Volume 2 • Number 6 • 1999
VALUE IN HEALTH

Estimating Endogenous Treatment Effects in Retrospective Data Analysis

Joseph Terza, PhD

Department of Economics, Pennsylvania State University, University Park, PA

ABSTRACT

Treatment effect estimation is one of the mainstays of the field of outcomes research. It is, for example, a key component in analyzing the cost-effectiveness of a proposed qualitative intervention. Some outcomes researchers are hesitant to use retrospective data for treatment effect estimation because of the potential endogeneity of the treatment variable. This is unfortunate, given the abundance and other advantages of retrospective data. Others who have used retrospective data have ignored the endogeneity problem, or have not recognized its potential for causing bias in their estimates. In this paper, an econometric method that is unbiased in the presence of endogeneity and therefore broadens the potential for use of retrospective data in the estimation of treatment effects is proposed. This two-stage method is also designed to accommodate nonlinearity in the relationship between the treatment variable and the outcome. An easy to apply GAUSS™ implementation of the estimator is offered.

Keywords: clinical trials, endogeneity, sample selection, two-stage estimation.

Introduction

Many analyses in the field of outcomes research are motivated by a desire to evaluate the influence of a qualitative variable (referred to as the treatment: d) on another variable (referred to as the outcome: y). The treatment is usually binary and its influence on the outcome is typically referred to as the treatment effect. Examples of applied contexts that require the estimation of treatment effects abound in both the public and private sectors. For instance, a key component of nearly every healthcare reform proposal is the extension of public health care insurance coverage to segments of the population that are not currently covered. Whether or not an individual citizen is granted public health insurance coverage is a qualitative policy variable and measuring its treatment effect on yearly healthcare expenditure levels should be of great interest to public policy makers. Similarly, as an important component of assessing the cost-effectiveness of pharmaceutical products, researchers must be able to accurately measure the effect of a drug on patient outcomes such as subsequent hospitalizations or physician visits.

In the present paper, an econometric method that can be used to estimate treatment effects from retrospective data is proposed. The method is designed to deal with two complications that typically arise in this context. First, because of the nonexperimental nature of the data, estimation is likely to be biased due to the presence of confounders. A confounder is a variable that both exerts influence on the outcome and is correlated with the treatment. In the health insurance example mentioned earlier, age would be a confounder because older individuals tend to spend more on healthcare and are more likely to possess health insurance coverage. Ignoring the effect of age would therefore cause an upward bias in the estimate of the health insurance treatment effect. Age, however, is an observable variable. Therefore, its effects can be accounted for in a regression modeling framework. The effect of age, or any other observable confounder, can be controlled by including it as an additional regressor in the model. The more technically troublesome case is the one involving unobservable confounders. If unobservable confounders are present, the treatment is said to be endogenous. Endogeneity is a particularly difficult problem because, due to the nonobservability of the confounder(s), direct application of the regression approach is not a feasible solution. Indeed it is the lack of econometric methods for dealing with the endogeneity problem that has greatly impeded the use of retrospective data for treatment effect estimation. This is unfortunate because retrospective

Address correspondence to: Joseph Terza, PhD, Department of Economics, Pennsylvania State University, University Park, PA 16802. E-mail: jvt@psu.edu

 1098-3015/99/$14.00/429 429–434

data is typically less costly to obtain than clinical trials data, and in many contexts is the only socially or ethically acceptable data option. Moreover, although experimental or clinical trials data are less likely to involve endogeneity, results obtained therefrom can only be used to analyze the efficacy of the treatment. On the other hand, estimates from retrospective data afford an assessment of the effectiveness of the treatment, which is often the research objective [1]. The econometric method proposed extends the regression approach in a way that yields unbiased estimates in the presence of endogeneity, and therefore broadens the potential for use of retrospective data in the estimation of treatment effects.

The second complicating factor in the estimation of a treatment effect is the possible nonlinearity of the relationship between the treatment and the outcome. For example, the treatment effect of a drug on subsequent doctor visits may differ depending on the frequency with which the individual consults his physician. Such nonlinearity can be accommodated through the appropriate nonlinear regression formulation. Nonlinearities in the modeling of treatment effects may also arise as a means of maintaining logical consistency when the outcome variable is restricted in range. For example, yearly health care expenditure is strictly nonnegative so a logically consistent regression model of the health insurance treatment effect must itself be nonnegative. The exponential functional form is often used for this purpose [2]. Another example is the case in which the outcome variable is itself a binary variable. Here a nonlinear binary response model like logistic regression or probit analysis is apropos. In addition to the elimination of bias due to endogeneity, the proposed treatment effect estimator is designed to accommodate nonlinearity in the relationship between the treatment and the outcome.

This paper details a generic nonlinear regression formulation, including a description of how unobservable confounders are accounted for in modeling the treatment effect. The estimator is developed and GAUSS™ computer software implementing the method is discussed.

Nonlinear Regression with an Endogenous Treatment Effect

The objective is to estimate the effect of a binary qualitative variable ($d = 1$ or 0) on a continuous or qualitative outcome variable (y) using retrospective data. For example, Treglia et al. [3] estimated the relative treatment effect that dothiepin (vs. fluoxetine—both antidepressants) has on healthcare resource utilization following the initiation of antidepressant therapy. The healthcare resources examined in the study included: general hospital accident and emergency room admissions, general hospital nonaccident and emergency room admissions, patients' visits to general practitioners, psychotherapy referrals, other general referrals to healthcare providers, the number of prescriptions for the antidepressant originally assigned, the number of follow-up antidepressant prescriptions, and counts of anxiolytic, hypnotic, and other prescriptions. One of the problems with using retrospective data is bias in the estimation of the treatment effect due to the presence of confounders—i.e., variables that influence the outcome and are correlated with the treatment. In the example, patient behavioral data observed in a retrospective survey are biased for the purpose at hand mainly because the surveyed individuals freely chose (under the advice/prescription of their physicians) to take dothiepin or fluoxetine. It may, for instance, be the case that older individuals are both more likely to visit the physician and more likely to take dothiepin, perhaps due to differences in side effects. A variable such as age is an observable confounder. Observable confounders can be dealt with by including them in a regression framework. It may also be that the relative treatment effect of dothiepin versus fluoxetine may differ with age and other observable confounders, so that the conventional linear regression specification in which the treatment effect is assumed to be constant will not suffice. Instead a nonlinear formulation of the regression is required. Moreover, the dependent variable in this example, i.e., the number of doctor visits, is limited in range and is nonnegative. A nonlinear form like exponential regression is often used to impose such range restrictions. These notions can be formalized, for cases involving observable confounders only, via a nonlinear regression model based on the following conditional mean assumption:

$$E[y \mid x, d] = J(x\beta + d\gamma) \qquad (1)$$

where x is the row vector of observable confounders, β is the conformable column vector of regression coefficients, γ is the regression coefficient of the treatment variable, and $J(\,)$ is a known nonlinear function. For instance, in the example discussed earlier, the dependent variable y is count valued (i.e., $y = 0, 1, 2, \ldots$) so Treglia et al. [3] implement the following version of (1):

$$E[y \mid x, d] = \exp\{x\beta + d\gamma\}$$

which is everywhere nonnegative in accordance with the range of y. Likewise, for binary response models (y = 0 or 1), the following would be used:

$$E[y \mid x, d] = F(x\beta + d\gamma)$$

where F( ) is a probability distribution function. This nonlinear form restricts the regression to the unit interval as would be required of the mean of any binary random variable.

In regression models arising from (1), the treatment effect of d is measured as

$$E_x\Big[E[y \mid x, d = 1] - E[y \mid x, d = 0]\Big] = E_x\Big[J(x\beta + \gamma) - J(x\beta)\Big]$$

where the x subscript denotes that the expectation is taken with respect to the observable confounders. The parameters of (1) can be easily and consistently (in the statistical sense) estimated from retrospective data by applying the nonlinear least squares (NLS) estimation method. The treatment effect would then be consistently estimated as:

$$\frac{\sum_{i=1}^{n}\Big[J(x_i\hat{\beta} + \hat{\gamma}) - J(x_i\hat{\beta})\Big]}{n} \tag{2}$$

where $\hat{\beta}$ and $\hat{\gamma}$ are the NLS estimates of β and γ, i denotes a particular sampled individual (i = 1, . . . , n), and n is the sample size.

If the treatment (d) is endogenous, then in addition to the observable confounders in x, the model involves unobservable confounders which will henceforth be represented by the scalar variable ν. In the example introduced above (in which d denotes whether or not dothiepin versus fluoxetine was prescribed, y denotes the number of doctor visits, and x denotes observable confounders like age) ν might include unobservable confounders like underlying disease severity, patient and/or physician preferences, and patient's prior compliance. Because it is a confounder, ν is correlated with d and exerts an influence on y, i.e., d is endogenous. The endogeneity of d is made explicit through the following extensions of the regression model characterized by (1):

$$d = I(z\alpha + \nu > 0) \tag{3}$$

and

$$E[y \mid w, d, \nu] = J(x\beta + d\gamma + \theta\nu) \tag{4}$$

where z denotes a row vector of observable variables that influence the treatment, α is the conformable column vector of unknown parameters, w is the vector of variables composed of the union of the elements of x and z, and I(C) denotes the indicator function whose value is 1 if condition C holds and 0 otherwise. Equation (3) captures the correlation between ν and d. It takes the form of the typical index/threshold assumption underlying a binary probit or logit model. The index in this case is $z\alpha + \nu$. This component of the model is designed in a way that if the index exceeds an arbitrary threshold then the value of d is observed to be equal to 1. As is conventional in binary response models of this type, the threshold is set equal to 0 for identification purposes. Equation (4) formalizes the influence of ν on y. If ν were instead an observable confounder, we would seek to include it as an additional regressor in the relevant nonlinear conditional mean regression formulation. Equation (4) is designed with this in mind. The version of expression (4) implemented by Treglia et al. [3] is

$$E[y \mid w, d, \nu] = \exp\{x\beta + d\gamma + \theta\nu\}.$$

This imposes the requisite nonnegativity constraint while directly accounting for the influence of the unobservable confounder ν. Note that if ν were observable, the parameters of (4) could be estimated via NLS as described earlier, and the treatment effect would be estimated as in (2). Unfortunately this is not the case, so an alternative approach must be found.

Given assumptions (3) and (4) it can be shown that:

$$E[y \mid w, d] = d\frac{\int_{-z\alpha}^{\infty} E[y \mid w, d = 1, \nu]h(\nu)d\nu}{1 - H(-z\alpha)} + (1 - d)\frac{\int_{-\infty}^{-z\alpha} E[y \mid w, d = 0, \nu]h(\nu)d\nu}{H(-z\alpha)}$$

$$= d\frac{\int_{-z\alpha}^{\infty} J(x\beta + d\gamma + \theta\upsilon)h(\nu)d\nu}{1 - H(-z\alpha)} + (1 - d)\frac{\int_{-\infty}^{-z\alpha} J(x\beta + d\gamma + \theta\upsilon)h(\nu)d\nu}{H(-z\alpha)} \tag{5}$$

where h( ) and H( ) respectively denote the pdf and cdf of ν conditional on w [henceforth denoted $(\nu \mid w)$]. The interpretation of equation (5) is intuitive. The first component of the expression to the right of the first equals sign is the conditional mean of y given the exogenous variables in w and d = 1. At the heart of this conditional mean is the expectation of y given w, d = 1, and ν (i.e., the term $E[y \mid w, d = 1, \nu]$). Using assumption (4), we replace $E[y \mid w, d = 1, \nu]$ with $J(x\beta + d\gamma + \theta\nu)$ which is the nonlinear regression specification that would be directly estimated via NLS if ν were observed. Unfortunately ν is not observed so the analysis must be conditioned on the limited, but useful, information we have regarding ν, mainly that if d = 1 then $\nu > -z\alpha$. The probability of this conditioning event, $1 - H(-z\alpha)$, is seen in the denominator, and because ν is unobserved, $J(x\beta + d\gamma + \theta\nu)$ must be integrated over the relevant range of ν to complete the desired conditional expectation. The formulation of the second component on the right-hand side of (5) can be similarly interpreted.

If an acceptable candidate for the distribution of $(\nu \mid w)$, can be found [i.e., if H( ), and therefore h( ) are specified], then (5) can be used as the basis for consistent NLS regression estimation of the parameters of the model, and the treatment effect can be estimated as in (2). The nonlinear regression specification that corresponds to (5) is

$$y = J^*(w, d, \beta, \gamma, \theta, \alpha) + e \tag{6}$$

where $J^*(w, d, \beta, \gamma, \theta, \alpha) = E[y \mid w, d]$ as expressed in (5), and $e \equiv y - J^*(w, d, \beta, \gamma, \theta, \alpha)$. In general $J^*(w, d, \beta, \gamma, \theta, \alpha)$ cannot be expressed in closed form. It can nevertheless be accurately and efficiently evaluated using numerical approximation. Appropriate quadrature methods are available for this purpose [4]. Alternatively, simulation methods can be used [5]. The GAUSS™ program described later in the paper uses quadrature to approximate the requisite integrals.

When y is a binary qualitative variable (y = 0 or 1):

$$J^*(w, d, \beta, \gamma, \theta, \alpha) = d\frac{\int_{-z\alpha}^{\infty} F(x\beta + d\gamma + \theta\nu)h(\nu)d\nu}{1 - H(-z\alpha)} + (1-d)\frac{\int_{-\infty}^{-z\alpha} F(x\beta + d\gamma + \theta\nu)h(\nu)d\nu}{H(-z\alpha)}$$

In [6], it is shown that for nonnegative regressions (e.g., count data models; y = 0, 1, 2, . . .), under the assumptions that J(q) = exp{q} and $(\nu \mid w)$ is standard normally distributed, the following closed-form expression for $J^*(w, d, \beta, \gamma, \theta, \alpha)$ obtains:

$$J^*(w, d, \beta, \gamma, \theta, \alpha) = \exp\{x\beta + d\gamma\} \times \left(d\frac{\Phi(\theta + z\alpha)}{\Phi(z\alpha)} + (1-d)\frac{1 - \Phi(\theta + z\alpha)}{1 - \Phi(z\alpha)} \right)$$

where $\Phi( )$ denotes the standard normal cumulative distribution function (cdf). It is interesting to note that for the linear case in which J(q) = q:

$$J^*(w, d, \beta, \gamma, \theta, \alpha) = x\beta + d\gamma + \theta\left(d\frac{\int_{-z\alpha}^{\infty} \nu\, h(\nu)\, d\nu}{1 - H(-z\alpha)} + (1-d)\frac{\int_{\infty}^{-z\alpha} \nu\, h(\nu)\, d\nu}{H(-z\alpha)} \right) \tag{7}$$

and if $(\nu \mid w)$ is assumed to be standard normally distributed, with $\phi( )$ denoting the standard normal probability density function (pdf), (7) becomes:

$$J^*(w, d, \beta, \gamma, \theta, \alpha) = x\beta + d\gamma + \theta\left[d\frac{\phi(z\alpha)}{\Phi(z\alpha)} - (1-d)\frac{\phi(z\alpha)}{1 - \Phi(z\alpha)} \right]$$

which is the regression specification suggested by Heckman [7]. If $(\nu \mid w)$ is assumed to be logisitically distributed (7) becomes:

$$J^*(w, d, \beta, \gamma, \theta, \alpha) = x\beta + d\gamma + \theta\left[d\frac{r(z\alpha)}{\Lambda(z\alpha)} - (1-d)\frac{r(z\alpha)}{1 - \Lambda(z\alpha)} \right]$$

where $\Lambda( )$ denotes the logistic cdf and

$$r(s) = \left[1n(1 + \exp\{s\})(1 + \exp\{s\}) - s\exp\{s\} \right][1 - \Lambda(s)]$$

This specification was first suggested by Hay [8].

The Two-Stage Estimator

In most applications, NLS estimation of (6) can be made easier by applying the following two-stage analog to the method suggested by Heckman [7]. In the first-stage, the appropriate binary response estimator [corresponding to the chosen specification of H( ), the cdf of $(\nu \mid w)$] is used to obtain $\hat{\alpha}$, a consistent estimate of α. In the second stage, es-

timates of β and θ can be obtained by applying the NLS method to:

$$y = J^*(w, d, \beta, \gamma, \theta, \hat{\alpha}) + e^* \qquad (8)$$

where

$$e^* \equiv J^*(w, d, \beta, \gamma, \theta, \alpha) - J^*(w, d, \beta, \gamma, \theta, \hat{\alpha}).$$

Using standard asymptotic results, it can be shown that this two-stage estimator is asymptotically normal, with asymptotic covariance matrix:

$$V = E[\, g_1' g_1 \,]^{-1} \Big\lfloor E[e^2 g_1' g_1 \,] + E[\, g_2' g_1 \,]' \mathrm{VAR}(\hat{\alpha}) E[\, g_2' g_1 \,] \Big\rfloor E[\, g_1' g_1 \,]^{-1}$$

where

$$b = \begin{bmatrix} \beta \\ \gamma \\ \theta \end{bmatrix}$$

$$g_1 = \frac{\partial J^*}{\partial b}$$

$$g_2 = \frac{\partial J^*}{\partial \alpha}$$

and $\mathrm{VAR}(\hat{\alpha})$ denotes the asymptotic covariance matrix of the first-stage estimator of α. In practice the following heteroskedasticity-consistent estimator of V can be used:

$$\hat{V} = (G_1' G_1)^{-1} \Big\lfloor G_1' \Psi^* G_1 + G_1' G_2 \hat{\mathrm{VAR}}\, G_2' G_1 \Big\rfloor (G_1' G_1)^{-1}$$

where G_1 and G_2 are, respectively, the $n \times (K_1 + 2)$ and $n \times K_2$ matrices whose typical rows are:

$$g_{1i} = \frac{\partial J_i^*}{\partial b} \text{ and } g_{2i} = \frac{\partial J_i^*}{\partial \alpha}$$

K_1 and K_2 are respectively the dimensions of x and z, Ψ^* is the $n \times n$ diagonal matrix whose i^{th} diagonal element is $\hat{e}_i^{*2}$, the squared residual (for the i^{th} sample member) from NLS estimation of (8), and $\hat{\mathrm{VAR}}$ denotes the estimated asymptotic covariance matrix of $\hat{\alpha}$.

The main competitor for this two-stage estimator would be a version of the generalized method of moments (GMM) estimator of Hansen [9]. The simplest and most familiar version of the GMM estimator is the instrumental variables (IV) method otherwise known as two-stage least squares (see Greene [10] for a discussion of the IV method). The problem with using IV estimation in this context is that it fails to take account of nonlinearity of the treatment effect. The only truly nonlinear version of the GMM approach that has been suggested in the present context is that of Mullahy [11]. There are, however, two important shortcomings of Mullahy's method vis-a-vis the method reported here. First, Mullahy's GMM estimator is designed specifically for the case in which the conditional mean regression model is exponential—i.e., it is useful for nonnegative models only. The formulation suggested here is generic in the sense that it applies to any nonlinear or limited dependent variable model that conforms to assumption (4). Secondly, Mullahy's model ignores important structural information and therefore is less efficient in the statistical sense than the method reported in this paper. Specifically, his GMM approach ignores the structural information supplied by assumption (3) used here. But (3) is a commonly accepted, and reasonable formulation for binary response models, e.g., conventional probit and logit models. The two-stage method reported here was compared in a more general setting to that of Mullahy and was found to be substantially more efficient even for large sample sizes [12]. For example, with 1000 simulated samples of size 20,000, the mean-square-error ratio (GMM/TSM) was 2.31 for cases in which endogeneity was relatively mild (the correlation between ν and y was low), and as large as 10.15 when endogeneity was more severe.

Treglia et al. [3] applied the two-stage method reported here to the estimation of the relative treatment effect of dothiepin versus fluoxetine on ten different resource utilization measures: accident and emergency (A&E) hospital visits; non-A&E hospital visits; doctor visits, general referrals; psychotherapy referrals; follow-up antidepressants; anxiolytic prescriptions; hypnotic prescriptions; other prescriptions; and the antidepressants considered in the study, dothiepin and fluoxetine. The specific versions of $J(\,)$ and $H(\,)$ that they used were $\exp\{\,\}$ and $\Phi(\,)$, respectively. For five of ten regressions the estimates of θ were significant at the 1% level. In these cases failing to control for the endogeneity of the treatment would have resulted in estimation bias. After correcting for endogeneity, the average individual who used dothiepin made 5.446 fewer doctor visits per year.

A GAUSS™ implementation of the two-stage estimator was developed which allows the user to choose from among the following four specifications for $J(\,)$: $J(q) = q$, the linear specification for conventional regression models; $J(q) = \exp(q)$, the

exponential specification for nonnegative models; $J(q) = \Phi(q)$, the prohibit, or $J(q) = \Lambda(q)$, the logit, for binary response models.

Possible choices for the distribution of $(v \mid w)$ are:

$$H(q) = \begin{matrix} \Phi(q) \\ \Lambda(q) \end{matrix} .$$

These selections are easily made through initial settings at the top of the program along with other user-supplied options. Application of the software requires minimal GAUSS™ programming experience.

Conclusion

The specter of endogeneity bias has caused outcomes researchers to avoid the use of retrospective data for the estimation of treatment effects. In this paper, an econometric method is proposed that, in addition to being unbiased in the presence of endogeneity, is designed to accommodate nonlinearity in the relationship between the treatment variable and the outcome. The two-stage estimator suggested here is analogous to the sample selection estimator proposed by Heckman [7] for the classical linear regression model. The estimator has been fully implemented in the GAUSS™ programming language for many of the most popular nonlinear regression specifications, and application of the software requires minimal GAUSS™ programming experience. Versions of the method that allow more flexibility in the functional form of the regression specification are currently under development.

References

1 Heckman JJ, Smith J. Assessing the case for social experiments. J Economic Perspectives 1995;9:85–100.
2 Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. J Health Econ 1998;17:247–95.
3 Treglia M, Neslusan CA, Dunn RL. Fluoxetine and dothiepin therapy in primary care and health resource utilization: evidence from the United Kingdom. Int J Psychiatry Clin Pract 1999;3:23–30.
4 Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes. Cambridge: Cambridge University Press, 1992.
5 Gourieroux C, Monfort A. Simulation-Based Econometric Methods. Oxford: Oxford University Press, 1996.
6 Terza JV. Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. J Econometrics 1998; 84:129–54.
7 Heckman JJ. Dummy endogenous variables in a simultaneous equation system. Econometrica 1978; 46:931–59.
8 Hay JW. Occupational choice and occupational earnings: selectivity bias in a simultaneous logit-OLS model. Rockville, MD: National Technical Information Service, 1980.
9 Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica 1982; 50:1029–54.
10 Greene WH. Econometric Analysis (3rd ed.). New York: Prentice Hall, 1997.
11 Mullahy J. Instrumental variable estimation of count data models: applications to models of smoking behavior. Rev Econ Stat 1997;79:586–93.
12 Neslusan CA, Terza JV. Exponential regression with endogenous polychotomous treatment effects: GMM vs. two-stage estimation [working paper]. University Park, PA: Pennsylvania State University Economics Department, 1997.