the conventional probit regression model. How do we make (5) useful for policy analysis? The answer is to integrate it with respect to  $\epsilon$  to obtain what we will henceforth refer to as the counterfactual *regression function*. Specifically,

$$J(\mathbf{x}_p^*, \mathbf{x}_o^*, \mathbf{x}_u^*, \tau) = \int_{\epsilon} H(\mathbf{x}_p^*, \mathbf{x}_o^*, \mathbf{x}_u^*, \epsilon, \tau) f_{\epsilon}(\epsilon) d\epsilon \quad (7)$$

where  $f_{\epsilon}(\epsilon)$  denotes the pdf of  $\epsilon$ . It follows from (5) and (7) that

$$E[y_{x_p^*}] = E[J(\mathbf{x}_p^*, \mathbf{x}_o^*, \mathbf{x}_u^*, \tau)] \quad (8)$$

where the outside expectation on the right hand side allows for the possibility that the counterfactually determined values of the confounders ( $x_o^*$  and  $x_u^*$ ) are random variables. In our probit regression example we obtain

$$\begin{aligned} J(\mathbf{x}_p^*, \mathbf{x}_o^*, \mathbf{x}_u^*, \tau) &= \int_{\epsilon} I(\mathbf{x}_p^* \gamma + \mathbf{x}_o^* \beta + \mathbf{x}_u^* \pi + \epsilon > 0) \phi(\epsilon) d\epsilon \\ &= \Phi(\mathbf{x}_p^* \gamma + \mathbf{x}_o^* \beta + \mathbf{x}_u^* \pi) \end{aligned} \quad (9)$$

where  $\phi(\ )$  and  $\Phi(\ )$  denote the standard normal pdf and cdf, respectively. Given (8), the policy effect in (1) can be restated as

$$PE = E[y_{x_{p2}}] - E[y_{x_{p1}}] = E[J(\mathbf{x}_{p2}, \mathbf{x}_o^*, \mathbf{x}_u^*, \tau) - J(\mathbf{x}_{p1}, \mathbf{x}_o^*, \mathbf{x}_u^*, \tau)] \quad (10)$$

In the context of our probit regression analysis (10) becomes

$$PE = E\left[\Phi(\mathbf{x}_{p2}\boldsymbol{\gamma} + \mathbf{x}_o^*\boldsymbol{\beta} + \mathbf{x}_u^*\boldsymbol{\pi}) - \Phi(\mathbf{x}_{p1}\boldsymbol{\gamma} + \mathbf{x}_o^*\boldsymbol{\beta} + \mathbf{x}_u^*\boldsymbol{\pi})\right]. \quad (11)$$

Now, given a consistent estimator ( $\hat{\boldsymbol{\tau}}$ ) of  $\boldsymbol{\tau}$  from the observable (survey) data, PE in (10) can be consistently estimated using

$$\hat{PE} = \sum_{i=1}^n \frac{1}{n} \left\{ J(\mathbf{x}_p = \mathbf{x}_{p2}, \mathbf{x}_{oi}, \mathbf{x}_{ui}, \hat{\boldsymbol{\tau}}) - J(\mathbf{x}_p = \mathbf{x}_{p1}, \mathbf{x}_{oi}, \mathbf{x}_{ui}, \hat{\boldsymbol{\tau}}) \right\}. \quad (12)$$

In (12), and throughout the remainder of the paper, we presume that  $\boldsymbol{\tau}$  has been consistently estimated. It is, of course, the non-observability of  $\mathbf{x}_u$  that embodies the endogeneity problem. Because the present paper focuses on bottom-line policy analysis, we leave the reader to consult other sources on regression estimation in the presence of endogeneity – instrumental variables, generalized method of moments, etc. Later in the paper, however, we discuss an application of the methods proposed herein and offer a consistent estimator of  $\boldsymbol{\tau}$  in the context of the application. Whatever the chosen econometric method, the consistency of the estimator of  $\boldsymbol{\tau}$  is based on the assumption that

$$E[y \mid \mathbf{x}_p, \mathbf{x}_o, \mathbf{x}_u] = J(\mathbf{x}_p, \mathbf{x}_o, \mathbf{x}_u, \boldsymbol{\tau}) \quad (13)$$

which is the multiple regression analog to condition (4) in the sense that (13) implies

$$E[y_{x_p^*}] = E\left[E[y \mid \mathbf{x}_p, \mathbf{x}_o, \mathbf{x}_u]\right]. \quad (14)$$

As does (4), condition (14) means that, assuming that the nonobservability of  $\mathbf{x}_u$  can be dealt with, the counterfactual expectation in which we are interested, can be estimated via survey data.

## 4. Subgroup Analyses, Partial Effects, and Asymptotics

### 4.1 Subgroup Analyses

It may also be of interest to estimate the policy effect, PE, as defined in (10) for subgroups of the population. Such subgroups can be defined in terms of values, or ranges of values, for the policy variable and/or the observable confounders. Let  $g(x_p, x_o)$  denote the subpopulation of interest.

We can then write the policy effect for that subgroup as

$$PE_{g(x_p, x_o)} = E_{g(x_p, x_o)} \left[ J(x_{p2}, x_o^*, x_u^*, \tau) - J(x_{p1}, x_o^*, x_u^*, \tau) \right] \quad (15)$$

with corresponding consistent estimator

$$\hat{PE}_{g(x_p, x_o)} = \sum_{i=1}^{n_{g(x_p, x_o)}} \frac{1}{n_{g(x_p, x_o)}} \left\{ J(x_p^* = x_{p2}, x_{oi}^*, x_u^*, \hat{\tau}) - J(x_p^* = x_{p1}, x_{oi}^*, x_u^*, \hat{\tau}) \right\} \quad (16)$$

where  $n_{g(x_p, x_o)}$  denotes the size of the sample drawn from subpopulation  $g(x_p, x_o)$ .

### 4.2 The Partial Derivative Version of the Policy Effect Estimator

It may be the case that the researcher has no specific values for the policy variable  $x_p$  in mind.

In such cases, the following partial derivative version of (10) may be used

$$PE = \lim_{\Delta x_p \rightarrow 0} \frac{E[y_{x_p}] - E[y_{x_p + \Delta x_p}]}{\Delta x_p} \quad (17)$$

$$= \lim_{\Delta x_p \rightarrow 0} \frac{E \left[ J(x_p, x_o^*, x_u^*, \tau) \right] - E \left[ J(x_p + \Delta x_p, x_o^*, x_u^*, \tau) \right]}{\Delta x_p} = \frac{\partial E \left[ J(x_p, x_o^*, x_u^*, \tau) \right]}{\partial x_p} \quad (18)$$

which, if  $J(\cdot)$  satisfies the premises of the dominated convergence theorem (see Bierens 1994, p.25), implies that

$$PE = E \left[ \frac{\partial J(x_p, x_o^*, x_u^*, \tau)}{\partial x_p} \right]. \quad (19)$$

The consistent estimator corresponding to (19) is

$$\hat{PE} = \sum_{i=1}^n \frac{1}{n} \left\{ \frac{\partial J(x_{pi}, x_{oi}, x_u, \hat{\tau})}{\partial x_p} \right\}. \quad (20)$$

#### 4.3 The Asymptotic Standard Error of the Policy Effect Estimator

Let us now consider the asymptotic distribution of the policy estimator in (12). For the sake of exposition we begin the discussion with the unrealistic supposition that the policy effect,  $pe_i$  is observable for each individual in the sample. If this were the case, a consistent and asymptotically normal estimator of the population policy effect would be

$$\hat{pe} = \frac{\sum_{i=1}^n pe_i}{n} \quad (21)$$

which is the optimizer of the following estimation objective function

$$\frac{\sum_{i=1}^n q^*(pe_i, PE)}{n} \quad (22)$$

where  $q^*(pe_i, PE) = (pe_i - PE)^2$ . In actuality, however,  $pe_i$  is not directly observable. Instead, we have

$$\text{pe}(\tau, \mathbf{x}_{oi}, \mathbf{x}_{ui}) = J(\mathbf{x}_p = \mathbf{x}_{p2}, \mathbf{x}_{oi}, \mathbf{x}_{ui}, \tau) - J(\mathbf{x}_p = \mathbf{x}_{p1}, \mathbf{x}_{oi}, \mathbf{x}_{ui}, \tau).$$

Therefore, in reality, the objective function in (22) should be written

$$Q_n(\tau, \text{PE}) = \frac{\sum_{i=1}^n q(\tau, \text{PE}, \mathbf{x}_{oi}, \mathbf{x}_{ui})}{n} \quad (23)$$

where  $q(\tau, \text{PE}, \mathbf{x}_{oi}, \mathbf{x}_{ui}) = (\text{pe}(\tau, \mathbf{x}_{oi}, \mathbf{x}_{ui}) - \text{PE})^2$ . In principle, estimates of  $\tau$  and PE could be directly obtained as the optimizers of the objective function (13). In general, however, we suggest using the two stage approach which is implicit in (12) and which we make explicit here. In the first stage,  $\tau$  is estimated as the optimizer of a first stage objective function of the form

$$Q_{1n}(\tau) = \frac{\sum_{i=1}^n q_1(\tau, y_i, \mathbf{x}_{pi}, \mathbf{x}_{oi}, \mathbf{x}_{ui})}{n} \quad (24)$$

where the exact form of  $q_1(\tau, y_i, \mathbf{x}_{pi}, \mathbf{x}_{oi}, \mathbf{x}_{ui})$  is determined by the extent to which the moments of the conditional distribution of  $y$  given  $x_p$ ,  $x_o$ , and  $x_u$  are specified, and how the nonobservability of  $x_{ui}$  is dealt with. Some insight into this issue will be given in the context of the application discussed in the next section. The second stage estimator of PE [given in (12)], can be equivalently represented as the optimizer of the following version of (23) with respect to PE

$$Q_n(\hat{\tau}, \text{PE}) = \frac{\sum_{i=1}^n q(\hat{\tau}, \text{PE}, \mathbf{x}_{oi}, \mathbf{x}_{ui})}{n} \quad (25)$$

where  $\hat{\tau}$  denotes the first-stage estimator of  $\tau$  obtained from (24). Rewriting the policy effect estimator (12) in this way as a two-stage estimator allows us to invoke standard asymptotic results

for two-stage estimators (see White, 1994, Chapter 6).

The following notational conventions will be maintained for a scalar function  $s$  of two vector arguments  $r$  and  $t$  (i.e.  $s = s(r, t)$  where  $s$  is a scalar and  $r$  and  $t$  are vectors):

$$\nabla_r s = \frac{\partial s}{\partial r}$$

and

$$\nabla_{rt} s = \frac{\partial^2 s}{\partial r \partial t}.$$

We also assume that the former is a row vector, and the latter is a matrix with row dimension equal to that of the first subscript on  $\nabla$ , and column dimension equal to that of the second subscript. Under the regularity conditions given in Theorem 6.11 of White (1994), the policy effect estimator (12) is consistent and

$$\sqrt{\frac{n}{\text{avar}(\hat{P}E)}} (\hat{P}E - P) \xrightarrow{d} n(0, 1) \quad (26)$$

where “ $d$ ” denotes convergence in distribution,  $n(0, 1)$  represents the standard normal variate, and

$$\begin{aligned} \text{avar}(\hat{P}E) &\equiv E[\nabla_{PE PE} q]^{-1} \left[ E[\nabla_{PE \tau} q] \text{AVAR}(\hat{\tau}) E[\nabla_{PE \tau} q]' \right. \\ &\quad - E[\nabla_{PE q} \nabla_{\tau} q_1] E[\nabla_{\tau \tau} q_1]^{-1} E[\nabla_{PE \tau} q]' - E[\nabla_{PE \tau} q] E[\nabla_{\tau \tau} q_1]^{-1} E[\nabla_{\tau} q_1' \nabla_{PE} q] \\ &\quad \left. + E[\nabla_{PE q}^2] \right] E[\nabla_{PE PE} q]^{-1} \end{aligned} \quad (27)$$

where  $\text{AVAR}(\hat{\tau})$  denotes the asymptotic covariance matrix of the first-stage estimator of  $\tau$ . Now

$$\begin{aligned} E[\nabla_{PE} \mathbf{q} \nabla_{\tau} \mathbf{q}_1] &= -2E[(\mathbf{pe}(\tau, \mathbf{x}_{oi}, \mathbf{x}_{ui}) - PE) \nabla_{\tau} \mathbf{q}_1] \\ &= -2E[(\mathbf{pe}(\tau, \mathbf{x}_{oi}, \mathbf{x}_{ui}) - PE) E[\nabla_{\tau} \mathbf{q}_1 | \mathbf{x}_{pi}, \mathbf{x}_{oi}, \mathbf{x}_{ui}]] \end{aligned}$$

but typically  $E[\nabla_{\tau} \mathbf{q}_1 | \mathbf{x}_{pi}, \mathbf{x}_{oi}, \mathbf{x}_{ui}] = \mathbf{0}$  so  $E[\nabla_{PE} \mathbf{q} \nabla_{\tau} \mathbf{q}_1] = \mathbf{0}$ . Moreover,

$$\nabla_{PE} PE \mathbf{q} = \mathbf{2}.$$

Therefore

$$\text{avar}(\hat{PE}) \equiv \frac{1}{4} [E[\nabla_{PE} \mathbf{q}] \text{AVAR}(\hat{\tau}) E[\nabla_{PE} \mathbf{q}]' + E[\nabla_{PE} \mathbf{q}^2]]. \quad (28)$$

Note that (28) differs from the standard  $\delta$ -method (see Greene, 2003, p. 674). The standard  $\delta$ -method treats the confounders as fixed (usually at the sample mean), and does not account for the inherent two-stage nature of “effects” estimators. We will revisit this issue in the context of the illustrative example in the next section.

## 5. An Application: Estimation of the Effect of Substance Abuse on Employment

### 5.1 The Model

Here we return to the example discussed in section 2 -- estimation of the policy effect of substance abuse (SA) on employment. In this case the health policy analyst is interested in how exogenous reductions in SA, perhaps effected by prevention and treatment policies, may cause improvements in individual employability. Recall that

$$y = \begin{cases} 1 & \text{if the individual is employed} \\ 0 & \text{otherwise} \end{cases}$$

and

$$x_p = \begin{cases} 1 & \text{if the individual is a substance abuser} \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the counterfactual model for policy analysis is given in (6) and (9) [repeated here for convenience of exposition]

$$y_{x_p^*} = I(x_p^* \gamma + x_o^* \beta + x_u^* \pi + \epsilon > 0) \quad (29)$$

and

$$\begin{aligned} J(x_p^*, x_o^*, x_u^*, \tau) &= \int_{\epsilon} I(x_p^* \gamma + x_o^* \beta + x_u^* \pi + \epsilon > 0) \phi(\epsilon) d\epsilon \\ &= \Phi(x_p^* \gamma + x_o^* \beta + x_u^* \pi). \end{aligned} \quad (30)$$

Given that the policy variable ( $x_p$ ) in this example is binary, the policy effect (11) becomes

$$PE = E \left[ \Phi(\gamma + x_o^* \beta + x_u^* \pi) - \Phi(x_o^* \beta + x_u^* \pi) \right]. \quad (31)$$

## 5.2 The Data

The data used in this illustrative application were taken from Wave 1 (1992) of the National Longitudinal Alcohol Epidemiologic Survey (NLAES). Wave 1 of the NLAES is composed of data for adults age 18 years and older taken from a survey of U.S. households over the period from October 1991 to November 1992. Two follow-up waves were planned using the same respondents,

but these waves have not yet been funded. The primary purpose of the NLAES is the collection of data on the incidence and prevalence of alcohol abuse and dependence and associated disabilities. Fortunately for the present purpose, similar information was collected for illicit drugs and pharmaceuticals.

The NLAES contains a wealth of information on relevant socio-economic-demographic (SED) control variables to be included in  $x_0$ , such as age, gender, race, marital status, geographic region of residence, whether or not the respondent lives in an urban setting, education level, the number of problematic health conditions from which an individual currently suffers, and the quarter in which the interview took place. The state-level unemployment rate for 1992 has also be included in  $x_0$ . This was obtained from the Statistical Abstract of the US. The NLAES includes indicators of current occupation from which the categorical employment status outcome vector ( $y$ ) was constructed. Market-eligible respondents who are unemployed or out of the labor force were categorized as not employed ( $y = 0$ ), all others were categorized as employed ( $y=1$ ). Because the objective is to estimate the effect of SA on employment, individuals less than 24 or more than 59 years of age, and full-time homemakers (mostly women) are excluded from the sample. These are groups of individuals who are likely to have low labor force participation rates. The young are typically in school, many older workers retire soon after age 59, and homemakers choose to remain out of the work force for important reasons such as child rearing.

In addition, the NLAES contains information on alcohol and drug use disorders based on the DSM-IV diagnosis criteria from which the binary SA variable ( $x_p$ ) will be coded. The individual will be defined as a substance abuser ( $x_p = 1$ ) if he meets the DSM-IV criteria for current abuse and/or dependence for any of the following: alcohol, marijuana, cocaine, methadone, hallucinogens, opiates,

sedatives, stimulants, tranquilizers, or heroine. The DSM-IV criteria are given in Table 1. For identification purposes, in our test for endogeneity we included the following instrumental variables: state-level apparent ethanol consumption; state-level beer and cigarette tax rates; and alcoholism of a biological parent. Definitions of the variables are given in Table 2, and descriptive statistics are displayed in Table 3.

### 5.3 Results

We first conducted a test of the endogeneity of the substance abuse variable ( $x_p$ ). We did so by estimating a bivariate probit model with  $y$  and  $x_p$  as the two outcome variable. The results, displayed in Table 4, yielded an estimate of the correlation coefficient that is not significantly different from zero. In the Appendix it is shown that a zero bivariate probit correlation coefficient is, in the context of the model defined in (29), tantamount to the parameter  $\pi$  being equal to zero. The results, therefore, yield no evidence of the endogeneity of the policy variable ( $x_p$ ). For this reason we impose the restriction that  $\pi = 0$ , and rewrite the model in (29) and (30) as

$$y_{x_p^*} = I(x_p^* \gamma + x_o^* \beta + \epsilon > 0) \quad (32)$$

and

$$\begin{aligned} J(x_p^*, x_o^*, \tau) &= \int_{\epsilon} I(x_p^* \gamma + x_o^* \beta + \epsilon > 0) \phi(\epsilon) d\epsilon \\ &= \Phi(x_p^* \gamma + x_o^* \beta). \end{aligned} \quad (33)$$

Accordingly we maintain the following conditional moment restriction which is the appropriate version of (4) in this context

$$E[y \mid \mathbf{x}_p, \mathbf{x}_o] = \Phi(\mathbf{x}_p \boldsymbol{\gamma} + \mathbf{x}_o \boldsymbol{\beta}). \quad (34)$$

The parameters of the model ( $\boldsymbol{\gamma}$  and  $\boldsymbol{\beta}$ ) were then estimated via simple probit analysis. The probit estimates are displayed in Table 5. The policy effect was estimated as

$$\hat{P}E = \sum_{i=1}^n \frac{1}{n} \left\{ \Phi(\hat{\boldsymbol{\gamma}} + \mathbf{x}_{oi} \hat{\boldsymbol{\beta}}) - \Phi(\mathbf{x}_{oi} \hat{\boldsymbol{\beta}}) \right\} = -.021 \quad (35)$$

where  $\hat{\boldsymbol{\gamma}}$  and  $\hat{\boldsymbol{\beta}}$  are the probit estimates of  $\boldsymbol{\gamma}$  and  $\boldsymbol{\beta}$ , respectively.

Let us now consider the computation of the standard error of  $\hat{P}E$ . It follows from (26), (28), and the consistency of the probit estimator of  $\boldsymbol{\gamma}$  and  $\boldsymbol{\beta}$ , that

$$\sqrt{\frac{n}{\text{avar}(\hat{P}E)}} (\hat{P}E - P) \xrightarrow{d} n(0, 1) \quad (36)$$

where

$$\text{avar}(\hat{P}E) \equiv \frac{1}{4} \left[ \left( \frac{\sum_{i=1}^n \nabla_{PE \tau} \hat{\mathbf{q}}_i}{n} \right) \left( \frac{\sum_{i=1}^n \hat{G}_i}{n} \right)^{-1} \left( \frac{\sum_{i=1}^n \nabla_{PE \tau} \hat{\mathbf{q}}_i'}{n} \right) + \left( \frac{\sum_{i=1}^n \nabla_{PE} \hat{\mathbf{q}}_i^2}{n} \right) \right]. \quad (37)$$

$$\nabla_{PE} \hat{\mathbf{q}}_i = -2 \left( \Phi(\hat{\boldsymbol{\gamma}} + \mathbf{x}_{oi} \hat{\boldsymbol{\beta}}) - \Phi(\mathbf{x}_{oi} \hat{\boldsymbol{\beta}}) - \hat{P}E \right)$$

$$\nabla_{PE \tau} \hat{\mathbf{q}}_i = \left[ \nabla_{PE \boldsymbol{\gamma}} \hat{\mathbf{q}}_i \quad \nabla_{PE \boldsymbol{\beta}} \hat{\mathbf{q}}_i \right]$$

$$\nabla_{PE \boldsymbol{\gamma}} \hat{\mathbf{q}}_i = -2 \phi(\hat{\boldsymbol{\gamma}} + \mathbf{x}_{oi} \hat{\boldsymbol{\beta}})$$

$$\nabla_{\mathbf{PE}\beta} \hat{q}_i = -2 \left( \phi(\hat{\gamma} + \mathbf{x}_{oi}\hat{\beta}) - \phi(\mathbf{x}_{oi}\hat{\beta}) \right) \mathbf{x}_{oi}.$$

and  $\hat{\mathbf{G}}_i$  denotes the probit hessian matrix evaluated at the estimated parameters using the observed value of  $\mathbf{x}_o$  for the  $i$ th individual. Using (36) and (37) we computed the asymptotic t-stat for  $\hat{\mathbf{P}}\mathbf{E}$  as

$$\frac{(\hat{\mathbf{P}}\mathbf{E} - \mathbf{P})}{\sqrt{\frac{\mathbf{avar}(\hat{\mathbf{P}}\mathbf{E})}{\mathbf{n}}}} = -3.52. \quad (38)$$

These results imply that substance abuse reduces the likelihood of employment by .021, and that result is statistically significantly different from zero.

It is also interesting to note that the t-stat in (38) differs from that which would have been obtained by the standard  $\delta$ -method computed with  $\mathbf{x}_o$  fixed at its sample mean. According to the standard  $\delta$ -method, the estimator of the asymptotic variance of  $\hat{\mathbf{P}}\mathbf{E}$  would be

$$\mathbf{avar}(\hat{\mathbf{P}}\mathbf{E}) = \hat{\mathbf{f}}(\bar{\mathbf{x}}_o) \left( \frac{\sum_{i=1}^{\mathbf{n}} \hat{\mathbf{G}}_i}{\mathbf{n}} \right)^{-1} \hat{\mathbf{f}}(\bar{\mathbf{x}}_o). \quad (39)$$

where  $\hat{\mathbf{f}}(\bar{\mathbf{x}}_o) = [\phi(\hat{\gamma} + \bar{\mathbf{x}}_o\hat{\beta}) \quad \phi(\bar{\mathbf{x}}_o\hat{\beta}) \bar{\mathbf{x}}_o]$ . The  $\delta$ -method formula in (39) is in error because it treats the confounders as fixed, and does not account for the inherent two-stage nature of  $\hat{\mathbf{P}}\mathbf{E}$ . Using (39) the errant version of the t-stat is

$$\frac{(\hat{\mathbf{P}}\mathbf{E} - \mathbf{P})}{\sqrt{\frac{\mathbf{avar}(\hat{\mathbf{P}}\mathbf{E})}{\mathbf{n}}}} = -3.74. \quad (40)$$

The use of the  $\delta$ -method results in a 7% error in the t-stat. In this case, the error is not large

enough to lead to false rejection of the null. One can, however, take this example as a cautionary note in support of the asymptotic variance formulation given in (28).

## 6. More Examples

Let us consider a few more common modeling contexts in which the proposed generic framework for health policy analysis can be applied.

### 6.1 The Linear Model

In this case, the fundamental equation (5) can be written as

$$\begin{aligned} y_{x_p^*} &= H(x_p^*, x_o^*, x_u^*, \epsilon, \tau) \\ &= x_p^* \gamma + x_o^* \beta + x_u^* \pi + \epsilon \end{aligned} \quad (41)$$

where  $\epsilon$  is distributed such that  $E[\epsilon] = 0$ . Under these conditions we obtain the counterfactual policy analytic regression (7) can be derived as

$$\begin{aligned} J(x_p^*, x_o^*, x_u^*, \tau) &= \int_{\epsilon} H(x_p^*, x_o^*, x_u^*, \epsilon, \tau) f_{\epsilon}(\epsilon) d\epsilon \\ &= x_p^* \gamma + x_o^* \beta + x_u^* \pi. \end{aligned} \quad (42)$$

Using (10) the health policy effect can in this case be written

$$PE = E[x_{p2} \gamma + x_o^* \beta + x_u^* \pi - x_{p1} \gamma + x_o^* \beta + x_u^* \pi] = (x_{p2} - x_{p1}) \gamma. \quad (43)$$

Given consistent estimates of the parameters, the corresponding health policy effect estimator (12) in this case is

$$\hat{PE} = (x_{p2} - x_{p1})\hat{\gamma} \quad (44)$$

where  $\hat{\gamma}$  is a consistent estimates of  $\gamma$ . Based on the following version of the conditional moment restriction (13)

$$E[y | x_p, x_o, x_u] = x_p\gamma + x_o\beta + x_u\pi \quad (45)$$

the conventional instrumental variables method can be used to obtain consistent estimates of the parameters.

## 6.2 A Flexible-Form Nonlinear Regression Model

Box and Cox (1964) suggest an extension of the classical linear regression model using the following transformation of the outcome variable  $y$

$$bc(y) = \begin{cases} \frac{y^\omega - 1}{\omega} & \omega \neq 0 \quad y \geq 0 \\ \ln(y) & \omega = 0 \quad y > 0 \end{cases} \quad (46)$$

The Box-Cox specification is attractive because it nests both the linear ( $\omega = 1$ ) and log-linear ( $\omega = 0$ ) models. Wooldridge (1992), however, highlights a number of well-known shortcomings of (46), and suggests the use of its inverse in regression modeling. The inverse Box-Cox transformation is

$$\text{ibc}(v, \eta) = \begin{cases} (\eta(v) + 1)^{\frac{1}{\eta}} & \eta \neq 0 \\ \exp(v) & \eta = 0. \end{cases} \quad (47)$$

The transformation in (47) is flexible in the sense that it subsumes both linear ( $\eta = 1$ ) and exponential ( $\eta = 0$ ) regression formulations.

Under the inverse Box-Cox formulation, the fundamental equation (5) can be written as

$$\begin{aligned} y_{x_p^*} &= H(x_p^*, x_o^*, x_u^* \epsilon, \tau) \\ &= (\eta(x_p^* \gamma + x_o^* \beta + x_u^* \pi) + 1)^{\frac{1}{\eta}} + \epsilon \end{aligned} \quad (48)$$

where  $\epsilon$  is distributed such that  $E[\epsilon] = 0$ . Under these conditions we obtain the counterfactual policy analytic regression (7) can be derived as

$$\begin{aligned} J(x_p^*, x_o^*, x_u^*, \tau) &= \int_{\epsilon} H(x_p^*, x_o^*, x_u^* \epsilon, \tau) f_{\epsilon}(\epsilon) d\epsilon \\ &= (\eta(x_p^* \gamma + x_o^* \beta + x_u^* \pi) + 1)^{\frac{1}{\eta}}. \end{aligned} \quad (49)$$

Using (10) the health policy effect can in this case be written

$$\text{PE} = E \left[ (\eta(x_{p2} \gamma + x_o^* \beta + x_u^* \pi) + 1)^{\frac{1}{\eta}} - (\eta(x_{p1} \gamma + x_o^* \beta + x_u^* \pi) + 1)^{\frac{1}{\eta}} \right] \quad (50)$$

Given consistent estimates of the parameters, the corresponding health policy effect estimator (12) in this case is

$$\hat{\text{PE}} = \sum_{i=1}^n \frac{1}{n} \left\{ (\hat{\eta}(x_{p2} \hat{\gamma} + x_{oi} \hat{\beta} + x_{ui} \hat{\pi}) + 1)^{\frac{1}{\hat{\eta}}} - (\hat{\eta}(x_{p1} \hat{\gamma} + x_{oi} \hat{\beta} + x_{ui} \hat{\pi}) + 1)^{\frac{1}{\hat{\eta}}} \right\} \quad (51)$$

where  $\hat{\eta}$ ,  $\hat{\gamma}$ ,  $\hat{\beta}$ , and  $\hat{\pi}$  are consistent estimates of  $\eta$ ,  $\gamma$ ,  $\beta$ , and  $\pi$ . Based on the following version of the conditional moment restriction (13)

$$E[y \mid x_p, x_o, x_u] = (\eta (x_p \gamma + x_o \beta + x_u \pi) + 1)^{\frac{1}{\eta}} \quad (52)$$

consistent estimates of the parameters can be obtained. Kenkel and Terza (2001) implement the method developed by McGeary and Terza (1998) for dealing with the nonobservability of  $x_u$  and obtaining consistent estimates of the parameters.

### 6.3 The Two-Part Model

In this case, the fundamental equation (5) can be written as

$$\begin{aligned} y_{x_p^*} &= H(x_p^*, x_o^*, x_u^*, \epsilon, \tau) \\ &= I(x_p^* \gamma_1 + x_o^* \beta_1 + x_u^* \pi_1 + \epsilon_1 > 0) \exp(x_p^* \gamma_2 + x_o^* \beta_2 + x_u^* \pi_2 + \epsilon_2) \end{aligned} \quad (53)$$

where  $\epsilon = [\epsilon_1 \ \epsilon_2]$  with  $\epsilon_1$  standard normally distributed and  $\epsilon_2$  distributed such that  $E[\exp(\epsilon_2)] = 1$ .

Under these conditions we obtain the counterfactual policy analytic regression (7) can be derived as

$$\begin{aligned} J(x_p^*, x_o^*, x_u^*, \tau) &= \int_{\epsilon} H(x_p^*, x_o^*, x_u^*, \epsilon, \tau) f_{\epsilon}(\epsilon) d\epsilon \\ &= \Phi(x_p^* \gamma_1 + x_o^* \beta_1 + x_u^* \pi_1) \exp(x_p^* \gamma_2 + x_o^* \beta_2 + x_u^* \pi_2). \end{aligned} \quad (54)$$

Using (10) the health policy effect can in this case be written

$$PE = E \left[ \Phi(x_{p2}^* \gamma_1 + x_o^* \beta_1 + x_u^* \pi_1) \exp(x_{p2}^* \gamma_2 + x_o^* \beta_2 + x_u^* \pi_2) \right]$$

$$- \Phi(x_{p1}^* \gamma_1 + x_o^* \beta_1 + x_u^* \pi_1) \exp(x_{p1}^* \gamma_2 + x_o^* \beta_2 + x_u^* \pi_2)]. \quad (55)$$

Using (10) the health policy effect can in this case be written

$$\begin{aligned} PE = E & \left[ \Phi(x_{p2} \gamma_1 + x_o^* \beta_1 + x_u^* \pi_1) \exp(x_{p2} \gamma_2 + x_o^* \beta_2 + x_u^* \pi_2) \right. \\ & \left. - \Phi(x_{p1}^* \gamma_1 + x_o^* \beta_1 + x_u^* \pi_1) \exp(x_{p1}^* \gamma_2 + x_o^* \beta_2 + x_u^* \pi_2) \right]. \end{aligned} \quad (56)$$

Given consistent estimates of the parameters, the corresponding health policy effect estimator (12) in this case is

$$\begin{aligned} \hat{PE} = \sum_{i=1}^n \frac{1}{n} & \left\{ \Phi(x_{p2} \hat{\gamma}_1 + x_{oi} \hat{\beta}_1 + x_{ui} \hat{\pi}_1) \exp(x_{p2} \hat{\gamma}_2 + x_{oi} \hat{\beta}_2 + x_{ui} \hat{\pi}_2) \right. \\ & \left. - \Phi(x_{p1} \hat{\gamma}_1 + x_{oi} \hat{\beta}_1 + x_{ui} \hat{\pi}_1) \exp(x_{p1} \hat{\gamma}_2 + x_{oi} \hat{\beta}_2 + x_{ui} \hat{\pi}_2) \right\} \end{aligned} \quad (57)$$

where  $\hat{\gamma}$ s,  $\hat{\beta}$ s, and  $\hat{\pi}$ s are consistent estimates of the  $\gamma$ s,  $\beta$ s, and  $\pi$ s. Based on the following version of the conditional moment restriction (13)

$$E[y | x_p, x_o, x_u] = \Phi(x_p \gamma_1 + x_o \beta_1 + x_u \pi_1) \exp(x_p \gamma_2 + x_o \beta_2 + x_u \pi_2) \quad (58)$$

Shea, Stuart, Briesacher, and Terza (2004) suggest a method for dealing with the nonobservability of  $x_u$ , and a corresponding method for obtaining consistent estimates of the parameters. Kenkel, Lin, Sakata, and Terza (2004) offer an alternative estimation method for the case in which  $y$  is categorical rather than continuous.

## 6.4 The Multinomial Logit Model

In this case, the mutually exclusive and collectively exhaustive outcome is represented by the vector  $y = [y_1 y_2 \dots y_j]$ , and for each  $j = 1, \dots, K$  the fundamental equation (5) can be written as

$$\begin{aligned} y_{j,x_p^*} &= H_j(x_p^*, x_o^*, x_u^* \epsilon, \tau) \\ &= I(y_j = \max\{y_r^0; r = 1, \dots, K\}) \end{aligned} \quad (59)$$

where  $y_r^0 = x_p^* \gamma_r + x_o^* \beta_r + x_u^* \pi_r + \epsilon_r$ ,  $y_r^0$  is the unobservable “indirect utility” index representing the proclivity to be observed in the  $r^{\text{th}}$  category, and  $\gamma_r$ ,  $\beta_r$ , and  $\pi_r$  are parameters. The vector  $\epsilon = [\epsilon_1, \dots, \epsilon_j]$  is assumed to be i.i.d. log-Weibull distributed. Note that if  $y_j = 1$  then  $y_r = 0$  for all  $r \neq j$ . Under these conditions we obtain the counterfactual policy analytic regression (7) can be derived as

$$J_1(x_p^*, x_o^*, x_u^*, \tau) = \int_{\epsilon} H_1(x_p^*, x_o^*, x_u^* \epsilon, \tau) f_{\epsilon}(\epsilon) d\epsilon = \frac{1}{1 + \sum_{r=2}^K \exp\{x_p^* \gamma_r + x_o^* \beta_r + x_u^* \pi_r\}}$$

and

$$J_m(x_p^*, x_o^*, x_u^*, \tau) = \int_{\epsilon} H_m(x_p^*, x_o^*, x_u^* \epsilon, \tau) f_{\epsilon}(\epsilon) d\epsilon = \frac{\exp\{x_p^* \gamma_m + x_o^* \beta_m + x_u^* \pi_m\}}{1 + \sum_{r=2}^K \exp\{x_p^* \gamma_r + x_o^* \beta_r + x_u^* \pi_r\}}$$

$$\text{(for } m = 2, \dots, K) \quad (60)$$

Using (10) the health policy effect can in this case be written

$$PE_1 = E \left[ \frac{1}{1 + \sum_{r=2}^K \exp\{x_{p2} \gamma_r + x_o^* \beta_r + x_u^* \pi_r\}} - \frac{1}{1 + \sum_{r=2}^K \exp\{x_{p1} \gamma_r + x_o^* \beta_r + x_u^* \pi_r\}} \right]$$

$$PE_m = E \left[ \frac{\exp\{x_{p2}\gamma_m + x_o^*\beta_m + x_u^*\pi_m\}}{1 + \sum_{r=2}^K \exp\{x_{p2}\gamma_r + x_o^*\beta_r + x_u^*\pi_r\}} - \frac{\exp\{x_{p1}\gamma_m + x_o^*\beta_m + x_u^*\pi_m\}}{1 + \sum_{r=2}^K \exp\{x_{p1}\gamma_r + x_o^*\beta_r + x_u^*\pi_r\}} \right] \quad (\text{for } m = 2, \dots, K) \quad (61)$$

Given consistent estimates of the parameters, the corresponding health policy effect estimator (12)

in this case is

$$\hat{PE}_1 = \sum_{i=1}^n \frac{1}{n} \left\{ \frac{1}{1 + \sum_{r=2}^K \exp\{x_{p2}\hat{\gamma}_r + x_{oi}\hat{\beta}_r + x_{ui}\hat{\pi}_r\}} - \frac{1}{1 + \sum_{r=2}^K \exp\{x_{p1}\hat{\gamma}_r + x_{oi}\hat{\beta}_r + x_{ui}\hat{\pi}_r\}} \right\}$$

$$\hat{PE}_m = \sum_{i=1}^n \frac{1}{n} \left\{ \frac{\exp\{x_{p2}\hat{\gamma}_m + x_{oi}\hat{\beta}_m + x_{ui}\hat{\pi}_m\}}{1 + \sum_{r=2}^K \exp\{x_{p2}\hat{\gamma}_r + x_{oi}\hat{\beta}_r + x_{ui}\hat{\pi}_r\}} - \frac{\exp\{x_{p1}\hat{\gamma}_m + x_{oi}\hat{\beta}_m + x_{ui}\hat{\pi}_m\}}{1 + \sum_{r=2}^K \exp\{x_{p1}\hat{\gamma}_r + x_{oi}\hat{\beta}_r + x_{ui}\hat{\pi}_r\}} \right\} \quad (\text{for } m = 2, \dots, K) \quad (62)$$

where  $\hat{\gamma}$ s,  $\hat{\beta}$ s, and  $\hat{\pi}$ s are consistent estimates of the  $\gamma$ s,  $\beta$ s, and  $\pi$ s. Based on the following version of the conditional moment restriction (13)

$$E[y_1 | x_p, x_o, x_u] = \frac{1}{1 + \sum_{r=2}^K \exp\{x_p\gamma_r + x_o\beta_r + x_u\pi_r\}}$$

and

$$E[y_m | x_p, x_o, x_u] = \frac{\exp\{x_p\gamma_m + x_o\beta_m + x_u\pi_m\}}{1 + \sum_{r=2}^K \exp\{x_p\gamma_r + x_o\beta_r + x_u\pi_r\}} \quad (\text{for } m = 2, \dots, K) \quad (63)$$

Terza (2002) suggest a method for dealing with the nonobservability of  $x_{it}$ , and a corresponding method for obtaining consistent estimates of the parameters.

## **7. Discussion**

This paper offers a generic and unified framework for the use of nonexperimentally based regression results for health policy analysis. It is hoped that this work will serve as a guide to researchers in applied health economics who seek to characterize and present their empirical results in a way that is useful to health policy analysts and health policy makers.

## Appendix

Assuming that all of the confounders (observable and unobservable) are accounted for in  $x_o$  and  $x_u$ , the specification of the model in (29) and (30) is completed by the following version of the conditional moment restriction (13)

$$E[y \mid x_p, x_o, x_u] = \Phi(x_p \gamma + x_o \beta + x_u \pi). \quad (\text{A-1})$$

In order to deal with the nonobservability of  $x_u$  and obtain a consistent estimate of  $\tau = [\gamma \ \beta' \ \pi]'$ , we formalize the correlation between  $x_p$  and  $x_u$  by assuming that

$$x_p = I(w\alpha + x_u > 0) \quad (\text{A-2})$$

where  $x_u$  is standard normally distributed and  $w$  is a vector of observable variables such that at least one of the elements of  $w$  is not included in  $x_o$ . In other words, the endogenous substance abuse outcomes also follow a conventional probit regression setup. Using (A-1) and (A-2) we can fully specify the joint conditional pdf of  $y$  and  $x_p$  given  $x_o$  and  $w$  as

$$f(y, x_p \mid x_o, w) = P_{11}^{yx_p} P_{10}^{y(1-x_p)} P_{01}^{(1-y)x_p} P_{00}^{(1-y)(1-x_p)} \quad (\text{A-3})$$

where

$$P_{11} = \int_{-w\alpha}^{\infty} \Phi(\gamma + x_o \beta + x_u \pi) \phi(x_u) dx_u$$

$$P_{10} = \int_{-\infty}^{-w\alpha} \Phi(x_o \beta + x_u \pi) \phi(x_u) dx_u$$

$$P_{01} = \int_{-w\alpha}^{\infty} [1 - \Phi(\gamma + x_o \beta + x_u \pi)] \phi(x_u) dx_u$$

and

$$P_{00} = \int_{-\infty}^{-w\alpha} [1 - \Phi(x_0\beta + x_u\pi)] \phi(x_u) dx_u.$$

Now make the following substitutions in  $P_{11}$ ,

$$\gamma = \frac{\omega}{\sqrt{1 - \rho^2}}$$

$$\beta = \frac{1}{\sqrt{1 - \rho^2}} \psi$$

and

$$\pi = \frac{\rho}{\sqrt{1 - \rho^2}}. \tag{A-4}$$

Rewriting  $P_{11}$  we obtain

$$P_{11} = \int_{-w\alpha}^{\infty} \Phi\left(\left(\frac{\omega}{\sqrt{1 - \rho^2}}\right) + x_0\left(\frac{1}{\sqrt{1 - \rho^2}} \psi\right) + x_u\left(\frac{\rho}{\sqrt{1 - \rho^2}}\right)\right) \phi(x_u) dx_u$$

$$= \int_{-w\alpha}^{\infty} \int_{-(\omega+x_0\psi)}^{\infty} \phi_2(\epsilon, x_u, \rho) d\epsilon dx_u$$

where  $\phi_2(\epsilon, x_u, \rho)$  denotes the standard bivariate normal pdf with correlation coefficient  $\rho$ . The expressions for  $P_{10}$ ,  $P_{01}$ , and  $P_{00}$  be similarly rewritten so that (A-3) can be reparameterized in terms of  $\omega$ ,  $\psi$ , and  $\rho$ ; and under this reparameterization, the relevant likelihood function conforms to the following conventional bivariate probit model with one of the equations being (A-2) and the other

being

$$y = I(x_p\omega + x_o\psi + \varepsilon > 0)$$

where  $x_u$  and  $\varepsilon$  are standard bivariate normally distributed with correlation coefficient  $\rho$ . It is clear from (A-4) that  $\pi = 0$  if and only if  $\rho = 0$ . This means that we can test for the endogeneity of  $x_p$  by estimating a conventional bivariate probit and testing the null hypothesis that  $\rho = 0$ .

## References

- Bierens, H. J. (1994): *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-section and Time Series Models*, New York: Cambridge University Press.
- Greene, W.H. (2003): *Econometric Analysis 5<sup>th</sup> Edition*, New Jersey: Prentice-Hall.
- Kenkel, D.S., Lin, T. F., Sakata, S., and Terza, J.V. (2004): "Econometric Analysis of Prenatal Advice as a Preventive Measure for Fetal Alcohol Syndrome," Working Paper, Center for Health Economic and Policy Studies, Medical University of South Carolina.
- Kenkel, D., Terza, J., "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect," *Journal of Applied Econometrics*, 16, (2001), 165-184.
- Landry, M. (1997). *Overview of Addiction Treatment Effectiveness*. Rockville, MD: Substance Abuse and Mental Health Services Administration.
- McGeary, K.A., and Terza, J.V. (1998): "Flexible Form Nonlinear Regression with Endogenous Switching," Working Paper, Center for Health Economic and Policy Studies, Medical University of South Carolina.
- Shea, D.G., Stuart, B.C., Briesacher, B., and Terza, J.V. (2004): "Prescription Drugs and Medicare: Risk Selection and Moral Hazard," Working Paper, Center for Health Economic and Policy Studies, Medical University of South Carolina.
- Terza, J.V. (2002): "Alcohol Abuse and Employment: A Second Look" *Journal of Applied Econometrics*, 17, 393-404.
- White, H. (1994): *Estimation, Inference and Specification Analysis*, New York: Cambridge University Press.
- Wooldridge, J. (1992): "Some Alternatives to the Box-Cox Regression Model," *International Economic Review*, 33, 935-955.

**Table 1: The DSM-IV Criteria**

**The American Psychiatric Association states that addiction is a maladaptive pattern of substance use, leading to clinically significant impairment or distress, as manifested by three (or more) of the following, occurring at any time in the same 12 month period.**

- 1. Tolerance, as defined by either of the following:**
  - A. A need for markedly increased amounts of the substance to achieve intoxication or desired effect.**
  - B. Markedly diminished effect with continued use of the same amount of the substance**
- 2. Withdrawal, as manifested by either of the following:**
  - A. The characteristic withdrawal syndrome for the substance**
  - B. The same (or a closely related) substance is taken to relieve or avoid withdrawal symptoms**
- 3. The substance is often taken in larger amounts or over a longer period than was intended**
- 4. There is a persistent desire or unsuccessful efforts to cut down or control substance use**
- 5. A great deal of time is spent in activities necessary to obtain the substance (e.g., visiting multiple doctors or driving long distances), use the substance (e.g., chain smoking), or recover from its effects**
- 6. Important social, occupational, or recreational activities are given up or reduced because of substance use**
- 7. The substance use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by the substance (e.g., current cocaine use despite recognition of cocaine-induced depression, or continued drinking despite recognition that an ulcer was made worse by alcohol consumption)**

**The preceding was reprinted from Landry (1997), Exhibit 2.1.**

**Table 2: Variable Definitions**

**Outcome Variable**

**y:** 1 if employed, 0 otherwise

**Policy Variable**

**x<sub>p</sub>:** 1 if substance abuser, 0 otherwise

**Variables Included in x and z**

**FEMALE:** 1 if female, 0 if male

**HLTHCOND:** Count of the number of health conditions that caused problems in the past year

**HHSIZE:** Count variable equal to the number of people in the household

**MARRIED:** 1 if married, 0 otherwise

**BLACK:** 1 if black, 0 otherwise

**ASIAN:** 1 if asian, 0 otherwise

**HISPANIC:** 1 if hispanic, 0 otherwise

**HIGHSCH:** 1 if a high school graduate only, 0 otherwise

**SOMECOLL:** 1 if some post secondary school education, 0 otherwise

**COLLEGE:** 1 if a college graduate or beyond, 0 otherwise

**MIDWEST, SOUTH, WEST:** 1 if resides in that region, 0 otherwise (Northeast excluded)

**URBAN:** 1 if living in an urban setting, 0 otherwise

**QTRINT2, QTRINT3, QTRINT4:** 1 if interview was conducted in that quarter, 0 otherwise  
(first quarter 1 excluded)

**UNEMPL92:** state unemployment rate for 1992

**AGE:** Age in years

**AGESQ:** age squared

**Instrumental Variables (Included in z Only)**

**DADALC:** 1 if biological father was an alcoholic, 0 otherwise

**MOMALC:** 1 if biological mother was an alcoholic, 0 otherwise

**ALCTAX:** State level alcohol tax

**ALCTAXSQ:** Alctax squared

**CIGTAX:** State level cigarette tax

**CIGTAXSQ:** Cigtax squared

**Table 3: Summary Statistics for the Data**

| <b>Variable</b> | <b>Mean</b> | <b>Min</b> | <b>Max</b> |
|-----------------|-------------|------------|------------|
| FEMALE          | .520        | 0          | 1          |
| HLTHCOND        | .442        | 0          | 9          |
| HHSIZE          | 2.819       | 1          | 14         |
| MARRIED         | .596        | 0          | 1          |
| BLACK           | .137        | 0          | 1          |
| ASIAN           | .026        | 0          | 1          |
| HISPANIC        | .065        | 0          | 1          |
| HIGHSCH         | .302        | 0          | 1          |
| SOMECOLL        | .274        | 0          | 1          |
| COLLEGE         | .302        | 0          | 1          |
| MIDWEST         | .250        | 0          | 1          |
| SOUTH           | .333        | 0          | 1          |
| WEST            | .209        | 0          | 1          |
| URBAN           | .740        | 0          | 1          |
| QTRINT2         | .085        | 0          | 1          |
| QTRINT3         | .278        | 0          | 1          |
| QTRINT4         | .360        | 0          | 1          |
| UNEMPL92        | .075        | .032       | .114       |
| AGE             | 38.465      | 24         | 59         |
| AGESQ           | 1567.51     | 576        | 3481       |
| DADALC          | .220        | 0          | 1          |
| MOMALC          | .069        | 0          | 1          |
| ALCTAX          | .226        | .02        | 1.05       |
| ALCTAXSQ        | .086        | .000       | 1.103      |
| CIGTAX          | .278        | .025       | .5         |
| CIGTAXSQ        | .090        | .001       | .25        |

**Table 4: Bivariate Probit Estimates for  
Substance Abuse/ Employment Model**

| Variable          | Substance Abuse            |              | Employment               |              |
|-------------------|----------------------------|--------------|--------------------------|--------------|
|                   | Dependent Variable = $x_p$ |              | Dependent Variable = $y$ |              |
| Variable          | Coefficient                | T-Statistics | Coefficient              | T-Statistics |
| CONSTANT          | -0.526                     | -2.32        | 1.335                    | 4.650        |
| SA ( $x_p$ )      |                            |              | -0.048                   | -0.12        |
| FEMALE            | -0.470                     | -19.72       | -0.018                   | -0.38        |
| HLTHCOND          | 0.009                      | 0.66         | -0.088                   | -6.07        |
| HHSIZE            | -0.057                     | -5.86        | -0.070                   | -6.24        |
| MARRIED           | -0.166                     | -6.06        | 0.326                    | 9.75         |
| BLACK             | -0.136                     | -3.64        | -0.307                   | -7.97        |
| ASIAN             | -0.581                     | -5.92        | 0.101                    | 0.96         |
| HISPANIC          | -0.260                     | -4.99        | -0.050                   | -0.88        |
| HIGHSCH           | -0.104                     | -2.61        | 0.308                    | 7.79         |
| SOMECOLL          | -0.086                     | -2.14        | 0.477                    | 11.4         |
| COLLEGE           | -0.083                     | -2.06        | 0.750                    | 16.44        |
| MIDWEST           | 0.085                      | 2.17         | -0.017                   | -0.38        |
| SOUTH             | 0.059                      | 1.43         | 0.118                    | 2.97         |
| WEST              | 0.104                      | 2.82         | 0.101                    | 2.34         |
| URBAN             | -0.008                     | -0.83        | -0.012                   | -1.09        |
| QTRINT2           | -0.035                     | -0.76        | -0.022                   | -0.43        |
| QTRINT3           | -0.042                     | -1.34        | 0.041                    | 1.12         |
| QTRINT4           | -0.011                     | -0.38        | 0.049                    | 1.42         |
| UNEMPL92          | 6.460                      | 5.51         | -9.111                   | -6.55        |
| AGE               | -0.024                     | -2.25        | 0.021                    | 1.7          |
| AGESQ             | 0.000                      | 1.06         | 0.000                    | -1.19        |
| DADALC            | 0.151                      | 5.47         |                          |              |
| MOMALC            | 0.253                      | 5.94         |                          |              |
| ALCTAX            | -0.318                     | -1.42        |                          |              |
| CIGTAX            | -0.110                     | -0.2         |                          |              |
| ALCTAXSQ          | 0.268                      | 1.2          |                          |              |
| CIGTAXSQ          | 0.821                      | 0.77         |                          |              |
| correlation coeff | -0.058                     | -0.271       |                          |              |

**Table 5: Simple Probit Estimates for Employment Model**

| <b>Variable</b> | <b>Coefficient</b> | <b>T-Statistics</b> |
|-----------------|--------------------|---------------------|
| CONSTANT        | 1.366              | 5.224               |
| SA ( $x_p$ )    | -0.156             | -3.806              |
| FEMALE          | -0.027             | -0.977              |
| HLTHCOND        | -0.088             | -6.057              |
| HHSIZE          | -0.071             | -7.020              |
| MARRIED         | 0.322              | 10.442              |
| BLACK           | -0.310             | -8.599              |
| ASIAN           | 0.091              | 0.926               |
| HISPANIC        | -0.056             | -1.057              |
| HIGHSCH         | 0.306              | 7.868               |
| SOMECOLL        | 0.475              | 11.449              |
| COLLEGE         | 0.748              | 16.481              |
| MIDWEST         | -0.015             | -0.351              |
| SOUTH           | 0.118              | 2.988               |
| WEST            | 0.103              | 2.451               |
| URBAN           | -0.012             | -1.112              |
| QTRINT2         | -0.023             | -0.442              |
| QTRINT3         | 0.040              | 1.103               |
| QTRINT4         | 0.048              | 1.415               |
| UNEMPL92        | -8.976             | -6.896              |
| AGE             | 0.021              | 1.679               |
| AGESQ           | 0.000              | -1.175              |