Estimating Average Treatment Effect by Nonlinear Endogenous Switching Regression With an Application in Botswana Fertility

Myoung-Jin Keay
South Dakota State University

This study explores the Average Treatment Effects (ATE) estimator proposed by Terza (2009)'s Nonlinear Full Endogenous Treatment (NFES) model, where count dependent and binary treatment variables are present. Asymptotic distribution of ATE estimators based on NFES model is provided to show that nonlinear estimators have additional terms in asymptotic variance of which magnitudes depend on population coefficient. Due to their presence, the asymptotic variance of nonlinear estimators can be either larger or smaller than the linear counterparts depending on the values of coefficients. It turns out that the nonlinear ATE estimators are more efficient than linear estimators when the ATE conditional on covariates has small variance. An application to Botswana fertility is given.

Keywords: endogenous switching, endogenous treatment, average treatment effect, count data, fertility

INTRODUCTION

In order to estimate the treatment effects of binary variable on count dependent variable, Terza (1998, 2008, 2009) proposed nonlinear models that take into account the nonlinear nature of dependent variable. As alternatives to those fully nonlinear models, a traditional linear regression model with probit treatment equation can also be used. Although it seems to be more sensible to apply the nonlinear models given count outcome variable, the previous literature have not clearly stated the advantages as well as disadvantages of using nonlinear outcome models rather than simply applying linear methods to estimate the treatment effects. While the linear models implemented by Heckman (1978)'s method is already well understood, large part of the statistical properties of Terza's nonlinear approaches are still unknown. The goal of this study is to explore the properties of nonlinear approaches to estimating the treatment effects and to give a guidance that might be useful to the empirical analyses.

Terza (1998) considers a model where the binary treatment variable shifts the intercept inside the exponential conditional mean function and provides estimating equations that can be implemented by using the observable variables. Also in later works, Terza (2008, 2009) extends the earlier model by incorporating the counterfactual framework where the treatment status puts the individual in a different regime. Following the terminology used in Terza (2009), the former model will be called throughout this paper "Nonlinear Endogenous Treatment Model" (NET), and the latter "Nonlinear Full Endogenous Switching Model" (NFES). As it will be shown in subsequent sections, NFES model is an extended version of NET in the sense that an appropriate restriction on coefficients along with a fairly weak assumption readily makes NFES and NET equivalent. While NFES is relatively new, NET has acquired wide popularity among empirical economists. For the last decade it has been applied to see the effect of founder CEO as incumbent on the active acquisition activity (Fahlenbrach, 2009), the effect of credit constraint on floating net aquaculture adoption in Indonesia (Miyata and Sawada, 2007), the effect of firm's voluntary pollution reduction program on pollution (Innes and Sam, 2008; Sam, 2010), the effect of duplicate coverage on the demand for health care in Germany (Vargas and Elhewaihi, 2007), the effect of illicit drug use on emergency room utilization (McGeary and French, 2000), the effect of physician advice on alcohol consumption (Kenkel and Terza, 2001), the effect of insurance on demand for health care (Koç, 2005), the effect of higher education on smocking (Miranda and Bratti, 2006), the effect of socio-economic factors on completed fertility (Miranda, 2003), the effect of Mexican families' migration in US on woman's domestic power (Parrado, Flippen and McQuiston, 2005; Parrado and Flippen, 2005), the effect of health maintenance organization plans on the health care expenditure in private sector (Shin and Moon, 2007) and the fertility differences between married and cohabiting couples (Zhang and Song, 2007) to name a few.

Since most studies enumerated above use the NET model to measure the effect of binary variables, the validity of their conclusions may be put into question unless the single regime restrictions are correct. One important exception is Koç (2005) where he estimates two different structural equations for each value of treatment variable. However, he mainly focuses on the equation in each regime and not paying full attention to comparing the values of dependent variables that might lead to ATE analysis. Although the first papers proposing the ATE estimator based on the NFES model is Terza (2008, 2009), it only proposes the possibility of such methodology in unifying framework with other nonlinear models without fully discussing nice properties of ATE estimator compared to traditional approaches. This study will show that the ATE estimators based on NFES model can have higher efficiency and smaller finite sample biases only under certain circumstances.

The rest of the paper is organized as follows. Section 2 introduces various switching regression models such as NFES, NET, LFES and LET and discuss how the ATE can be identified for each model. Section 3 characterizes the asymptotic biases when the methods being used does not reflect the true population. Section 4 describes the various estimation methods for NFES model. In Section 5, the proposed approach is applied to a real data set to estimate ATE and Section 6 presents the concluding remarks.

MODEL

In what follows the term nonlinear is exclusively reserved to describe the nature of dependent variable of structural equation. In this count dependent variable setting, nonlinear models will use the linear index transformed by exponential function as their conditional expectation function. On the other hand the linear models will be constructed as if the dependent variable were continuous.

Nonlinear Models

The "Nonlinear Endogenous Treatment Model" (NET) first proposed by Terza (1998) is as follow.

$$E[y|x,w,\epsilon] = \exp(\alpha + x\beta + \gamma w + \epsilon)$$
$$w = 1[z\delta + v > 0],$$

where x are covariates, w is binary treatment variable and ϵ is unobserved heterogeneity. The vector of covariates x and the vector of exogenous variables z are all assumed to be independent with the structural and selection errors. Usually x is the subset of z. The value of treatment variable, i.e. either one or zero, is determined by a binary choice model such as probit. The treatment equation tells that the value of w is determined by the exogenous variables z and the selection error v. When their sum is greater than zero, w is equal to one, and zero otherwise. If w is determined purely randomly as in randomized experiment, then it will be independent with the unobserved heterogeneity ϵ and the regression will become very simple and straightforward. However, when w is correlated with the unobserved heterogeneity, then a usual estimation that does not control for the correlated error might suffer from an endogeneity problem for the estimation of γ. For example, when the number of children a woman has at the time of observation is set as a dependent variable y, it will be determined by her age and marriage status and so on that constitute the covariates x. The dependent variable will also be affected by the education status w that is either zero or one depending on whether she has education at all. Since the education status is determined by an individual's utility maximization, the factor that affects w might also affect y creating an endogeneity. Terza (1998) suggests an estimating equation in the form of conditional mean function with a correction term that is conditioned only on the observables.

The above model, however, is restrictive in that it supposes a constant semi-elasticity of dependent variable with respect to the treatment across all the individuals in population. This is related to the fact that the coefficient on covariates and the unobserved heterogeneity are invariant under different treatment status. The model that extends the above one is proposed by Koç (2005) and Zhang and Song (2007) as below.

$$E[y_g|x, w, \epsilon_1, \epsilon_0] = E[y_g|x, \epsilon_g] = \exp(\alpha_g + x\beta_g + \epsilon_g), \qquad g = 0,1 \tag{1}$$
$$w = 1[z\delta + v > 0],$$

where different coefficients on covariates and unobserved heterogeneity depending on the treatment status are allowed for. In other words, the treatment status puts an individual in a different regime; if $w = 1$, then she is in regime 1 with the outcome y_1 and similarly for the other regime. Presumably each individual has her y_0 and y_1 for each treatment status but one of them is not observed. The way to recover the unobserved counterfactual will be discussed later on for estimation, but for the time being let's focus on the population model itself. If those two outcome variables are known, then $y_{i1} - y_{i0}$ would be an individual treatment effect. Since it might be different from person to person, we might want to know the averaged individual treatment effect $E(y_{i1} - y_{i0})$ that is the so-called Average Treatment Effect (ATE). Incidentally the individual semi-elasticity can be computed by $(y_{i1} - y_{i0})/y_{i0}$ that might not be constant across individuals either. This is the extended Terza model that will be called throughout this paper "Nonlinear Full Endogenous Switching Model" (NFES). The quantity of interest will then be the ATE that captures the causal effect of treatment.

Returning to (1), the first equality in the upper equation tells that the conditional expectations of dependent variables for each regime depend neither on switching variable w nor on unobservables for other regime. The exclusion of w is particularly important; once the covariates and the unobservables ϵ_g are controlled for, the knowledge about realized regime does not provide any additional information on the conditional expectation of dependent variables. In other words, the equality assumes the ignorability (Rubin, 1978) or unconfoundedness (Imbens, 2005) of w conditional on covariates and unobservables.

Although the treatment equation in (1) is expressed by a binary choice model, it is also possible to use the linear probability model that is essentially a linear projection of w on z. However, in the present model, the fact that the endogenous variable is binary is not neglected so that an appropriate binary choice model is used. The treatment equation that describes the regime switching mechanism can be modeled by any binary choice model, but here let us assume that it is governed by probit model for the sake of simplicity. The robustness of this assumption will also be discussed later. Now let the errors in outcome and treatment equation be denoted by ϵ and v and follow trivariate normal distribution as below.

$$\begin{bmatrix}\epsilon_0\\ \epsilon_1\\ v\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\sigma_0^2 & & \rho_0\sigma_0\\ & \sigma_1^2 & \rho_1\sigma_1\\ & & 1\end{bmatrix}\right)$$

This assumption becomes sufficient condition for each error to follow normal distribution. If there is no correlation between ϵ and v, then the regime switching becomes entirely random. Unless the covariances are equal to zero, the regime choice will be determined by each individual's own idiosyncrasies that create correlation between w and ϵ. Heckman correction can be used to solve this endogeneity problem in linear model where the dependent variable is continuous; the difference between Heckman corrected linear model and current one is that the latter allows for noncontinuous outcome distribution with exponential CEF while

Heckit presupposes a continuous structural error of which the conditional expectation is expressed as a linear function of v. Nevertheless the basic situation is more or less the same.

Under the above assumption the ATE can be identified as below (Terza, 2009).

$$\begin{aligned} ATE &= E[y_1 - y_0] = E(E[y_1|x] - E[y_0|x]) \\ &= E[\exp(\alpha_1 + \sigma_1^2/2 + x\beta_1) - \exp(\alpha_0 + \sigma_0^2/2 + x\beta_0)] \end{aligned} \tag{2}$$

Thus an estimate can be computed by using the sample analogue method.

The NFES model discussed so far nests NET model shown in the very beginning of this section. By putting restrictions $\beta_0 = \beta_1$ and $\epsilon_0 = \epsilon_1$ the two outcome equations in NFES can be combined to be written as

$$E[y|x, w, \epsilon] = \exp(\alpha_0 + (\alpha_1 - \alpha_0)w + x\beta + \epsilon),$$

where $y = y_0 + w(y_1 - y_0)$. The NET model, although having been claimed as a switching regression in Terza (1998), does not clearly incorporate the two distinct regimes; the regime changes according to the value of the binary variable, but switching is expressed only by shifting the intercept term inside the exponential function. In linear model, it is similar to the case where the coefficients of covariates for two regimes are identical except for the intercept. Thus it is recommended to run the NFES model first; it is preferable unless test rejects the hypothesis of $\beta_1 = \beta_0$. In ET model, the parameter of interest is usually the coefficient on w, i.e. α_1 of which interpretation is the semi-elasticity of y with respect to the treatment variable. This is distinct from ATE that we are in many cases interested; ATE must be computed as in equation (2).

Linear Models

Angrist(2001, 2010) and Angrist and Pischke(2009) have pointed out that the linear model is sufficiently good for estimating the marginal effect of a model with binary dependent variable. Angrist and Pischke (2009) also maintain the validity of such approach even for the general limited dependent variable models on the grounds that the linear coefficient can provide the linear projection coefficients that might be very close to the actual causal effect. In line with that approach, the above endogenous switching model can be expressed in linear form as below despite the nonlinear nature of count dependent variables.

$$\begin{aligned} &y_g = \mu_g + x\beta_g + u_g, \qquad g = 0,1 \\ &w = 1[z\delta + v > 0] \end{aligned} \tag{3}$$

Let the explanatory variables be demeaned, then the ATE is $E[y_1 - y_0] = \mu_1 - \mu_0$. We call this model “Linear Full Endogenous Switching Model” (LFES) as a linear counterpart of NFES. As NFES model nests the NET, LFES does it for “Linear Endogenous Treatment Model” (LET) under the restriction that $\beta_1 = \beta_0$ and $u_1 = u_0$, whereby the coefficient on w becomes the ATE that is constant across all individuals. The treatment equation is modeled as probit as usual.

When the true model is such that the outcome variable is nonnegative, the structural equations in LFES model cannot be viewed as the error form of conditional expectation. Rather it is the linear projection of y on covariates and therefore $E(y_g) = \mu_g$ since all the covariates are already demeaned. The ATE is the difference between the two intercepts for each regime. One way to identify these intercepts and ATE is by using the Heckman correction method (Heckman, 1978) with one additional assumption. In order to be able to write the correction term as inverse Mill’s ratio, the minimal assumption required is that $E(u|v) = \rho v$ (Olsen, 1980). Under this assumption along with the probit treatment equation, the intercept μ_g for each regime can be identified and so is the ATE. The difference between the nonlinear and linear approach to estimating ATE is that the former finds $E[y_g|x]$ and then take their average for whole population to get the ATE, whereas the latter directly finds $E[y_g]$ without bothering to model the conditional mean on covariates.

ESTIMATION

Various estimation methods for NFES models are presented in this section. Based on the estimating equations in Terza(1998), the estimation methods for NFES are discussed below.

Quasi-Maximum Likelihood Estimator

Unless the distributional assumption used in FIML are correct, the FIML estimator might not be consistent; this is a cost of FIML in exchange for efficiency. By the way there is another method called Quasi-Maximum Likelihood Estimator(QMLE) that trades the efficiency with robustness by using weaker condition that only the conditional expectation function (CEF) is correctly specified. As long as the used likelihood is in the class of linear exponential family, and the CEF is correctly specified, the estimator is consistent even if the whole likelihood function is not correctly specified (Gourieroux, Monfort and Trognon, 1984). Given the model in equation (1), a natural way to estimate might be running QMLE or Nonlinear Least Squares (NLS) by using the $E(y_g|x,\epsilon_g)$. However, it does not give an estimable equation due to the ignorance of ϵ_g; the unobserved variable needs to be removed by integrating out from the conditioning set of that CEF. By using the fact that ϵ and v are correlated, one can construct $E(y|z,v)$.

$$E(y_g|z,v) = \exp(\alpha_g + \frac{1}{2}\sigma_g^2(1-\rho_g^2) + x\beta_g + \rho_g\sigma_g v)$$

Conditional on z, v determines the value of w. Since z,w makes a sparser σ-field than z,v does, by law of iterated expectation,

$$E(y_g|z,w) = \exp(\alpha_g + \frac{1}{2}\sigma_g^2(1-\rho_g^2) + x\beta_g)E[\exp(\rho_g\sigma_g v)|z,w]$$

Thus $E(y|z,w)$ can be expressed by using only the observable variables z,w. Then the estimating equation is obtained as

$$E(y|z,w) = w\cdot[\exp(\alpha_1 + \tfrac{\sigma_1^2}{2} + x\beta_1)\tfrac{\Phi(z\delta+\rho_1\sigma_1)}{\Phi(z\delta)}] + (1-w)\cdot[\exp(\alpha_0 + \tfrac{\sigma_0^2}{2} + x\beta_0)\tfrac{\Phi(-(z\delta+\rho_0\sigma_0))}{\Phi(-z\delta)}]. \quad (4)$$

The detailed derivation of the above estimating equation can be found in Appendix B. One can run a QML estimation using the above CEF. A distributional assumption on y is needed as in FIML; the difference is that FIML models y_g to follow certain distribution with $E(y_g|z,\epsilon_g)$ as CEF, whereas QMLE does it with $E(y|z,w)$. The integration does not appear in Poisson likelihood based on $E(y|z,w)$ because the unobservable was already got rid of and the correction term does that role instead. Both FIML and QMLE relies on correctly specified conditional mean for consistent estimation of parameters. However, the conditional mean in QMLE, i.e., $E(y|z,w)$, is expressed by all observable variables that makes the QMLE likelihood simpler than FIML. On can run a QMLE by using a conditional distribution with the mean $E(y|z,w)$. Specifically two step method can be employed where the first stage probit estimates are substituted in the correction terms. It does not, however, have to be carried out sequentially by two steps; they can be estimated by a single step procedure where all the necessary parameters for ATE are separately identified. Keay (2010) and Hellström and Nordström (2008) have shown that the single step ML method for estimating ATE in linear endogenous switching model is relatively less efficient in finite sample; it will be examined in the sequel whether that is still the case in this nonlinear model with count dependent variable.

Nonlinear Least Squares Estimator

The above QML method is run by using a likelihood in linear exponential family based on the condition that the conditional mean function is correctly specified. By the way given the correctly specified conditional mean function it is also possible to use Nonlinear Least Squares (NLS) method. This NLS can also be viewed as a method of moment estimator. Let's write the equation in additive form with the CEF.

$$y = E[y|z,w] + e,$$

where by definition $E(e|z,w) = 0$. NLS seeks an estimate that minimizes $E[y - E(y|z,w)]^2$. The NLS is consistent because the estimate is such that it satisfies $\sum\ (dE(y|z,w)/d\theta \times e) = 0$, that can be viewed as a sample analogue of $E[dE(y|z,w)/d\theta \times e] = 0$. Since the conditional mean contains the correction terms, it should be estimated through the first stage probit.

We have seen above that the NLS gives consistent estimator based on the GMM argument by using the law of iterated expectation. However, by applying the optimal GMM concept, one can find even more efficient GMM estimator. This can be done by dividing the instrument by conditional variance, i.e., the estimator using $E(dE(y|z,w)/d\theta \cdot var[y|z,w]^{-1} \times e) = 0$ moment condition is more efficient than the above one. By the way, this is equivalent to the NLS applied on the error form equation of which both sides were divided by the square root of conditional variance. Therefore the optimal GMM is equivalent to Weighted Nonlinear Least Squares (WNLS) estimator. The estimation will be carried out by three steps. The correction term is estimated in the first step and the structural parameters are estimated in the second step from which the conditional variance is estimated. The last third step again estimates the structural parameters by using the conditional variance estimated in the earlier step. Terza (1998) has proposed two approaches to estimating the conditional variance. Among those, the regression based method is computationally easier and will be used here. Terza (1998) shows that

$$\text{var}[y|z,w] = w\delta_1(\delta_1(\exp(\sigma_1^2)L_{1,2} - L_1^2) + L_1) + (1-w)\delta_0(\delta_0(\exp(\sigma_0^2)L_{0,2} - L_0^2) + L_0), \tag{5}$$

where $\delta_g = \exp(\alpha_g + \sigma_g^2/2 + x\beta_g)$, $L_{1,2} = \Phi(z\delta + 2\rho_1\sigma_1)/\Phi(z\delta)$, $L_1 = \Phi(z\delta + \rho_1\sigma_1)/\Phi(z\delta)$, $L_{0,2} = \Phi(-z\delta - 2\rho_0\sigma_0)/\Phi(-z\delta)$, and $L_0 = \Phi(-z\delta - \rho_0\sigma_0)/\Phi(-z\delta)$. Regression based method estimates the σ_g^2 that will be used to compute the conditional variance for WNLS.

ASYMPTOTIC DISTRIBUTION

The asymptotic distribution of FIML estimator is straightforward. Given the likelihood function in proposition 1, the score and hessian will be constructed as usual. If the multivariate normal assumption is correct and so is the likelihood function, then the asymptotic variance will be simplified. The disadvantage of FIML is that the parameters are not consistent any more when the likelihood function is misspecified.

Now consider the WNLS estimator. The objective function is $(y - E[y|z,w])^2/2 \cdot var[y|z,w]$, where $E[y|z,w]$ and $var[y|z,w]$ are from equation (4) and (5). Ignoring the first stage error, the asymptotic distribution can be written under the condition $var(y|z,w) = v(z,w,\gamma)$ as (See Wooldridge, 1997)

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\longrightarrow} \text{N}(0, \quad [E(h(z,w,y,\theta))]^{-1}),$$

where

$$E[h(z,w,y,\theta)] = E[\frac{\nabla_\theta m(z,w,\theta)\nabla_\theta m(z,w,\theta)\prime}{v(z,w,\gamma)}], \tag{6}$$

and $m(z,w,\theta) = E[y|z,w]$.

Now consider the asymptotic distribution of PQMLE. The likelihood function is constructed using the Poisson distribution with the conditional mean in equation (4). Then the asymptotic distribution is

$$\sqrt{n}(\hat{\theta}-\theta_0)\underset{d}{\longrightarrow}\mathrm{N}(0,\quad E[h(y|z,w,\theta_0)]^{-1}E[s(y|z,w,\theta_0)s(y|z,w,\theta_0)']E[h(y|z,w,\theta_0)]^{-1}),$$

where

$$E[s(y|z,w,\theta_0)s(y|z,w,\theta_0)'] = E[\frac{\nabla_\theta m(z,w,\theta)(y_i-m(z,w,\theta))}{\mathrm{qvar}(y_i)}\cdot\frac{(y_i-m(z,w,\theta))\nabla_\theta m(z,w,\theta)'}{\mathrm{qvar}(y_i)}]$$
$$E[h(y|z,w,\theta_0)] = -E[\tfrac{\nabla_\theta m(z,w,\theta)\nabla_\theta m(z,w,\theta)'}{\mathrm{qvar}(y_i)}]$$

The denominator qvar is the variance implied by the used distribution function in QML. For WNLS the denominator of the expected Hessian was the conditional variance of y, whereas qvar, that of expected Hessian and score for PQML, is the variance implied from the distribution used for quasi-likelihood, i.e. the conditional mean for Poisson QMLE. The asymptotic variance of PQMLE can be simplified under the condition

$$Var[y|z,w]=\sigma^2\cdot\mathrm{qvar}, \tag{7}$$

This condition says that the true conditional variance is proportional to the variance implied in the quasi-likelihood. Generalized Conditional Information Matrix Equality (GCIME) holds under this condition that gives

$$\sqrt{n}(\hat{\theta}-\theta_0)\underset{d}{\longrightarrow}\mathrm{N}(0,\quad -\sigma^2[E(h(y|z,w,\theta))]^{-1}).$$

By plugging (9) in (8), it is obvious that the two asymptotic distribution for WNLS and PQML are equivalent. Having said that, without the condition (9), PQML might be less efficient than the WNLS. Of course this conclusion is true in as much as the first stage estimation error is ignored.

Now consider our model with the estimating equation as in (7). Although the dependent variable y_g conditional on x and ϵ_g follows the Poisson distribution with the mean $E(y_g|x,\epsilon_g)=\exp(\alpha_g+x\beta_g+\epsilon_g)$, it does not necessarily mean that y_g conditional on x and w follows Poisson distribution with the mean $E(y_g|z,w)=\exp(\alpha_g+\sigma_g^2/2+x\beta_g)\Phi(f(z\delta))/\Phi(z\delta)$. To see this point, mean and variance conditional on z,w are

$$E[y|z,w]=w\delta_1L_1+(1-w)\delta_0L_0$$
$$\mathrm{var}[y|z,w]=w\delta_1(\delta_1(\exp(\sigma_1^2)L_{1,2}-L_1^2)+L_1)+(1-w)\delta_0(\delta_0(\exp(\sigma_0^2)L_{0,2}-L_0^2)+L_0)$$

It is obvious that they are neither same nor proportional by a constant. Therefore the condition for GCIME is not satisfied and the PQML is asymptotically less efficient than the WNLS. Nevertheless, it should also be noted that the first stage estimation error is ignored in the asymptotic distribution above, that the small sample behavior can be different. The asymptotic distribution of the estimators for structural parameters that accounts for first stage error is straightforward with additional terms on the score functions. Then this adjustment make it impossible to use GCIME and creates a sandwich form variance matrices both for WNLS and PQMLE.

The discussion so far has been about the structural coefficients inside the exponential function. When our quantity of interest is ATE, which is a nonlinear function of structural parameters, the asymptotic approximation of the variance matrix can be obtained by delta method. Recall that the ATE is estimated as in equation (2). Let the nonlinear functions $g_1(x,\theta)$ and $g_0(x,\theta)$ are continuous and differentiable at θ_0 and the derivatives with respect to θ be denoted by G_1 and G_0. Keay (2022) shows that the asymptotic distribution of ATE estimator in equation (2) is

$$\sqrt{N}(\widehat{ATE} - ATE) \to_d N(0, V),$$

where

$$V = E[T]^2 + (G_1 - G_0)A_0^{-1}B_0A_0^{-1}(G_1 - G_0)'$$

and

$$g_1(x, \theta_0) \equiv \exp(x, \beta_1) \ , \qquad g_0(x, \theta_0) \equiv \exp(x, \beta_0)$$

$$T \equiv g_1(x, \theta_0) - g_0(x, \theta_0) - (E[g_1(x, \theta_0)] - E[g_0(x, \theta_0)])$$

In the above, T is the demeaned ATE conditional on x, which is a population property that is not related to particular estimator being used. Therefore the asymptotic variance of a nonlinear ATE estimator is effectively determined by the asymptotic variance of structural parameters and the covariance between T and $(G_1 - G_0)A_0^{-1}s_i(\theta_0)$. In other words, larger asymptotic variance of structural parameters gives larger asymptotic variance for nonlinear ATE estimator unless it is in such a way that the covariance of T and $(G_1 - G_0)A_0^{-1}s_i(\theta_0)$ becomes larger to cancel out at the same time. This issue will be discussed shortly. Incidentally if the structural parameters are estimated by two-step method, the terms B_0 and $s_i(\theta_0)$ can be easily adjusted by using the result from two-step M-estimator (See Wooldridge, 2010). Now let's discuss the asymptotic distribution of LFES estimator that uses the first stage probit and second stage OLS. The ATE is the difference of intercepts as shown in equation (3). Each of the intercepts can be identified by using the Heckman correction method for each regime. By using the first stage probit index estimates that are used by both the regimes, the second stage estimating equation can be written in a single equation as

$$E[y|z, w] = \mu_0 + (\mu_1 - \mu_0)w + x\beta_0 + wx(\beta_1 - \beta_0) + \rho_1 w\lambda(z\delta) + \rho_2(1 - w)\lambda(-z\delta),$$

where $\lambda(\cdot)$ is the inverse Mill's ratio, i.e. the ratio of normal pdf to the normal cdf. A nice feature in LFES model is that the ATE is not estimated as a nonlinear function of parameters, but a parameter itself. Therefore the asymptotic variance of ATE, i.e. the coefficient on w is now straightforward.

Now how do the asymptotic variances of ATEs in LFES and NFES compare? Since they are estimated from different likelihood functions, it is hard to compare the variances directly and one has to resort to a numerical comparison for a particular estimation case. However, one aspects of nonlinear ATE estimator is that its asymptotic variance contains the term $E[T]^2$. The asymptotic variance of nonlinear ATE estimators can increase as $E[T]^2$ increases while that of linear estimators does not. Practically, nonlinear estimator might have higher variance when there are large differences of conditional ATE among individuals. In the next section, we discuss an application where $E[T]^2$ is not large by nature, so the nonlinear estimator is more efficient than the linear ones.

BOTSWANA FERTILITY

Primary education may increase the human capital and lifetime wage and thereby increase the opportunity cost of having a child (Becker and Barro, 1988; Barro and Becker, 1989), and it may help reduce the child's mortality rate and hence let mothers have fewer children to reach a desired level of family size (Lam and Duryea, 1999; Schultz, 1994a,b). Other than that an enhanced literacy can help them use contraceptive method more effectively (Rosenzweig and Schultz, 1985, 1989). Based on those theoretical background, we are interested on how much the primary education reduces the number of children in Botswana. The sign of the effect is certainly presumed to be negative. Moreover, those who got primary education may have better health information for their children, which may possibly reduce the child mortality.

The data used in this empirical analysis is from Wooldridge (2010, Chapter 21). The variables description and descriptive statistics are given in Table 1 and 8. There was a huge increase of enrollment rate in Botswana during 1970s. The female enrollment rate in early 1970s were roughly 60% and kept increasing for the whole decade until it reached nearly 100% in 1980(UNESCO, 2011). Due to that increase in enrollment, in 1989, the year this data set was collected, more than half the total female population had at least seven years of primary education. Thus this data set captures the ideal time point where there were even amount of control and treatment groups.

The dependent variables under analysis are $children$ (number of living children), ceb (number of total children born) and $mort$ (number of dead children) and the covariates are age, $agesq$, $evermarr$ (ever married), $urban$ (living in urban area), $electric$ (has electricity), tv (has a TV) and $radio$ (has a radio). The variable of interest, i.e. the treatment variable is $educ7$ (finished primary education) and the instrument variable is $frsthalf$ (born in first half of year). The correlation between $educ7$ and $frsthalf$ is -.106. We are interested in the effect of women's primary education on the number of children that she ever has(ceb) and that of living children($children$). Although we are trying all the linear and nonlinear methods for estimating the ATE of education on fertility, the nonlinear estimators are expected to perform better in two reason: First, the outcome variable is typical count variable with small natural numbers and thus modeling the conditional mean as exponential function is well justified. Second, the ATE conditional on covariates might not be substantially different. In other words, we would not assume neither substantial difference of causal effects across different age groups nor any particular time trend.

Tables 3, 4 and 5 display regression results for various models and estimation methods with $children$ as the dependent variable. In what follows the regime with primary education will be called regime one with a subscript 1 and the regime without it will be regime zero with a subscript 0. In Table 3 presents the estimation results for linear models. The ATE estimates of LFES(Heckit) is -1.552 but not statistically significant. Although LET(Heckit) and 2SLS differ only in the first stage regression, the estimates of LET(Heckit) is almost twice as large as the 2SLS estimate. The LFES(Heckit), LET(Heckit) and 2SLS give $\widehat{ATE}$ with a lot larger magnitude OLS does, which might be an evidence of endogeneity. It is, however, very hard to get any meaningful conclusion just by seeing the linear regression results: the only consistent estimator LFES(Heckit) fails to give significant result, and other estimators of which estimates are significant do not seem to agree with one another.

Table 4 shows the results of NET estimators. We already know that the NET model does not identity the true ATE unless the single regime restriction is true. Indeed the ATE_{NET} estimates are substantially smaller than the ones from other estimators. It was also pointed out in Section 3 that each estimator does not even agree with each other under wrong restriction, which is well demonstrated here; the magnitude of PQMLE and NLS estimates are very different and they seem to head to different places. The results show that the ATE_{NET} estimate by PQMLE is close to zero and not significant. Although only NLS gives an estimate weakly significant at 10% level, the magnitude is relatively smaller than those of linear models; it estimates that the primary education reduces the number of children by no more than 0.68. The coefficient on *educ7* is the semi-elasticity because it is inside the exponential function. The PQML estimate for the semi-elasticity does not give any evidence of effectiveness of primary education. Only the NLS estimate is weakly significant reporting roughly 30% decrease of living children. The big difference in the estimation results indicates that the NET model might be misspecified.

Table 5 lists the results of NFES estimators. The NFES estimates report that the primary education reduces 0.8(PQML) or 1.2(NLS) children. It is worth mentioning that standard error of NFES estimates are a lot smaller than those of other estimators, due to which all the three NFES estimates are significant at 1% level. What is particularly interesting is the fact that the NFES estimates support the validity of 2SLS estimate by providing similar values. As Angrist and Evans(1998) and Angrist and Pischke(2009) point out, the linear IV methods must give a consistent estimator of the LATE. If LATE and ATE are not substantially different, then the similar results of the two approaches suggest that the nonlinear model is valid. The main benefit of the NFES is that it provides the ATE with an improved efficiency.

The estimated regime one (with primary education) averages $\sum \widehat{children}_1/N$ for three estimators are 1.264(NLS), 1.499(2PQML), 1.482(1PQML) and those of regime zero (without primary education) $\sum \widehat{children}_0/N$ are 2.488(NLS), 2.340(2PQML), 2.312(1PQML); from those values one can compute the semi-elasticities, i.e. -0.49(NLS), -0.36(2PQML), and -0.36(1PQML). All those estimates are greater in absolute value than the ones from NET model. From these, it becomes more obvious that the NET estimators give us information that looks very much different from what was provided by other estimators. Lastly we can directly test the restriction put on the NET model. One may use the Wald test of $H_0: \beta_1 = \beta_0$. The p-values for 2PQMLE and NLS are 0.000 and that of 1PQMLE is 0.001 implying that there actually exist two regimes. Since 2QMLE and NLS use two-step procedure, the asymptotic variance approximation has to account for the first stage error. One of the advantages of single-step 1PQMLE is that such first stage error is not present and the inference is straightforward. Although there is slight difference in the p-values, such trivial difference is not thought to be of any practical importance. All the above results unequivocally show that the NET model is not an appropriate model to be used to describe this data set. We can also test the endogeneity by checking the covariance between v and ϵ_g. Ignoring NET model, all the two regime estimators show that the regime one covariance is significantly positive, whereas the one at regime zero, slightly negative, is not statistically different from zero. Overall the use of two regime endogenous switching model is well justified.

CONCLUDING REMARK

The main contribution of this study is to clarify the asymptotic distribution of the ATE estimator based on NFES model. Unlike other structural parameters, the ATE estimates are computed by a nonlinear function of the parameter estimates. The estimation error therefore comes both from the error in parameter estimation and also from the computation of ATE by the parameter estimates. The asymptotic distribution reveals that each factor can be written additive separably. The theory predicts that the efficiency of nonlinear ATE estimator is not taken for granted as in many other nonlinear cases. The application shows an example in which this nonlinear methodology can be successfully used. A nonlinear method is expected to be perform better if the variance of ATE conditional on covariates are not substantial as in the Botswana fertility example.

REFERENCES

Angrist, J. (2001). Estimations of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice. *Journal of Business and Economic Statistics*, *19*, 216.

Angrist, J. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Angrist, J., & Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton. Princeton University Press.

Barro, R., & Becker, G. (1989). Fertility Choice in a Model of Economic Growth. *Econometrica*, *57*(2), 481–501.

Becker, G., & Barro, R. (1988). A Reformulation of the Economic Theory of Fertility. *Quarterly Journal of Economics*, *103*(1), 1–25.

Bratti, M., & Miranda, A. (2010). Non-pecuniary returns to higher education: The effect on smoking intensity in the UK. *Health Economics*.

Choi, P., & Min, I. (2009). Estimating endogenous switching regression model with a flexible parametric distribution function: Application to Korean housing demand. *Applied Economics*, *41*(23), 3045–3055.

Fahlenbrach, R. (2009). Founder-CEOs, Investment Decisions, and Stock Market Performance. *Journal of Financial and Quantitative Analysis*, *44*(2), 439–466.

Gourieroux, C., Monfort, A., & Trognon, C. (1984). Pseudo-Maximum Likelihood Methods: Theory. *Econometrica*, *52*, 681–700.

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. *Annals of Economic and Social Measurement*, *5*, 475–492.

Hellström, J., & Nordström, J. (2008). A count data model with endogenous household specific censoring: The number of nights to stay. *Empirical Economics*, *35*, 179–192.

Hirano, K., Imbens, G., Rubin, D., & Zhou, X.H. (2000). Assessing the Effect of an Influenza Vaccine in an Encouragement Design. *Biostatistics*, *1*(1), 69–88.

Holland, P. (1986). Statistics of Causal Inference. *Journal of the American Statistical Association*, *81*, 945–960.

Imbens, G. (2005). Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*.

Imbens, G., & Wooldridge, J. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Innes, R., & Sam, A. (2008). Voluntary Pollution Reductions and the Enforcement of Environmental Law: An Empirical Study of the 33/50 Program. *Journal of Law and Economics*, *51*(2), 271–296.

Keay, M.J. (2010). *Alternative Estimators of Average Treatment Effect under Misspecification and Weak Instrumental Variables*. Department of Economics, Michigan State University.

Keay, M.J. (2022). An Exponential Endogenous Switching Regression with Correlated Random Coefficients. *Econometrics*, *10*(1).

Kenkel, D., & Terza, J. (2001). The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect. *Journal of Applied Econometrics*, *16*(2), 165–184.

Koç, Ç. (2005). Health-Specific Moral Hazard Effects. *Southern Economic Journal*, *72*(1), 98–118.

Lam, D., & Duryea, S. (1999). Effects of schooling on fertility, labor supply and investments in children, with evidence from Brazil. *Journal of Human Resources*, *34*(1), 160–192.

Masuhara, H. (2008). Semi-nonparametric count data estimation with an endogenous binary variable. *Economics Bulletin*, *3*(42), 1–13.

McGeary, K., & French, M. (2000). Illicit Drug Use and Emergency Room Utilization. *Health Services Research*, *35*(1), 153–169.

Mealli, F., Imbens, G., Ferro, S., & Biggeri, A. (2004). Analyzing a Randomized Trial on Breast Self-Examination with Noncompliance and Missing Outcomes. *Biostatistics*, *5*(2), 207–222.

Miranda, A. (2003). *Socio-economic characteristics, completed fertility, and the transition from low to high order parities in Mexico*. University of Warwick.

Miyata, S., & Sawada, Y. (2007). *Learning, Risk, and Credit in Households' New Technology Investments: The Case of Aquaculture in Rural Indonesia*.

Parrado, E., & Flippen, C. (2005). Migration and Gender among Mexican Women. *American Sociological Review*, *70*(4), 606–632.

Parrado, E., Flippen, C., & McQuiston, C. (2005). Migration and Relationship Power among Mexican Women. *Demography*, *42*(2), 347–372.

Partha, D., & Trivedi, P. (2004). *Specification and Simulated Likelihood Estimation of a Non-normal Treatment-Outcome Model with Selection: Application to Health Care Utilization*. Department of Economics, Indiana University.

Romeu, A., & Vera-Hernández, M. (2005). Counts with an endogenous binary regressor: A series expansion approach. *Econometrics Journal*, *8*, 1–22.

Rosenzweig, M., & Schultz, T. (1985). The demand and supply of births and its life-cycle consequences. *American Economic Review*, *75*(5), 992–1015.

Rosenzweig, M., & Schultz, T. (1989). Schooling, information, and nonmarket productivity: Contraceptive use and its effectiveness. *International Economic Review*, *30*(2), 457–477.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Education Psychology*, *66*, 688–701.

Rubin, D. (1978). Bayesian Inference for Causal Effects. *Annals of Statistics, 6,* 34–58.

Sam, A. (2010). Impact of government-sponsored pollution prevention practices on environmental compliance and enforcement: Evidence from a sample of US manufacturing facilities. *Journal of Regulatory Economics*, *37*, 266–286.

Schultz, T. (1994a). Human capital, family planning, and their effects on population growth. *American Economic Review*, *84*(2), 255–260.

Schultz, T. (1994b). Studying the impact of household economic and community variables on child mortality. *Population and Development Review*, *10*(0), 215–235.

Shin, J., & Moon, S. (2007). Do HMO plans reduce health care expenditure in the private sector? *Economic Inquiry*, *45*(1), 82–99.

Terza, J. (1998). Estimating Count Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects. *Journal of Econometrics, 84,* 129–154.

Terza, J. (1999). Estimating Endogenous Treatment Effects in Retrospective Data Analysis. *Value in Health*, *2*(6), 429–434.

Terza, J. (2009). Parametric nonlinear regression with endogenous switching. *Econometric Reviews, 28*(6), 555–580.

Terza, J., Bradford, D., & Dismuke, C. (2008). The Use of Linear Instrumental Variables Methods in Health Services Research and Health Economics: A Cautionary Note. *Health Services Research, 43*(3).

UNESCO. (2011). Retrieved from http://stats.uis.unesco.org/unesco/TableViewer/tableView.aspx?ReportId=3674

Vargas, M., & Elhewaihi, M. (2007). *What is the impact of duplicate coverage on the demand for health care in Germany?*

Wooldridge, J. (1997). Quasi-Likelihood Methods for Count Data. In M.H. Pesaran & P. Schmidt (Eds.), *Handbook of Applied Econometrics 2* (pp. 352–406). Oxford: Blackwell.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT: Cambridge, MA.

Zhang, J., & Song, X. (2007). *Fertility Differences between Married and Cohabiting Couples: A Switching Regression Analysis*. [Discussion paper series]. Institute for the Study of Labor.

APPENDIX

TABLE 1
VARIABLES DESCRIPTION

children	number of living children
ceb	children ever born
mort	number of dead children
educ7	$= 1$ if educ ≥ 7
age	age in years
agesq	age^2
evermarr	$= 1$ if ever married
urban	$= 1$ if live in urban area
electric	$= 1$ if has electricity
tv	$= 1$ if has tv
radio	$= 1$ if has radio
frsthalf	$= 1$ if mnthborn ≤ 6

TABLE 2
DESCRIPTIVE STATISTICS

Variable	Mean	Std. Dev.	Min	Max
children	2.267828	2.222032	0	13
ceb	2.441642	2.406861	0	13
mort	.1738133	.5113953	0	7
educ7	.5556065	.4969553	0	1
age	27.40518	8.685233	15	49
agesq	826.46	526.9232	225	2401
evermarr	.4767255	.4995153	0	1
urban	.5166246	.4997808	0	1
electric	.1402019	.3472363	0	1
tv	.0929112	.2903413	0	1
radio	.7017665	.457535	0	1
frsthalf	.5404724	.4984164	0	1

TABLE 3
LINEAR REGRESSION RESULTS: DEPENDENT VARIABLE $children$

Variable	Linear Models				
	OLS	2SLS	LET(Hkt)	LFES(Hkt)	
educ7	-0.398 ***	-1.185 *	-2.232 ***		-1.552
	(0.046)	(0.691)	(0.432)		(0.979)
				R1	R0
age	0.272 ***	0.262 ***	0.249 ***	0.251 ***	0.384 ***
	(0.019)	(0.021)	(0.021)	(0.030)	(0.049)
agesq	-0.002 ***	-0.002 ***	-0.002 ***	-0.003	-0.003
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
evermarr	0.694 ***	0.610 ***	0.499 ***	0.194 **	0.930 ***
	(0.054)	(0.096)	(0.080)	(0.098)	(0.198)
urban	-0.246 ***	-0.178 **	-0.088	0.206 **	-0.478 ***
	(0.047)	(0.078)	(0.066)	(0.088)	(0.172)
electric	-0.337 ***	-0.233 **	-0.094	0.197	-0.512
	(0.074)	(0.114)	(0.098)	(0.126)	(0.347)
tv	-0.330 ***	-0.155	0.078	0.467 ***	-0.563
	(0.085)	(0.182)	(0.124)	(0.142)	(0.698)
radio	0.027	0.153	0.322 ***	0.620 ***	-0.052
	(0.053)	(0.126)	(0.099)	(0.129)	(0.294)
constant	2.489 ***	2.926 ***	3.508 ***		1.878 **
	(0.035)	(0.385)	(0.246)		(0.943)
cov(ϵ, v)			1.108 ***	2.550 ***	-0.524
			(0.257)	(0.454)	(0.905)
R2	0.586	0.563	0.589		0.595
sigma	1.431	1.471	1.427		1.417

TABLE 4
NONLINEAR ENDOGENOUS TREATMENT (NET) WITH 1 REGIME RESULTS: DEPENDENT VARIABLE *children*

Variable	NET (1 Regime)		
	1PQML	2PQML	NLS
ATE	-0.103	-0.120	-0.681 *
	(0.363)	(0.324)	(0.394)
educ7	-0.046	-0.053	-0.303 *
	(0.158)	(0.142)	(0.172)
age	0.340 ***	0.340 ***	0.265 ***
	(0.009)	(0.009)	(0.012)
agesq	-0.004 ***	-0.004 ***	-0.003 ***
	(0.000)	(0.000)	(0.000)
evermarr	0.326 ***	0.325 ***	0.291 ***
	(0.030)	(0.029)	(0.033)
urban	-0.101 ***	-0.101 ***	-0.085 ***
	(0.024)	(0.023)	(0.027)
electric	-0.162 ***	-0.161 ***	-0.135 ***
	(0.044)	(0.042)	(0.051)
tv	-0.203 ***	-0.201 ***	-0.108 *
	(0.061)	(0.058)	(0.070)
radio	-0.015	-0.014	0.035
	(0.032)	(0.030)	(0.037)
constant	-5.514 ***	-5.507 ***	-4.059 ***
	(0.085)	(0.077)	(0.240)
cov(ϵ, v)	-0.060	-0.056	0.099
	(0.094)	(0.085)	(0.104)
L-likelihood	-1088.24	1283.41	8597.94

TABLE 5
NONLINEAR FULL ENDOGENOUS SWITCHING REGRESSION (NFES) WITH 2 REGIMES RESULTS: DEPENDENT VARIABLE *children*

Variable	NFES (2 Regimes)					
	1PQML		2PQML		NLS	
ATE		-0.830 ***		-0.841 ***		-1.224 ***
		(0.318)		(0.322)		(0.375)
	_ R1	R0	R1	R0	R1	R0
age	0.412 ***	0.294 ***	0.410 ***	0.293 ***	0.333 ***	0.239 ***
	(0.019)	(0.015)	(0.018)	(0.015)	(0.022)	(0.017)
agesq	-0.006 ***	-0.003 ***	-0.005 ***	-0.003 ***	-0.004 ***	-0.003 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
evermarr	0.268 ***	0.312 ***	0.270 ***	0.309 ***	0.219 ***	0.293 ***
	(0.045)	(0.042)	(0.044)	(0.042)	(0.050)	(0.045)

urban	0.006	-0.129 ***	0.003	-0.127 ***	0.031	-0.111 ***
	(0.041)	(0.034)	(0.040)	(0.033)	(0.045)	(0.036)
electric	-0.070	-0.155 *	-0.074	-0.150 *	-0.028	-0.142 *
	(0.057)	(0.079)	(0.057)	(0.077)	(0.065)	(0.085)
tv	-0.060	-0.121	-0.063	-0.111	0.075	-0.062
	(0.079)	(0.155)	(0.078)	(0.153)	(0.089)	(0.174)
radio	0.067	0.000	0.063	0.004	0.169 **	0.021
	(0.061)	(0.046)	(0.060)	(0.045)	(0.069)	(0.049)
constant	-6.843 ***	-4.728 ***	-6.813 ***	-4.697 ***	-5.671 ***	-3.677 ***
	(0.101)	(0.153)	(0.098)	(0.146)	(0.092)	(0.169)
cov(ϵ, v)	0.390 ***	-0.081	0.373 **	-0.062	0.656 ***	-0.009
	(0.178)	(0.169)	(0.169)	(0.159)	(0.202)	(0.178)
L-likelihood		-1067.75		1303.73		8521.37

Note: R0=no education regime, R1=education regime. All the covariates are demeaned. All the figures in the parenthesis are bootstrap standard errors. *: significant at 10%, **: 5%, ***: 1%