ESSAYS ON AVERAGE TREATMENT EFFECTS

By

Myounggin Keay

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Economics

2012

ABSTRACT

ESSAYS ON AVERAGE TREATMENT EFFECTS

By

Myounggin Keay

This dissertation consists of three essays on estimating average treatment effects (ATE) under counterfactual framework. In Chapter 1, I compare the performances of single-step and two-step estimators for estimating the ATE in a linear model when treatment assignment depends on unobservables. Recent advances in computing technology have enabled the extensive use of single-step estimators, such as Limited Information Maximum Likelihood (LIML), instead of 2SLS. In this study I make clear that there are two kinds of single-step estimators for estimating ATE. LIML-type estimator is the one which uses the control function method, on which the two-step method is also based, whereas FIML-type estimator directly uses the joint distribution of underlying errors or endogenous variables. I find that the relative asymptotic efficiency between two-step Heckit and single-step LIML cannot be determined in general. However, the relative efficiency of single-step LIML with respect to two-step Heckit is decreasing as the sample size increases, implying that if the asymptotic variances are same, then single-step LIML is less efficient in finite samples. On the other hand the FIML estimator tends to have very small finite sample variances, but it is less robust to misspecification. Newey-type series estimators are also considered for correcting the misspecification of error distributions, but it turned out that cost is greater than the benefit. Under weak many instrument cases, the advantage of LIML in terms of median bias was not as strong as in the linear models.

Chapter 2 explores the ATE estimator proposed by Terza (2009)'s Nonlinear Full En-

dogenous Treatment (NFES) model, where count dependent and binary treatment variables are present. When the true conditional mean function takes the form of exponential function, the Heckit-type linear method, while it can be a good approximation, is inconsistent for the true ATE since it is derived under the assumption of linear conditional mean. The asymptotic distribution of nonlinear estimators have additional terms in asymptotic variance of which magnitudes depend on population coefficient. Due to their presence, the asymptotic variances of nonlinear estimators can be either larger or smaller than the linear counterparts depending on the values of coefficients. It turns out that they tend to have small variances when the variance of conditional ATE are small. And Monte Carlo experiments show that they are fairly robust to various distributional misspecifications. In summary, nonlinear ATE estimators are robust and consistent with small variance when the treatment effects are not substantially different across individuals. An application to Botswana fertility is given where the treatment is seven years of education with the dependent variable fertility.

Chapter 3 presents a method for estimating ATE for the case that the dependent variable is count variable and the coefficients of covariates are random variables which are correlated with the binary treatment variable. The identifying assumptions are given and the estimating equation based on them is derived. Simulations show that, in large samples, it has usually smaller biases and larger variances than the linear methods have. An application on Botswana fertility is given with same variables as in Chapter 2.

Copyright by
MYOUNGGIN KEAY
2012

To Jiyoung and Un-hyeng

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Jeff Wooldridge. Without his support and guidance, this dissertation would never have been able to come to existence. I also extend my appreciation to Todd Elder, Peter Schmidt, Tim Vogelsang and Hira Koul. Their comments and insights have helped me a great deal whenever I was stuck on obstacles. I also thank Otávio Bartalotti, Guojun Chen, Cheol Keun Cho, Do Won Kwak, Jin Young Lee, Shengwu Shang, Valentin Verdier, Yali Wang and all other my collegues at Michigan State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Alternative Estimators of Average Treatment Effect under Misspecification and Weak Instrumental Variables

1.1 Introduction

Although single-step estimators such as LIML has long been recognized as a good alternative to the traditional 2SLS (Anderson *et al.*, 1982), the computational burden has prevented them from being widely used in applied researches. Rather, people have preferred two-step methods or, for some nonlinear models, have sought to find appropriate two-step estimation procedure in order to circumvent the computational difficulty (Greene, 1998; Maddala, 1986). Nevertheless recent new findings about LIML along with the advances in computing technology have reinvigorated researchers' interest on those single-step estimators. Since the controversial Angrist and Kruger (1991), we are now more familiar to the properties or advantages of LIML estimators especially when there are many weak instrumental variables (Bekker, 1994; Staiger-Stock, 1997; Flores-Lagunes, 2007). Extending our attention to the

broad class of single-step estimators to which LIML belongs, we see that there are some more advantages of single-step estimators other than in weak many IV's: First, it reduces the sampling error by skipping the first stage. Second, as in bivariate probit models, there are some cases where there is no proper consistent two-step method (Wooldridge, 2010); single-step estimator usually exists under some appropriate distributional assumptions.

One of the leading examples of single-step estimator is LIML in linear simultaneous models. In LIML, after reduced form transformation, both the structural equation and linear projection of endogenous explanatory variable are jointly estimated by maximum likelihood method . Although 2SLS and LIML give different estimates, they can be interpreted as two different opeartional version of same control function (CF) method; LIML is identical to the single-step CF method, while 2SLS is to the two-step CF in linear models. If CF method is available for a particular nonlinear model, then each single-step and two-step CF estimating procedure can be understood as a generalized LIML and 2SLS respectively in such nonlinear models (Wooldridge, 2007). In line with this idea, LIML can be viewed as a joint ML estimation method based on the CF approach; such approach provides greater generality since LIML can then be applied to the nonlinear models as well as the usual linear simultaneous equations models.

In linear simultaneous models LIML and FIML are mechnically the same; the only difference is that FIML is a term used when the whole equations in the system are under consideration, whereas LIML is used for a single equation in the system (Hayashi, 2000). In the context of endogenous switching regression, FIML indicates an estimation method through direct modeling of endogenous variables without resorting to control functions (Maddala, 1986). If the selection equation is viewed as a structural equation, such method can be called FIML because the absence of selection variable in the other structural equations makes

the reduced form transformation unnecessary. Therefore there are essentially two kinds of single-step estimators at hand, which are the ones with and without control function in the estimating equation. In this article, the former will be called LIML and the later FIML.

Even though LIML was fully robust in linear models, one of the shortcomings of the single-step estimators in nonlinear models is that they usually require strong distributional assumptions for constructing the likelihood functions. They can be inefficient or even inconsistent unless the distributions are correctly specified. In order to address this problem, one can also consider some distribution-free methods such as nonparametric or semiparametric approaches. The objective of this article is to propose those various estimators and then to investigate their performances for the estimation of average treatment effect (ATE) under counterfactual settings. One of the most common situations requiring two-step estimation might be the case where an explanatory variable is endogenous. In such a situation, a nonlinearity can easily be brought by assuming a binary endogenous explanatory variable. Therefore the discussion starts from estimating ATE in endogenous switching regression models with linear structural equations.[1] As it will be clearer later, the partial effects of binary endogenous variables is a special case of ATE under some parameter restrictions. Although Angrist (2001) argues that the linear model is sufficiently fine for the partial effect estimation, this article attempts to justify the use of nonlinear models, which will make the whole argument more meaningful.

Section 2 describes the model under consideration. Here the relation between the partial effects of binary endogenous variable and the ATE under counterfactual setting will be clarified. Section 3 will provide a detailed description of those four estimators i.e. two-

[1] The case of nonlinear structural model is the topic of the second chapter of this dissertation.

step Heckit, single-step (Q)LIML, single-step (Q)FIML and Series estimators. Section 4 derives the asymptotic distributions of two-step and single-step estimators with the same CF estimating equation. As it turned out, the asymptotic variance of one estimator is not always greater or smaller than that of the other. Threfore the asymptotic analysis does not provide any guideline as to which estimator is more efficient. Section 5 briefly describes the data generating process used for the simulation and Section 6 discusses the results. The focus of the analysis will be on the comparison between single-step and two-step estimators mainly in terms of consistency and efficiency.

1.2 Model

As it was briefly discussed in introduction, the main objective of this paper is to compare the estimators for the partial effects of binary variables or ATEs depending on the model specifications. Suppose that we are given a model with a binary endogenous variable. Although the partial effect can easily be estimated given a set of proper instrumental variables, the main focus here is how to use the fact that the underlying endogenous variable is binary. Although it is perfectly legitimate to use usual IV procedure without paying too much attention to the binary nature of the variable, I attempt to make use of such information in order to make a better estimation. Therefore the discussion starts from the generic problem of estimating partial effect of an endogenous explanatory variable. As it will be clarified below, this issue can obtain generality by extending it to the topic of ATE. Below it will be shown that they use common estimating equation although there are slight differences between the ATE and the partial effects of binary endogenous explanatory variable. To state more properly it will be shown that estimating the partial effects of binary variables is a special case of estimating

the ATE.

Consider a model as below.

$$\begin{aligned} y &= \mu + \tau w + \beta x + u \\ w &= \mathbf{z}\delta + \xi, \end{aligned} \tag{1.1}$$

In the above equations, w is the binary endogenous explanatory variable where τ is the partial effect, x is exogenous covariates, and $\mathbf{z}$ is a set of all exogenous variables containing x which are uncorrelated with u. In order for $\mathbf{z}$ to be proper IV, it must be $\delta \neq 0$. There is absolutely no problem in estimating the τ by the usual IV procedure. If the endogeneity of w is interpreted as an existence of unobserved variables c included in the error, then the partial effect τ can be written as $E(y|x, c, w = 1) - E(y|x, c, w = 0)$.

Now we want to estimate better by somehow using the additional information, i.e. the binary nature of w. One natural possibility is to use any binary choice model for the reduced form equation instead of linear probability model. A clear account of ATE is warranted at this point before a further discussion. Suppose a counterfactual setting with a binary choice variable; there exist two regimes for each unit of observation. Let the response variable only in regime one is observed under $w = 1$ while that in regime zero is observed under $w = 0$. The exogenous variables x are always observed. The value of w is affected by the values of response variables for each regime, which creates correlation between response variables in each regime and w. Thus the binary variable w gets an endogeneity, and the model under such environment is sometimes called endogenous switching regression. A classic example of such model is found in job training program analysis. In such model the binary variable w denotes whether an individual takes the program or not, and the response variables y_1 and y_0 denote the wages with and without participation in the program (for recent survey

on program evaluation, see Imbens and Wooldridge (2009)). It is not hard to imagine those two regimes for each particular individual although those two response variables are not observed in the real world at the same time; in line with such an idea the model is called counterfactual (Rubin, 1974). For a particular individual the difference between y_1 and y_0 is called the treatment effect. The expectation of the treatment effects over the whole population is called the average treatment effect, i.e. $E(y_1 - y_0)$. Thus it is obvious that this ATE is not generally same as the partial effect of the binary treatment variable $E(y|x, c, w = 1) - E(y|x, c, w = 0)$.

An estimable equation for estimating the ATE is derived below. Since there are two regimes,

$$\begin{aligned} y_0 &= \mu_0 + x\beta_0 + u_0 \\ y_1 &= \mu_1 + x\beta_1 + u_1. \end{aligned} \tag{1.2}$$

In the above equations, the relation between x and y are assumed to be linear. For now let us use another assumption that $\beta_0 = \beta_1$ and $u_0 = u_1$, which implies that the treatment has an intercept shifting effect only. In other words, the treatment shifts the response variable by exactly same magnitude for all the individuals in population. In addition to that let $E(x) = 0$. This is without loss of generality because the intercept can be adjusted so that $E(x) = 0$ be true. By those manipulations, it is easy to see that the ATE can be expressed as

$$E(y_1 - y_0) = \mu_1 - \mu_0$$

In equations (1.2) y_1 is denoted as the response variable when $w = 1$ while y_0 is when $w = 0$. In fact the observable response variable is y_0 when $w = 0$ and so on. Let us denote the

observed response variable as y, then

$$y = y_0 + (y_1 - y_0)w$$

Therefore using the above equation, the equations (1.2) can be incorporated into a single equation.

$$y = \mu_0 + (\mu_1 - \mu_0)w + \beta x + u \tag{1.3}$$

Thus the partial effect of w becomes the ATE under the restriction above. Notice that the equation (1.3) is same as the equation (1.1). Remembering that the restrictions $\beta_0 = \beta_1$ and $u_0 = u_1$ were put in equation (1.2) to obtain equation (1.3), we can easily see that the equation (1.1) is just a special case of those models with counterfactual settings. Although we started from a special case of binary endogenous explanatory variable in the beginning, it will be discussed under the broad settings of ATEs for the rest of this article.

Returning to the equation (1.2), let us consider how to identify the ATE in a model without any restrictions by using the fact that the treatment is binary. To do that there has to be at least more than one IV for w. Rewriting the equation (1.2) with the reduced form equation,

$$\begin{aligned} y_0 &= \mu_0 + x\beta_0 + u_0 \\ y_1 &= \mu_1 + x\beta_1 + u_1 \\ w &= 1[\mathbf{z}\delta + \xi > 0] \end{aligned} \tag{1.4}$$

It can be easily seen that once we drop one of the regimes in the above equation, then it simply becomes the well-known Heckman correction method (Heckman, 1976). Therefore we can use the Heckman correction model twice for each regime in order to consistently estimate the coefficients in each regime. Of course all the assumptions for Heckman correction model

are also required in this model: $\mathbf{z}$ is mean independent of u_g where $g = 0, 1$, $E(u_g|\xi)$ is linear in ξ, and ,although not necessary for the identification itself, w follows probit model so that the inverse Mill's ratio can be used. Therefore the estimating equation can be written as

$$\begin{aligned} y &= \mu_0 + (\mu_1 - \mu_0)w + x\beta_0 + wx(\beta_1 - \beta_0) + u_0 + (u_1 - u_0)w \\ &= \mu_0 + \tau w + x\beta_0 + wx(\beta_1 - \beta_0) \\ &\quad + \rho_1 w\lambda(\mathbf{z}\delta) + \rho_0(1-w)\lambda(-\mathbf{z}\delta) + e_0 + (e_1 - e_0)w, \end{aligned} \tag{1.5}$$

where $\lambda(\cdot)$ is inverse Mill's ratio, and τ is denoted as the ATE. Estimating procedure is basically same as Heckman correction model; in the first stage estimate the selection equation by probit and put the estimated parameters $\widehat{\delta}$ for δ and run the second stage regression just like usual ordinary least squares. If one believes that $\beta_0 = \beta_1$ and $u_0 = u_1$, then the estimating equation simply becomes

$$y = \mu_0 + \tau w + x\beta_0 + \rho\Big(w\lambda(\mathbf{z}\delta) - (1-w)\lambda(-\mathbf{z}\delta)\Big) + e_0. \tag{1.6}$$

Running two-stage regression of the above equation (1.6) is an alternative estimation method to the usual IV regression with linear probability model for selection. Nevertheless, following discussions will be based on the equation (1.5) with greater generality rather than (1.6) .

1.3 Estimation

1.3.1 Two-step Heckit and Single-Step Quasi-LIML

So far we have seen the relationship between partial effect of a dummy endogenous variable and ATE in the previous section. In what follows the equation (1.5) will be used as a basic estimating equation for most part of the discussion.

In equation (1.5) the errors e_g are the difference between the structural errors u_g and the correction terms for each regime. We needed three important assumptions to be able to write the model as in equation (1.5). The first two are the conditional mean independence of $\mathbf{z}$ and u_g and the linear conditional expectation assumption between the errors in each equation i.e. $E(u_g|\mathbf{z}, \xi) = E(u_g|\xi) = \rho\xi$. A sufficient condition for linearity is the trivariate normal distribution between u_g and ξ; under such assumption the coefficient ρ_g will then be the covariance between u_g and ξ. However, the linear conditional expectation will be enough for writing the equation (1.5). Here the distribution of u_g can be flexible; it doesn't have to follow any particular distribution as far as it is continuous. The next assumption is that ξ follows normal distribution. This is more important than the former ones since it warrants the use of inverse Mill's ratio in the structural equation as well as the use of probit for the first stage estimation. Without it, the above representation is invalid particularly for the inverse Mill's ratio, for which case we can possibly consider semiparametric estimation methods instead of using the inverse Mill's ratio for correction terms. And again the first stage estimation can also be run by semiparametric binary choice model, which will be clarified later. The following three propositions summarize the above discussion.

Proposition 1.1. Given equation (1.4), it can be shown that

$$E(y_1|\mathbf{z}, w = 1) = \mu_1 + x\beta_1 + \frac{1}{\Pr(w = 1)} \int \int_{-\mathbf{z}\gamma}^{\infty} u_1 p(u_1, \xi|\mathbf{z}) d\xi du$$

Proposition 1.2. Given equation (1.4), if $E(u_1|\mathbf{z}, \xi) = E(u_1|\xi) = \rho\xi$, then

$$E(y_1|\mathbf{z}, w = 1) = \mu_1 + x\beta_1 + \rho \int_{-\mathbf{z}\gamma}^{\infty} \xi p(\xi|\xi > -\mathbf{z}\gamma) d\xi$$

Corollary 1.1. If $\xi \sim N(0, 1)$ in addition to the assumption in proposition 1.2, then

$$E(y_1|\mathbf{z}, w = 1) = \mu_1 + x\beta_1 + \rho\lambda(\mathbf{z}\gamma),$$

where $\lambda(\cdot)$ is the inverse Mill's ratio. Moreover, if u_1 and ξ follow bivariate normal, then $\rho = \text{cov}(u_1, \xi)/\text{var}(\xi)$.

One can use the two-step Heckit approach discussed in the previous section in order to estimate ATE using equation (1.5). Easy to implement as it is, it could be inefficient after accounting for the first stage sampling error. If we are willing to make some further assumptions, then there actually exists a single-step ML estimation method using the likelihood function for e_g, which can be constructed as below. Let us call the composite error as $e = e_0 + (e_1 - e_0)w$ and assume that it follows standard normal distribution and that the selection mechanism is probit. Then a joint density function of y and w can be written as

below.

$$
\begin{aligned}
& p(y, w|\mathbf{z}) = p(y|w, \mathbf{z}) \cdot p(w|\mathbf{z}) \\
= \; & \frac{1}{\sqrt{2\pi \cdot \mathrm{var}(e|w, \mathbf{z})}} \\
& \cdot \exp\left(- \frac{\{y - \mu_0 - \tau \cdot w - x\beta_0 - wx(\beta_1 - \beta_0) - \rho_1 w\lambda(\mathbf{z}\delta) + \rho_2(1-w)\lambda(-\mathbf{z}\delta)\}^2}{2 \cdot \mathrm{var}(e|w, \mathbf{z})} \right) \\
& \cdot [\Phi(\mathbf{z}\delta)]^w \cdot [1 - \Phi(\mathbf{z}\delta)]^{1-w}, \qquad (1.7)
\end{aligned}
$$

where inverse Mill's ratio is compactly denoted as $\lambda(\cdot)$. The conditional variance of each e_g is given by Johnson and Kotz (1972) as

$$
\begin{aligned}
\mathrm{var}(e_0|w = 0, \mathbf{z}) &= \sigma_0^2 - \sigma_{0\xi}^2 \lambda(-\mathbf{z}\delta)\{-\mathbf{z}\delta + \lambda(-\mathbf{z}\delta)\} \\
\mathrm{var}(e_1|w = 1, \mathbf{z}) &= \sigma_1^2 + \sigma_{1\xi}^2 \lambda(\mathbf{z}\delta)\{-\mathbf{z}\delta + \lambda(\mathbf{z}\delta)\},
\end{aligned}
$$

where $\sigma_g^2 \equiv \mathrm{var}(u_g)$ and $\sigma_{g\xi} \equiv \mathrm{cov}(u_g, \xi)$. All the parameters are estimable by a ML estimation with equation (1.7). One of the advantages of this approach is that there is no more sampling error to account for in the first stage estimation, which greatly simplifies the computation of the standard errors. Let us now consider the normality assumption of e which is one of two assumptions used in equation (1.7). In fact our knowledge about e is in fact very limited, although the normality assumption on e was used in order to construct a likelihood function. The composite error e was obtained by subtracting the Heckman correction terms from u. Although we can maintain that u follows normal by the usual central limit theorem argument, it is very hard to maintain that it still follows normal even after it is purged of the elements causing the endogeneity. Other than that, one can surely suspect the possibility of heteroscedasticity because the correction terms are the functions

of explanatory variables and again e also is.

However, even for the case where the likelihood is not correctly specified, we can expect to have at least a consistent estimator as long as the conditional expectation of y is correctly specified (Gourieroux *et al.*, 1984). Without caring about the conditional variances, let us just use a standard normal distribution for $p(y|w, \mathbf{z})$. Taking natural log, equation (1.7) can be written as

$$\begin{aligned} & \ell(y, w|\mathbf{z}) \\ = & -\frac{\{y - \mu_0 - \tau \cdot w - x\beta_0 - wx(\beta_1 - \beta_0) - \rho_1 w\lambda(\mathbf{z}\delta) + \rho_2(1-w)\lambda(-\mathbf{z}\delta)\}^2}{2} \\ & + w \cdot \ln[\Phi(\mathbf{z}\delta)] + (1-w) \cdot \ln[1 - \Phi(\mathbf{z}\delta)] \end{aligned} \tag{1.8}$$

Now suppose that the assumptions in Proposition 1.2. and Corollary 1.1. are all satisfied. Then it can be seen that the true parameter values maximize the expectation of log-likelihood function in equation (1.8). If we label the true parameter value by "o" subscript, then δ_o will surely maximize the expectation of probit log-likelihood function. Also under the assumption that $E(y|w, \mathbf{z}) = \mu_{0o} - \tau_o \cdot w - x\beta_{0o} - wx(\beta_{1o} - \beta_{0o}) - \rho_{1o} w\lambda(\mathbf{z}\delta_o) + \rho_{2o}(1-w)\lambda(-\mathbf{z}\delta_o)$, the true parameters ($\mu_{0o}$, τ_o, β_{go}, ρ_{go}, δ_o) will maximize the expectation of the first term in (1.8). The latter statement is true because the normal density belongs to the linear exponential family. Since (μ_{0o}, τ_o, β_{go}, ρ_{go}, δ_o) maximize the first term in (1.8) and δ_o does the same for the second and third terms, the whole set of parameters maximize $E\ell(y, w|\mathbf{z})$, which guarantees the consistency. The above argument makes it clear that consistency depends on the fact that the selection error is normally distributed, since it enables us to write the correction terms as the inverse Mill's ratio, which is a part of conditional expectation function of y. In other words, although e does not have to follow standard normal distribution, the selection

Table 1.1: Relationship among Estimators

	linear	nonlinear
single-step CF	LIML	(Q)LIML
two-step CF	2SLS	Heckit

error has to be normally distributed in order to ensure consistency.

A general form of correction terms without the normality but with linearity assumption on ξ is shown in the proposition 1.2. Also if the conditional expectation $E(u_g|\xi)$ is a nonlinear function of ξ containing terms of higher degrees, then some additional correction terms might be needed. As it will be discussed in the following section, one can consider using nonparametric or semiparametric methods in such cases where it is hard to determine the validity of those assumptions.

The estimation method using the above log-likelihood function (1.8) will be called quasi-LIML or QLIML. A qualifier quasi- is used here due to the ignorance of the correct distribution of e. It is LIML since the likelihood function is constructed by the joint density function of the response and endogenous explanatory variables as in the linear simultaneous equation case. The only difference between the QLIML used here and the one in linear model is that the former uses joint density of normal and Bernouilli while the later uses multivariate normal.

1.3.2 Series Estimator

According to proposition 1.2 and corollary 1.1, the correction term will take the form of inverse Mill's ratio when $E(u_g|\xi)$ is linear in ξ and ξ is normally distributed. If ξ is not, the correction term will not take the form of inverse Mill's ratio any more. If $E(u_g|\xi)$ is a

nonlinear function of ξ, then the correction term will have some more additional terms as well as the leading inverse Mill's ratio. Although it is still possible to derive the appropriate formulae for correction terms for each distributional assumption, we usually don't have any knowledge about the population whenever the failure of the assumptions is suspected; in such cases one can instead use some distribution-free methods. Rather than the fully parametric approach, a semiparametric method can also be applied for the correction terms. Here I will follow the methods proposed by Powell, Newey and Walker (1990), Newey (1994) and Newey (2009). The discussions in those papers are mainly about the semiparametric estimation of sample selection models where the β's in equation (1.2) are identified. They are semiparametric because the linear index $\mathbf{z}\gamma$ is always used as an argument of control function.

The estimation will be carried out by two steps. In the first stage, the selection equation is estimated by the nonparametric binary choice models as in Powell, Stock, and Stoker (1989), Ichimura (1993) and Klein and Spady (1993). I use Klein-Spady estimator in this paper since it is the most efficient one among those. Given a first stage estimate, a series estimator can be constructed in the second stage as below.

$$y = \mu_0+\tau\cdot w+x\beta_0+w\cdot x(\beta_1-\beta_0)+w\Big(\sum_{p=1}^{P}\rho_p\psi(\mathbf{z}\widehat{\delta})^p\Big)+(1-w)\Big(\sum_{p=1}^{P}\eta_p\psi(-\mathbf{z}\widehat{\delta})^p\Big)+\epsilon_0+w\cdot(\epsilon_1-\epsilon_0) \tag{1.9}$$

The errors appeared here are not the same as e_g in Heckit or in QLIML. They are essentially the sums of the series terms from $P+1$ to infinity. The function ψ is monotone transformation which makes the estimate less sensitive to outliers (Newey, 1994). For ψ, Newey (2009) proposes three monotone functions: identity, inverse Mill's ratio and CDF of standard normal distribution.

In Heckman correction method the correction term in observed subpopulation is the inverse Mill's ratio $\lambda(\mathbf{z}\gamma)$ under the usual assumptions, whereas that in the other unobserved subpopulation is $-\lambda(-\mathbf{z}\gamma)$. Then the unconditional expectation of correction term over the whole population is equal to zero; otherwise the unconditional expectation of error in structural equation won't be zero causing the intercept estimator inconsistent. This condition is not automatically satisfied as in the original Heckman correction model when we apply series estimator with powers of degree more than two and with a function $\psi(\cdot)$ other than the inverse Mill's ratio. Therefore an adjustment that makes the expectation of the correction terms over the whole population be zero is required for a consistent estimation of the intercept and ATE; it can be done by subtracting the sample means of the correction terms.

1.3.3 Quasi-FIML

Before considering this estimator, it should be emphasized that this is not in the class of control function approaches as those previous three estimators. The basic motivation of using control function approach is that we suspect unobserved variables hidden in error which are correlated with one of the regressors in the structural equation. Since it is unobserved, by definition, the distribution that it follows is unknown. Indeed we haven't put any distributional assumptions for u_g or e_g in the two-step Heckman model. Even though we used normal assumption for e_g in QLIML, we do not claim that it is truly normal by labeling it as QLIML. Therefore putting distributional assumption directly to the structural error u_g is not in line with our spirit of viewing this problem. However, if the normality assumption happens to be true, then a FIML estimator that exploits the joint density function becomes a very attractive alternative estimator with correctly specified likelihood. If it is not correctly specified, then the FIML with wrong distributional assumption on the errors becomes

a quasi-FIML or QFIML that can still be used for estimating the ATE. Below is a description of FIML: suppose the joint distribution of u_g and ξ is multivariate normal as below.

$$\begin{bmatrix} u_0 \\ u_1 \\ \xi \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{10} & \sigma_{1\xi} \\ & \sigma_1^2 & \sigma_{0\xi} \\ & & 1 \end{bmatrix} \right)$$

The variance of the selection error is set to be one since the coefficients in probit model can only be estimated up to a scale factor. Under this assumption, we want to derive the joint distribution of y and w conditional on the exogenous variables $\mathbf{z}$. We can easily transform u into y since it is linear. The transformation of ξ into w can be done by the equation below.

$$f(u, w|\mathbf{z}) = \left(\int_{-\mathbf{z}\delta}^{\infty} g(u_1, \xi)\ d\xi \right)^w \cdot \left(\int_{-\infty}^{-\mathbf{z}\delta} h(u_0, \xi)\ d\xi \right)^{1-w} \tag{1.10}$$

Those functions g and h are marginal density functions for u_g and ξ conditional on $\mathbf{z}$. For example g can be obtained by integrating out irrelevant u_0 from the joint distribution for all three variables. Here we can see that the trivariate normal assumption is sufficient condition for our purpose; the above likelihood function shows that the pairwise bivariate distribution assumption for each u_g and ξ is all that is necessary since we are hardly interested in estimating σ_{01}. The integrals on the right hand side of the above equation were needed in order to get the marginal distributions for u_g for the event where $w = 1$ and $w = 0$ respectively. It can be shown that the above density function can be written as

$$f(u, w|\mathbf{z}) = \left[\Phi\left(\frac{\mathbf{z}\delta + (\sigma_{1a}/\sigma_1^2)u_1}{\sqrt{1 - (\sigma_{1a}/\sigma_1)^2}} \right) \cdot \frac{\phi\left(\frac{u_1}{\sigma_1}\right)}{\sigma_1} \right]^w \cdot \left[\left\{ 1 - \Phi\left(\frac{\mathbf{z}\delta + (\sigma_{1a}/\sigma_1^2)u_0}{\sqrt{1 - (\sigma_{1a}/\sigma_1)^2}} \right) \right\} \cdot \frac{\phi\left(\frac{u_0}{\sigma_0}\right)}{\sigma_0} \right]^{1-w} \tag{1.11}$$

Table 1.2: Distributional assumptions

	QFIML	QLIML	Heckit	Series
u	normal	none	none	none
ξ	normal	normal	normal	none
e	NA	none	none	NA

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the CDF and PDF of standard normal distribution respectively. By using (1.2) the above joint density can easily be transformed to $f(y, w|\mathbf{z})$. The estimator using the likelihood function given above is called FIML (Maddala, 1986). In linear simultaneous equations models FIML basically uses the joint distribution of all endogenous variables conditional on exogenous ones; the only difference in this nonlinear setting is that the joint distribution of endogenous variables are constructed without reduced form transformation. Since there is no guarantee that the errors follow trivariate normal, it should be properly called "quasi"-FIML with a qualifier. The ATE is estimated simply as the difference of estimated intercepts in equation (1.2). If the true distribution is trivariate normal, then this estimator becomes FIML and will be consistent and efficient. However, if that is not the case, then it is neither consistent nor efficient. Therefore FIML is less robust than Heckit or QLIML under various violations of distributional assumptions.

In sum there are the four available estimators for ATE using instrumental variables in counterfactual setting. Those estimators were constructed under different distributional assumptions which are summarized below. NA indicates the underlying error term is not in the model. Even though not in the table, it has to be noted that there is linear conditional expectation condition between u and ξ for the Heckit and QLIML.

1.4 Asymptotic Variances

In this section the asymptotic distributions of Heckit and QLIML, which are applications of two-step and single-step CF method respectively, are given. The asymptotic distribution of QFIML is straightforward and that of the series estimator in the following discussion is derived by using essentially same method as in the two-step Heckit. In what follows a common structural log-likelihood function for both single-step and two-step methods will be used. The standard normal distribution will be used as the quasi-likelihood of structural equation for QLIML estimation. Heckit uses OLS at second step; computing the least square is equivalent to using the standard normal distribution for e. Therefore one can simply use a common structural log-likelihood function to derive the asymptotic distribution of the two-step Heckit as well as single-step QLIML. To make the discussion as simple as possible, the parameters θ_1 and θ_2 are assumed to be scalars.

Let us first consider the asymptotic distribution of QLIML. Let $q^1(y|w, \mathbf{z}; \theta_1, \theta_2)$ be the log-likelihood or objective function to be considered for the structural equation and $q^2(w|\mathbf{z}; \theta_2)$ for the reduced form equation. In this particular endogenous switching regression setup, $\theta_1 = (\tau, \beta')'$ and $\theta_2 = \delta$, but again the discussion below assumes that they are scalar for the sake of simplicity. Since the two sets of parameters are estimated together, the log-likelihood function for QLIML is expressed in additive form as $q^1(y|w, \mathbf{z}; \theta_1, \theta_2) + q^2(w|\mathbf{z}; \theta_2)$.

Let $q_j^i \equiv \partial q^i/\partial\theta_j$ and $q_{jk}^i \equiv \partial^2 q^i/\partial\theta_j\partial\theta_k$. Then

$$\begin{aligned} B_0 &= \begin{pmatrix} E(q_1^1)^2 & E(q_1^1 q_2^1) \\ E(q_1^1 q_2^1) & E(q_2^1)^2 + E(q_2^2)^2 \end{pmatrix} \\ A_0 &= \begin{pmatrix} E(q_{11}^1) & E(q_{12}^1) \\ E(q_{21}^1) & E(q_{22}^1) + E(q_{22}^2) \end{pmatrix} \end{aligned}$$

and $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d (0, A_0^{-1}B_0A_0^{-1})$. The asymptotic distribution of θ_1 is the first diagonal element of the sandwich form as below.

$$\begin{aligned} &\text{Avar}\sqrt{n}(\widehat{\theta}_{1n} - \theta_{10}) \\ &= \frac{E(q_1^1)^2K^2 - E(q_{12}^1)\Big(2E(q_1^1q_2^1)K - E(q_{12}^1)[E(q_2^1)^2 + E(q_2^2)^2]\Big)}{\Big(E(q_{11}^1)K - [E(q_{12}^1)]^2\Big)^2}, \end{aligned}$$

where $K = E(q_{22}^1) + E(q_{22}^2)$ and $\widehat{\theta}_{1n}$ denotes QLIML estimator of θ_1.

A two-step approach maximizes $q^1(y|w, \mathbf{z}; \theta_1, \widehat{\theta}_2)$ given the first stage estimates $\widehat{\theta}_2$. Then it can be shown by using the result in Wooldridge (2010, Chapter 12) that

$$\text{Avar}\sqrt{n}(\widetilde{\theta}_{1n} - \theta_{10}) = \frac{1}{[E(q_{11}^1)]^2}\left(E(q_1^1)^2 + \frac{[E(q_{12}^1)]^2[E(q_2^2)^2]^2}{[E(q_{22}^2)]^2}\right)$$

where $\widetilde{\theta}_{1n}$ is the second stage estimate of θ_{10}. The only case where those two asymptotic variances are same is when $E(q_{12}^1) = 0$, where one can estimate equation by equation without any interaction between θ_1 and θ_2 possibly through some control functions. Since our model generally contains the interaction terms, the presence of nonzero $E(q_{12}^1)$ should be understood as a main cause differentiating the asymptotic variances of single-step and two-

step methods. The magnitudes of Avar$\sqrt{n}(\widehat{\theta}_{1n} - \theta_{10})$ and Avar$\sqrt{n}(\widetilde{\theta}_{1n} - \theta_{10})$ cannot be determined in general. However, holding other terms fixed, higher value of $E(q_{12}^1)$ makes the asymptotic variance of single-step estimator decrease since its denominator and numerator are a polynomial of degree four and two respectively. Thus under a higher value of $E(q_{12}^1)$, the denominator would dominate the numerator yielding smaller asymptotic variance. On the other hand the asymptotic variance of two-step estimator would increase smoothly for higher value of $E(q_{12}^1)$. However, if $E(q_{12}^1)$ is close to zero, then the fraction can possibly be subject to an abrupt change even by a small change in $E(q_{12}^1)$ making comparison more difficult.

1.5 Simulation Design

Monte Carlo simulations were carried out in order to compare the finite sample performances of the four estimators under consideration. The data generating process is

$$
\begin{aligned}
x &\sim \text{uniform}[-10, 10] \\
y_1 &= 3 + \ \ 0.1x + u_0 \\
y_0 &= 2 + 0.15x + u_1 \\
w &= 1[1.4 - 0.05x - 0.3z + \xi \geq 0] \\
z &\sim \text{Binomial}(1, 1/2),
\end{aligned}
$$

where u_g and ξ can follow normal, centered $t(5)$ and centered $\chi^2(5)$ with variances all normalized as unity.[2] Their correlations are set 0.0, 0.4, 0.5 and 0.6. The ATE, the difference

[2] First and second moments alone do not uniquely determine the joint distribution of nonnormal random variables. The stata code for this simulation will be provided upon request.

of the intercepts for each regime, is equal to one. To see the effect of weak instrumental variable, a set of 10 randomly created binary variables was used. Because the nonlinear selection function reduces the information in the domain through the indicator function, the effect of weak IV is not very perceivable unless a large number of weak IVs is put in the equation. To measure the quality of IVs as a whole, the concentration parameter is used, i.e. $E(\delta' Q' Q \delta)/L$ for $y^* = 1.4 - .05x - .3z + \xi$ where the i-th row of Q is $(1, x_i, z_i)$ and L is the number of instrumental variables. Then for the observation number $n = 100$, the concentration parameter for y^* is 10.575 when a relevant IV is used, and 1.0575 when ten irrelevant IVs were added.

Vella and Verbeek (1999) already compared the performances of IV and CF under various error distributions, but what they essentially compared was the two-step CF methods with and without using the binary nature of treatment variable in the present context. And they just investigated what they called *restricted* CF which is the single-regime model such as equation (1.6). The contribution of this paper is that (1) both two-regime and one-regime models are considered, (2) both single-step and two-step estimation methods are investigated, and (3) the misspecifications of both selection and structural errors are allowed for.

1.6 Simulation Results

As it was mentioned in the previous sections, we focus on those four estimators, i.e. Heckit, QLIML, Series and QFIML, the estimable equations of which are equations (1.5), (1.8), (1.9) ad (1.11). In addition to those two-regime estimations, we also discuss one-regime procedures i.e. IV estimator and Heckit with single regime, which are (1.1) and (1.6). In order to distinguish Heckits in those two settings, it will be referred to as Heckit1 and Heckit2

for regime one and two respectively. Heckit1 is essentially same as Heckit2 except for the fact that the restriction that the partial effects of the covariates across two regimes are equal was used for Heckit1. Also the IV estimator using equation (1.1) will be called linear IV estimation or LIV lest one get confused with instrumental variable. Thus the difference between LIV and Heckit1 is that the selection equation for LIV is the linear probability model while for Heckit1 is probit. Of course we can think of the optimal IV estimation using probit, but it will not be considered here. LIV is two-step estimation method; when the same model is estimated by single-step LIML, then it will be written as LLIML in order to distinguish that from QLIML above. Thus the focus is on the comparison between linear and nonlinear control function methods i.e. LIV and Heckit1 for one-regime estimation. For two-regime estimation, the focus is on the comparisons among Heckit1, QLIML, Series, and QFIML. We discuss strong IV results from section 6.1 through 6.4. Sections 6.5 and 6.6 are devoted for weak IV.

1.6.1 Comparison between LIV and Heckit1

To begin with, let us investigate the behavior of LIV and Heckit1 with strong IV under various misspecifications. Although Vella and Verbeek (1999) dealt with this issue, the discussion here is more extensive. The results are summarized in Table B.1 through Table B.4. For each simulation session, the seven summary statistics are provided: Monte Carlo mean, standard deviation, root mean squared error (RMSE), and median. The simulation was run for different correlation values such as 0.0, 0.4, 0.5 and 0.6, which makes it easy to find out patterns of behavior of estimators if any. Since there are three possible error specifications for both structural and selection errors, we have nine combinations to consider. However, it turned out that the cases of heavier tails generated by $t(5)$ distribution is very similar to the

normal cases and their results are omitted in the tables[3]. The horizontal rows are for three possible selection error specifications and the vertical columns are for structural error.

Let us first compare LIV and Heckit1 in terms of bias. One can see that those one-regime models are not particularly bad despite the true data generating process has two regimes. The two estimators also converge well to the true parameter value, which is 1, under sufficiently large samples. Heckit1 is valid under three assumptions which were already made clear in Proposition 1.2 and Corollary 1.1. Although the exogenous variable z is independent in the simulation design, the other two assumptions are designed to violated except for in normal-normal case, where Heckit1 is truly legitimate. The conditional expectation will not be linear unless both of the errors follow same distribution. Particularly if one of them is skewed while the other one is symmetric, then the conditional expectation can be nonlinear. Given the independent z, the other two assumptions fail unless the error combination is normal-normal. If those assumptions are not satisfied, then the control functions in the form of inverse Mill's ratio will be misspecified ultimately causing an inconsistency. The results in Tables B.1-B.4 clearly show this: regardless of correlation, when the selection errors are skewed, the finite sample biases of Heckit1 are greater than those of LIV. Such large biases are found only when the selection errors are skewed. Moreover, they are persistent even when the sample size is sufficiently large. Also it shows that the results under misspecified structural error are not as sensitive as those under misspecified selection errors. Since large finite sample biases are common among the cases with skewed selection errors, it appears that it is not so much caused by nonlinearity of $E(u|\xi)$ as caused by nonnormality of ξ. Nevertheless the results are also affected by nonlinearity of $E(u|\xi)$. [4] One can also see the

[3] The results for $t(5)$ will be provided upon request.

[4] Even though I did not include the results in this article, I also created an error generating process where each of them is normal but the conditional expectation is not linear in order

behaviors of estimates under various correlation values; given symmetric selection errors, as the correlation becomes larger, the relative magnitudes of Heckit1 biases compared to those of LIV decrease. However, under skewed selection errors, the Heckit1 biases are larger than those of LIV regardless of correlations. To summarize, LIV has smaller bias under skewed selection errors while Heckit1 does under symmetric ones. LIV is robust under misspecification.

For efficiency, Tables B.1-B.4 show that Heckit1 has smaller Monte Carlo standard deviation than the LIV does, which is true irrespective of the error specifications. Also in terms of RMSE, except for some cases where the selection errors are skewed, the RMSE of Heckit1 is smaller than that of LIV overall. Therefore we have good reason to use nonlinear models such as Heckit1 unless a nonsymmetric selection error is suspected.

IV estimators do not necessarily have finite moments, which is reflected by the considerable magnitudes of standard errors under small sample sizes. Therefore, as suggested by Angrist et al. (1999), it would be very useful to see the median or MAD as well as those moment estimators such as mean and standard deviation. In terms of mean, Heckit1 has a smaller bias than LIV except for the cases under skewed selection errors. In terms of median, on the contrary, the tendency is just the opposite: one can see that in most cases the median biases of LIV are smaller than those of Heckit1, and, particularly, they are for all the cases under skewed selection errors. Nevertheless one can see Heckit1 is more efficient than LIV in terms of MAD, which is in agreement with the results for standard deviation.

to see the pure effect of nonlinearity. The results show that nonlinearity of $E(u|\xi)$ causes larger finite sample bias for Heckit1.

1.6.2 Comparison between Heckit2 and QLIML

Let us now compare the performances of Heckit2 and QLIML. Those two estimators use same estimating equation; the difference is that Heckit2 is estimated by two-step procedure while QLIML is by single-step. It is already known that in linear models with just identification 2SLS and LIML give numerically same estimates(Anderson *et al.*, 1982). If IV or 2SLS is understood as a special case of control function method, then the comparison between Heckit2 and QLIML is essentially a comparison between 2SLS and LIML with nonlinear selection function. However, as it was already discussed in previous sections, unlike linear models, when the selection equation is nonlinear, it is impossible to determine in general whether the asymptotic variance of one estimator is same or greater than the other. Therefore it is not very meaningful to discuss the efficiency simply by comparing the Monte Carlo standard deviations of two estimators under some sample sizes; it would be very likely to see the similar tendency also in finite samples if the asymptotic variance of one estimator is substantially larger or smaller than the other. Rather it would be more interesting to find out how the relative efficiencies, i.e. the ratio of MSE's, behave under different sample sizes.

For consistency of those two estimators, it is essential that the conditional expectation of y on x and w be correctly specified(Gourieroux *et al.*, 1984). It is basically whether the conditional expectation on x, w can be correctly specified by the inverse Mill's ratio once it is taken for granted that the conditional expectation on x is linear. It is again based on the assumption that the selection equation is probit. If true model is not probit, then both Heckit2 and LIML might be inconsistent. Although it can be well expected that the two estimators might be inconsistent under skewed selection errors, it turned out that they converge well to the true parameter value even under such misspecifications. Therefore

misspecification of selection error does not seem to be very critical.

The simulation results for Heckit2 and QLIML are also in Tables B.1-B.4 and Figure D.1 displays the cumulative density functions only for the case of $\rho = 0.6$ and $n = 1000$. According to the results, there are some cases where the Monte Carlo means are not converging to the true value monotonically. It can be explained by two ways: First, theoretically, in the definitions of any kinds of convergence including the convergence in probability, the behaviors of early terms of sequence are not very important; also the monotonicity is never considered in weak law of large numbers or in uniform weak law of large numbers. The second reason is practically more relevant. Since IV estimators do not necessarily have finite moments, the IV estimators can produce many estimates of large magnitude on which the numerical values of mean or standard deviation can be very sensitive. Thus it should not be a major concern even though the sequence of mean is not monotonically converging to the true value.

Let us now discuss the efficiencies of those two estimators. It can be seen that the Monte Carlo standard deviation of QLIML is very large compared to Heckit2, which is more so under small sample sizes. In addition to that, although QLIML is very inefficient under small sample sizes, it nearly catches up with Heckit2 very quickly. Table C.1 shows the relative efficiencies of some selected pair of estimators including QLIML with respect to Heckit2. The relative efficiencies were computed as the MSE of each relevent estimators divided by the MSE of Heckit1 or Heckit2. Thus if the value is greater than one, then it implies that the MSE of Heckit1 or Heckit2 is relatively smaller. It can be seen from Table C.1 that the relative efficiencies are decreasing as the sample sizes increase. In other words, the single-step QLIML is less efficient particularly under small sample sizes.

On the contrary, the estimated asymptotic variances for actual estimation give mislead-

ing information: a lot of cases the estimated asymptotic variances of single-step QLIML are smaller than those of two-step Heckit2. Table C.2 tabulates the ratio of estimations with higher estimated asymptotic variances for two-step Heckit2 out of total 1000 repetitions.[5] According to Table C.2, roughly 50-60% of total estimations produce misleading information about the true sampling distribution; overall the single-step QLIML underestimates its asymptotic variance

In so far as the median bias of Heckit2 and quasi-LIML, it is hard to find any evidence in favor of either estimator; the magnitudes of their median biases are more or less the same. Like the standard deviation, MAD also shows that QLIML is more dispersed than Heckit2.

1.6.3 Comparison between Heckit2 and Series estimator.

A series estimator is expected to solve the following three major problems: allowing for nonlinear $E(u|\xi)$, consistent estimation of index under nonprobit settings, and thus correction of the misspecified control function.

There are various forms of series estimator, but here the three leading ones suggested by Newey (2009) were used. According to the results presented in Tables A.1-A.12, one can see that the ones with inverse Mill's ratio have the smallest MSE. As for the degree of power series of correction terms, the estimator that includes only the terms of degree one is the best in terms of MSE; it is because larger number of control function terms interacted by w creates severer multicollinearity. Therefore the only difference between best Series estimator and Heckit2 is that the former uses Klein-Spady semiparametric estimator instead of probit in the first stage. They are similar in that they use inverse Mill's ratio of degree

[5] This simulation was designed to generate same pseudo-random numbers across different estimators enabling a direct comparison among them.

one. According to Proposition 1.2, under the linearity of $E(u|\xi)$, the control or correction terms can be expressed as some function of index with single term. Corollary 1.1 implies that normality of ξ lets us to write the correction terms as inverse Mill's ratios. Without the linearity of $E(u|\xi)$, the correction terms should be expressed not as a single term, but as a power series of some functions of index. However, Tables A show that such methods create multicollinearity due to the interaction with w, implying that the series estimators are not good alternatives for the cases where $E(u|\xi)$ is nonlinear. Unfortunately, among the three motivations of using series estimators, the only one that the Newey type estimators can address is the second one, that is the consistent estimation of index.

The Series estimator is expected to do better under the skewed selection error, but there is no particular evidence that the Series estimator gives smaller mean or median biases than Heckit2 does; in summary, the KS semiparametric estimator combined with inverse Mill's ratio does not do better than the usual Heckit2 does in terms of bias.

There is a tendency that the MSE of series estimator is greater than that of Heckit2. However, under small sample sizes, there are some cases where the MSE of series estimator is smaller than the Heckit2. Such phenomena are more often found as the correlation becomes higher.

1.6.4 QFIML

Heckit2, LIML and Series estimators are easy to compare with each other since they all use or based on same estimating equation. On the contrary QFIML is different in that the structural and selection errors are directly modeled without resorting to control function method. QFIML uses trivariate normal assumption; therefore if the specification is true then its asymptotic variance becomes simpler by information equality and QFIML becomes

FIML. On the other hand QFIML fails to be a consistent estimator if the error distribution is misspecified. Tables B.1-B.4 show this point: when the structural error is nonnormal, then the sequence of Monte Carlo mean is passing through the true parameter value and converging to somewhere else. Under normal structural error, QFIML converges to the true value relatively well. In other words although a misspecified selection error does not have strong effect, a misspecified structural error does cause inconsistency. One interesting thing is that the inconsistency is not so serious under zero correlation.

For efficiency, QFIML has the smallest variance among those four estimators. Such fact is also reflected and even exaggerated in the asymptotic variance approximations; the estimated asymptotic variances are so small that the 95% coverage rates are abnormally smaller than other estimators. The coverage rate converges to zero as the sample size grows bigger. This is also true even when the errors are correctly specified. Therefore QFIML estimator is not a desirable choice for test unless bootstrap standard error is used. To summarize QFIML fails to be consistent under misspecified structural errors, and has extremely small coverage rates.

1.6.5 Comparison between LIV and Heckit1 under weak IV

All the results for weak IV are presented in Tables B.5-B.8. The cumulative distribution functions for the case of $\rho = 0.6$ with $n = 1000$ are shown in Figure D.2. From the table, it can be seen that the behaviors of LIV and Heckit1 are not very different from those under strong IV except for in terms of bias. Under both strong and weak IV, if the selection error is skewed, then the linear model has smaller bias. However, under weak IV, even if the selection error is symmetric, then the nonlinear model does not have any noticeable advantage over linear ones as it used to have under strong IV. Thus the linear model is more robust under

weak IV case.

Now let us discuss the results on RMSE. First, under symmetric selection error, the nonlinear model still has advantage over the linear model in terms of RMSE. It is due to the fact that the nonlinear model is more efficient than the linear one in terms of standard deviation. Second, on the other hand, under skewed selection error, the efficiency of nonlinear model is not strong enough to make the nonlinear model more advantageous than the linear one in terms of RMSE.

1.6.6 Comparison between LIV and L-LIML and between Heckit2 and QLIML under weak IV

As it was already discussed, we have LIV and L-LIML for linear models and as their nonlinear counterpart Heckit2 and QLIML. Unless the correlation is zero, under all circumstances the median biases of L-LIML are smaller than those of LIV regardless of sample size and error specification. On the other hand, the relation of the median biases of Heckit2 and QLIML are not as simple as in linear models. Particularly, when $\rho = 0.4$, the median biases of QLIML are greater than those of Heckit2 under all circumstances and it is still true in many cases when $\rho = 0.5$. Nevertheless it shows a tendency of smaller QLIML median bias than Heckit2 as the correlation becomes larger.

1.7 Conclusion

In finite samples, especially when the sample size is small, QLIML is relatively less efficient than the two-step Heckit2 suggesting that there is no efficiency gain by running single-step procedure which mimics two-step. QFIML is more efficient than Heckit2; however, it is not

robust to distributional misspecification. In terms of efficiency, Heckit2 is a middle ground between those two single-step estimators. Nevertheless Heckit2 is more preferable since it is not only easy to compute but also robust to arbitrary misspecifications. Newey-type Series estimators were expected to do relatively better under misspecification, but turned out not to be good alternatives.

Under weak many instruments, QLIML shows better performance in terms of median bias only when the degree of endogeneity is strong. Under weak endogeneity, there is no clear evidence of such an advantage.

Chapter 2

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

2.1 Introduction

In order to estimate the treatment effects of binary variable on count dependent variable, Terza (1998, 2009) proposed nonlinear models that take into account the limited dependent variables. 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's (1978) 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 the 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. Section 5 is devoted to Monte Carlo simulation and discusses the results and implications. In Section 6, the proposed approach is applied to a real data set to estimate ATE and Section 7 presents the concluding remarks.

2.2 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.

2.2.1 Nonlinear Models

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

$$\begin{aligned} E[y|x, w, \epsilon] &= \exp(\alpha + \mathbf{x}\beta + \gamma w + \epsilon) \\ w &= 1[\mathbf{z}\delta + v > 0], \end{aligned}$$

where $\mathbf{x}$ is $1 \times K$ vector of covariates, w is binary treatment variable and ϵ is unobserved heterogeneity. The vector of covariates $\mathbf{x}$ and the vector of exogenous variables $\mathbf{z}$ are all assumed to be independent with the structural and selection errors. Usually $\mathbf{x}$ is the subset of $\mathbf{z}$. 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 $\mathbf{z}$ 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.

$$\begin{aligned} E[y_g|\mathbf{x}, w, \epsilon_1, \epsilon_0] &= E[y_g|\mathbf{x}, \epsilon_g] = \exp(\alpha_g + \mathbf{x}\beta_g + \epsilon_g), \qquad g = 0, 1 \\ w &= 1[\mathbf{z}\delta + v > 0], \end{aligned} \tag{2.1}$$

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 us 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 (2.1), 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 (2.1) 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 $\mathbf{z}$. 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 implication is that it can be viewed as a structural equation[1]. 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

[1] In this respect the one step estimator that simultaneously maximizes the outcome as well as treatment equation is called Full Information Maximum Likelihood estimator.

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).

$$\begin{aligned} ATE &= E[y_1 - y_0] = E\Big(E[y_1|x] - E[y_0|x]\Big) \\ &= E\Big[\exp(\alpha_1 + \sigma_1^2/2 + x\beta_1) - \exp(\alpha_0 + \sigma_0^2/2 + x\beta_0)\Big] \end{aligned} \tag{2.2}$$

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

$$\widehat{ATE} = N^{-1}\sum_{i=1}^{N}\Big[\exp(\alpha_1\widehat{+\sigma_1^2}/2 + x\widehat{\beta}_1) - \exp(\alpha_0\widehat{+\sigma_0^2}/2 + x\widehat{\beta}_0)\Big].$$

As it will be shown in equation (2.4), the composite intercepts $\alpha_g + \sigma_g^2/2$ are identified

without separately identifying α_g and σ_g^2. The term $\widehat{\alpha_g + \sigma_g^2/2}$ are the estimates of the composite intercepts. Incidentally, the Average Treatment Effects on the Treated (See p. 906, Wooldridge, 2010) is computed by

$$\widehat{ATT} = \left(\sum_{i=1}^{N} w_i \right)^{-1} \sum_{i=1}^{N} w_i \Big[\exp(\widehat{\alpha_1 + \sigma_1^2}/2 + x\widehat{\beta}_1) - \exp(\widehat{\alpha_0 + \sigma_0^2}/2 + x\widehat{\beta}_0) \Big].$$

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

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

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 (2.2).

2.2.2 Linear Models

Angrist (2001, 2010) and Angrist and Pischke (2009) have pointed out that in many cases a linear model may be 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.

$$\begin{aligned} y_g &= \mu_g + x\beta_g + u_g, \qquad g = 0, 1 \\ w &= 1[z\delta + v > 0] \end{aligned} \tag{2.3}$$

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 outcome equations, i.e. the equations of which dependent variable is y_g, 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. Then the ATE is the difference between the two intercepts for each regime. The problem is that

there is no known identifying strategy of those intercepts when the true model is exponential. For example, one can try using the Heckman correction method (Heckman, 1978) for the outcome equation. However, since the minimum condition is that $E[u|\mathbf{x}, v] = E[u|v] = \rho v$ (Olsen, 1980) and the model does not satisfy the first equality under the exponential conditional mean, the LFES estimator with Heckman correction does not identify the true ATE. Although u and $\mathbf{x}$ are orthogonal by linear projection, they are not mean independent under the exponential conditional mean assumption.

Another alternative linear approach is the 2SLS that does not explicitly model the two distinct regimes. As it was already shown by Angrist et al. (1996), the coefficient on w in the regression without any other covariates will identify the Local Average Treatment Effect (LATE). With covariates, the coefficient on w identifies the weighted averages of LATE for each covariate cell (Hirano et al., 2000; Mealli et al., 2004). Although LATE will generally be different from ATE, Angrist and Pischke (2009) support its usefulness on the grounds that it does not require any distributional assumptions and its estimates end up being very similar to ATE estimated by nonlinear modeling. In later sections, we will see the validity of this claim through simulations.

Instead of applying the 2SLS method, one can make use of the fact that the endogenous treatment variable is binary. Thus the probit model can be used for the first stage. Let us call this the "Linear Endogenous Treatment (LET)" model and it is obvious that it is a special case of LFES model with the restriction that $\beta_1 = \beta_0$ and $u_1 = u_0$ in equation (2.3).

2.3 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.

2.3.1 Full Information Maximum Likelihood

When a binary choice model is used for regime switching, the treatment equation can be regarded as another structural equation. Specifically we assume that the treatment variable follows probit model and the outcome variables do Poisson distribution. Let the one-step ML estimation method under these assumptions be called Full Information Maximum Likelihood (FIML).

One of the advantages of FIML is that it is the most efficient estimator achieving the Cramér-Rao bound with correctly specified likelihood function. Also unlike other estimators that will be discussed below, it estimates all the parameters separately in a single step. (In other estimators the structural intercept and covariance between ϵ and v are not separately identified or, even if they are, it requires further steps to do that.) Despite those advantages it is still computationally burdensome; it might not numerically converge to any meaningful solution depending on specific data set, and, if ever, it usually takes a lot of time to make the bootstrapping very awkward. In the case when it needs numerical integration by using Gauss-Hermite quadrature, the error from approximation can be substantial depending on the population. The likelihood of FIML estimator is as below. The assumption used is again that ϵ and v follow trivariate normal.

Proposition 2.1. The joint density function of y and w conditional on the exogenous

variables is

$$f(y,w|\mathbf{z}) = \left[\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\sqrt{2}\sigma_1\zeta_1)\Phi^*(\sqrt{2}\sigma_1\zeta_1)\ \exp(-\zeta_1^2)d\zeta_1\right]^w \cdot \left[\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\sqrt{2}\sigma_0\zeta_0)\Big(1-\Phi^*(\sqrt{2}\sigma_0\zeta_0)\Big)\exp(-\zeta_0^2)d\zeta_0\right]^{1-w},$$

where

$$\Phi^*(\sqrt{2}\sigma_g\zeta_g) = \Phi^*(\epsilon) = F\left[\frac{\mathbf{z}\gamma + (\sigma_{ga}/\sigma_g^2)\epsilon}{\sqrt{1-(\sigma_{ga}/\sigma_g)^2}}\right]$$

Proof The basic proof was given in Terza (1998) and its extention to the NFES model is given in Appendix A. ■

The function $f(\cdot)$ denotes the used conditional distribution; it is usually either Poisson or NegBinII. The above equation is a function of ζ that is being integrated out; the integration will be computed by Gauss-Hermite quadrature method in actual estimation.

2.3.2 Quai-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 (2.1), a natural way to estimate might be running QMLE

or Nonlinear Least Squares (NLS) by using the $E(y_g|\mathbf{x}, \epsilon_g)$. 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 $E(y|\mathbf{z}, v)$.

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

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

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

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

$$\begin{aligned} E(y|\mathbf{z}, w) &= w \cdot \left[\exp\left(\alpha_1 + \frac{\sigma_1^2}{2} + \mathbf{x}\beta_1\right)\frac{\Phi(\mathbf{z}\delta + \rho_1\sigma_1)}{\Phi(\mathbf{z}\delta)}\right] \\ &\quad + (1-w) \cdot \left[\exp\left(\alpha_0 + \frac{\sigma_0^2}{2} + \mathbf{x}\beta_0\right)\frac{\Phi(-(\mathbf{z}\delta + \rho_0\sigma_0))}{\Phi(-\mathbf{z}\delta)}\right], \end{aligned} \tag{2.4}$$

where the composite intercepts and β_g are identified. As was already seen in equation (2.2) these parameters are sufficient for identifying ATE. A 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 $E(y_g|\mathbf{z}, \epsilon_g)$ as CEF, whereas QMLE does it with $E(y|\mathbf{z}, w)$. The integration does not appear in Poisson likelihood based on $E(y|\mathbf{z}, w)$

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., $E(y|\mathbf{z}, w)$, is expressed by all observable variables that makes the QMLE likelihood simpler than FIML. One can run a QMLE by using a conditional distribution with the mean $E(y|\mathbf{z}, w)$. 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 (Wooldridge, 2011). However, as Hellström and Nordström (2008) and Chapter 1 have shown, 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.

2.3.3 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 us write the equation in additive form with the correctly specified conditional expectation as

$$y = E[y|\mathbf{z}, w] + e,$$

where $E(e|\mathbf{z}, w) = 0$ by definition. The true parameter values are such that it minimizes $E[y - E(y|\mathbf{z}, w)]^2$ and the estimation can be performed by sample analogue. Since the conditional mean contains the correction terms, it should be estimated through the first stage probit. Under some regularity conditions this approach gives a consistent estimator. See Wooldridge (2010, Chapter 12) for detailed discussion.

On the other hand this NLS estimator can also be interpreted as method of moment estimator. From the minimization problem the true parameters satisfy the first order condition $E(dE(y|\mathbf{z}, w)/d\theta \times e) = 0$ which can be a moment condition with the instrument vector $dE(y|\mathbf{z}, w)/d\theta$. By the fact that $E(e|\mathbf{z}, w) = 0$ along with law of iterated expectation, it can be easily shown that any function of instruments can also be a good instrument. Therefore we have infinitely many possible instruments that can serve to increase the asymptotic efficiency of parameter estimators.

2.3.4 Weighted Nonlinear Least Squares Estimator

We have seen above that NLS with correctly specified conditional mean gives consistent estimator under some regularity conditions. Also it became clear that the NLS estimator can also be viewed as a method of moment estimator, from which one can construct infinitely many valid instruments and moment conditions. In order to avoid poor finite sample performances due to overidentifying restrictions (Altonji and Segal, 1996; Ziliak, 1997), one can resort to the optimal IV approach (See Wooldridge (2010, Chapter 8)). This can be done in this case by dividing the instrument by conditional variance; the estimator using $E(dE(y|\mathbf{z}, w)/d\theta \times var[y|\mathbf{z}, w]^{-1} \times e) = 0$ as moment condition is more efficient than NLS in the previous section.

One can also draw almost same conclusion by using the weighted nonlinear least squares

(WNLS) approach, in which both sides of the error form equation is divided by the square root of an arbitrary function $v(\mathbf{z}, w, \gamma)$ with nuisance parameter γ. Then the first order condition becomes $E(dE(y|z,w)/d\theta \times v(\mathbf{z}, w, \gamma)^{-1} \times e) = 0$. This WNLS estimator is consistent under correctly specified conditional mean and identification condition; see Wooldridge (2010, p. 411). Note that the difference between moment condition for optimal IV approach and first order condition in WNLS is whether the original instrument is divided by $var[y|\mathbf{z}, w]$ or by an arbitrary function $v(\mathbf{z}, w, \gamma)$. Under the condition $\sigma^2 v(\mathbf{z}, w, \gamma) = var[y|\mathbf{z}, w]$ for some constant σ^2, the generalized information matrix equality (GIME) holds and the WNLS estimator becomes efficient; if $\sigma^2 = 1$ then this essentially restates the efficiency result from optimal IV. What if $\sigma^2 v(\mathbf{z}, w, \gamma) \neq var[y|z, w]$? Then the inference has to be made robust due to the failure of GIME. Having said that, the consistency result still holds under the assumption of identification condition and correctly specified conditional mean. Under the assumption that $y|\mathbf{z}, w, \epsilon$ follows Poisson distribution, Terza (1998) found correctly specified conditional variance $var[y|\mathbf{z}, w]$. If should also be noted that if the Poisson assumption fails, then the conditional variance will be misspecified and robust inference is called for.

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 will be used in this paper since it is computationally easier to implement. The derivation for NFES model under the assumption that $y|\mathbf{z}, w, \epsilon$ follows Poisson is given in

the Appendix C that turns out to be

$$\text{var}[y|\mathbf{z},w] = w\delta_1\Big(\delta_1(\exp(\sigma_1^2)L_{1,2}-L_1^2)+L_1\Big)+(1-w)\delta_0\Big(\delta_0(\exp(\sigma_0^2)L_{0,2}-L_0^2)+L_0\Big), \quad (2.5)$$

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. If the Poisson assumption is true, then the GIME is applicable. Otherwise we need robust inference. In any case, the parameter estimators are consistent under the regularity conditions given above.

2.4 Asymptotic Distributions

The asymptotic distribution of FIML estimator is straightforward. Given the likelihood function in proposition 2.1, 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 $(y-E[y|\mathbf{z},w])^2/2\cdot var[y|\mathbf{z},w]$, where $E[y|\mathbf{z},w]$ and $var[y|\mathbf{z},w]$ are from equation (2.4) and (2.5). Ignoring the first stage error, the asymptotic distribution can be written under the condition $var(y|\mathbf{z},w) = v(\mathbf{z},w,\gamma)$ as (See Wooldridge, 1997)

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow[d]{} \text{N}\Big(0, \quad [E\big(h(\mathbf{z},w,y,\theta)\big)]^{-1}\Big),$$

where

$$E[h(\mathbf{z}, w, y, \theta)] = E\left[\frac{\nabla_\theta m(\mathbf{z}, w, \theta)\nabla_\theta m(\mathbf{z}, w, \theta)'}{v(\mathbf{z}, w, \gamma)}\right], \tag{2.6}$$

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

Now consider the asymptotic distribution of PQMLE. The likelihood function is constructed using the Poisson distribution with the conditional mean in equation (2.4). Then the asymptotic distribution is

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

where

$$\begin{aligned} E[s(y|\mathbf{z}, w, \theta_0)s(y|\mathbf{z}, w, \theta_0)'] &= E\left[\frac{\nabla_\theta m(\mathbf{z}, w, \theta)(y_i - m(\mathbf{z}, w, \theta))}{\mathrm{qvar}(y_i)}\right. \\ &\qquad \left. \times \frac{(y_i - m(\mathbf{z}, w, \theta))\nabla_\theta m(\mathbf{z}, w, \theta)'}{\mathrm{qvar}(y_i)}\right] \\ E[h(y|\mathbf{z}, w, \theta_0)] &= -E\left[\frac{\nabla_\theta m(\mathbf{z}, w, \theta)\nabla_\theta m(\mathbf{z}, w, \theta)'}{\mathrm{qvar}(y_i)}\right] \end{aligned}$$

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

$$Var[y|\mathbf{z}, w] = \sigma^2 \cdot \mathrm{qvar}, \tag{2.7}$$

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

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow[d]{} \mathrm{N}\Big(0, \quad -\sigma^2[E\big(h(y|\mathbf{z}, w, \theta)\big)]^{-1}\Big).$$

By plugging (2.7) in (2.6), it is obvious that the two asymptotic distribution for WNLS and PQML are equivalent. Having said that, without the condition (2.7), 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 (2.7). Although the dependent variable y_g conditional on $\mathbf{x}$ and ϵ_g follows the Poisson distribution with the mean $E(y_g|\mathbf{x}, \epsilon_g) = \exp(\alpha_g + \mathbf{x}\beta_g + \epsilon_g)$, it does not necessarily mean that y_g conditional on $\mathbf{x}$ and w follows Poisson distribution with the mean $E(y_g|\mathbf{z}, w) = \exp(\alpha_g + \sigma_g^2/2 + \mathbf{x}\beta_g)\Phi(f(\mathbf{z}\delta))/\Phi(\mathbf{z}\delta)$. To see this point, mean and variance conditional on $\mathbf{z}, w$ are

$$\begin{aligned} E[y|\mathbf{z}, w] &= w\delta_1 L_1 + (1-w)\delta_0 L_0 \\ \mathrm{var}[y|\mathbf{z}, w] &= w\delta_1\Big(\delta_1(\exp(\sigma_1^2)L_{1,2} - L_1^2) + L_1\Big) + (1-w)\delta_0\Big(\delta_0(\exp(\sigma_0^2)L_{0,2} - L_0^2) + L_0\Big) \end{aligned}$$

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 (2.2). Terza (2009) derived the following result.

Proposition 2.2. Along with the regularity conditions for Uniform Weak Law of Large Numbers and asymptotic normality, suppose that the nonlinear functions $g_1(x,\theta)$ and $g_0(x,\theta)$ are continuous and differentiable at θ_0 and that their first derivative with respect to θ satisfies all the conditions in Uniform Weak Law of Large Numbers for objective function. Let the expectation of their derivatives with respect to θ be denoted by G_1 and G_0. Then the asymptotic distribution of ATE estimator in equation (2.2) is

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

where

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

and

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

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

Proof See Terza (2009). ■

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. 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, Chapter 12).

Under this setting if the two exponential terms, i.e. $g_1(x, \theta_0)$ and $g_1(x, \theta_1)$, are substantially different along the values of covariates x, then the variance of conditional ATE will also be large. Therefore one can have a large variance of ATE estimator to the extent that the values of β_1 and β_0 are different. A lesson from this argument is that ATE estimator by NFES might become less accurate when there are substantial inequality of the treatment effect from person to person. On the other hand, there also exist the set of parameters with which the variance of ATE estimator by NFES can be arbitrarily small. The extreme case is when the conditional mean for each regime are identical, where the variance of T as well as the covariance term becomes zero.

The above discussion shows that the consistent NFES method can have larger or smaller variance depending on the coefficient values. As the conditional mean of two regimes are similar, then one can expect that ATE estimator by NFES might be very efficient, but when they are very much different, then it might have larger variance.

2.5 Monte Carlo Simulation

As it was already discussed in earlier sections, the NFES and LFES models are thought to be consistent for ATE. Therefore the purpose of Monte Carlo simulation is to compare the linear and nonlinear models for estimating ATE. Incidentally other alternative estimators

including LET, estimated by 2SLS and Heckit, and NET model that has acquired wide popularity needs to be examined in terms of consistency.

Specifically the main objectives of simulation study are: (1) to present the possibility that NFES estimators can be more efficient than LFES under small variance of T, (2) to find out how severe the bias in one-regime model is compared to that in two-regime model, and (3) to check how robust the NFES model is to various distributional misspecifications. In addition to that, it might also be interesting (4) to clarify the advantage of one-step estimations over two-step ones in nonlinear models, (5) to compare Poisson QMLE and WNLS in terms of efficiency, and (6) to find out the advantage of FIML estimator.

2.5.1 Data Generating Processes

For each simulation session, the number of replication is 1000, and the sample sizes are 1000, 3000 and 5000. There will be five Data Generating Processes (DGP). For all the DGPs the following setup will commonly be used.

$$\begin{aligned} x &\sim \text{uniform}[-5,5] \\ z &\sim \text{Binomial}(1,1/2) \\ w &= 1[.15 - 0.05x - 0.3z + v \geq 0] \\ E(y_1|x,\epsilon_1) &= \exp(\beta_{10} + \beta_{11}x + \epsilon_1) \\ E(y_0|x,\epsilon_0) &= \exp(0.1 + \beta_{01}x + \epsilon_0), \end{aligned}$$

where the errors follow trivariate normal distribution. The treatment equation is designed so that the numbers of observation for each regime are approximately same. The population

ATE is set equal to one. For the errors, the covariance between ϵ_1 and ϵ_0 was set equal to 0.5 and the variances of all the errors are set equal to one. What is important is the covariances between ϵ_g and v, for which 0.4, 0.5 and 0.6 were used. Below are the descriptions of data generating processes used for each session.

DGP 0 : $\beta_{10} = 0.564, \quad \beta_{11} = 0.01, \quad \beta_{01} = 0.1, \quad y_g \sim \text{Poisson}(E[y_g|x, \epsilon_g])$

$\epsilon_g, v \sim$ trivariate normal

This is the ideal case for the NFES estimators. Since the data generating process is nonlinear, the value of population ATE cannot be computed in closed form; rather the claimed true value is computed numerically. Using Stata®, I tried to find the value of the intercept in regime one such that the difference between one and the numerically computed $\sum(y_1 - y_0)/N$ is less than 0.001 with a million observation from the above data generating process. The value of intercept in regime one, i.e. 0.564 was found in this way. The dependent variables y_1 and y_0 follow Poisson distribution for given conditional mean. The other DGP's for misspecified distribution will be slight modifications of this basic [DGP 0].

There are three distributional assumptions in the above data generating process: 1) the dependent variable follows Poisson, 2)the unobservables in conditional mean function follow the normal distribution, and 3)the selection error follows standard normal, i.e. the treatment is probit. It might not be very surprising even if the NFES estimator performs well under [DGP 0] on which the model is based. To be fully reliable and useful the nonlinear models have to show their validity under misspecified distributions also. This is particularly important because it has been a major source of critics in favor of simpler linear models that the nonlinear models rely on strong distributional assumption (Angrist and Pischke, 2010). There have been some papers addressing this issue: Romeu and Vera-Hernández (2005)

use flexible functional form for the conditional probability of counts by using polynomial Poisson expansions. Masuhara (2007) models the joint distribution of errors by Hermite polynomials. Choi and Min (2009) use Johnson's S_U-normal distribution that is shown to outperform the normal model in terms of consistency. And Deb and Trivedi (2004) use latent factor structure and simulated likelihood methods to deal with nonnormality. Although those methods provide very nice alternatives to the models based on normal assumption, it will be shown below that the nonlinear model with normality still in many cases performs better than the linear IV methods.

To check the robustness of NFES, I designed some other data generating processes such that the above distributional assumptions are violated one by one.

$$\textbf{DGP 1}: \beta_{10} = 0.564, \quad \beta_{11} = 0.01, \quad \beta_{01} = 0.1, \quad y_g \sim \text{uniform}[0, 2E[y_g|x, \epsilon_g]]$$
$$\epsilon_g, v \sim \text{trivariate normal}$$

In [DGP 1], discrete uniform distribution was employed in place of the Poisson. It is set such that the left and right end points of the support are zero and two times the conditional mean.

$$\textbf{DGP 2}: \beta_{10} = 0.564, \quad \beta_{11} = 0.01, \quad \beta_{01} = 0.1, \quad y_g \sim \text{Poisson}(E[y_g|x, \epsilon_g])$$
$$\epsilon_g \sim \text{bivariate normal}, \quad v \sim (\chi^2(5) - 5)/\sqrt{10}$$

In [DGP 2], for the standard normal selection error, a skewed distribution $(\chi^2(5) - 5)/\sqrt{10}$ was used[2]. The mean and standard deviation are the same as the standard normal. This

[2] A correlated trivariate distribution in which two of the errors follow standard normal and the other follows centered χ^2 distribution can be properly modeled by copula density functions. Nevertheless the purpose of present simulation is not finding a multivariate distribution with parameters that uniquely determines a specific multivariate distribution. In DGP used above, each random numbers were generated first by creating many standard normal random numbers as "basis" variables. The dependence structure is then created

probit assumption was important since it was this very assumption that made it possible to write the correction terms as the fractions of two normal cumulative density function. The conditional mean $E[y|x,z]$ is misspecified without this.

DGP 3 : $\beta_{10} = 0.516, \quad \beta_{11} = 0.01, \quad \beta_{01} = 0.1, \quad y_g \sim \text{Poisson}(E[y_g|x,\epsilon_g])$

$$\epsilon_g \sim (\chi^2(20) - 20)/\sqrt{40}, \quad v \sim \text{normal}$$

In [DGP 3], the case where the structural errors follow $(\chi^2(20)-20)/\sqrt{40}$ instead of standard normal is considered. When the structural errors are not normally distributed as in [DGP 0], it directly changes the distribution of dependent variables and thus the population ATE. To ensure that the ATE is equal to one, the constant term in regime 1 is changed to 0.516. Normality of ϵ_g is used in the process of getting rid of ϵ_g from the conditioning set. Thus a violation of this assumption causes the composite intercept misspecified. Only under this assumption, we could obtain a nice result that the composite intercepts in estimating equation $E[y_g|\mathbf{x},w]$ and that in $E[y_g|\mathbf{x}]$ are identical. Therefore a violation of normality will make ATE estimator inconsistent. The degrees of freedom in chi-square were chosen so that the comparison between various estimators easy and meaningful to the extent that they are shown to be consistent converging to the true value. It was found that the results for structural errors were more sensitive to distributional misspecification than it is for selection error. Also when the degree of freedom is as high as that used in selection, a lot of cases the PQML estimator fails to converge to give nice estimates. One can see from the results below that this choice of degrees of freedom is indeed good for comparison purpose without too many extreme outcomes.

by using some common "basis" such that the linear combination gives intended correlation structure. The errors in [DGP2] and [DGP3] are generated in this way. Stata codes will be provided upon request.

In [DGP 0], the variance of individual treatment effect, i.e. $T \equiv g_1(x,\theta_0) - g_0(x,\theta_0) - \Big(E[g_1(x,\theta_0)] - E[g_0(x,\theta_0)]\Big)$ was approximately 18.9. The asymptotic distribution in earlier section says that the variance of ATE estimator is an increasing function of $E[T^2]$. [DGP 4] is designed to show this.

DGP 4 : $\beta_{10} = 0.865, \quad \beta_{11} = 0.001, \quad \beta_{01} = 0.35, \quad y_g \sim \text{Poisson}(E[y_g|x,\epsilon_g])$

$\epsilon_g, v \sim$ trivariate normal

In the above process, $E[T^2]$ is approximately 52.5. In order to create higher variance of individual treatment, the vertical distance between two conditional mean function at a fixed x has been adjusted without changing the true ATE value.

2.5.2 Main Results

The results for [DGP 0] is tabulated for all the discussed estimators except for FIML in Tables E.1-E.3 for each correlation. The FIML results are separately given in Table E.4. For other DGPs, only 2SLS, LFES (Heckit), NFES (2PQMLE) and NFES (WNLS), which are the focus of our discussion, are tabulated in Tables E.5-E.7.

First, compare LFES with NFES under various estimation methods with small T. Tables E.1-E.3 present the [DGP 0] results with $E[T^2] = 18.9$ that are relevent for this purpose. For any correlation, it can be seen that NFES are extremely inefficient and their RMSEs are very large compared to LFES under the sample size 1000. This is predicted by the asymptotic distribution; the variance of T in nonlinear estimator is still prevalent in small sample. However, such constant term rapidly disappears as the sample size increases. Also the results in terms of median and mean absolute deviation (MAD) is not too bad; the medians of NFES are closer to one than those of linear models. In addition to that, although the Monte Carlo

standard deviations of NFES were very large, the degrees of dispersion measured by MAD are smaller than those of linear models, from which one can conclude that the large amount of RMSE was caused by some outlying estimates with extreme values. Therefore the measures that are less affected by the extreme values give more favorable figures for NFES. At any rate, such extreme estimates disappear as the sample sizes grow larger. With $n = 5000$, the mean approaches to the true parameter value sufficiently close, and their Monte Carlo standard deviations are smaller than those of linear models let alone median and MAD. In sum, ruling out seemingly absurd estimates, when the variance of T is small, NFES model does better than the linear models in terms of both consistency and efficiency particularly as the sample size grows larger. On the contrary, the results under $E[T^2] = 52.5$ in Table E.8 looks very different. Even under large sample size, NFES estimators are less efficient than LFES. As seen in Proposition 2.2, the asymptotic variance of NFES estimators can be arbitrarily large as the variance of T diverges to infinity. However, the results in Tables E.1-E.3 imply that there can be the cases where NFES becomes more efficient than linear estimators when the individual treatment effect is not substantially different from person to person.

Incidentally it has been claimed that LATE and ATE obtained by linear IV and nonlinear methods respectively might not be substantially different (Angrist and Evans, 1998; Angrist, 2001; Angrist and Pischke, 2009). Nevertheless the simulation results indicate that their finite sample distributions and the estimates can be very different with each other. This issue will also be revisited in next section.

Second, we have seen in the above that the NFES estimator is a very useful alternative to LFES when the variance of T is small. However, it must be shown further that the performances of NFES estimators are not very sensitive to the assumptions on which it is based.

If the said nice properties of NFES are valid only under the ideal assumptions, its usefulness will be much limited. The results of additional simulations under wrong distributions mentioned in previous subsection is tabulated in Tables E.5-E.7. In those tables, only PQML, WNLS, and linear estimators that are still thought to be consistent were considered.

To begin with, see the case where the discrete uniform distribution was used in the place of Poisson, i.e., under given $E[y_g|\mathbf{x}, \epsilon_g]$, the support is $[0, 2E[y_g|\mathbf{x}, \epsilon_g]]$. Since the distributional misspecification on dependent variable does not affect the consistency results of PQML and WNLS, the point of interest should be the efficiency of NFES compared to linear models. The results in Table E.5 show that the performances of linear and nonlinear models are not very different from the ones in correct distribution. Regardless of ρ, NFES model is better than the linear models in terms of efficiency as well as consistency. As in the correct distribution case, despite some extreme values of estimates in small samples, the performance of NFES measured by median and MAD is still better than the linear models. Ruling out the seemingly unacceptable estimates, it can be said that NFES model is still more efficient than linear models even in small samples. Therefore NFES is robust about the distributional assumption on dependent variable.

Now consider the case where the true treatment equation is not probit. To that end, a skewed random variable using $\chi^2(5)$ was used. This is important because the correction terms that we used for PQML and WNLS were derived under the probit assumption. When violated, then the conditional mean function becomes misspecified and they lose the consistency property. However, the results in Table E.6 show that their behaviors are not very different from the cases under correct distribution. Although the performances of NFES in terms of mean and variance are worse than the linear estimators in small samples, they are better in terms of median and MAD. In larger samples they are a lot better by all criteria.

Although a departure from probit misspecifies the conditional mean, the impact is almost negligible.

Lastly, consider the cases where the distributions for unobservables are skewed. The advantage of normal distribution assumption was that it leads to same intercept both in estimating equation (2.7) and in equation (2.2) and thus the estimates from switching regression could directly be used for ATE estimation. Therefore under misspecification, the switching regression estimates do not give correct information for ATE, which leads to inconsistency. The results for misspecified unobservables are given in Table E.7, where the Monte Carlo mean of NFES model is farther away than the other cases as expected. However, it should be noted that it is not only the nonlinear model but also the linear models that are under strain with misspecified errors. Monte Carlo means for both linear and nonlinear models are also farther away from the true value and their deviation is also noticeable. In other words, skewed error moves the means away from the true value and makes the estimates less accurate by increasing the variance. However, the NFES model tends to have smaller RMSE in larger samples mainly due to its advantage in efficiency. The efficiency gain is particularly notable when it is measured by MAD; in larger samples the MAD value is merely about half as much as the ones of linear models. The WNLS is even more efficient than the PQMLE.

Third, from the results in Tables E.1-E.3 and E.5-E.6, the performances of 2SLS and NFES estimators are more or less the same. As the correlation gets higher, their finite sample biases grow bigger. However, 2SLS is diverging faster than the NFES, which implies that the NFES estimators are not only robust to distributional misspecification on y and v, but also suffering less to the higher degree of endogeneity than 2SLS is. The only case of concern is when the structural error ϵ is misspecified, where both 2SLS and NFES are diverging even faster than in the previous cases. However, unlike the above cases, NFES

are affected more than 2SLS under higher correlation. Nevertheless due to the advantage in efficiency the RMSE of NFES in larger sample is still smaller until the correlation is up to 0.5. Even with 0.6, the RMSE of WNLS is still smaller. In sum, NFES performs better than the 2SLS unless the structural errors are both misspecified and highly correlated with the selection error at the same time.

To better understand the simulation results, the sampling distributions for some selected estimators and DGP are graphically shown in figures. Figure G.1 displays the results under ideal cases [DGP 0] where the normality conditions hold. Among many estimators with different sample sizes, only 2SLS, NET(WNLS) and NFES(WNLS) with 5000 sample sizes for different correlation values are listed. The blue line juxtaposed with the histogram is nonparametric PDF estimate generated by Epanechnikov kernel with an optimal half-width. It can be seen from the figure that the NET estimator[3] behaves poorly in terms of both consistency and efficiency as anticipated in earlier sections. Even under large observation of 5000, sampling mean is very different from those of other estimators. Under multivariate normality, NFES is clearly consistent and efficient in large samples. For misspecified cases [DGP 1] and [DGP 2], it is clearly noticeable, in Figure G.2 and Figure G.3, that the NFES estimator is consistent and more efficient than 2SLS at large sample sizes. Figure G.4 depicts the sampling distribution with skewed structural error [DGP 3]. Although it creates bigger finite sample biases for those two estimators, the efficiency loss of 2SLS is more pronounced than in other misspecified cases. An important lesson from these figures is that the violation of distributional assumptions on which nonlinear models rely can also cause disadvantage for 2SLS that is not explicitly seen in traditional asymptotic analysis.

[3] The estimators based on NET model will simply be called NET estimators. Similarly for NFES estimators.

2.5.3 Some Other Results

In addition to the main results, here are some interesting, but off the main topic results. First, examine how large the biases of NET estimators are when the true model is of NFES. It can be seen in Tables E.1-E.3 that neither by mean nor by median do the NET estimators converge to the true value. It seems that the asymptotic biases of 1PQML and WNLS estimators are considerable; it is clear that a single-regime estimation is worse than simple linear models when the true model is of two-regime. Moreover, the seemingly large difference of Monte Carlo means of 2PQMLE and NLS reflects that those two estimators are not even estimating same agreed upon parameters. Gourieroux, Monfort and Trognon (1984) show that the true parameters are consistently estimated when the conditional mean is correctly specified. Also under the same condition, the NLS estimator also does the same thing since it is the sample analogue of appropriately defined moment conditions. By those argument, correctly specified conditional mean guarantees the convergence of those two estimators to a same quantity. In other words, a substantial difference even with large sample size is indicative of conditional mean misspecification. Therefore any parameter restriction in M-estimator should be used with caution.

Second, it can be seen in Tables E.1-E.3 for nonlinear estimators, the one-step Quasi-LIML, here NFES(1PQML) and NET(1PQML), is less efficient than the two-step methods, here NFES(2PQML) and NET(2PQML), in small samples. Such result was already reported in Chapter 1 for linear models where the dependent variable in structural equation was continuous; present simulation also shows that the finding is still valid when the dependent variable is nonnegative count variable. The results support the claim that two-step estimation procedure gives more efficient estimator than one-step.

Third, compare the 2PQML and WNLS for NFES model in Tables E.1-E.3. As it was already mentioned above, the asymptotic distribution of QMLE and WNLS are very similar. Under the condition $var(u|x,w) = \sigma^2 qvar$, their asymptotic distributions are the same. Simulation results show that the performances of those two estimators are indeed very similar. Nevertheless it is also clearly noticeable that WNLS is slightly more efficient than the 2PQML, which may be due to the fact that the GCIME does not hold for the 2PQML estimator.

Forth, according to the results for FIML that were tabulated in Table E.4, the FIML estimator is considerably more efficient than the others. Nevertheless it does not seem to converge well to the true parameter value, which may have resulted from the quadrature approximation error. To make this point clearer, another set of simulations using two different numbers of abscissas, i.e. 8 and 16 were run; more abscissas clearly help the mean approach to the true value. Incidentally the variances are not affected by number of abscissas. One can have more accurate FIML estimator by increasing number of abscissas, but it would take considerable amount of time in actual applications. One other feature that makes the FIML less attractive is that the PQML and WNLS quickly catches up FIML in terms of efficiency as the sample size grows bigger. Although FIML is the most efficient among all other estimators, the RMSE for $n = 5000$ is larger than PQML and WNLS due to the inaccuracy. On the other hand under small sample sizes the linear estimators do as nicely as FIML. Also using FIML becomes even less attractive when the standard error has to be found by bootstrapping; it may take too much time to complete a single session of bootstrap. There are other disadvantages too; since FIML heavily relies on the distributional assumption, any violation of them will cause the estimator inconsistent. In sum, FIML does not have any clear advantage over PQML and WNLS.

2.6 Empirical Application

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. Then we can also expect that the difference between *ceb* and *children* might be smaller for those who got the primary education. In this case the treatment effect would be greater when *ceb* was used as the dependent variable.

The data used in this empirical analysis is from Wooldridge (2010, Chapter 21). The variables description and descriptive statistics are given in Tables E.9 and E.10. 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 *educ*7 (finished primary education) and the instrument variable is $frsthalf$ (born in first half of year). The correlation between *educ*7 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.

Table E.12 presents 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 E.12 first four columns present 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.

The next three columns present 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. Also for semi-elasticity, the PQML estimate does not give any evidence of effectiveness of primary education. Again only the NLS estimate is weakly significant reporting roughly 30% decrease of living children. These results seem to mimic the behavior shown in simulation of last section and it can be an evidence that the NET model is inappropriate.

The last three (double) columns in Table E.12 list 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. LFES(Heckit) being consistent under probit selection assumption, it can have higher finite sample bias than the 2SLS when the assumption is violated as shown in Table E.5. Now can we say with greater certainty that ATE estimates cannot be substantially different from the LATE by 2SLS as was claimed by Angrist and Evans (1998) and Angrist and Pischke (2009) for bivariate probit case? As it was already seen in the simulation results, being similar under weak endogeneity, they start to diverge as

the degree of endogeneity becomes higher. Therefore the fact that 2SLS and NFES estimates are similar would be indicative of relatively weak correlation between selection error v and structural error ϵ, rather than the validity of the above claim.

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[4]. 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.

The above discussion was about treatment effect on the number of living children that reveals the difference in the desired number of children for each education group. Another interesting aspect can be the treatment effect on child mortality and those educated mothers

[4] 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.

are expected to have fewer dead children(Lam and Duryea, 1999; Schultz, 1994a,b). Thus the treatment effect might be negative. Since a direct estimation yields extreme outlying estimates for NFES model[5], an alternative indirect way of estimation procedure is used. To that end, another regression with the dependent variable *ceb* is run and presented in Table E.13. Then the expected child death at regime g for each individual is computed as $\widehat{mort}_{ig} = \widehat{ceb}_{ig} - \widehat{children}_{ig}$. The treatment effect for each observation is then computed as $\widehat{mort}_{i1} - \widehat{mort}_{i0}$ and their average through the whole population becomes the ATE on mortality presented in Table E.14. For example, the estimated regime one averages, i.e. $\sum \widehat{ceb}_{i1}/N$ are 1.402(NLS), 1.697(2PQML) and 1.683(1PQML), whereas those of regime zero $\sum \widehat{ceb}_{i0}/N$ are 2.890(NLS), 2.680(2PQML) and 2.657(1PQML). The differences in estimated regime zero averages, i.e. $\sum \widehat{ceb}_{i0}/N - \sum \widehat{children}_{i0}/N$ are 0.401(NLS), 0.340(2PQML) and 0.345(1PQML) implying that the mothers without primary education lose on average 0.4 children. For those with primary education the quantity is $\sum \widehat{ceb}_{i1}/N - \sum \widehat{children}_{i1}/N$ of which numerical values are 0.138(NLS), 0.198(2PQML), and 0.201(1PQML). The implication is that the child mortality rate is reduced roughly by 0.14 to 0.26 per mother by primary education. Although the ATE's on mortality for nonlinear estimators are not significant, they are still more efficient than the linear ones with an exception of LET(Heckit); the bootstrap standard deviations of nonlinear estimators are smaller than the linear ones.

[5] The 2SLS estimator gives insignificant -.002(.208). Unlike the general results, the NFES estimators do not give a reasonable estimate; two-step QML estimate is 25.821 with bootstrap standard deviation 784296.5 by 50 replications. It turns out that predicted values for dependent variable in regime 1, i.e. $\widehat{mort}_{i1}$ are very volatile; about 10% of the observations have more than 10 predicted child's deaths and about 1% have more than 100. Strangely such phenomenon does not occur in regime zero.

2.7 Conclusion

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 with a covariance term. The theory predicts that the efficiency of nonlinear ATE estimator is not taken for granted as in many other nonlinear cases. However, when the nonlinear estimators are actually more efficient, simulations show not only that it tends to have smaller finite sample bias, but also that its performance under misspecified model is not too bad compared with the linear IV or LFES models. 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 conditional ATE are not substantial as in the Botswana fertility example.

Chapter 3

A Two-Regime CRC Model with Nonnegative Dependent Variable

3.1 Introduction

The Correlated Random Coefficient (CRC) Models that were first introduced by Heckman and Vytlacil (1998) provided an alternative way to model the individual heterogeneity. The main focus having been on clarifying the conditions for identifying the parameters of interest under the presence of CRC since its inception (Wooldridge 1997, 2003, 2005; Card 2001), more recent development is focusing on the extension to the cases where the support of dependent variables are of limited nature. Wooldridge (2007) has suggested a method for estimating Average Partial Effects (APE) where the count dependent variable as well as random coefficients correlated with variable of interest are present. Cases of interest in this article are very similar to it except for the fact that a binary variable with counterfactual causal model is considered as a starting point. Specifically the goal of this paper is to provide an estimation method for Average Treatment Effects (ATE) when the dependent variable is count variable and all the covariates have CRC's that are correlated with the binary variable. This paper is organized as follows: In Section 2, previous discussion on CRC model will be reviewed. Section 3 discusses various CRC models proposed so far in the context of ATE.

Section 4 is the core of this article where the ATE estimation method for Nonlinear Two-regime CRC model will be provided. Section 5 discusses two specification tests; the test for endogeneity is designed for detecting the correlatedness of the random coefficients and the model selection test is for choosing the model between the ones with and without random coefficients. The method in this article is dependent on some distributional assumptions and their sensitivity on the estimation performance will be examined by Monte Carlo simulations in Section 6. Section 7 is an empirical application of this method and the performance of other competing estimator will be compared. Lastly, Section 8 is concluding remarks.

3.2 Previous Literature

Random coefficients typically arise from a model where an unobserved variable interacts with one or more observed variables. It then calls for different approaches due to the random nature of the coefficients on the observed variables. The randomness of the coefficients makes the conventional partial effect be random, and the quantities of interest are usually the means of the random coefficients. In other words, we would like to estimate not the partial effect but the average partial effect. If a regressor is mean independent from the unobserved variables, then the APE of the regressor can be consistently estimated by using OLS (See Amemiya, 1985).

Let the unobserved heterogeneity denoted q and the $1 \times K$ vector of exogenous variables $\mathbf{x}$. Then for the case where $E[q|\mathbf{x}] \neq 0$, even if $E[q|\mathbf{x}]$ can be appropriately modeled as a function of $\mathbf{x}$, the APE of a particular variable in $\mathbf{x}$ cannot be consistently estimated as long as the function contains the $\mathbf{x}$ with degree one. In such cases, q needs to be expressed by some proxy variables that does not contain $\mathbf{x}$ to consistently estimate the APE of a variable in $\mathbf{x}$.

Now let us consider a CRC model where $E(q|\mathbf{x}) \neq 0$ and proxy variables are not available as follows. In CRC, no correlation means mean independence, not the usual orthogonality.

$$y = \alpha_0 + (\alpha_1 + q_1)x + u$$

Assume for now that the regressor of interest is exogenous, i.e. orthogonal to the structural error. Then the model will have the regressor itself inside the composite error, i.e. $q_1 x + u$. Unlike from the mean independent case, this composite error does not vanish by using conditional expectation operation and renders the regressor effectively endogenous. Heckman and Vytlacil (1998) and Wooldridge (2003) have shown that under some assumptions, the instrumental variables method can be used to consistently estimate α_1.

In sum, when there are uncorrelated random coefficients, OLS gives a consistent estimator under the mean independence of q conditional on regressor. For CRC models, under the violation of mean independence, even a regressor that is independent of the structural error u in the population effectively becomes endogenous due to the random part of the coefficient. In what follows, our assumption is that the regressor is independent with the structural error for the sake of simplicity. Too restrictive the independent regressor as above might seem to be, the solutions for the CRC usually work for the nonindependent cases also (See Wooldridge, 2010).

It must be noted that given a particular regressor x, the coefficients on other regressors can be both random and correlated with x. The first case to consider is the model where the CRC with respect to a specific regressor is only on that regressor, i.e. the coefficients on all other covariates are assumed to be constants. The problem with this case is that a usual IV conditions are inadequate; even if the mean independence between IV and structural error

and between IV and CRC are established, the orthogonality between that IV and composite error is not guaranteed. In order for the IV to be effective, it has to satisfy an additional condition that $E[q_1 x|z]$ is constant (Wooldridge, 2003).[1] Although this condition is not very strong, its validity may be put in question when x is not continuous(Card, 2001; Wooldridge, 2005).

For the cases where the conditions for IV are not met, there is still another method that uses control function approaches with stronger assumptions. Heckman (1976) provides a solution for binary regressor, whereas Garen (1984) does it for continuous one. See the equation below.

$$y_1 = \eta_1 + \mathbf{z}_1\delta_1 + a_1 y_2 + y_2\mathbf{z}_1\gamma_1 + u_1, \tag{3.1}$$

where $\mathbf{z}_1$ is $1 \times L_1$ vector of exogenous variables and δ_1 and γ_1 are $L_1 \times 1$ coefficients vectors. The common convention is that Greeks are constant and Romans random. Regressors may or may not have random coefficients which may or may not be correlated with the structural error. The variables $\mathbf{z}_1$ that have constant coefficients without any correlation with u_1 will be called exogenous variable throughout this article. And any variable that is correlated with structural error or at least has correlated random coefficient will be called endogenous variable; for instance in the above equation the endogenous variable y_2 not only has correlation with u_1, but also has CRC.

One can also enrich the model by introducing some interaction terms between exogenous and endogenous variables such as $y_2\mathbf{z}_1$ in equation (2.1). Let us put some restrictions here; suppose that y_2 is binary endogenous variable and that the coefficient of the interaction

[1] Although this condition is for the IV estimation in CRC models, it is still needed for a random coefficient that is uncorrelated with the regressor. Even if $E[q_1 x] = 0$, it is not necessarily that $E[q_1 x|z]$ is constant. Again no correlation in CRC is not about the orthogonality but mean independence.

term γ_1 is constant. Then this model simply becomes Linear Full Endogenous Switching Regression or LFES model (Keay, 2011) with two regimes between which an individual can switch. In this model all the coefficients in each regime are nonrandom. For this LFES it is also possible to introduce CRC to the exogenous variables for each equation (Wooldridge, 2007). Or can one equivalently construct such two-regime CRC model, i.e. a two regime regression model where the covariates in each equation have random coefficients correlated with y_2, by directly introducing CRC not only to the endogenous variable but also to the exogenous variables as well as the interaction. Here is an equivalence result: two-regime CRC model can be constructed by putting CRC to the all regressors in the equation including the interaction terms.

The discussion so far was about the models with continuous structural error and hence continuous dependent variable. In this paper I will provide a model where the dependent variable takes nonnegative or more specifically natural numbers. Such cases arise very often in real life such as when the dependent variable is commuting frequency (Terza, 1998) or number of child (Keay, 2011). In addition to the nonlinearity in the first stage binary choice, another nonlinearity to the dependent variable of the structural equation will be taken care of. Although Terza (2009) has already considered such nonlinear model with two regimes without CRC, its extension with CRC on all exogenous variables has not been proposed so far. In the following, the former model will be called Nonlinear Full Endogenous Switching Regression or NFES model and the later be called Nonlinear Two-regime CRC or NTCRC model. The linear version of Two-regime CRC model was already proposed by Wooldridge (2007) that will be called Linear Two-regime CRC or LTCRC model.

In what follows Poisson distribution will be used to model the nonnegative response in the structural model. Incidentally an ensuing natural question is whether all the results in the

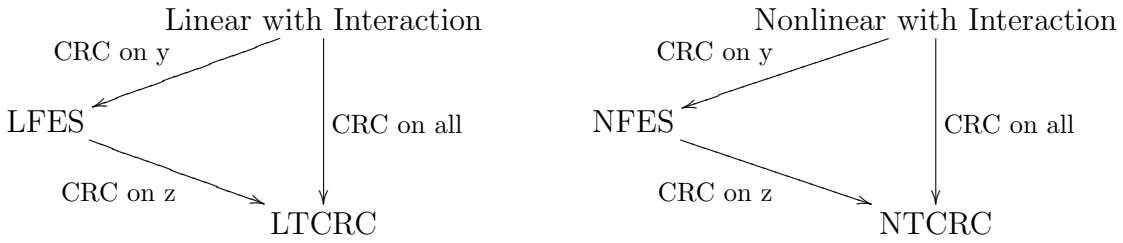

Figure 3.1: Analogy Structure

linear model is valid to the same extent. Of course the model itself does not undergo much changes although some quantities of interest will not be identified, which will be discussed in the following sections.

The problem of NFES approach is that the intercept term inside the exponential function in the structural equation is not identified by itself alone; instead Terza (2009) estimates the average treatment effect for nonlinear model and such approach will be employed also for the two-regime CRC model in this paper. As it was already mentioned above, the CRC on binary endogenous variable makes two-regime switching regression model. The objective of this paper will be to set out a NTCRC model by bringing once again the CRC to the two-regime nonlinear switching model, where the average treatment effect will be identified. Incidentally this is equivalent to the model where all the variables including the interaction have CRC. The above discussion is summarized in the analogy scheme in Figure 3.1.

3.3 Various Models of CRC

3.3.1 Continuous Endogenous Variable

Let the variable with which random coefficient is correlated be denoted w. By the unobserved heterogeneity argument the variable w is usually regarded as endogenous, which is also the

case here. The CRC can be either only on w or on all other variables. In population regression model, the variable w and all other covariates can be interacted, which may create many possible combinations of situations: CRC only on w with no interaction, CRC only on w with interaction, CRC on all variables with no interaction and finally CRC on all variables with interaction. Consider a regression model as follows.

$$y = \mathbf{x}\beta + \tau w + \gamma q + q\mathbf{x}\delta + u,$$

where $\mathbf{x}$ is a $1 \times K$ vector of exogeneous variables and β and δ are $K \times 1$ vectors. The variables q and w are assumed to be scalar. The former is unobserved variable that is correlated with the latter, but interacts with $\mathbf{x}$. One of the characteristics of this model is that the variable w will be correlated with the random coefficients of $\mathbf{x}$, i.e. $\beta + q\delta$ and, at the same time, will be an endogenous variable in traditional sense because it is correlated with the composite error $\gamma q + u$. If an interaction term between $\mathbf{x}$ and w is included in this model and also if that interaction term is again interacted with the unobserved variable q, then, given that w is binary, the resulting model will be the two-regime CRC model that will be discussed below.

$$y = \mathbf{x}(\beta + \delta_0 q) + w(\tau + \delta_1 q) + w\mathbf{x}(\kappa + \delta_2 q) + \gamma q + u$$

A random coefficient model with an endogenous variable with which the coefficients on an exogenous variable is correlated is created in such manner. This model turns out to be a very good description of education-fertility behavior which will be discussed in later sections. In addition to that, this model is very general in that many other previous models can be treated as its special cases. In case w is endogenous binary variable, a restriction $\delta_0 = \delta_2 = 0$ makes the model as an ordinary two-regime switching model and also an additional restriction

$\delta_1 = \kappa = 0$ will make it as Endogenous Treatment Model as in Terza (1998, 2009). There are also many other possibilities that are nonetheless special cases of the model given above.

Let us focus on a continuous w here deferring discussion of binary case in subsequent sections. With count dependent variable, suppose the conditional expectation can be modeled by an exponential function as link function. This approach explicitly considers the dependent variable of the structural model as count variable that does not take negative value. As the simplest case consider a model of CRC only on w without interaction as below.

$$\begin{aligned} E(y_1|\mathbf{z}, w, a_1, r_1) &= \exp(\mathbf{x}\delta_1 + a_1 w + r_1) \\ w &= \mathbf{z}\delta_2 + v_2, \end{aligned}$$

where $\mathbf{x}$ and $\mathbf{z}$ are $1 \times L_1$ and $1 \times L$ vectors of exogenous variables with $\mathbf{x} \subset \mathbf{z}$ that are independent with all the random variables generated in model. The coefficient vectors δ_1 and δ_2 are $L_1 \times 1$ and $L \times 1$ respectively. Also assume that

$$\begin{aligned} a_1 &= \alpha_1 + d_1 \\ d_1 &= \psi v_2 + e_d \qquad v_2 \perp\!\!\!\perp e_d \\ r_1 &= \theta v_2 + e_r \qquad v_2 \perp\!\!\!\perp e_r \end{aligned}$$

In the first equation, $\alpha_1 = E[a_1]$ and $d_1 = a_1 - \alpha_1$. The second and third equations are basically the linear projections of each random variable on v_2. However, we need a stronger condition that the orthogonal decompositions are not only uncorrelated but also independent in order to facilitate the identification. The regressor w suffers from endogeneity unless $\psi = \theta = 0$. Along with those assumptions, we put multivariate normal assumption on

the random variables (d_1, r_1, v_2). Under these assumptions, Wooldridge (2007) has derived a conditional mean function as below.

$$E(y_1|\mathbf{z}, v_2) = \exp\left(\mathbf{x}\delta_1 + \alpha_1 w + \psi v_2 w + \theta v_2 + \frac{\sigma_r^2 + 2\sigma_{dr}w + \sigma_d^2 w^2}{2}\right)$$

Although the above equation is estimable once v_2 is estimated in the first stage regression, the semi-elasticity of w, i.e. α_1 is not identified due to the fraction term, and this is still the case even if w is exogenous, where $\psi = 0,\ \theta = 0$. In order to identify α_1, it is required that the random coefficient on w is uncorrelated not only with w but also with r_1. Although the semi-elasticity is not identified, one can still identify the APE of w over z, w and v_2. From the above equation, the partial effect of w is

$$\frac{\partial E(y_1|\mathbf{z}, w, v_2)}{\partial w} = \exp\left(\mathbf{x}\delta_1 + \alpha_1 w + \psi v_2 w + \theta v_2 + \frac{\sigma_r^2 + 2\sigma_{dr}w + \sigma_d^2 w^2}{2}\right) \cdot (\alpha_1 + \sigma_{dr} + \psi v_2 + \sigma_d^2 w).$$

By taking the average over the whole population, APE of w is identified. Wooldridge (2007) also extends to discuss the case where the coefficients of other covariates are also correlated with w. The results are basically the same; APE is identified although the semi-elasticity is not.

3.3.2 Binary Endogenous Variable

Recall Figure 3.1. It should be noted that we are already familiar with the models with CRC only on y, which are equivalent to the two-regime endogenous switching regression (See Maddala, 1986). For the linear and nonlinear cases, one can use the correction methods proposed by Heckman (1978) and Terza (1998) respectively. As an extension in linear model,

the one with CRC not only on endogenous but also on all other variables, i.e. Linear Two-regime CRC model (hereafter LTCRC), was explored by Wooldridge (2007). Consider a regression model as follows.

$$y_1 = \mathbf{x}d_1 + a_1 y_2 + \mathbf{x}y_2 g_1 + u,$$

where $\mathbf{x}$ is $1 \times K$ vector of exogenous covariates and d_1 and g_1 are $K \times 1$ vectors of CRC's. Also we have a vector of all exogenous variables $\mathbf{z}$ such that $\mathbf{x} \subset \mathbf{z}$. As in Chapter 2, it is assumed that $\mathbf{x}$ is demeaned without loss of generality. Wooldridge (2007) shows that under the assumptions

$$\begin{aligned} E(a_1|\mathbf{z}, v_2) &= \alpha_1 + \varphi_1 v_2 \\ E(d_1|\mathbf{z}, v_2) &= \delta_1 + \psi_1 v_2 \\ E(g_1|\mathbf{z}, v_2) &= \xi_1 + \omega_1 v_2, \end{aligned}$$

the estimating equation is

$$E(y_1|\mathbf{z}, y_2) = \mathbf{x}\delta_1 + \alpha_1 y_2 + y_2\mathbf{x}\xi_1 + \rho_1 h_2(y_2, \mathbf{z}\delta_2) + \varphi_1 h_2(y_2, \mathbf{z}\delta_2)y_2 + h_2(y_2, \mathbf{z}\delta_2)\mathbf{x}\psi_1 + h_2(y_2, \mathbf{z}\delta_2)y_2\mathbf{x}\omega_1,$$

where

$$h_2(y_2, \mathbf{z}\delta_2) = y_2\lambda(\mathbf{z}\delta_2) - (1 - y_2)\lambda(-\mathbf{z}\delta_2),$$

where $\lambda(\cdot)$ is inverse Mill's ratio.

From this model and estimating equation, one can identify the ATE of y_2 on y_1, i.e. α_1. In the next section, the main contribution of this article, the Nonlinear Two-regime CRC

(NTCRC) model and its estimation method will be discussed.

3.4 Nonlinear Two-Regime CRC Model

3.4.1 Model

In what follows we allow for a binary case of the variable of interest w, which calls for an analysis of ATE under counterfactual framework. Under this setting, the first stage binary selection equation is assumed to follow probit model, which makes the estimation more efficient. For the continuous w, the average of the semi-elasticity of w, i.e. α_1 was not identified, while the APE was. Analogously, although the average semi-elasticity of w is not identified, ATE will be. Consider a model in terms of conditional expectation functions given below.

$$\begin{aligned} E(y_0|\mathbf{z}, w, a_0, b_0, a_1, b_1) &= E(y_0|\mathbf{x}, a_0, b_0) = \exp(a_0 + \mathbf{x}b_0) \\ E(y_1|\mathbf{z}, w, a_0, b_0, a_1, b_1) &= E(y_1|\mathbf{x}, a_1, b_1) = \exp(a_1 + \mathbf{x}b_1) \\ w &= 1[\mathbf{z}\gamma + v > 0], \end{aligned} \tag{3.2}$$

where a_g are scalar errors, $\mathbf{x}$ is $1 \times K$ vector of covariates, and b_g are $K \times 1$ random coefficient vectors. The $\mathbf{z}$ is the vector of all the available exogenous variables such that $\mathbf{x} \subset \mathbf{z}$. For those random coefficients, our quantities of interest will be the means of random variables. In what follows, we will use notation $\text{var}(x) \equiv \sigma^2(x)$ and $\text{cov}(x, y) \equiv \sigma(x, y)$. Also the regime subscript will be suppressed whenever obvious. Here are the assumptions for further discussion.

Assumptions

1. **Unconfoundedness**: As was already stated in equation (3.2), w is redundant conditional on a and b that summarize all the information about the determined regime, i.e.

$$E(y_1|\mathbf{z}, w, a_0, b_0, a_1, b_1) = E(y_1|\mathbf{x}, a_1, b_1),$$

and same for Regime 0. Although the main focus is on w, this equation also assumes the same thing for the unobserved heterogeneity for the other regime.

2. **Multivariate Normality**: For each regime in equation (3.2), we have

$$\begin{aligned} a &= \alpha + e \\ b &= \beta + d, \end{aligned}$$

where α and β are the means of a and b. The vector (e_0, d_0, e_1, d_1, v) of which dimension is $(2K+3) \times 1$ follows multivariate normal, i.e.

$$(e_0, d_0, e_1, d_1, v)' \sim N(\mathbf{0}, \mathbf{V}),$$

where $\mathbf{V}$ is the appropriate variance-covariance matrix.

Let the linear projections of e and d on v as

$$\begin{aligned} e &= \sigma(e, v)v + \epsilon \\ d &= \sigma(d, v)v + \delta. \end{aligned}$$

Note that b, d, δ and $\sigma(d, v)$ are $K \times 1$ vectors. Specifically, $\sigma(d, v) =$

$[\sigma(d_1, v), \cdots, \sigma(d_K, v)]'$, where $b_j = \beta_j + d_j$ is the j-th element in b. Since v and ϵ, δ are orthogonal, they are independent under multivariate normality. Another implication of the above assumption is that the selection equation follows probit model.

3. **Conditional Independence**: v and $\epsilon + \mathbf{x}\delta$ are independent conditional on $\mathbf{z}, w$, i.e.,

$$v \perp\!\!\!\perp \epsilon + \mathbf{x}\delta \mid \mathbf{z}, w.$$

Under these assumptions, one can obtain the following lemma.

Lemma 3.1. Under the assumptions 2 and 3,

$$E[\exp(e_1 + \mathbf{x}d_1)|\mathbf{z}, w = 1]$$

$$= \exp\left[\left(\sigma^2(e_1) + \sum_{j=1}^{K}\sigma^2(d_{1j})x_j^2 + 2\sum_{j=1}^{K}\sigma(e_1, d_{1j})x_j + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_{1j}, d_{1r})x_j x_r\right)\Big/2\right]$$
$$\times \frac{\Phi\left(z\gamma + \sigma(e_1, v) + \sum_{j=1}^{K}\sigma(d_{1j}, v)x_j\right)}{\Phi(z\gamma)}$$

for Regime 1, and

$$E[\exp(e_0 + \mathbf{x}d_0)|\mathbf{z}, w = 0]$$

$$= \exp\left[\left(\sigma^2(e_0) + \sum_{j=1}^{K}\sigma^2(d_{0j})x_j^2 + 2\sum_{j=1}^{K}\sigma(e_0, d_{0j})x_j + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_{0j}, d_{0r})x_j x_r\right)\Big/2\right]$$
$$\times \frac{\Phi\left(-z\gamma - \sigma(e_0, v) - \sum_{j=1}^{K}\sigma(d_{0j}, v)x_j\right)}{\Phi(-z\gamma)}$$

for Regime 0.

Proof Consider the Regime 1 only. For notational simplicity the regime subscripts are suppressed. Derivation for Regime 0 is almost identical. Note that

$$\begin{aligned}
E[\exp(e+\mathbf{x}d)|\mathbf{z},w=1] &= E\Big[\exp\Big(\sigma(e,v)v+\epsilon+\mathbf{x}\sigma(d,v)v+\mathbf{x}\delta\Big)\Big|\ \mathbf{z},w=1\Big] \\
&= E\Big[\exp\Big((\sigma(e,v)+\mathbf{x}\sigma(d,v))v\Big)\exp(\epsilon+\mathbf{x}\delta)\Big|\ \mathbf{z},w=1\Big] \\
&= E\Big[\exp\Big((\sigma(e,v)+\mathbf{x}\sigma(d,v))v\Big)\Big|\ \mathbf{z},w=1\Big] \\
&\qquad\qquad \times E[\exp(\epsilon+\mathbf{x}\delta)|\mathbf{z},w=1] \\
&= E\Big[\exp\Big((\sigma(e,v)+\mathbf{x}\sigma(d,v))v\Big)\Big|\ \mathbf{z},w=1\Big] \\
&\qquad\qquad \times E[\exp(\epsilon+\mathbf{x}\delta)|\mathbf{z}] \\
&= \frac{\Phi\Big(z\gamma+\sigma(e,v)+\mathbf{x}\sigma(d,v)\Big)}{\Phi(z\gamma)}\exp\left(\frac{(\sigma(e,v)+\mathbf{x}\sigma(d,v))^2}{2}\right) \\
&\qquad\qquad \times E[\exp(\epsilon+\mathbf{x}\delta)|\mathbf{z}]
\end{aligned}$$

For the second term on RHS,

$$\begin{aligned}
\Big(\sigma(e,v)+\mathbf{x}\sigma(d,v)\Big)^2 &= \Big(\sigma(e,v)+\sum_{j=1}^{K}\sigma(d_j,v)x_j\Big)^2 \\
&= \sigma^2(e,v)+2\sum_{j=1}^{K}\sigma(e,v)\cdot\sigma(d_j,v)x_j+\sum_{j=1}^{K}\sigma^2(d_j,v)x_j^2 \\
&\quad +\sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_j,v)\cdot\sigma(d_r,v)x_jx_r.
\end{aligned}$$

Now consider the third term on RHS. From the normality of $\epsilon + \mathbf{x}\delta$ that is guaranteed by the multivariate normal assumption,

$$E[\exp(\epsilon + \mathbf{x}\delta)|\mathbf{z}] = \exp\Big(\text{var}[\epsilon + \mathbf{x}\delta \mid \mathbf{x}]\Big/2\Big).$$

Now that

$$\begin{aligned}
\text{var}[\epsilon + \mathbf{x}\delta \mid \mathbf{x}] &= \sigma^2(\epsilon) + \sigma^2(\mathbf{x}\delta) + 2\sigma(\epsilon, \mathbf{x}\delta) \\
&= \sigma^2(\epsilon) + \sum_{j=1}^{K} \sigma^2(\delta_j) x_j^2 + \sum_{j=1}^{K} \sum_{r \neq j} \sigma(\delta_j, \delta_r) x_j x_r + 2 \sum_{j=1}^{K} \sigma(\epsilon, \delta_j) x_j.
\end{aligned}$$

Collecting those terms we have the stated result. ■

Remark Terza (1998) directly used the multivariate normal property in order to solve the conditional expectation, i.e. $E[\exp(\epsilon)|\mathbf{z}, v] = \exp\left(\rho\sigma v + \frac{1}{2}\sigma^2(1-\rho^2)\right)$ was derived under the multivariate normal assumption between ϵ and v. Although such approach is also applicable here, it does not give a convenient expression for estimating the average treatment effect. In the above derivation I used the linear projections of e and d on v that give un estimating equation of which the coefficients can be directly used to find the ATE. As will be discussed below, the structural parameters such as β is not independently identified, but the coefficients on x and x^2 from the estimating equation derived by the linear projection coincides with those coefficients of the estimating equation for ATE.

Now let us derive an estimating equation of the model in (3.2). Note that by assumptions 1 and 2,

$$\begin{aligned}
E[y|\mathbf{z},w,a_0,b_0,a_1,b_1] &= (1-w)E[y_0|\mathbf{z},w,a_0,b_0,a_1,b_1] + wE[y_1|\mathbf{z},w,a_0,b_0,a_1,b_1] \\
&= (1-w)E[y_0|\mathbf{x},a_0,b_0] + wE[y_1|\mathbf{x},a_1,b_1] \\
&= (1-w)\exp\Big(\alpha_0 + e_0 + \mathbf{x}(\beta_0+d_0)\Big) \\
&\qquad + w\exp\Big(\alpha_1 + e_1 + \mathbf{x}(\beta_1+d_1)\Big) \\
E[y|\mathbf{z},w] &= (1-w)\exp(\alpha_0+\mathbf{x}\beta_0)E[\exp(e_0+\mathbf{x}d_0)|\mathbf{z},w=0] \\
&\qquad + w\exp(\alpha_1+\mathbf{x}\beta_1)E[\exp(e_1+\mathbf{x}d_1)|\mathbf{z},w=1]
\end{aligned}$$

By Lemma 3.1.

$$\begin{aligned}
= \; & w\cdot\exp\Bigg[\Big(2\alpha_1+\sigma^2(e_1)+\sum_{j=1}^K\sigma^2(d_{1j})x_j^2+2\sum_{j=1}^K[\beta_1+\sigma(e_1,d_{1j})]x_j \\
& \qquad +\sum_{j=1}^K\sum_{r\neq j}\sigma(d_{1j},d_{1r})x_jx_r\Big)\Big/2\Bigg]\times\frac{\Phi\Big(z\gamma+\sigma(e_1,v)+\sum_{j=1}^K\sigma(d_{1j},v)x_j\Big)}{\Phi(z\gamma)} \\
& +(1-w)\exp\Bigg[\Big(2\alpha_0+\sigma^2(e_0)+\sum_{j=1}^K\sigma^2(d_{0j})x_j^2+2\sum_{j=1}^K[\beta_0+\sigma(e_0,d_{0j})]x_j \\
& +\sum_{j=1}^K\sum_{r\neq j}\sigma(d_{0j},d_{0r})x_jx_r\Big)\Big/2\Bigg]\times\frac{\Phi\Big(-z\gamma-\sigma(e_0,v)-\sum_{j=1}^K\sigma(d_{0j},v)x_j\Big)}{\Phi(-z\gamma)} \qquad (3.3)
\end{aligned}$$

The above equation identifies only $\sigma^2(d_j)$, $\sigma(e,v)$ and $\sigma(d_j,v)$ separately, and the average of semi-elasticity, i.e. β is not identified due to the nuisance parameters $\sigma(e,d_j)$'s. It can be estimated by two step by using the probit model for the selection equation; given the estimated index $\mathbf{z}\gamma$, the above equation can be estimated either by NLS or Quasi-ML

under Poisson distribution. Or can they still be estimated simultaneously by a single-step method. See Chapter 2 for detailed discussion of estimation procedure and their asymptotic distribution. Since the structural parameters α and β are not separately identified, our interest should lie in average treatment effects.

3.4.2 Identification of ATE

Note that the ATE is $E[y_1 - y_0]$. One of the easiest way to identify this is by using the law of iterated expectation, i.e. $E[y_1 - y_0] = EE[y_1 - y_0|\mathbf{x}]$ as long as the expectation conditional on x can be derived. To that end, the following lemma is derived.

Lemma 3.2. Under the assumptions given above, the following result holds.

$$E[\exp(e+\mathbf{x}d)|\mathbf{x}] = \exp\left[\left(\sigma^2(e)+\sum_{j=1}^{K}\sigma^2(d_j)x_j^2+2\sum_{j=1}^{K}\sigma(e,d_j)x_j+\sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_j,d_r)x_jx_r\right)\Big/2\right]$$

Proof Note that

$$\begin{aligned} E(e+\mathbf{x}d|\mathbf{x}) &= E(e|\mathbf{x}) + E(\mathbf{x}d|\mathbf{x}) = 0 \\ \text{var}(e+\mathbf{x}d|\mathbf{x}) &= \text{var}(e|\mathbf{x}) + \text{var}(\mathbf{x}d|\mathbf{x}) + 2\text{cov}(e,\mathbf{x}d|\mathbf{x}) \\ &= \sigma^2(e) + \sum_{j=1}^{K}\sigma^2(d_j)x_j^2 + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_j,d_r)x_jx_r + 2\sum_{j=1}^{K}\sigma(e,d_j)x_j. \end{aligned}$$

For any fixed value of $\mathbf{x}$, $e + \mathbf{x}d$ is normally distributed since it is a combination of multivariate normal random variables as was mentioned in assumption 2. Since the mean and variance of normal random variables are already obtained, the mean of its log-normal variable is trivially found. ■

The ATE conditional on $\mathbf{x}$ is

$$E[y_1 - y_0|\mathbf{x}] = \exp(\alpha_1 + \mathbf{x}\beta_1)E[\exp(e_1 + \mathbf{x}d_1)|\mathbf{x}] - \exp(\alpha_0 + \mathbf{x}\beta_0)E[\exp(e_0 + \mathbf{x}d_0)|\mathbf{x}].$$

Thus from the above lemma,

$$\begin{aligned} &E[y_1 - y_0|\mathbf{x}] \\ &= \exp\left[\left(2\alpha_1 + \sigma^2(e_1) + \sum_{j=1}^{K}\sigma^2(d_{1j})x_j^2 + 2\sum_{j=1}^{K}[\beta_1 + \sigma(e_1, d_{1j})]x_j + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_{1j}, d_{1r})x_j x_r\right)\Big/2\right] \\ &\quad - \exp\left[\left(2\alpha_0 + \sigma^2(e_0) + \sum_{j=1}^{K}\sigma^2(d_{0j})x_j^2 + 2\sum_{j=1}^{K}[\beta_0 + \sigma(e_0, d_{0j})]x_j + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(d_{0j}, d_{0r})x_j x_r\right)\Big/2\right]. \end{aligned} \tag{3.4}$$

Note the similarity of this equation with the estimating equation in (3); except for the correction functions, the expressions inside the exponential function are identical which makes the identification of ATE possible. The estimator will be the sample average of equation (3.4) over the values of $\mathbf{x}$. The fact that each parameter is not identified does not make any problems for ATE identification. The asymptotic distribution of this ATE estimator is essentially same as was presented in Terza (2008).

3.5 Specification Test

3.5.1 Tests for Endogeneity

In the previous section, we have derived the estimating equation (3.3) and ATE conditional on covariates (3.4). In these equations no restrictions were imposed and an estimation of the model without restriction, where the number of identifiable composite parameters is no less than $(K+4)(K+1)/2$, might cause difficulty in numerical optimization. One can apply the Lagrange Multiplier (LM) test in order to test for the presence of *correlated* random coefficients. Another test called Variable Addition Test (VAT) is also available which is asymptotically equivalent to LM but easier to apply. This test constructs a conditional mean by adding appropriately defined variables to create the likelihood function of which the score under restriction is same as the one used in LM test (See Wooldridge, 2011). The actual test is performed by Wald test on the significance of coefficients of the added variables. Revisit the estimating equation (3.3). Inside the exponential function we have each covariate, the square of each covariate and their cross products. Let the vector of these functions of covariates denoted $\widetilde{\mathbf{x}}$ and rewrite the estimating equation for Regime 1 as follows.

$$E[y|\mathbf{z}, w=1] = \exp(\widetilde{\mathbf{x}}\theta_1) \times \frac{\Phi\big(\mathbf{z}\gamma + \theta_2 + \mathbf{x}\theta_3\big)}{\Phi(\mathbf{z}\gamma)}, \tag{3.5}$$

where $\theta_2 = \sigma(e_1, v)$ and $\theta_3 = \sigma(d_1, v)$. As was already mentioned, θ_2 is scalar, θ_3 is $K \times 1$ vector and θ_1 is $(K+2)(K+1)/2 \times 1$ vector. The estimating equation for Regime 0 is the same as above except for the minus sign inside the $\Phi(\cdot)$ in correction function.

In order to perform VAT, we need to construct a conditional mean function of which score is identical to the equation (5) under restriction. Let the restriction or equivalently

the null hypothesis

$$H_0 : \theta_2 = \theta_3 = 0.$$

In other words, there is no correlation between selection error v and any other errors in the structural equations. An LM test can be applied that does not require an estimation of complicated model without restriction. The VAT is a device that facilitates this LM test by using an auxiliary regression for Regime 1

$$\exp\Big(\widetilde{\mathbf{x}}\theta_1 + \theta_2\lambda(\mathbf{z}\gamma) + \theta_3\lambda(\mathbf{z}\gamma)\mathbf{x}\Big), \tag{3.6}$$

where $\lambda(\cdot)$ is inverse Mill's ratio. The auxiliary regression for Regime 0 will have $-\mathbf{z}\gamma$ for $\mathbf{z}\gamma$. It can be easily verified that the likelihood and score functions derived from (3.6) under restrictions are same as the one from (3.5). Therefore the score tests on the significance of coefficients from the above auxiliary regression is equivalent to the LM tests on (3.5). Thus the Wald tests for (3.6) will give a simple and asymptotically equivalent way to test the null without estimating the model without restriction. In actual test, $\lambda(\mathbf{z}\gamma)$ has to be estimated through the first stage regression.

3.5.2 Model Selection Test

What we have considered above is whether the already given random coefficients are correlated with the selection error or not. Even when the null hypothesis is not rejected, it does not warrants our coming back to the Nonlinear Full Endogenous Switching (NFES) Regression provided by Terza (1998). In current model, presence of CRC is assumed both in a and b in equation (3.2). As a natural consequence one also assumes the multivariate normality

between those errors and v. On the contrary, the coefficient b is assumed to be constant in NFES and the multivariate normal joint error distribution is assumed just between a and v. It should be emphasized that the VAT tests for the endogeneity under the assumption that b is random, and thus a failure to reject null hypothesis does not imply an acceptance of NFES. A practical consequence of this observation is that one has to include the square and cross products of the covariates, which were not included in NFES, even when there turns out to be no correlation between those random coefficients. Then how can one perform a test of which the alternative is NFES model? The NFES model assumes that the coefficient b in equation (3.2) is constant and thus $d = 0$ as well as $\sigma^2(d_j) = \sigma(\ \cdot\ , d_j) = 0$. From equation (3.3), the null hypothesis will be the zero restriction on the coefficients of square and cross product terms inside the exponential functions and the coefficients on x_j inside the correction functions.

3.6 Monte Carlo Simulation

The purpose of the simulation is two-fold: First, it will compare the performances of this nonlinear ATE estimator with the linear method proposed by Wooldridge (2007). This will show the advantage of using the nonlinear model for estimating ATE. Second, it will see the impact of violations of the assumption that were used in order to derive the estimating equation in the previous section.

3.6.1 Data Generating Processes

The derivation of the estimating equations depends on the three assumptions listed in the above section. In this section, we examine the performances of the ATE estimator under

violations of the first and second assumptions only. In order to check the robustness of those assumption, a series of simulations are run by using the following data generating processes (DGPs).

$$
\begin{aligned}
x &\sim \text{uniform}[15, 50] \\
z &\sim \text{Binomial}(1, 1/2) \\
w &= 1[1.4 - 0.05x - 0.3z + v \geq 0] \\
E(y_1|x, z, e_1, d_1) &= \exp\Big(0.15 + e_0 + (0.02 + d_0)x\Big) \\
E(y_0|x, z, e_0, d_0) &= \exp\Big(0.1 + e_1 + (0.0059 + d_1)x\Big),
\end{aligned}
$$

where y_1 and y_0 follow Poisson distribution with the mean specified above. This DGP was designed to make the population ATE approximately equal to unity.

There are three sessions of simulations; DGP 1 deals with the ideal case where the errors follow multivariate normal distribution and $v \perp\!\!\!\perp \epsilon + x\delta$. The joint distribution of errors for DGP 1 is

$$
\begin{bmatrix} e_0^* \\ d_0^* \\ e_1^* \\ d_1^* \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & 0 & 0 & \rho \\ & 1 & 0 & 0 & \rho \\ & & 1 & \rho & \rho \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix} \right).
$$

This variance-covariance matrix is in fact correlation matrix since the variances are all equal to unity. For the simulation, $e_0 = e_0^*/100$, $e_1 = e_1^*/100$, $d_0 = d_0^*/100$, and $d_1 = d_1^*/100$ were used. Also for the correlation values ρ, 0.3 and 0.5 were used. This is the case where the

above estimator might perform the best, since the actual DGP conforms the assumptions on which the estimating equation is based.

DGP 2 generates a process where the linear projection errors ϵ and δ are uncorrelated but dependent with v. For a given v, ϵ is designed to be either the same value of v or $-v$ by a half probability for each. Then the scattergram will look like a X letter, where the covariance of the two variables becomes zero, but very strongly dependent with each other.

Another important assumption for Lemma 3.1 is that v follows normal distribution, which brought the normal cumulative distribution function $\Phi(\cdot)$ in the equation. In order to check the robustness of the estimator on that assumption, DGP 3 uses $\chi^2(5)$ distribution to create skewed errors. The variance-covariance matrix looks just the same as in the above equation. However, we here drop the multivariate normal assumption; all the *marginal* distributions of the errors follow some adjusted χ^2 distribution with the predetermined correlation[2].

3.6.2 Simulation Results

Tables F.1 and F.2 list the simulation results for $\rho = 0.3$ and 0.5 respectively. For each table, each column for I, II and III lists the results for the DGP I, II and III both for linear and nonlinear estimator. Let us first consider the performances of nonlinear estimator. In either Table F.1 or F.2, the column I for DGP 1 shows that the ATE estimator is performing well as expected. The results show that there are nontrivial amount of outlying values under the sample size 1000. It is usually the case that the numerical optimization does not work

[2] There might be infinitely many such joint distribution and the multivariate χ^2 distribution, which is not used here, is just one of them. The reason why multivariate χ^2 distribution is not used is that 1) it is hard to generate the DGP 3 by Stata, and that 2) what matters is just the failure of the normality for each marginal distribution and its impact on the estimation. In this DGP 3, I generated χ^2 distributions by first generating many basis standard normal distributions. The correlations were created by using some common basis standard normal distribution. The Stata code will be provided upon request.

well in smaller sample size and gives many dubious estimates. In the Table F.1 and F.2, the outlying values that are greater or smaller than the maximum and minimum values for the sample size of 3000 were discarded from the generated data. Roughly about 8 to 10% of the total data is removed in this way. However, whole data are kept for median and MAD calculation which are not sensitive to the extreme values.

There seems to be no discernible difference between $\rho = 0.3$ and 0.5 The column for DGP 2 shows that for those two tables the results do not show any clear sign of the adverse impact due to the violation of the independence assumption. Therefore we can conclude that the independence assumption is more or less practically negligible.

The last column is for DGP 3, where the multivariate normal assumption is violated. In this case, it is clearly visible that the estimator suffers from larger variance and bias even under 5000 observations. It happens because we used the fact that $E[\exp(u)] = \exp(\sigma^2/2)$ under normality in Lemma 3.1. If the normality fails, then the whole estimating equation will be misspecified causing inconsistency.

A comparison between linear and nonlinear estimators shows that the nonlinear estimator has larger variance particularly when the sample size is small. However, it is also clear that the linear estimators do not show any sign of consistency. The nonlinear estimator keeps approaching the true value, i.e. one, from sample size of 3000 to 5000, the linear estimator does not show any convergence as the sample sizes grow. Another interesting aspect is that the skewed distribution of errors not only affects the nonlinear estimator, but also the linear one too. This is particularly true in Table F.2; in column III, the linear estimator is actually moving away from the true value one. In sum, although having larger variance, the nonlinear estimator is better than the linear one in terms of consistency. It is preferable to use the nonlinear ATE estimator when the sample size is sufficiently large.

3.7 An application: the effect of elementary school education on fertility in Botswana

In this section, the ATE estimator derived above will be applied to the Botswana fertility data set from Wooldridge (2010, Chapter 21). Detailed descriptions of the data set are already given in Chapter 2 and thus are omitted here. Before the regression results are presented, it warrants some justification at this point why the data can be analyzed by the CRC model framework. Consider a regression equation as below.

$$children = \beta_0 + b_1 age + \beta_2 agesq + u$$

For the time being let us focus on the motivation of CRC modeling and ignore other covariates that might also be present in the structural equation. The interpretation of β_1 is the marginal birth per year that is determined by the opportunity cost of childbearing (Becker and Barro, 1988; Barro and Becker, 1989). Since education increases human capital, higher education would lead to lower b_1. For now just assume β_2 is constant. Then b_1 is a negative function of years of education. Thus the above equation becomes

$$children = \beta_0 + b_1(educ)age + \beta_2 agesq + u.$$

It is assumed in the above model that *agesq* does not have the CRC. In other words, all the heterogeneity of age are assumed to be absorbed into *age* only. This assumption helps us dispense with age^4 that might possibly cause much trouble in exponential regression. Now suppose *educ* is continuous. Then for each value of *educ*, there are uncountably many

regression equations with constant coefficient. Let us divide the support of *educ* by using threshold 7, i.e. the duration of primary education in Botswana to have a model below.

$$\begin{aligned} children_0 &= \beta_{00} + b_{01}(educ)age + \beta_{02}agesq + u_0 \\ children_1 &= \beta_{10} + b_{11}(educ)age + \beta_{12}agesq + u_1 \end{aligned}$$

The upper equation is defined on $\{\omega | educ(\omega) < 7\}$ whereas the lower one on the complement set.

Next suppose *educ* is not available but only *educ*7 is. The information reduction is done by the following rule.

$$\begin{aligned} educ7 = 1 \quad \text{if} \quad educ > 7 \\ educ7 = 0 \quad \text{if} \quad educ < 7, \end{aligned}$$

where $educ = z\delta + v$. Then the above model can be written as

$$\begin{aligned} children_0 &= \beta_{00} + b_{01}(v)age + \beta_{02}agesq + u_0 \\ children_1 &= \beta_{10} + b_{11}(v)age + \beta_{12}agesq + u_1 \\ educ7 &= 1[z\delta + v > 0] \end{aligned}$$

Considering the nonnegativity of *children*,

$$\begin{aligned} E[children_0|age, u_0] &= \exp(\beta_{00} + b_{01}(v)age + \beta_{02}agesq + u_0) \\ E[children_1|age, u_1] &= \exp(\beta_{10} + b_{11}(v)age + \beta_{12}agesq + u_1) \\ educ7 &= 1[z\delta + v > 0] \end{aligned}$$

It is expected that there are negative correlations between b_{g1} and v; in other words $\sigma_{d_g v}$s in equation (3.3) might be negative. Higher education can reduce fertility either by increasing opportunity cost or by providing people with more information on, say, contraception. Although omitted for simplicity so far, other variables such as *evermarr*, *urban*, and *electric*, which might have CRC's too, will also be included in the following regression. Therefore all the variables but *agesq* have the CRC's and their square and cross product will be included in the estimating equations.

Before running regression using NTCRC model, it might be a good idea to test for endogeneity by VAT. The test statistic for both the regime that follows $\chi^2(10)$ is 46.76 with p-value .0000. The test statistics for Regime 1 and 0 that follow $\chi^2(5)$ are 26.13 with p-value .0001 and 20.64 with p-value .0009. Thus the null hypothesis of no endogeneity is rejected and we will run a regression with full-blown model.

The regression results for LFES and NFES models are summarized in Tables F.3 and F.4. First and second columns in Table F.3 show the results by OLS and 2SLS. Next columns are the ones by LFES estimated by Heckman's correction method. NFES model is estimated by using both Poisson Quasi-ML and NLS methods. The results for NTCRC model are provided separately in Table F.4. Although both NLS and Poisson QMLE is available also for NTCRC, the estimate for NLS does not give a reasonable value; the estimated ATE for

NLS is -5.28 and thus only the QMLE results are listed. All the standard errors in Tables F.3 and F.4 are bootstrap standard errors. The bootstrap ran 500 replications. Among those the bootstrap standard error for ATE in NTCRC model was particularly large. Since the maximum number of children in the data set was 13, it might be highly unlikely that the absolute value of ATE is greater than 13. By that reasoning, all the data with estimated ATE over 13 were discarded to get the bootstrap standard error presented in the table. In Table F.4, each basic covariate is numerically labeled such as age (1), evermarr (2) and so forth. Using that numbers cov(1)=cov(δ_1, v), cov(2)=cov(δ_2, v), and so on.

Let us now see the estimation results for NTCRC model. As it was already expected in theory, there are indeed, although insignificant, negative correlations between the selection error and the coefficients of *age* for both the regimes. The estimated ATE is -1.020 which is very close to the estimates from NFES model in Table F.3. Let us now see the LTCRC estimates in Table F.4. The ATE estimate is substantially large and highly significant, which is indicative of inconsistency of the estimator. Remember that both LTCRC and NTCRC estimating equations are based on the conditional expectation and thus these two estimators cannot be consistent at the same time. As long as we are convinced that the effect of education on fertility is negative, the lesson is that the inconsistency of LTCRC can be substantial, which supports the usefulness of NTCRC model provided in this article.

Finally one can perform the model selection test discussed in Section 5.2. The null hypothesis that this model is in fact NFES is that the coefficients of all the bilateral products as well as the covariances between δ and v are equal to zero. The Wald statistics of this null hypothesis for Regime 1, Regime 0 and both the regimes are 106.74 (df=10), 339.63 (df=10) and 602.32 (df=20) with p-values all equal to .0000, which justifies the model specification as NTCRC.

3.8 Conclusion

We have so far considered the endogenous switching regression where there is a random coefficient correlated with the switching variable under the count dependent variable. The count dependent variable requires a nonlinear modeling using the exponential conditional mean function. Although it is impossible to identify the means of each CRC, we have seen that the ATE, which might be more interesting, can be identified. Simulations show that this ATE estimator by nonlinear two-regime CRC model performs well with large sample size.

APPENDICES

Appendix A

Simulation Results for Series Estimator

Note to Tables A.1-12:

Each table presents the series estimator results for specified correlation and selection error. Simulations were run for $\rho = .0, .4, .5$ and .6. For each correlation, three types of selection error were considered. For a table, each column shows the monotone functions, i.e. identity, inverse Mill's ratio, and normal CDF used for binary choice index. Subcolumns $p = 1, 2$ and 3 specify the degree of polynomial of monotone function. Each row shows the distribution used for u for each sample size.

Table A.1: Simulation results for series estimators with $\rho = 0.0$, ξ: normal

monotone functions			identity			IMR			normal CDF		
degrees of polynomial			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	0.905	0.672	0.184	1.238	-4.003	-4.536	0.002	-9.141	-2.676
		s. d.	2.497	5.768	18.913	13.086	160.863	2144.130	15.011	207.018	2245.681
		RMSE	2.499	5.777	18.931	13.088	160.940	2144.137	15.044	207.266	2245.684
	obs=500	mean	1.009	1.010	1.018	1.104	1.470	2.085	0.956	0.997	0.117
		s. d.	0.177	0.180	0.292	1.760	9.016	40.975	2.212	8.300	46.724
		RMSE	0.177	0.180	0.292	1.763	9.028	40.989	2.213	8.300	46.732
	obs=1000	mean	1.002	1.001	1.002	1.030	0.870	0.758	1.006	0.861	0.650
		s. d.	0.070	0.070	0.072	0.732	2.567	18.714	1.101	2.712	24.053
		RMSE	0.070	0.070	0.072	0.733	2.570	18.715	1.101	2.716	24.056
u: $t(5)$	obs=100	mean	0.890	0.668	-0.235	1.673	-2.652	-66.012	-0.380	-8.688	47.569
		s. d.	2.048	4.759	28.807	22.520	139.004	1954.183	26.305	161.743	1967.634
		RMSE	2.050	4.771	28.833	22.530	139.052	1955.331	26.342	162.033	1968.185
	obs=500	mean	1.000	1.001	1.009	1.121	1.694	1.946	0.941	0.932	0.402
		s. d.	0.188	0.186	0.309	2.398	15.246	38.925	3.741	11.778	44.746
		RMSE	0.188	0.186	0.310	2.401	15.262	38.936	3.742	11.778	44.750
	obs=1000	mean	1.000	1.000	1.000	1.013	0.891	0.751	0.988	0.821	0.650
		s. d.	0.074	0.075	0.077	0.739	2.419	17.300	0.951	2.682	22.353
		RMSE	0.074	0.075	0.077	0.739	2.421	17.302	0.951	2.688	22.356
u: $\chi^2(5)$	obs=100	mean	1.047	1.066	66.547	0.664	-0.076	-16.218	1.412	1.304	23.753
		s. d.	1.665	2.042	1592.114	10.238	30.095	415.140	12.484	26.002	518.243
		RMSE	1.665	2.043	1593.462	10.244	30.114	415.497	12.490	26.004	518.742
	obs=500	mean	1.002	1.000	0.918	0.940	1.242	-2.349	1.040	-4.611	1.482
		s. d.	0.108	0.141	2.064	1.753	4.146	48.550	2.220	142.931	67.768
		RMSE	0.108	0.141	2.066	1.754	4.153	48.666	2.220	143.041	67.770
	obs=1000	mean	0.997	0.997	0.998	1.008	1.062	0.387	0.973	1.011	1.909
		s. d.	0.071	0.072	0.073	0.997	2.668	16.232	1.403	2.848	21.341
		RMSE	0.071	0.072	0.073	0.997	2.669	16.243	1.403	2.848	21.361

Table A.2: Results for $\rho = 0.0$, ξ: $t(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.462	1.470	1.215	0.537	8.813	18.942	-1.648	1.447	-4.180
		s. d.	11.395	11.423	13.376	5.228	122.882	580.978	74.963	83.649	524.571
		RMSE	11.404	11.433	13.378	5.248	123.130	581.255	75.010	83.650	524.597
	obs=500	mean	1.000	1.001	1.000	1.021	1.183	3.112	0.969	1.150	-0.667
		s. d.	0.095	0.095	0.102	0.996	2.973	31.698	1.305	3.011	36.946
		RMSE	0.095	0.095	0.102	0.996	2.979	31.769	1.306	3.015	36.984
	obs=1000	mean	1.001	1.000	1.001	0.966	0.872	0.903	1.043	0.942	0.613
		s. d.	0.064	0.065	0.067	0.563	1.473	9.677	0.586	1.870	13.924
		RMSE	0.064	0.065	0.067	0.564	1.478	9.678	0.588	1.871	13.930
u: $t(5)$	obs=100	mean	1.251	1.118	2.949	0.711	0.669	-0.795	0.975	-0.022	-24.906
		s. d.	3.150	2.484	39.619	4.924	33.028	1066.294	8.052	33.815	2086.952
		RMSE	3.160	2.487	39.667	4.933	33.030	1066.296	8.052	33.830	2087.113
	obs=500	mean	0.959	0.962	0.997	0.991	1.054	2.382	1.036	1.192	2.860
		s. d.	2.136	2.109	1.384	1.261	3.918	28.998	1.589	3.125	106.597
		RMSE	2.136	2.109	1.384	1.261	3.918	29.031	1.589	3.131	106.614
	obs=1000	mean	0.995	0.995	0.995	0.999	1.050	0.995	0.986	1.041	0.938
		s. d.	0.069	0.070	0.072	0.530	1.426	9.878	0.566	1.843	14.212
		RMSE	0.069	0.070	0.073	0.530	1.427	9.878	0.566	1.843	14.213
u: $\chi^2(5)$	obs=100	mean	1.727	1.624	1.502	0.530	-6.811	-15.234	2.146	-2.217	19.497
		s. d.	12.393	10.564	10.555	10.995	173.140	262.276	34.055	154.943	392.817
		RMSE	12.415	10.582	10.567	11.005	173.316	262.778	34.075	154.977	393.252
	obs=500	mean	1.002	1.001	1.000	1.010	0.891	2.720	1.014	0.863	-1.823
		s. d.	0.121	0.125	0.128	0.990	3.382	27.108	1.669	3.076	33.335
		RMSE	0.121	0.125	0.128	0.990	3.384	27.162	1.669	3.079	33.455
	obs=1000	mean	0.996	0.996	0.997	1.011	0.957	0.937	0.983	0.908	0.815
		s. d.	0.069	0.069	0.072	0.543	1.475	9.421	0.562	1.926	13.655
		RMSE	0.069	0.069	0.072	0.543	1.476	9.421	0.562	1.928	13.656

Table A.3: Results for $\rho = 0.0$, ξ: $\chi^2(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	-0.190	-0.139	4.089	0.819	2.004	5.072	1.272	3.022	-14.460
		s. d.	23.632	25.422	114.077	6.847	42.215	831.961	7.037	56.951	903.143
		RMSE	23.662	25.447	114.119	6.849	42.227	831.971	7.042	56.987	903.275
	obs=500	mean	1.027	1.032	1.025	1.048	1.289	2.570	0.956	1.175	-0.249
		s. d.	2.031	2.024	2.064	1.285	5.439	35.175	1.964	4.368	45.661
		RMSE	2.031	2.024	2.064	1.286	5.446	35.210	1.965	4.371	45.678
	obs=1000	mean	0.789	0.808	0.790	0.963	0.900	0.704	1.044	1.072	0.845
		s. d.	5.320	4.891	5.353	0.629	2.695	18.370	0.655	2.589	23.248
		RMSE	5.325	4.894	5.357	0.630	2.696	18.373	0.656	2.590	23.249
u: $t(5)$	obs=100	mean	0.753	0.535	0.048	1.108	-9.782	-7.016	1.195	-9.550	34.038
		s. d.	6.102	7.119	18.836	6.519	175.169	1084.230	8.013	186.925	976.187
		RMSE	6.107	7.134	18.860	6.520	175.501	1084.260	8.016	187.222	976.746
	obs=500	mean	1.027	1.022	1.028	0.984	1.196	-1.794	1.075	1.094	4.281
		s. d.	1.071	1.021	1.048	1.848	5.267	102.992	1.914	4.474	120.464
		RMSE	1.071	1.022	1.048	1.848	5.271	103.030	1.915	4.475	120.508
	obs=1000	mean	0.990	-1.301	-1.298	0.978	1.148	1.030	0.970	1.126	0.977
		s. d.	0.155	56.188	56.188	0.721	2.230	18.927	0.757	2.648	25.814
		RMSE	0.155	56.235	56.235	0.721	2.235	18.927	0.758	2.651	25.814
u: $\chi^2(5)$	obs=100	mean	1.502	1.480	1.448	1.201	0.357	-26.260	0.628	-1.474	14.801
		s. d.	14.174	14.044	16.133	4.734	53.792	469.023	8.254	51.984	753.007
		RMSE	14.183	14.052	16.139	4.738	53.796	469.815	8.262	52.043	753.133
	obs=500	mean	0.846	0.860	1.818	0.880	2.901	-0.413	1.084	3.467	7.182
		s. d.	2.177	2.259	23.846	1.861	49.073	72.928	2.138	53.878	107.651
		RMSE	2.182	2.264	23.860	1.865	49.110	72.941	2.140	53.934	107.828
	obs=1000	mean	0.994	0.998	1.000	0.979	1.057	0.765	0.965	1.162	1.735
		s. d.	0.207	0.199	0.211	0.915	3.777	17.255	0.880	3.305	23.844
		RMSE	0.207	0.199	0.211	0.915	3.777	17.257	0.880	3.309	23.855

Table A.4: Results for $\rho = 0.4$, ξ: normal

monotone functions			identity			IMR			normal CDF		
degrees of polynomial			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.585	1.550	1.850	1.307	-1.812	-244.475	1.860	-0.862	98.108
		s. d.	1.435	1.571	10.140	5.927	63.999	4349.198	8.261	66.486	3190.363
		RMSE	1.550	1.665	10.175	5.935	64.061	4356.120	8.306	66.512	3191.841
	obs=500	mean	1.656	1.656	1.657	0.957	0.705	-0.264	2.376	2.823	4.548
		s. d.	0.211	0.205	0.204	1.796	6.261	32.845	2.813	4.549	40.649
		RMSE	0.689	0.688	0.688	1.796	6.268	32.870	3.131	4.900	40.803
	obs=1000	mean	1.653	1.653	1.653	0.945	0.949	1.195	2.296	2.512	2.830
		s. d.	0.063	0.064	0.067	1.440	5.583	17.378	3.409	4.274	22.629
		RMSE	0.656	0.657	0.657	1.441	5.583	17.379	3.647	4.534	22.703
u: $t(5)$	obs=100	mean	1.602	1.577	1.800	1.335	-1.399	-218.822	1.455	-1.487	78.668
		s. d.	1.279	1.441	7.871	5.523	58.693	3561.923	9.688	62.218	3133.514
		RMSE	1.414	1.553	7.911	5.534	58.742	3568.700	9.699	62.268	3134.476
	obs=500	mean	1.654	1.655	1.655	0.990	0.830	0.255	2.342	2.798	4.030
		s. d.	0.186	0.178	0.181	1.695	5.922	35.032	2.826	4.555	42.870
		RMSE	0.680	0.679	0.680	1.695	5.925	35.040	3.129	4.897	42.977
	obs=1000	mean	1.654	1.655	1.655	0.935	0.943	1.534	2.301	2.504	2.436
		s. d.	0.067	0.068	0.072	1.120	5.478	18.083	3.116	3.948	23.089
		RMSE	0.657	0.658	0.659	1.122	5.478	18.091	3.377	4.225	23.134
u: $\chi^2(5)$	obs=100	mean	1.630	1.647	1.643	0.557	2.528	54.997	1.883	2.775	-51.650
		s. d.	0.669	0.709	0.846	13.044	32.373	1091.634	17.769	32.364	1110.114
		RMSE	0.919	0.960	1.063	13.051	32.409	1092.969	17.791	32.412	1111.362
	obs=500	mean	1.554	1.555	1.552	0.954	0.681	-0.374	2.471	2.601	7.852
		s. d.	1.465	1.469	1.528	1.930	11.416	56.784	5.579	6.516	152.745
		RMSE	1.566	1.570	1.625	1.931	11.421	56.800	5.769	6.710	152.899
	obs=1000	mean	1.617	1.618	1.618	0.888	1.013	-0.488	2.290	2.712	5.325
		s. d.	0.069	0.069	0.071	0.943	2.861	20.391	1.436	2.837	25.032
		RMSE	0.621	0.622	0.622	0.950	2.861	20.445	1.931	3.313	25.402

Table A.5: Results for $\rho = 0.4$, ξ: $t(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.409	2.225	2.288	0.560	0.223	6.892	2.466	3.017	-12.984
		s. d.	2.607	18.243	19.057	5.776	18.087	711.156	8.557	16.688	1016.387
		RMSE	2.639	18.284	19.100	5.793	18.104	711.180	8.682	16.809	1016.483
	obs=500	mean	1.598	1.600	1.600	0.875	0.830	0.825	2.348	2.796	3.758
		s. d.	0.086	0.087	0.092	1.128	2.996	14.825	1.330	2.956	20.996
		RMSE	0.604	0.607	0.607	1.135	3.000	14.826	1.894	3.459	21.176
	obs=1000	mean	1.604	1.606	1.605	0.965	1.015	1.373	2.244	2.647	2.821
		s. d.	0.054	0.055	0.057	0.504	1.392	8.595	0.523	1.852	12.566
		RMSE	0.607	0.609	0.608	0.505	1.392	8.603	1.349	2.478	12.697
u: $t(5)$	obs=100	mean	1.751	1.782	1.720	1.101	7.428	-10.197	1.515	5.061	11.073
		s. d.	15.885	16.054	15.538	5.884	80.305	691.627	9.981	61.300	747.279
		RMSE	15.903	16.073	15.555	5.884	80.562	691.718	9.995	61.435	747.347
	obs=500	mean	1.650	1.652	1.653	0.875	1.210	1.918	2.238	2.803	3.455
		s. d.	0.891	0.873	0.856	1.206	5.961	26.496	3.160	4.053	30.221
		RMSE	1.102	1.090	1.076	1.213	5.965	26.512	3.394	4.436	30.321
	obs=1000	mean	1.602	1.604	1.604	0.945	1.136	1.697	2.249	2.835	3.044
		s. d.	0.067	0.068	0.071	0.565	1.612	10.620	0.583	2.011	15.660
		RMSE	0.605	0.608	0.608	0.568	1.617	10.642	1.379	2.722	15.793
u: $\chi^2(5)$	obs=100	mean	1.774	1.799	2.166	0.825	2.110	-40.554	2.742	5.056	72.784
		s. d.	1.966	2.097	7.903	4.384	50.278	767.518	6.341	58.716	1266.234
		RMSE	2.113	2.244	7.988	4.387	50.291	768.642	6.576	58.856	1268.267
	obs=500	mean	1.696	1.698	1.695	0.893	0.766	0.666	2.476	2.818	3.713
		s. d.	0.115	0.119	0.119	1.105	3.736	15.999	2.016	3.012	21.298
		RMSE	0.705	0.708	0.705	1.111	3.744	16.003	2.499	3.518	21.470
	obs=1000	mean	1.690	1.691	1.690	0.942	0.807	1.360	2.446	2.653	2.657
		s. d.	0.061	0.062	0.063	0.520	1.468	8.368	0.529	1.848	12.220
		RMSE	0.693	0.693	0.693	0.523	1.481	8.376	1.540	2.479	12.332

Table A.6: Results for $\rho = 0.4$, ξ: $\chi^2(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.726	1.225	4.124	1.235	0.229	30.498	2.139	0.775	24.472
		s. d.	3.445	12.631	48.106	5.758	26.793	1313.152	5.254	30.090	1332.741
		RMSE	3.521	12.633	48.208	5.763	26.804	1313.483	5.376	30.091	1332.948
	obs=500	mean	0.807	0.825	0.697	0.860	1.041	-1.822	2.357	2.733	2.109
		s. d.	19.300	19.152	22.329	1.501	7.222	61.044	1.903	4.446	69.006
		RMSE	19.301	19.153	22.331	1.507	7.222	61.109	2.338	4.772	69.015
	obs=1000	mean	1.602	1.606	1.606	0.904	0.390	0.960	2.366	2.528	2.314
		s. d.	0.251	0.248	0.251	0.856	6.148	17.520	0.875	3.334	24.728
		RMSE	0.652	0.654	0.656	0.861	6.178	17.520	1.622	3.667	24.763
u: $t(5)$	obs=100	mean	0.036	0.220	0.271	1.478	0.467	7.669	2.376	4.986	-12.510
		s. d.	34.923	33.714	52.038	5.905	65.448	690.188	9.930	59.551	814.498
		RMSE	34.937	33.723	52.043	5.924	65.451	690.220	10.025	59.684	814.610
	obs=500	mean	1.227	1.252	1.051	0.837	0.913	-0.252	2.428	3.008	5.698
		s. d.	9.042	8.805	10.061	1.600	5.490	38.116	1.710	5.476	45.458
		RMSE	9.045	8.808	10.061	1.608	5.491	38.137	2.228	5.832	45.700
	obs=1000	mean	0.855	0.872	0.830	0.874	0.648	-0.647	2.351	2.774	2.880
		s. d.	18.134	17.918	18.865	0.866	5.677	39.131	0.972	2.808	33.121
		RMSE	18.135	17.918	18.866	0.875	5.688	39.166	1.664	3.322	33.174
u: $\chi^2(5)$	obs=100	mean	1.454	1.521	1.707	1.160	1.187	-6.888	2.169	2.538	-40.546
		s. d.	3.380	3.689	7.703	6.246	21.007	1121.906	6.961	19.943	1167.326
		RMSE	3.411	3.726	7.736	6.248	21.008	1121.934	7.059	20.002	1168.065
	obs=500	mean	1.233	1.248	1.243	0.783	0.877	-1.680	2.413	2.818	10.402
		s. d.	9.881	9.540	9.832	2.428	9.586	50.114	2.987	5.788	135.818
		RMSE	9.884	9.544	9.835	2.437	9.587	50.185	3.304	6.067	136.143
	obs=1000	mean	1.615	1.621	1.627	0.902	0.860	-0.128	2.341	2.694	5.406
		s. d.	0.129	0.137	0.150	0.757	2.054	16.384	0.681	2.356	22.334
		RMSE	0.628	0.636	0.645	0.764	2.059	16.423	1.504	2.902	22.765

Table A.7: Results for $\rho = 0.5$, ξ: normal

monotone functions			identity			IMR			normal CDF		
degrees of polynomial			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.751	1.715	1.997	1.308	-1.748	-223.241	2.225	-0.054	71.763
		s. d.	1.401	1.517	9.293	5.795	66.855	3905.942	8.544	69.193	2887.926
		RMSE	1.590	1.677	9.347	5.803	66.911	3912.374	8.631	69.201	2888.793
	obs=500	mean	1.821	1.822	1.823	0.940	0.681	-0.167	2.715	3.226	4.965
		s. d.	0.222	0.216	0.214	1.869	6.120	32.301	2.960	4.529	40.405
		RMSE	0.851	0.850	0.850	1.870	6.128	32.322	3.421	5.046	40.599
	obs=1000	mean	1.816	1.817	1.817	0.925	0.938	1.166	2.640	2.944	3.502
		s. d.	0.060	0.060	0.064	1.420	5.778	17.438	3.503	4.298	22.576
		RMSE	0.818	0.819	0.819	1.422	5.778	17.439	3.867	4.717	22.714
u: $t(5)$	obs=100	mean	1.771	1.746	1.898	1.353	-1.332	-211.368	1.807	-0.759	61.653
		s. d.	1.184	1.338	5.350	5.525	60.928	3223.614	9.346	64.255	2838.222
		RMSE	1.413	1.532	5.425	5.537	60.973	3230.602	9.381	64.279	2838.870
	obs=500	mean	1.821	1.822	1.822	0.974	0.812	0.415	2.673	3.180	4.326
		s. d.	0.196	0.186	0.189	1.750	5.686	33.790	3.034	4.551	41.661
		RMSE	0.844	0.842	0.843	1.751	5.689	33.795	3.464	5.046	41.794
	obs=1000	mean	1.817	1.818	1.818	0.918	0.939	1.506	2.645	2.941	3.133
		s. d.	0.065	0.066	0.070	1.106	5.655	18.523	3.158	3.929	23.448
		RMSE	0.820	0.821	0.821	1.109	5.655	18.530	3.561	4.382	23.544
u: $\chi^2(5)$	obs=100	mean	3.642	-152.907	-152.886	0.779	3.737	41.465	2.791	7.904	-19.124
		s. d.	54.878	3751.113	3751.115	5.864	64.437	888.309	7.768	78.898	1295.865
		RMSE	54.941	3754.269	3754.270	5.868	64.495	889.230	7.972	79.199	1296.021
	obs=500	mean	1.772	1.772	1.772	0.906	0.239	-61.222	2.842	2.999	29.286
		s. d.	0.104	0.105	0.113	2.089	7.406	1422.404	3.171	5.429	527.206
		RMSE	0.779	0.779	0.781	2.091	7.445	1423.764	3.668	5.785	527.964
	obs=1000	mean	1.765	1.767	1.771	0.933	1.817	-72.234	2.636	3.944	-73.574
		s. d.	0.066	0.071	0.108	0.981	16.911	1492.984	1.243	15.817	1969.699
		RMSE	0.768	0.770	0.779	0.983	16.931	1494.779	2.055	16.088	1971.110

Table A.8: Results for $\rho = 0.5$, ξ: $t(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.306	2.090	2.152	0.556	0.455	-3.359	2.801	3.397	-0.139
		s. d.	7.878	11.556	12.338	5.503	21.766	879.925	8.941	19.344	1156.982
		RMSE	7.883	11.607	12.392	5.521	21.773	879.936	9.121	19.492	1156.983
	obs=500	mean	1.753	1.756	1.757	0.851	0.789	0.739	2.697	3.249	4.499
		s. d.	0.082	0.083	0.088	1.132	3.022	14.768	1.336	2.888	20.764
		RMSE	0.757	0.761	0.762	1.142	3.029	14.770	2.160	3.661	21.057
	obs=1000	mean	1.759	1.762	1.761	0.949	1.016	1.397	2.569	3.082	3.437
		s. d.	0.050	0.051	0.052	0.511	1.368	8.331	0.529	1.809	12.175
		RMSE	0.761	0.764	0.763	0.513	1.368	8.341	1.656	2.758	12.417
u: $t(5)$	obs=100	mean	2.265	2.313	2.238	1.085	0.694	8.374	2.778	3.693	1.471
		s. d.	11.937	12.426	10.230	5.171	26.726	407.675	7.736	20.710	410.407
		RMSE	12.004	12.495	10.305	5.172	26.728	407.741	7.937	20.885	410.407
	obs=500	mean	1.755	1.759	1.764	0.943	1.104	12.207	2.575	3.202	-7.942
		s. d.	0.100	0.104	0.145	1.011	2.692	268.626	1.018	2.993	292.329
		RMSE	0.761	0.766	0.777	1.013	2.694	268.860	1.876	3.716	292.466
	obs=1000	mean	1.755	1.758	1.759	0.913	1.067	14.594	2.587	3.175	-8.899
		s. d.	0.065	0.067	0.069	0.639	2.300	342.987	0.738	2.671	337.136
		RMSE	0.758	0.761	0.762	0.645	2.300	343.256	1.751	3.445	337.281
u: $\chi^2(5)$	obs=100	mean	2.110	2.119	5.009	0.961	2.997	-32.207	2.458	60.007	20.749
		s. d.	6.737	6.601	70.993	4.437	40.355	887.756	5.877	1359.200	476.175
		RMSE	6.828	6.695	71.106	4.437	40.405	888.377	6.055	1360.480	476.584
	obs=500	mean	1.859	1.861	1.859	0.871	0.419	-0.556	2.917	3.209	5.544
		s. d.	0.159	0.167	0.172	1.420	5.204	21.634	1.956	3.350	29.005
		RMSE	0.874	0.877	0.876	1.426	5.236	21.690	2.739	4.013	29.359
	obs=1000	mean	1.862	1.863	1.863	0.944	0.771	0.537	2.800	3.063	4.313
		s. d.	0.058	0.058	0.060	0.558	1.432	7.954	0.603	1.703	11.422
		RMSE	0.864	0.865	0.865	0.561	1.450	7.967	1.898	2.675	11.892

Table A.9: Results for $\rho = 0.5$, ξ: $\chi^2(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.820	1.379	3.770	1.261	0.199	20.936	2.421	1.174	34.170
		s. d.	2.751	11.212	41.059	5.494	26.939	1211.318	5.171	29.922	1349.311
		RMSE	2.871	11.218	41.153	5.500	26.951	1211.482	5.363	29.923	1349.719
	obs=500	mean	1.555	1.577	1.439	0.831	0.947	-1.624	2.706	3.133	3.035
		s. d.	4.775	4.544	7.980	1.605	7.436	54.468	2.089	4.489	60.785
		RMSE	4.807	4.580	7.992	1.614	7.436	54.532	2.697	4.970	60.819
	obs=1000	mean	1.758	1.763	1.763	0.877	0.320	0.833	2.713	2.962	3.072
		s. d.	0.258	0.255	0.258	0.849	6.540	17.694	0.884	3.347	24.876
		RMSE	0.800	0.804	0.806	0.858	6.575	17.695	1.928	3.879	24.962
u: $t(5)$	obs=100	mean	0.254	0.447	1.164	1.498	0.300	0.825	2.665	5.009	-8.262
		s. d.	33.037	31.418	47.442	5.905	60.848	648.799	9.067	55.814	760.331
		RMSE	33.046	31.423	47.442	5.926	60.852	648.799	9.218	55.958	760.387
	obs=500	mean	1.335	1.364	1.227	0.790	0.855	-0.637	2.802	3.505	6.911
		s. d.	10.238	9.983	10.567	1.641	5.729	35.654	1.796	5.696	43.801
		RMSE	10.243	9.990	10.569	1.655	5.731	35.692	2.544	6.223	44.198
	obs=1000	mean	0.997	1.015	0.966	0.845	0.541	-0.616	2.723	3.206	3.572
		s. d.	18.615	18.424	19.523	0.862	5.734	36.839	0.918	2.803	32.939
		RMSE	18.615	18.424	19.524	0.876	5.752	36.874	1.952	3.567	33.039
u: $\chi^2(5)$	obs=100	mean	2.585	2.533	4.956	1.197	3.327	5.477	2.480	4.877	4.803
		s. d.	19.918	19.806	45.317	6.037	52.412	1014.042	6.409	46.194	1058.208
		RMSE	19.981	19.865	45.490	6.040	52.463	1014.052	6.578	46.356	1058.215
	obs=500	mean	1.758	1.760	1.748	0.711	-0.487	-0.813	2.871	2.357	4.825
		s. d.	0.244	0.295	0.476	1.832	21.893	27.023	2.506	21.745	36.775
		RMSE	0.797	0.816	0.886	1.855	21.943	27.084	3.128	21.787	36.973
	obs=1000	mean	1.755	1.764	1.763	0.845	0.883	0.263	2.656	3.040	4.997
		s. d.	0.296	0.300	0.309	1.157	4.033	17.755	1.302	3.414	24.746
		RMSE	0.811	0.820	0.823	1.168	4.035	17.771	2.107	3.977	25.066

Table A.10: Results for $\rho = 0.6$, ξ: normal

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.915	1.879	2.143	1.300	-1.608	-193.518	2.605	0.895	38.808
		s. d.	1.342	1.436	8.169	5.694	69.991	3321.385	9.142	72.123	2551.367
		RMSE	1.624	1.684	8.249	5.702	70.040	3327.076	9.282	72.123	2551.647
	obs=500	mean	1.982	1.983	1.983	0.922	0.673	0.004	3.038	3.603	5.272
		s. d.	0.238	0.232	0.231	1.948	6.047	31.575	3.186	4.544	39.956
		RMSE	1.011	1.010	1.010	1.949	6.055	31.591	3.782	5.237	40.184
	obs=1000	mean	1.974	1.975	1.975	0.906	0.927	1.122	2.973	3.367	4.184
		s. d.	0.055	0.056	0.059	1.366	5.835	17.307	3.507	4.235	22.248
		RMSE	0.975	0.977	0.976	1.369	5.836	17.308	4.025	4.851	22.475
u: $t(5)$	obs=100	mean	1.943	1.918	2.008	1.363	-1.144	-192.521	2.197	0.241	35.246
		s. d.	1.052	1.185	3.545	5.512	64.164	2704.632	9.076	66.997	2508.761
		RMSE	1.412	1.499	3.686	5.524	64.200	2711.547	9.154	67.002	2508.995
	obs=500	mean	1.985	1.986	1.987	0.958	0.819	0.686	2.987	3.526	4.452
		s. d.	0.225	0.214	0.215	1.815	5.594	31.594	3.362	4.624	39.343
		RMSE	1.010	1.009	1.010	1.816	5.597	31.596	3.905	5.269	39.494
	obs=1000	mean	1.976	1.977	1.977	0.902	0.933	1.423	2.984	3.378	3.910
		s. d.	0.063	0.064	0.067	1.058	5.529	18.686	3.014	3.744	23.460
		RMSE	0.978	0.980	0.979	1.062	5.529	18.691	3.608	4.435	23.639
u: $\chi^2(5)$	obs=100	mean	1.955	1.922	2.132	1.121	-4.785	4.184	3.445	0.366	17.115
		s. d.	0.764	1.123	4.645	5.532	152.666	782.850	14.623	152.234	1219.060
		RMSE	1.223	1.453	4.781	5.534	152.775	782.856	14.826	152.236	1219.167
	obs=500	mean	1.917	1.923	1.923	0.840	0.659	-8.181	3.224	3.797	23.476
		s. d.	0.217	0.250	0.253	2.353	9.886	348.054	3.657	8.948	401.758
		RMSE	0.942	0.956	0.957	2.358	9.892	348.175	4.280	9.375	402.387
	obs=1000	mean	1.917	1.918	1.918	0.861	0.827	-0.399	3.029	3.584	9.358
		s. d.	0.066	0.066	0.069	0.819	3.603	32.080	1.385	3.346	87.408
		RMSE	0.919	0.920	0.920	0.831	3.607	32.111	2.457	4.228	87.807

Table A.11: Results for $\rho = 0.6$, ξ: $t(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.446	2.157	2.219	0.562	0.498	-11.457	3.171	3.872	11.510
		s. d.	7.859	9.976	10.750	5.293	23.983	1027.405	9.272	20.461	1310.067
		RMSE	7.871	10.043	10.819	5.311	23.988	1027.481	9.523	20.662	1310.109
	obs=500	mean	1.912	1.916	1.917	0.828	0.735	0.709	3.057	3.702	5.147
		s. d.	0.077	0.078	0.083	1.160	3.163	14.463	1.432	2.830	20.154
		RMSE	0.915	0.920	0.921	1.172	3.174	14.466	2.507	3.912	20.576
	obs=1000	mean	1.918	1.921	1.920	0.931	1.013	1.409	2.904	3.528	4.083
		s. d.	0.045	0.045	0.047	0.524	1.344	8.027	0.541	1.754	11.680
		RMSE	0.919	0.922	0.921	0.529	1.344	8.038	1.980	3.077	12.080
u: $t(5)$	obs=100	mean	1.143	1.088	1.189	0.728	2.240	3.905	2.845	4.253	51.089
		s. d.	19.862	21.975	19.205	4.783	18.476	358.296	6.649	16.360	1135.995
		RMSE	19.863	21.976	19.206	4.791	18.518	358.307	6.901	16.680	1137.099
	obs=500	mean	1.913	1.916	1.916	0.772	0.933	1.523	3.125	3.963	4.964
		s. d.	0.121	0.121	0.126	1.499	4.518	25.017	2.401	4.036	31.489
		RMSE	0.921	0.924	0.925	1.516	4.518	25.022	3.207	5.007	31.737
	obs=1000	mean	1.900	1.903	1.903	0.903	0.862	0.986	2.893	3.412	4.494
		s. d.	0.065	0.065	0.068	0.739	5.396	20.919	0.767	5.591	21.474
		RMSE	0.902	0.905	0.905	0.746	5.398	20.919	2.042	6.089	21.756
u: $\chi^2(5)$	obs=100	mean	1.994	2.570	-14.227	0.941	1.218	-17.108	3.017	20.454	20.701
		s. d.	0.596	13.838	396.369	4.994	16.735	495.433	6.849	414.109	535.525
		RMSE	1.159	13.926	396.661	4.994	16.736	495.763	7.140	414.566	535.888
	obs=500	mean	2.033	2.035	2.036	0.898	0.822	0.082	3.162	3.603	5.496
		s. d.	0.092	0.095	0.098	1.498	3.789	15.135	2.345	3.285	21.537
		RMSE	1.037	1.039	1.040	1.501	3.793	15.163	3.190	4.191	22.001
	obs=1000	mean	2.035	2.053	2.052	0.925	0.960	0.085	3.161	3.725	6.038
		s. d.	0.056	0.410	0.410	0.563	4.975	25.166	0.584	5.150	44.941
		RMSE	1.036	1.130	1.129	0.568	4.975	25.183	2.238	5.827	45.223

Table A.12: Results for $\rho = 0.6$, ξ: $\chi^2(5)$

monotone functions degrees of polynomial			identity			IMR			normal CDF		
			$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=3$
u: normal	obs=100	mean	1.978	1.641	3.899	1.297	0.203	6.196	2.702	1.417	45.284
		s. d.	3.035	8.456	40.539	4.568	25.212	1097.418	4.901	27.891	1357.101
		RMSE	3.189	8.480	40.642	4.578	25.224	1097.430	5.188	27.894	1357.823
	obs=500	mean	2.358	2.384	2.252	0.803	0.832	-1.308	3.067	3.521	4.062
		s. d.	10.869	11.195	7.897	1.721	7.599	46.393	2.277	4.513	48.607
		RMSE	10.954	11.280	7.996	1.732	7.601	46.450	3.076	5.170	48.704
	obs=1000	mean	1.923	1.930	1.930	0.849	0.265	0.721	3.080	3.415	3.858
		s. d.	0.248	0.244	0.247	0.818	6.557	17.481	0.911	3.255	24.531
		RMSE	0.956	0.962	0.962	0.832	6.598	17.483	2.270	4.052	24.697
u: $t(5)$	obs=100	mean	0.571	0.746	2.299	1.476	0.254	-5.936	2.959	5.113	-3.693
		s. d.	27.921	26.659	39.932	5.926	55.876	643.289	8.291	50.636	729.581
		RMSE	27.925	26.660	39.953	5.946	55.881	643.326	8.519	50.803	729.596
	obs=500	mean	1.463	1.494	1.460	0.745	0.801	-1.061	3.182	4.001	8.114
		s. d.	11.088	10.818	10.984	1.689	6.117	32.627	1.899	6.071	41.818
		RMSE	11.098	10.830	10.993	1.708	6.120	32.692	2.893	6.772	42.419
	obs=1000	mean	1.182	1.200	1.145	0.819	0.442	-0.535	3.106	3.633	4.312
		s. d.	18.243	18.064	19.343	0.861	5.515	32.703	0.908	2.765	32.058
		RMSE	18.244	18.065	19.344	0.880	5.543	32.739	2.293	3.817	32.229
u: $\chi^2(5)$	obs=100	mean	-7.626	-2.993	-5.618	1.078	0.476	-17.813	2.942	6.034	18.191
		s. d.	237.699	251.249	264.834	10.765	109.312	826.419	13.383	120.664	608.739
		RMSE	237.856	251.281	264.917	10.765	109.313	826.633	13.523	120.769	608.982
	obs=500	mean	1.856	1.885	1.881	0.700	0.331	0.582	3.173	4.010	6.456
		s. d.	0.886	0.909	0.893	2.492	8.895	31.604	2.064	6.382	48.811
		RMSE	1.232	1.269	1.254	2.510	8.920	31.606	2.997	7.057	49.115
	obs=1000	mean	1.895	1.923	1.922	0.789	0.782	1.053	3.073	3.790	5.538
		s. d.	0.163	0.191	0.211	0.928	2.266	15.294	0.850	2.507	25.215
		RMSE	0.910	0.942	0.946	0.951	2.277	15.294	2.241	3.751	25.620

Appendix B

Simulation Results for All Other Estimators

Note to Tables B.1-18:

Each table presents the all the results but series estimator for specified correlation and IV. Simulations were run for $\rho = .0, .4, .5$ and .6. For a table, each column specifies the estimator categorized under different types of distributions for u. Each row shows the distribution used for selection error ξ for each sample size. M, S, R, Md denote mean, Monte Carlo standard deviation, RMSE and median respectively.

Table B.1: Simulation Results: Strong IV and $\rho = 0.0$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	0.803	0.976	0.809	0.135	0.975	4.440	1.126	1.124	0.060	1.006
		S	18.556	7.140	7.539	18.502	1.431	108.561	6.028	6.146	26.560	1.546
		R	18.557	7.140	7.542	18.522	1.431	108.616	6.030	6.148	26.577	1.546
		Md	1.022	1.006	0.935	0.921	0.926	0.988	1.004	0.972	0.938	0.985
	500	M	1.040	1.046	1.041	1.019	1.030	1.102	1.034	1.026	1.024	1.032
		S	1.327	1.528	1.264	3.349	0.601	2.413	1.092	1.213	3.539	0.877
		R	1.328	1.529	1.265	3.349	0.602	2.415	1.092	1.214	3.539	0.877
		Md	1.004	1.010	1.002	1.022	1.017	1.064	1.023	1.035	0.991	1.014
	1000	M	1.012	1.015	1.011	0.974	0.995	1.002	1.015	1.010	0.993	1.017
		S	0.681	0.662	0.649	1.367	0.497	0.752	0.669	0.664	1.198	0.911
		R	0.681	0.662	0.650	1.367	0.497	0.752	0.669	0.664	1.198	0.911
		Md	1.024	1.045	1.032	0.995	1.006	1.011	1.012	1.020	1.002	0.995
ξ: $\chi^2(5)$	100	M	0.732	1.955	0.909	0.032	1.027	-2.554	2.144	1.092	1.623	0.986
		S	14.543	4.979	6.455	23.840	1.449	99.600	5.654	5.975	24.003	1.537
		R	14.546	5.069	6.455	23.860	1.450	99.663	5.769	5.975	24.011	1.537
		Md	1.005	1.396	1.003	0.960	1.007	1.346	1.422	1.056	1.096	0.835
	500	M	1.036	1.603	0.985	0.909	0.991	0.893	1.555	0.966	0.936	0.732
		S	1.398	1.394	1.302	3.765	0.573	1.455	1.386	1.085	2.806	0.981
		R	1.399	1.519	1.303	3.766	0.573	1.459	1.493	1.086	2.807	1.017
		Md	1.078	1.399	1.008	0.957	1.011	1.104	1.363	0.936	1.012	0.759
	1000	M	1.052	1.411	1.003	0.975	1.021	1.017	1.390	0.965	0.892	0.689
		S	0.631	0.665	0.618	1.788	0.499	0.708	0.686	0.624	1.756	0.925
		R	0.633	0.781	0.618	1.788	0.499	0.708	0.789	0.625	1.759	0.976
		Md	1.048	1.350	1.033	1.019	1.032	1.116	1.310	0.932	0.985	0.631

Table B.2: Results for Strong IV and $\rho = 0.4$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	1.259	1.361	1.195	-1.082	1.294	0.320	1.171	0.847	-1.217	1.205
		S	24.316	9.359	8.948	25.431	1.307	13.649	4.706	6.192	23.647	1.444
		R	24.318	9.366	8.950	25.516	1.340	13.666	4.709	6.194	23.751	1.458
		Md	1.203	1.218	1.194	1.136	1.280	1.202	1.250	1.211	1.192	1.208
	500	M	0.816	0.895	0.881	0.626	1.245	0.745	0.936	0.881	0.671	0.925
		S	1.742	1.583	1.353	3.221	0.548	3.649	1.729	1.310	2.687	0.822
		R	1.751	1.586	1.358	3.243	0.600	3.658	1.730	1.315	2.707	0.825
		Md	1.051	1.060	1.080	1.031	1.237	1.000	1.026	1.007	1.031	1.048
	1000	M	0.951	0.967	0.960	0.900	1.059	0.930	0.943	0.918	0.816	0.602
		S	0.690	0.663	0.648	0.928	0.619	0.815	0.678	0.695	1.557	0.827
		R	0.691	0.663	0.649	0.933	0.621	0.818	0.680	0.700	1.568	0.918
		Md	1.032	1.018	1.028	0.980	1.088	0.975	0.998	0.978	1.011	0.211
ξ: $\chi^2(5)$	100	M	5.185	2.227	1.156	-1.879	1.325	1.384	2.491	1.365	-0.404	1.175
		S	71.770	4.895	5.504	24.664	1.381	43.334	8.577	9.176	25.317	1.555
		R	71.892	5.046	5.506	24.832	1.418	43.336	8.705	9.183	25.356	1.565
		Md	1.229	1.432	1.113	0.969	1.255	1.239	1.566	1.196	1.046	1.105
	500	M	0.687	1.502	0.854	0.740	1.249	0.918	1.525	0.899	0.667	0.805
		S	6.310	1.580	1.430	2.620	0.529	2.085	1.318	1.405	3.649	0.866
		R	6.318	1.658	1.437	2.633	0.585	2.086	1.419	1.408	3.664	0.887
		Md	1.102	1.393	1.022	1.008	1.195	1.129	1.409	1.062	0.996	0.757
	1000	M	1.046	1.367	0.971	0.902	1.139	1.021	1.323	0.937	0.761	0.482
		S	0.639	0.616	0.646	1.067	0.539	0.631	0.601	0.671	1.858	0.842
		R	0.640	0.717	0.647	1.071	0.557	0.631	0.683	0.674	1.874	0.988
		Md	1.090	1.333	0.996	0.944	1.114	1.058	1.317	0.997	0.960	0.075

Table B.3: Results for Strong IV and $\rho = 0.5$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	0.979	1.357	1.158	-1.361	1.374	0.695	1.498	0.608	-0.633	1.361
		S	24.720	9.161	8.891	28.617	1.273	37.964	9.773	15.743	23.151	1.361
		R	24.720	9.168	8.892	28.714	1.327	37.965	9.786	15.748	23.208	1.408
		Md	1.251	1.233	1.254	1.128	1.365	1.214	1.269	1.162	1.312	1.274
	500	M	0.775	0.869	0.853	0.621	1.280	0.681	0.901	0.829	0.567	0.998
		S	1.864	1.589	1.413	2.870	0.512	3.592	2.078	2.201	3.691	0.664
		R	1.878	1.595	1.421	2.895	0.583	3.606	2.081	2.207	3.717	0.664
		Md	1.052	1.075	1.087	1.024	1.243	1.016	1.062	1.012	0.999	0.652
	1000	M	0.935	0.952	0.945	0.877	1.154	0.864	0.921	0.877	0.808	0.707
		S	0.715	0.670	0.663	0.900	0.393	1.030	0.704	0.778	1.624	0.469
		R	0.718	0.671	0.665	0.909	0.422	1.039	0.708	0.787	1.635	0.553
		Md	1.030	1.025	1.028	0.975	1.117	1.011	1.019	0.993	1.004	0.518
ξ: $\chi^2(5)$	100	M	4.799	2.193	1.128	-0.612	1.395	0.647	1.756	0.866	-1.120	1.169
		S	59.319	4.692	5.354	33.681	1.303	23.916	6.539	8.393	24.404	1.459
		R	59.440	4.841	5.355	33.720	1.362	23.918	6.583	8.394	24.496	1.468
		Md	1.265	1.483	1.165	1.153	1.306	1.215	1.569	1.227	1.263	1.017
	500	M	0.648	1.464	0.819	0.909	1.290	0.853	1.459	0.875	0.826	0.769
		S	7.227	1.549	1.501	3.030	0.490	3.762	1.598	1.431	4.111	0.799
		R	7.236	1.617	1.512	3.031	0.570	3.765	1.663	1.436	4.115	0.832
		Md	1.093	1.390	1.025	1.009	1.232	1.053	1.334	0.949	0.982	0.396
	1000	M	1.031	1.342	0.950	0.850	1.190	0.968	1.317	0.898	0.835	0.509
		S	0.644	0.591	0.651	1.291	0.388	0.759	0.695	0.747	1.346	0.628
		R	0.645	0.683	0.653	1.300	0.432	0.760	0.764	0.754	1.356	0.797
		Md	1.098	1.324	0.992	0.966	1.121	1.044	1.310	0.982	0.962	0.221

Table B.4: Results for Strong IV and $\rho = 0.6$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	0.646	1.340	1.101	-0.581	1.407	1.536	1.481	0.884	0.024	1.361
		S	25.625	8.788	8.717	20.334	1.251	31.611	10.474	6.833	20.848	1.361
		R	25.628	8.794	8.717	20.395	1.316	31.615	10.485	6.834	20.871	1.983
		Md	1.325	1.329	1.292	1.166	1.368	1.235	1.251	1.217	1.244	1.274
	500	M	0.736	0.844	0.827	0.595	1.295	0.819	0.918	0.799	0.573	0.998
		S	1.988	1.600	1.475	2.543	0.476	2.794	1.623	1.607	4.106	0.664
		R	2.005	1.608	1.485	2.575	0.560	2.800	1.625	1.620	4.128	0.440
		Md	1.068	1.073	1.065	1.028	1.267	1.017	1.068	1.017	1.044	0.652
	1000	M	0.920	0.938	0.931	0.897	1.137	0.894	0.941	0.841	0.844	0.555
		S	0.749	0.686	0.688	1.090	0.360	0.903	0.740	0.902	1.573	0.767
		R	0.754	0.689	0.691	1.095	0.385	0.909	0.742	0.916	1.580	0.786
		Md	1.027	1.021	1.016	0.988	1.084	1.021	1.032	0.972	0.991	0.506
ξ: $\chi^2(5)$	100	M	4.452	2.216	1.145	-0.533	1.484	-2.554	2.671	0.619	0.710	1.243
		S	48.623	4.660	4.814	26.714	1.288	99.600	9.893	17.531	27.404	1.626
		R	48.746	4.816	4.817	26.758	1.375	99.663	10.033	17.535	27.406	1.644
		Md	1.324	1.550	1.199	1.295	1.383	1.346	1.790	1.232	1.231	1.030
	500	M	0.628	1.435	0.791	0.899	1.323	0.893	1.501	0.740	0.653	0.769
		S	7.991	1.548	1.562	2.946	0.463	1.455	0.999	1.624	4.086	0.753
		R	8.000	1.608	1.576	2.948	0.565	1.459	1.118	1.644	4.101	0.620
		Md	1.085	1.357	1.018	1.036	1.255	1.104	1.489	0.945	0.942	0.476
	1000	M	1.014	1.319	0.930	0.827	1.189	1.017	1.435	0.887	0.747	0.504
		S	0.652	0.566	0.659	1.155	0.389	0.708	0.606	0.744	1.709	0.575
		R	0.652	0.649	0.663	1.168	0.433	0.708	0.746	0.752	1.728	0.576
		Md	1.088	1.326	0.987	0.937	1.100	1.116	1.431	0.978	0.894	0.335

Table B.5: Results for Weak IV and $\rho = 0.0$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	0.980	10.114	0.983	0.691	0.963	1.018	-1.436	1.014	2.490	0.998
		S	0.638	172.074	0.636	32.420	1.096	0.640	93.693	0.630	24.057	1.007
		R	0.638	172.315	0.637	32.421	1.096	0.640	93.724	0.630	24.103	1.007
		Md	0.976	0.981	1.002	1.021	0.964	0.999	1.048	0.984	0.853	0.961
	500	M	0.995	0.857	0.994	1.253	0.992	1.017	2.092	1.018	0.988	0.981
		S	0.545	6.247	0.537	6.483	0.735	0.524	77.712	0.518	5.271	0.582
		R	0.545	6.249	0.537	6.488	0.735	0.524	77.720	0.518	5.271	0.583
		Md	0.990	0.979	0.993	0.974	0.956	1.005	1.016	1.004	1.190	0.993
	1000	M	0.976	0.923	0.973	0.906	0.963	1.014	0.762	1.010	1.053	1.057
		S	0.453	1.501	0.450	2.831	0.569	0.434	6.072	0.429	2.668	0.803
		R	0.453	1.503	0.451	2.832	0.571	0.434	6.077	0.429	2.669	0.805
		Md	0.967	0.939	0.973	0.972	0.968	1.001	0.995	1.005	1.038	1.008
ξ: $\chi^2(5)$	100	M	1.035	-10.261	1.069	0.005	0.995	0.978	-3.914	1.009	0.948	0.910
		S	0.613	265.733	0.614	24.815	1.094	0.630	223.518	0.623	30.667	0.983
		R	0.614	265.971	0.618	24.835	1.094	0.631	223.572	0.624	30.667	0.987
		Md	1.031	1.047	1.068	0.946	0.976	0.991	0.842	1.012	1.211	0.874
	500	M	1.013	0.934	1.150	1.232	0.987	1.033	0.716	1.164	1.076	0.938
		S	0.547	8.995	0.548	5.896	0.707	0.537	13.047	0.535	6.180	0.664
		R	0.547	8.995	0.568	5.900	0.707	0.538	13.050	0.559	6.180	0.667
		Md	0.996	0.954	1.128	0.958	0.992	1.009	1.006	1.143	1.101	1.034
	1000	M	1.046	1.113	1.231	1.156	1.012	1.031	1.052	1.222	0.791	0.728
		S	0.452	5.328	0.460	3.455	0.553	0.457	5.238	0.453	3.327	0.825
		R	0.454	5.329	0.515	3.459	0.553	0.458	5.238	0.505	3.333	0.869
		Md	1.037	1.065	1.196	1.109	1.004	1.029	1.032	1.202	0.874	0.748

Table B.6: Results for Weak IV and $\rho = 0.4$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	1.586	2.836	1.588	-3.068	1.586	1.536	-2.592	1.539	-2.606	1.433
		S	0.633	35.243	0.606	20.161	1.022	0.597	240.707	0.592	25.049	0.941
		R	0.863	35.291	0.844	20.568	1.178	0.802	240.734	0.800	25.307	1.035
		Md	1.578	1.438	1.587	0.224	1.575	1.508	1.283	1.523	-0.266	1.391
	500	M	1.400	0.037	1.398	-0.585	1.330	1.383	0.670	1.387	-1.625	1.192
		S	0.526	24.103	0.521	5.314	0.677	0.515	9.581	0.513	6.592	0.778
		R	0.661	24.123	0.655	5.546	0.753	0.642	9.586	0.643	7.095	0.801
		Md	1.407	1.138	1.404	0.250	1.307	1.401	1.116	1.407	-0.122	1.394
	1000	M	1.268	1.130	1.264	-0.057	1.206	1.277	1.006	1.276	-0.315	0.899
		S	0.433	6.115	0.428	2.832	0.527	0.440	3.613	0.436	3.314	0.767
		R	0.509	6.117	0.503	3.023	0.566	0.520	3.613	0.516	3.565	0.774
		Md	1.282	1.014	1.274	0.566	1.167	1.287	1.019	1.280	0.286	1.103
ξ: $\chi^2(5)$	100	M	1.587	2.641	1.617	-3.805	1.549	1.586	1.557	1.610	-1.096	1.523
		S	0.594	34.063	0.579	23.632	1.033	0.633	51.806	0.638	40.577	1.066
		R	0.835	34.103	0.846	24.116	1.170	0.863	51.809	0.883	40.631	1.187
		Md	1.588	1.471	1.615	0.115	1.491	1.578	1.350	1.612	0.331	1.510
	500	M	1.432	1.638	1.540	-0.501	1.398	1.399	-0.065	1.508	-0.406	0.957
		S	0.530	23.301	0.532	5.135	0.675	0.528	25.444	0.529	7.720	0.871
		R	0.684	23.310	0.758	5.350	0.783	0.662	25.467	0.733	7.847	0.872
		Md	1.417	1.135	1.524	0.315	1.321	1.433	1.117	1.519	-0.074	1.102
	1000	M	1.310	0.994	1.460	0.279	1.230	1.295	0.700	1.455	-0.084	0.635
		S	0.428	8.085	0.426	2.615	0.501	0.435	9.323	0.435	3.917	0.793
		R	0.529	8.085	0.627	2.713	0.551	0.525	9.328	0.629	4.065	0.873
		Md	1.310	1.092	1.444	0.578	1.152	1.320	1.063	1.453	0.400	0.483

Table B.7: Results for Weak IV and $\rho = 0.5$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	1.729	3.130	1.734	-3.201	1.721	1.686	-2.573	1.685	-0.896	1.595
		S	0.592	62.997	0.588	21.170	0.998	0.600	182.319	0.589	28.523	0.930
		R	0.939	63.033	0.941	21.583	1.231	0.911	182.354	0.904	28.586	1.104
		Md	1.720	1.555	1.741	0.232	1.623	1.704	1.564	1.713	0.460	1.571
	500	M	1.497	0.508	1.495	-0.280	1.388	1.488	0.541	1.496	-0.625	1.209
		S	0.514	10.178	0.510	5.184	0.649	0.520	7.570	0.518	8.002	0.771
		R	0.715	10.190	0.710	5.339	0.756	0.713	7.584	0.717	8.166	0.799
		Md	1.498	1.152	1.494	0.463	1.333	1.505	1.132	1.515	0.151	1.350
	1000	M	1.335	0.838	1.332	0.350	1.229	1.349	0.900	1.358	0.042	0.894
		S	0.428	3.494	0.423	2.425	0.492	0.427	5.431	0.423	3.894	0.708
		R	0.544	3.498	0.537	2.510	0.543	0.552	5.432	0.554	4.010	0.716
		Md	1.353	1.009	1.350	0.683	1.165	1.354	1.038	1.361	0.425	0.463
ξ: $\chi^2(5)$	100	M	1.726	3.822	1.754	-2.987	1.692	1.747	1.056	1.771	1.409	1.639
		S	0.573	169.952	0.566	27.520	1.022	0.593	17.854	0.592	37.053	1.009
		R	0.925	169.976	0.942	27.807	1.234	0.954	17.854	0.972	37.055	1.194
		Md	1.731	1.506	1.753	0.771	1.623	1.740	1.615	1.747	0.973	1.594
	500	M	1.501	4.126	1.608	0.403	1.417	1.492	0.421	1.599	0.790	0.972
		S	0.506	77.381	0.493	5.296	0.652	0.545	8.400	0.546	8.326	0.828
		R	0.712	77.444	0.782	5.329	0.774	0.734	8.420	0.811	8.328	0.829
		Md	1.494	1.124	1.598	0.572	1.332	1.515	1.099	1.615	0.501	0.860
	1000	M	1.366	1.141	1.509	0.702	1.244	1.354	1.039	1.500	0.905	0.714
		S	0.410	4.313	0.402	2.793	0.472	0.417	2.638	0.415	4.161	0.726
		R	0.550	4.315	0.648	2.809	0.532	0.547	2.639	0.650	4.162	0.780
		Md	1.369	1.063	1.499	0.707	1.172	1.364	1.075	1.501	0.567	0.336

Table B.8: Results for Weak IV and $\rho = 0.6$

	obs		u: normal					u: $\chi^2(5)$				
			IV	Heckit1	Heckit2	QLIML	QFIML	IV	Heckit1	Heckit2	QLIML	QFIML
ξ: normal	100	M	1.870	2.304	1.837	-0.322	1.848	1.836	1.918	1.837	1.730	1.695
		S	0.571	13.195	0.612	16.511	0.969	0.617	44.461	0.612	27.031	0.948
		R	1.041	13.259	1.037	16.563	1.288	1.039	44.470	1.037	27.041	1.176
		Md	1.879	1.702	1.826	0.885	1.716	1.837	1.665	1.826	0.965	1.635
	500	M	1.590	1.115	1.589	0.222	1.421	1.618	0.601	1.628	1.247	1.251
		S	0.498	6.704	0.495	5.268	0.625	0.513	21.976	0.511	7.572	0.750
		R	0.773	6.705	0.769	5.325	0.754	0.803	21.979	0.810	7.576	0.791
		Md	1.602	1.155	1.590	0.724	1.388	1.621	1.231	1.633	0.939	1.279
	1000	M	1.401	0.892	1.397	0.529	1.231	1.441	1.013	1.452	0.990	0.979
		S	0.421	2.709	0.416	2.149	0.458	0.431	4.355	0.429	3.972	0.668
		R	0.581	2.711	0.575	2.200	0.513	0.617	4.355	0.623	3.972	0.668
		Md	1.423	1.008	1.420	0.776	1.136	1.448	1.057	1.461	0.718	0.587
ξ: $\chi^2(5)$	100	M	1.890	0.524	1.910	1.180	1.804	1.922	2.613	1.963	1.934	1.755
		S	0.557	120.541	0.545	15.472	0.942	0.617	86.358	0.618	37.442	1.112
		R	1.050	120.542	1.060	15.473	1.238	1.110	86.373	1.145	37.454	1.344
		Md	1.894	1.771	1.902	1.495	1.724	1.936	1.728	1.993	2.310	1.700
	500	M	1.593	1.180	1.696	1.464	1.472	1.610	0.280	1.805	3.106	1.000
		S	0.501	12.399	0.492	4.972	0.636	0.491	14.177	0.574	8.526	0.871
		R	0.777	12.401	0.852	4.994	0.792	0.783	14.196	0.989	8.783	0.871
		Md	1.612	1.144	1.703	1.079	1.357	1.612	1.118	1.813	1.150	0.624
	1000	M	1.431	0.902	1.570	0.895	1.244	1.437	0.645	1.645	1.035	0.674
		S	0.404	1.596	0.389	2.999	0.422	0.426	3.557	0.422	5.245	0.683
		R	0.591	1.599	0.690	3.000	0.487	0.610	3.574	0.771	5.246	0.756
		Md	1.441	1.066	1.566	0.797	1.152	1.472	1.031	1.652	0.564	0.371

Appendix C

Other Tables in Chapter 1

Table C.1: Relative Efficiencies (Ratio of MSEs)

			Heckit1 vs LIV		Heckit2 vs QLIML		Heckit2 vs Series		Heckit2 vs QFIML	
ρ	ξ	obs.	u: normal	u: $\chi^2(5)$	u: normal	u: $\chi^2(5)$	u: normal	u: $\chi^2(5)$	u: normal	u: $\chi^2(5)$
0	normal	100	6.03	18.69	3.01	2.78	0.04	0.06	6.75	324.48
		500	7.01	8.51	1.94	2.09	0.23	0.52	0.75	4.89
		1000	4.43	3.26	1.27	2.26	0.59	1.89	1.06	1.26
	$\chi^2(5)$	100	13.66	16.15	1.13	0.63	0.05	0.07	8.23	298.49
		500	8.36	6.69	0.97	2.95	0.19	0.88	0.85	0.95
		1000	8.36	7.92	1.04	2.14	0.65	2.44	0.66	0.81
0.4	normal	100	8.13	14.70	0.44	4.44	0.02	0.06	6.74	8.42
		500	5.70	4.24	1.75	2.16	0.20	0.49	1.22	4.47
		1000	2.07	5.02	4.93	1.84	0.92	1.72	1.09	1.44
	$\chi^2(5)$	100	20.34	7.62	1.10	0.46	0.07	0.03	202.95	24.78
		500	3.36	6.77	1.10	3.00	0.17	0.40	14.52	2.16
		1000	2.74	7.72	1.77	1.28	0.74	2.15	0.80	0.86
0.5	normal	100	10.43	2.17	0.43	0.14	0.02	0.01	7.27	15.05
		500	4.15	2.83	1.73	0.90	0.17	0.09	1.39	3.00
		1000	1.87	4.31	4.57	1.56	0.40	0.49	1.14	2.15
	$\chi^2(5)$	100	39.64	8.52	1.05	0.52	0.06	0.03	150.73	13.20
		500	4.02	8.20	1.14	1.67	0.14	0.34	20.02	5.13
		1000	3.96	3.23	1.72	2.40	0.44	1.12	0.89	0.99
0.6	normal	100	5.47	9.33	0.43	0.66	0.02	0.08	8.49	9.09
		500	3.01	6.49	1.72	2.12	0.14	0.07	1.56	2.97
		1000	2.51	2.98	3.92	0.82	0.31	0.74	1.20	1.50
	$\chi^2(5)$	100	30.86	2.44	0.90	0.38	0.08	0.01	102.44	98.67
		500	3.50	6.22	1.21	2.33	0.13	0.14	24.76	1.70
		1000	3.11	5.28	1.57	1.60	0.43	0.59	1.01	0.90

Table C.2: Ratio of estimations with $\widehat{Avar}_{QLIML} < \widehat{Avar}_{Two-step}$

ρ	ξ	obs.	u: normal	u: $t(5)$	u: $\chi^2(5)$
0.0	normal	100	0.54	0.56	0.53
		500	0.54	0.52	0.53
		1000	0.52	0.52	0.54
	ξ: $t(5)$	100	0.53	0.52	0.53
		500	0.57	0.55	0.56
		1000	0.56	0.55	0.53
	$\chi^2(5)$	100	0.55	0.56	0.56
		500	0.58	0.58	0.57
		1000	0.55	0.56	0.54
0.4	normal	100	0.48	0.48	0.51
		500	0.53	0.53	0.56
		1000	0.53	0.54	0.54
	ξ: $t(5)$	100	0.53	0.53	0.51
		500	0.52	0.54	0.51
		1000	0.53	0.54	0.58
	$\chi^2(5)$	100	0.54	0.52	0.55
		500	0.56	0.58	0.59
		1000	0.55	0.58	0.55
0.5	normal	100	0.48	0.48	0.51
		500	0.53	0.55	0.54
		1000	0.54	0.55	0.56
	ξ: $t(5)$	100	0.52	0.52	0.51
		500	0.57	0.54	0.54
		1000	0.56	0.56	0.54
	$\chi^2(5)$	100	0.53	0.56	0.51
		500	0.58	0.58	0.56
		1000	0.56	0.58	0.61
0.6	normal	100	0.48	0.48	0.54
		500	0.54	0.55	0.54
		1000	0.56	0.56	0.57
	ξ: $t(5)$	100	0.52	0.54	0.53
		500	0.56	0.54	0.56
		1000	0.57	0.57	0.58
	$\chi^2(5)$	100	0.53	0.55	0.55
		500	0.59	0.59	0.57
		1000	0.60	0.60	0.59

Appendix D

Figures in Chapter 1

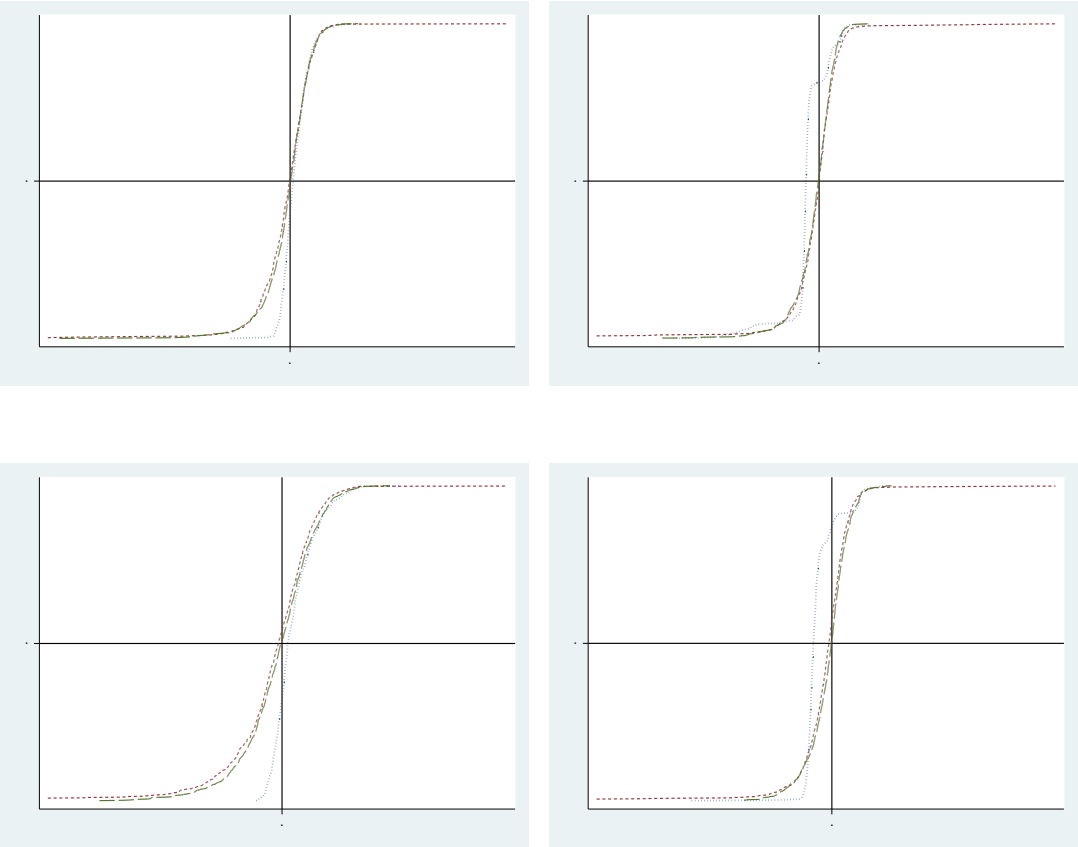

Figure D.1: Sample Cumulative Density Functions under Strong IV with $\rho = 0.6$ and obs=1000.

Note: The upper row is for normal ξ and lower row is for chi-squared ξ. The left column is for normal u and right column is for chi-squared u. The dotted, short-dashed and long-dashed line represent FIML, LIML and Heckit estimators respectively.

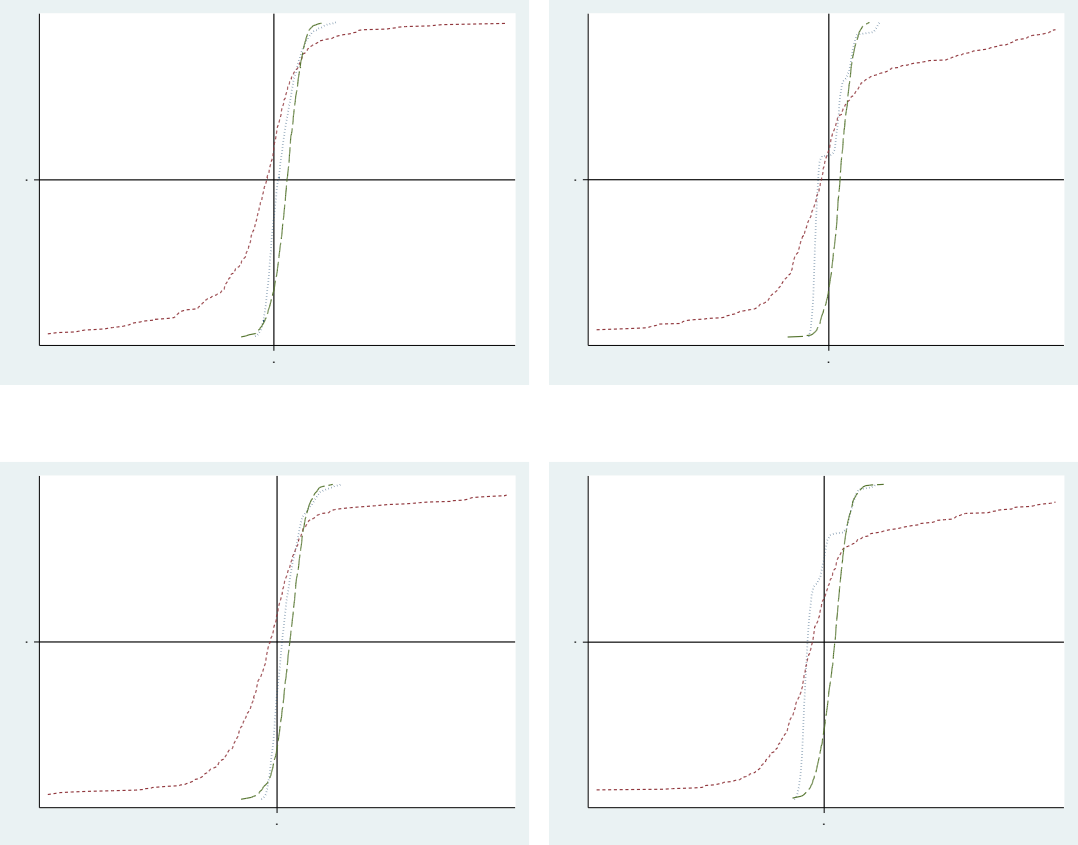

Figure D.2: Sample Cumulative Density Functions under Weak IV with $\rho = 0.6$ and obs=1000.

Note: Graphs are placed by same way as in Figure D.1.

Appendix E

Proofs in Chapter 2

E.1 Proof of Proposition 1

Although the basic proof is presented in Terza (1998), the following generalizes it to the two regime setting.

Lemma 1

$$\int_{-\mathbf{z}\gamma}^{\infty} g(v|\epsilon)dv = F\left[\frac{\mathbf{z}\gamma + (\sigma_{1a}/\sigma_1^2)\epsilon}{\sqrt{1-(\sigma_{1a}/\sigma_1)^2}}\right]$$

Proof First imagine the joint density function of v and ϵ. Then the expression is the probability of $v \in [-\mathbf{z}\gamma, \infty)$, i.e. $d = 1$ conditional on ϵ.
A linear projection can be written as

$$v = \frac{\sigma_{1a}}{\sigma_1^2}\epsilon + e \qquad \text{with} \quad e \sim N\left(0, 1 - \left(\frac{\sigma_{1a}}{\sigma_1}\right)^2\right)$$

Then

$$\begin{aligned} w &= 1[\mathbf{z}\gamma + (\sigma_{1a}/\sigma_1^2)\epsilon + e > 0] \\ &= 1[e > -\mathbf{z}\gamma - (\sigma_{1a}/\sigma_1^2)\epsilon] \\ &= 1\left[\frac{e}{\sqrt{1-(\sigma_{1a}/\sigma_1)^2}} > \frac{-\mathbf{z}\gamma - (\sigma_{1a}/\sigma_1^2)\epsilon}{\sqrt{1-(\sigma_{1a}/\sigma_1)^2}}\right] \end{aligned}$$

Therefore

$$\mathbf{P}(w=1|\epsilon) = \int_{-\mathbf{z}\gamma}^{\infty} g(v|\epsilon)dv = F\left[\frac{\mathbf{z}\gamma + (\sigma_{1a}/\sigma_1^2)\epsilon}{\sqrt{1-(\sigma_{1a}/\sigma_1)^2}}\right] \equiv \Phi^*(\epsilon)$$

as was to be shown. ■

Proposition The joint density function of y and d conditional on the exogenous variables

is

$$f(y,w|\mathbf{z}) = \left[\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\sqrt{2}\sigma_1\zeta_1)\Phi^*(\sqrt{2}\sigma_1\zeta_1)\ \exp(-\zeta_1^2)d\zeta_1\right]^w$$
$$\cdot\left[\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\sqrt{2}\sigma_0\zeta_0)\Big(1-\Phi^*(\sqrt{2}\sigma_0\zeta_0)\Big)\exp(-\zeta_0^2)d\zeta_0\right]^{1-w}$$

Proof Note that

$$\begin{aligned} f(\epsilon_1,w=1) &= \int_{-\mathbf{z}\gamma}^{\infty} f(\epsilon_1,v)dv \\ f(\epsilon_1,w=0) &= \int_{-\infty}^{-\mathbf{z}\gamma} f(\epsilon_1,v)dv \\ f(\epsilon_0,w=1) &= \int_{-\mathbf{z}\gamma}^{\infty} f(\epsilon_0,v)dv \\ f(\epsilon_0,w=0) &= \int_{-\infty}^{-\mathbf{z}\gamma} f(\epsilon_0,v)dv \end{aligned}$$

Let $f(\cdot|\mathbf{z}) \equiv g(\cdot)$.

$$\begin{aligned} \int_{\mathbb{R}} g(y_0|w=0,\epsilon_0)\Big(\int_{-\infty}^{-\mathbf{z}\gamma} g(\epsilon_0,v)\ dv\Big)\ d\epsilon_0 &= \int_{\mathbb{R}} g(y_0|w=0,\epsilon_0)\cdot g(\epsilon_0,w=0)\ d\epsilon_0 \\ &= \int_{\mathbb{R}} g(y_0,w=0,\epsilon_0)\ d\epsilon_0 \\ &= g(y_0,w=0) \end{aligned}$$

$$\begin{aligned} \int_{\mathbb{R}} g(y_1|w=1,\epsilon_1)\Big(\int_{-\mathbf{z}\gamma}^{\infty} g(\epsilon_1,v)\ dv\Big)\ d\epsilon_1 &= \int_{\mathbb{R}} g(y_1|w=1,\epsilon_1)\cdot g(\epsilon_1,w=1)\ d\epsilon_1 \\ &= \int_{\mathbb{R}} g(y_1,w=1,\epsilon_1)\ d\epsilon_1 \\ &= g(y_1,w=1) \end{aligned}$$

Thus we have

$$\begin{aligned} g(y,w) &\equiv g(y_1,w=1)^w\cdot g(y_0,w=0)^{1-w} \\ &= \left(\int_{\mathbb{R}} g(y_1|w=1,\epsilon_1)\Big(\int_{-\mathbf{z}\gamma}^{\infty} g(\epsilon_1,v)\ dv\Big)\ d\epsilon_1\right)^w \\ &\quad\times\left(\int_{\mathbb{R}} g(y_0|w=0,\epsilon_0)\Big(\int_{-\infty}^{-\mathbf{z}\gamma} g(\epsilon_0,v)\ dv\Big)\ d\epsilon_0\right)^{1-w} \end{aligned}$$

Recovering the original notation

$$
\begin{aligned}
f(y,w|\mathbf{z}) &= \left[\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\epsilon_1)\Big(\int_{-\mathbf{z}\gamma}^{\infty} f(\epsilon_1,v|\mathbf{z})\ dv\Big)\ d\epsilon_1\right]^{w} \\
&\qquad \cdot \left[\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\epsilon_0)\Big(\int_{-\infty}^{-\mathbf{z}\gamma} f(\epsilon_0,v|\mathbf{z})\ dv\Big)\ d\epsilon_0\right]^{1-w} \\
&= \left[\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\epsilon_1)\Big(\int_{-\mathbf{z}\gamma}^{\infty} f(\epsilon_1|\mathbf{z})f(v|\epsilon_1,\mathbf{z})\ dv\Big)\ d\epsilon_1\right]^{w} \\
&\qquad \cdot \left[\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\epsilon_0)\Big(\int_{-\infty}^{-\mathbf{z}\gamma} f(\epsilon_0|\mathbf{z})f(v|\epsilon_0,\mathbf{z})dv\Big)\ d\epsilon_0\right]^{1-w} \\
&= \left[\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\epsilon_1)f(\epsilon_1|\mathbf{z})\Big(\int_{-\mathbf{z}\gamma}^{\infty} f(v|\epsilon_1,\mathbf{z})\ dv\Big)\ d\epsilon_1\right]^{w} \\
&\qquad \cdot \left[\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\epsilon_0)f(\epsilon_0|\mathbf{z})\Big(\int_{-\infty}^{-\mathbf{z}\gamma} f(v|\epsilon_0,\mathbf{z})dv\Big)\ d\epsilon_0\right]^{1-w} \\
&= \left[\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\epsilon_1)f(\epsilon_1|\mathbf{z})\Phi^*(\epsilon_1)\ d\epsilon_1\right]^{w} \\
&\qquad \cdot \left[\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\epsilon_0)f(\epsilon_0|\mathbf{z})\Big(1-\Phi^*(\epsilon_0)\Big)d\epsilon_0\right]^{1-w}
\end{aligned}
$$

Let

$$\zeta \equiv \frac{\epsilon}{\sqrt{2}\sigma}, \qquad \epsilon \sim N(0,\sigma^2) \tag{E.1}$$

Then

$$
\begin{aligned}
f(y,w|\mathbf{z}) &= \left[\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} f(y_1|\mathbf{z},w=1,\sqrt{2}\sigma_1\zeta_1)\Phi^*(\sqrt{2}\sigma_1\zeta_1)\ \exp(-\zeta_1^2)d\zeta_1\right]^{w} \\
&\qquad \times \left[\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} f(y_0|\mathbf{z},w=0,\sqrt{2}\sigma_0\zeta_0)\Big(1-\Phi^*(\sqrt{2}\sigma_0\zeta_0)\Big)\exp(-\zeta_0^2)d\zeta_0\right]^{1-w}
\end{aligned}
$$

as was to be shown. ■

E.2 Derivation of Estimating Equation for NFES Model

In what follows the regime subscripts were omitted for simplicity. For each regime

$$\begin{aligned} E(y|\mathbf{z},v) &= \exp(\alpha+\mathbf{x}\beta)E[\exp(\epsilon)|\mathbf{z},v] \\ &= \exp(\alpha+\mathbf{x}\beta)\exp\Big(\rho\sigma v+\frac{1}{2}\sigma^2(1-\rho^2)\Big) \\ &= \exp\left(\alpha+\frac{1}{2}\sigma^2(1-\rho^2)+\mathbf{x}\beta+\rho\sigma v\right) \end{aligned}$$

The second equation derived by the result in Appendix A in Terza (1998).
Note that for regime 1,

$$\begin{aligned} E[\exp(\rho\sigma v)|v>-\mathbf{z}\delta] &= \int_{-\mathbf{z}\delta}^{\infty}\exp(\rho\sigma v)p(v|v>-\mathbf{z}\delta)dv \\ &= \int_{-\mathbf{z}\delta}^{\infty}\exp(\rho\sigma v)\frac{p(v)}{\mathbf{P}(v>-\mathbf{z}\delta)}dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\int_{-\mathbf{z}\delta}^{\infty}\exp(\rho\sigma v)p(v)dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\int_{-\mathbf{z}\delta}^{\infty}\exp(\rho\sigma v)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{v^2}{2}\Big)dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\int_{-\mathbf{z}\delta}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{2\rho\sigma v-v^2}{2}\Big)dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\int_{-\mathbf{z}\delta}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{(\rho\sigma)^2}{2}\Big)\exp\Big(-\frac{(\rho\sigma)^2}{2}\Big)\exp\Big(\frac{2\rho\sigma v-v^2}{2}\Big)dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\exp\Big(\frac{(\rho\sigma)^2}{2}\Big)\int_{-\mathbf{z}\delta}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{(v-\rho\sigma)^2}{2}\Big)dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\exp\Big(\frac{(\rho\sigma)^2}{2}\Big)\int_{-(\mathbf{z}\delta+\rho\sigma)}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{v^2}{2}\Big)dv \\ &= \frac{1}{\Phi(\mathbf{z}\delta)}\exp\Big(\frac{(\rho\sigma)^2}{2}\Big)\Phi(\mathbf{z}\delta+\rho\sigma) \end{aligned}$$

By the same reasoning it can be shown that for regime 0,

$$E[\exp(\rho\sigma v)|v<-\mathbf{z}\delta]=\frac{1}{\Phi(-\mathbf{z}\delta)}\exp\Big(\frac{(\rho\sigma)^2}{2}\Big)\Phi(-(\mathbf{z}\delta+\rho\sigma))$$

Since $\mathcal{G}(\mathbf{z},v)\subset\mathcal{G}(\mathbf{z},w)$, by law of iterated expectation,

$$\begin{aligned} E(y_1|\mathbf{z},w) &= \exp\left(\alpha_1+\frac{1}{2}\sigma_1^2(1-\rho_1^2)+\mathbf{x}\beta_1\right)E[\exp(\rho_1\sigma_1 v)|\mathbf{z},w] \\ E(y_1|\mathbf{z},w=1) &= \exp\left(\alpha_1+\frac{1}{2}\sigma_1^2(1-\rho_1^2)+\mathbf{x}\beta_1\right)E[\exp(\rho_1\sigma_1 v)|v>-\mathbf{z}\delta] \\ &= \exp\Big(\alpha_1+\frac{\sigma_1^2}{2}+\mathbf{x}\beta_1\Big)\frac{\Phi(\mathbf{z}\delta+\rho_1\sigma_1)}{\Phi(\mathbf{z}\delta)} \end{aligned}$$

Similarly

$$\begin{aligned} E(y_0|\mathbf{z}, w=0) &= \exp\left(\alpha_0 + \frac{1}{2}\sigma_0^2(1-\rho_0^2) + \mathbf{x}\beta_0\right) E[\exp(\rho_0\sigma_0 v)|v < -\mathbf{z}\delta] \\ &= \exp\left(\alpha_0 + \frac{\sigma_0^2}{2} + \mathbf{x}\beta_0\right) \frac{\Phi(-(\mathbf{z}\delta + \rho_0\sigma_0))}{\Phi(-\mathbf{z}\delta)} \end{aligned}$$

Therefore the estimating equation is,

$$\begin{aligned} E(y|\mathbf{z}, w) &= w \cdot \left[\exp\left(\alpha_1 + \frac{\sigma_1^2}{2} + \mathbf{x}\beta_1\right) \frac{\Phi(\mathbf{z}\delta + \rho_1\sigma_1)}{\Phi(\mathbf{z}\delta)}\right] \\ &\quad + (1-w) \cdot \left[\exp\left(\alpha_0 + \frac{\sigma_0^2}{2} + \mathbf{x}\beta_0\right) \frac{\Phi(-(\mathbf{z}\delta + \rho_0\sigma_0))}{\Phi(-\mathbf{z}\delta)}\right], \end{aligned}$$

where $y = wy_1 + (1-w)y_0$.

E.3 Derivation of Conditional Variance for WNLS Estimator

It was shown in Appendix A. in Terza (1998) that

$$\begin{aligned} E[\exp(c_g)|v] &= \exp\left(\rho_g\sigma_g v + \frac{1}{2}\sigma_g^2(1-\rho_g^2)\right) \\ \text{var}[\exp(c_g)|v] &= \exp\left(2\rho_g\sigma_g v + 2\sigma_g^2(1-\rho_g^2)\right) - \exp\left(2\rho_g\sigma_g v + \sigma_g^2(1-\rho_g^2)\right) \end{aligned}$$

Let all the expectations below are conditional on z. Then

$$\begin{aligned} E[\exp(c_g)^2|w] &= E\Big[E[\exp(c_g)^2|v]\Big|w\Big] \\ &= E\Big[\text{var}[\exp(c_g)|v]\Big|w\Big] + E\Big[E[\exp(c_g)|v]^2\Big|w\Big] \end{aligned}$$

Thus we need to find out the expressions for the two terms on the RHS.

$$\begin{aligned} E[\text{var}(\exp(c_g)|v)|w] &= E\left[\exp\left(2\rho_g\sigma_g v + 2\sigma_g^2(1-\rho_g^2)\right) - \exp\left(2\rho_g\sigma_g v + \sigma_g^2(1-\rho_g^2)\right)\Big|w\right] \\ &= \exp\left[2\sigma_g^2(1-\rho_g^2)\right] \cdot E\left[\exp\left(2\rho_g\sigma_g v\right)|w\right] \\ &\qquad - \exp\left[\sigma_g^2(1-\rho_g^2)\right] \cdot E\left[\exp\left(2\rho_g\sigma_g v\right)|w\right] \\ &= \left(\exp\left[2\sigma_g^2(1-\rho_g^2)\right] - \exp\left[\sigma_g^2(1-\rho_g^2)\right]\right) \\ &\qquad \times E\left[\exp\left(2\rho_g\sigma_g v\right)|w\right] \end{aligned} \tag{E.2}$$

Also

$$\begin{aligned} E[E(\exp(c_g)|v)^2|w] &= E\Big[\exp\Big(2\rho_g\sigma_g v + \sigma_g^2(1-\rho_g^2)\Big)\Big|w\Big] \\ &= \exp\Big(\sigma_g^2(1-\rho_g^2)\Big)\cdot E\big[\exp(2\rho_g\sigma_g v)|w\big] \end{aligned}$$

Therefore

$$E[\exp(c_g)^2|w] = \exp\big[2\sigma_g^2(1-\rho_g^2)\big]\cdot E\big[\exp(2\rho_g\sigma_g v)|w\big]$$

By the way

$$\begin{aligned} E[\exp(c_g)|w] &= E\big[E(\exp(c_g)|v)\big|w\big] \\ &= E\Big[\exp\Big(\rho_g\sigma_g v + \frac{1}{2}\sigma_g^2(1-\rho_g^2)\Big)\Big|w\Big] \\ &= \exp\Big(\frac{1}{2}\sigma_g^2(1-\rho_g^2)\Big)\cdot E\big[\exp(\rho_g\sigma_g v)|w\big] \end{aligned}$$

Therefore

$$\begin{aligned} \mathrm{var}[\exp(c_g)|w] &= E[\exp(c_g)^2|w] - E[\exp(c_g)|w]^2 \\ &= \exp\big[2\sigma_g^2(1-\rho_g^2)\big]\cdot E\big[\exp(2\rho_g\sigma_g v)|w\big] \\ &\qquad - \exp\Big(\sigma_g^2(1-\rho_g^2)\Big)\cdot E\big[\exp(\rho_g\sigma_g v)|w\big]^2 \end{aligned}$$

From the results in Appendix A,

$$\begin{aligned} \mathrm{var}[\exp(\epsilon_1)|w=1] &= \exp(\sigma_1^2)\Big\{\exp(\sigma_1^2)\frac{\Phi(z\delta+2\rho_1\sigma_1)}{\Phi(z\delta)} - \Big(\frac{\Phi(z\delta+\rho_1\sigma_1)}{\Phi(z\delta)}\Big)^2\Big\} \\ \mathrm{var}[\exp(\epsilon_0)|w=0] &= \exp(\sigma_0^2)\Big\{\exp(\sigma_0^2)\frac{\Phi(-z\delta-2\rho_0\sigma_0)}{\Phi(-z\delta)} - \Big(\frac{\Phi(-z\delta-\rho_0\sigma_0)}{\Phi(-z\delta)}\Big)^2\Big\} \end{aligned}$$

Also

$$\begin{aligned} \mathrm{var}\Big[E[y|z,w,\epsilon_1,\epsilon_0]\Big|z,w\Big] &= \mathrm{var}\Big[w\exp(\alpha_1+x\beta_1+\epsilon_1)+(1-w)\exp(\alpha_0+x\beta_0+\epsilon_0)\Big|z,w\Big] \\ &= w\exp(\alpha_1+x\beta_1)^2\mathrm{var}[\exp(\epsilon_1)|z,w] \\ &\quad +(1-w)\exp(\alpha_0+x\beta_0)^2\mathrm{var}[\exp(\epsilon_0)|z,w] \\ &\quad +2w(1-w)\exp(\alpha_1+x\beta_1)\exp(\alpha_0+x\beta_0) \\ &\qquad\qquad \times \mathrm{cov}[\exp(\epsilon_1),\exp(\epsilon_0)|z,w] \\ &= w\exp(\alpha_1+x\beta_1)^2\mathrm{var}[\exp(\epsilon_1)|z,w=1] \\ &\qquad + (1-w)\exp(\alpha_0+x\beta_0)^2\mathrm{var}[\exp(\epsilon_0)|z,w=0] \end{aligned}$$

$$\begin{aligned} &= w\exp(\alpha_1+\sigma_1^2/2+x\beta_1)^2\Big\{\exp(\sigma_1^2)\frac{\Phi(z\delta+2\rho_1\sigma_1)}{\Phi(z\delta)} - \Big(\frac{\Phi(z\delta+\rho_1\sigma_1)}{\Phi(z\delta)}\Big)^2\Big\} \\ &\quad +(1-w)\exp(\alpha_0+\sigma_0^2/2+x\beta_0)^2\Big\{\exp(\sigma_0^2)\frac{\Phi(-z\delta-2\rho_0\sigma_0)}{\Phi(-z\delta)} - \Big(\frac{\Phi(-z\delta-\rho_0\sigma_0)}{\Phi(-z\delta)}\Big)^2\Big\} \end{aligned}$$

Let $\delta_g = \exp(\alpha_g + \sigma_g^2/2 + x\beta_g)$, $L_{1,2} = \dfrac{\Phi(z\delta + 2\rho_1\sigma_1)}{\Phi(z\delta)}$, $L_1 = \dfrac{\Phi(z\delta + \rho_1\sigma_1)}{\Phi(z\delta)}$, $L_{0,2} = \dfrac{\Phi(-z\delta - 2\rho_0\sigma_0)}{\Phi(-z\delta)}$, and $L_0 = \dfrac{\Phi(-z\delta - \rho_0\sigma_0)}{\Phi(-z\delta)}$. Then the last equation can be simply written as

$$\mathrm{var}\Big[E[y|z,w,\epsilon_1,\epsilon_0]\Big|z,w\Big] = w\delta_1^2\Big(\exp(\sigma_1^2)L_{1,2} - L_1^2\Big) + (1-w)\delta_0^2\Big(\exp(\sigma_0^2)L_{0,2} - L_0^2\Big)$$

And

$$\begin{aligned}
\mathrm{var}[y|z,w,\epsilon_1,\epsilon_0] &= \mathrm{var}[wy_1 + (1-w)y_0|z,w,\epsilon_1,\epsilon_0] \\
&= w\mathrm{var}[y_1|z,w,\epsilon_1,\epsilon_0] + (1-w)w\mathrm{var}[y_0|z,w,\epsilon_1,\epsilon_0] \\
&\qquad + 2w(1-w)\mathrm{cov}[y_1,y_0|z,w,\epsilon_1,\epsilon_0] \\
&= wE[y_1|z,w,\epsilon_1] + (1-w)E[y_0|z,w,\epsilon_0]
\end{aligned}$$

Taking $E[\cdot|z,w]$ at both sides,

$$\begin{aligned}
E\big[\mathrm{var}[y|z,w,\epsilon_1,\epsilon_0]\big|z,w\big] &= wE[y_1|z,w=1] + (1-w)E[y_0|z,w=0] \\
&= w\delta_1 L_1 + (1-w)\delta_0 L_0
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathrm{var}[y|z,w] &= \mathrm{var}\Big[E[y|z,w,\epsilon_1,\epsilon_0]\Big|z,w\Big] + E\big[\mathrm{var}[y|z,w,\epsilon_1,\epsilon_0]\big|z,w\big] \\
&= w\delta_1\Big(\delta_1(\exp(\sigma_1^2)L_{1,2} - L_1^2) + L_1\Big) + (1-w)\delta_0\Big(\delta_0(\exp(\sigma_0^2)L_{0,2} - L_0^2) + L_0\Big)
\end{aligned}$$

Appendix F

Tables in Chapter 2

Table F.1: Simulation Results for $rho = 0.4$

		Linear Models			NET (1 Regime Nonlinear)			
		2SLS	LET(Hkt)	LFES(Hkt)	1PQML	2PQML	NLS	WNLS
n=1000	mean	0.726	0.742	0.747	-18813768	-1037.586	-6220.699	7.289E+13
	mc. st. dev	2.463	2.448	2.404	378300000	30871.585	186544.680	2.055E+15
	RMSE	2.478	2.462	2.417	378800000	30889.051	186648.410	2.056E+15
	median	0.825	0.845	0.855	-91.134	-0.296	0.747	1.177
	MAD	1.453	1.446	1.444	94.588	3.387	2.257	1.757
n=3000	mean	0.804	0.820	0.818	-2389.450	-4.619	-2.516	-0.347
	mc. st. dev	1.255	1.254	1.256	19778.908	16.242	17.923	18.548
	RMSE	1.270	1.267	1.269	19922.838	17.187	18.264	18.597
	median	0.814	0.804	0.809	-187.915	-0.239	0.616	0.270
	MAD	0.809	0.798	0.806	188.648	1.881	1.407	1.208
n=5000	mean	0.892	0.904	0.904	-904.755	-1.712	-0.857	-0.340
	mc. st. dev	0.980	0.982	0.982	5179.628	6.637	10.024	3.833
	RMSE	0.986	0.987	0.986	5258.226	7.170	10.194	4.061
	median	0.907	0.913	0.906	-220.859	0.033	0.495	0.065
	MAD	0.647	0.630	0.628	192.470	1.388	1.034	0.981

Note: The true ATE=1, The number of Monte Carlo repetition is 1000. LET(Hkt): LET model estimated by two-step Heckit, 1PQML: one step Poisson Quasi-Maximum Likelihood estimator, RMSE=root mean squared error, MAD=mean absolute deviation

Table F.1 (cont'd)

		NFES (2 Regime Nonlinear)			
		1PQML	2 PQML	NLS	WNLS
n=1000	mean	-0.528	-0.084	-3.555	-0.100
	mc. st. dev	17.753	14.587	105.187	18.089
	RMSE	17.818	14.628	105.286	18.122
	median	0.909	0.984	1.096	0.971
	MAD	1.132	1.160	1.198	1.138
n=3000	mean	0.856	0.885	0.879	0.926
	mc. st. dev	1.487	1.322	1.444	1.298
	RMSE	1.494	1.327	1.449	1.300
	median	0.934	0.937	1.004	0.958
	MAD	0.679	0.674	0.706	0.649
n=5000	mean	0.983	0.980	0.973	0.996
	mc. st. dev	0.933	0.935	0.987	0.912
	RMSE	0.933	0.935	0.988	0.912
	median	1.027	1.022	1.001	1.044
	MAD	0.539	0.540	0.594	0.528

Table F.2: Simulation Results for $rho = 0.5$

		Linear Models			NET (1 Regime Nonlinear)			
		2SLS	LET(Hkt)	LFES(Hkt)	1PQML	2PQML	NLS	WNLS
n=1000	mean	0.620	0.634	0.635	-2.097E+11	-107.190	-70.525	1.E+07
	mc. st. dev	2.512	2.517	2.522	6.902E+12	1678.496	1008.897	3.E+08
	RMSE	2.541	2.544	2.549	6.905E+12	1681.979	1011.429	3.E+08
	median	0.784	0.820	0.789	-90.775	-0.391	0.779	0.878
	MAD	1.467	1.457	1.461	93.716	3.289	2.149	1.821
n=3000	mean	0.748	0.763	0.761	-4252.613	-7.376	-3.185	-1.604
	mc. st. dev	1.254	1.255	1.256	56506.153	37.336	18.650	17.555
	RMSE	1.279	1.277	1.279	56666.026	38.264	19.113	17.747
	median	0.735	0.770	0.772	-151.285	-0.702	0.292	0.028
	MAD	0.809	0.803	0.803	133.619	2.212	1.452	1.330
n=5000	mean	0.835	0.847	0.847	-622.799	-2.529	-1.083	0.221
	mc. st. dev	0.986	0.988	0.988	2763.448	8.182	10.301	26.549
	RMSE	0.999	1.000	1.000	2832.979	8.910	10.510	26.561
	median	0.881	0.888	0.888	-191.688	-0.260	0.517	-0.121
	MAD	0.638	0.640	0.631	131.812	1.558	1.084	1.112

Table F.2 (cont'd)

		NFES (2 Regime Nonlinear)			
		1PQML	2 PQML	NLS	WNLS
n=1000	mean	-4.675	-4.115	-19.700	-1.291
	mc. st. dev	101.905	79.327	383.617	28.446
	RMSE	102.063	79.491	384.175	28.538
	median	0.942	1.002	1.141	1.010
	MAD	1.071	1.128	1.189	1.121
n=3000	mean	0.811	0.820	0.777	0.852
	mc. st. dev	1.364	1.337	1.494	1.320
	RMSE	1.377	1.349	1.510	1.328
	median	0.932	0.936	0.950	0.954
	MAD	0.651	0.646	0.721	0.625
n=5000	mean	0.935	0.940	0.937	0.954
	mc. st. dev	0.952	0.946	1.004	0.920
	RMSE	0.954	0.948	1.006	0.921
	median	1.016	1.017	1.017	1.029
	MAD	0.550	0.546	0.559	0.536

Table F.3: Simulation Results for $rho = 0.6$

		Linear Models			NET (1 Regime Nonlinear)			
		2SLS	LET(Hkt)	LFES(Hkt)	1PQML	2PQML	NLS	WNLS
n=1000	mean	0.589	0.610	0.615	-3178.000	-83.600	-430.3381	2933.343
	mc. st. dev	2.444	2.424	2.417	10050.000	828.368	7701.3678	80317.211
	RMSE	2.478	2.455	2.447	10050.000	832.677	7713.4375	80370.723
	median	0.743	0.725	0.761	-73.208	-0.760	0.50252408	0.818
	MAD	1.500	1.500	1.483	76.071	3.636	2.3827269	1.897
n=3000	mean	0.690	0.703	0.703	-1579.442	-9.535	-5.6418819	326.938
	mc. st. dev	1.269	1.269	1.267	17617.457	43.824	29.77737	9798.644
	RMSE	1.306	1.304	1.301	17688.204	45.073	30.509119	9804.063
	median	0.724	0.723	0.727	-134.971	-1.165	0.33116622	-0.149
	MAD	0.830	0.843	0.846	112.208	2.613	1.4962865	1.456
n=5000	mean	0.777	0.787	0.788	-411.791	-3.600	-1.1125125	-1.055
	mc. st. dev	1.012	1.015	1.015	1144.600	8.520	7.3739959	5.262
	RMSE	1.036	1.037	1.037	1216.761	9.683	7.6706274	5.649
	median	0.830	0.835	0.855	-158.018	-0.705	0.53132847	-0.315
	MAD	0.655	0.658	0.656	104.547	1.930	1.0969538	1.192

Table F.3 (cont'd)

		NFES (2 Regime Nonlinear)			
		1PQML	2 PQML	NLS	WNLS
n=1000	mean	-2.296	-362.847	-4313.352	-143.539
	mc. st. dev	39.603	11411.335	126080.480	4209.908
	RMSE	39.740	11417.134	126154.270	4212.388
	median	0.983	0.998	1.170	1.188
	MAD	1.113	1.165	1.176	1.150
n=3000	mean	0.717	0.725	0.670	0.782
	mc. st. dev	1.650	1.595	1.842	1.517
	RMSE	1.674	1.618	1.871	1.533
	median	0.945	0.943	0.989	0.974
	MAD	0.685	0.688	0.706	0.664
n=5000	mean	0.901	0.907	0.893	0.923
	mc. st. dev	0.976	0.973	1.054	0.954
	RMSE	0.981	0.978	1.059	0.957
	median	0.975	0.986	0.993	1.007
	MAD	0.556	0.547	0.596	0.539

Table F.4: Simulation Results for FIML estimator

	correlation	0.4		0.5		0.6	
	# of abscissas	8	16	8	16	8	16
n=1000	mean	2.143	1.336	2.365	1.395	2.511	1.422
	mc. st. dev	1.342	1.525	1.325	1.524	1.630	1.606
	RMSE	1.763	1.561	1.902	1.574	2.222	1.661
	median	2.014	1.430	2.246	1.523	2.376	1.593
	MAD	0.737	0.725	0.726	0.747	0.749	0.843
n=3000	mean	1.888	1.348	2.234	1.557	2.302	1.670
	mc. st. dev	1.048	0.975	1.121	1.032	1.074	0.904
	RMSE	1.374	1.035	1.667	1.173	1.687	1.125
	median	1.880	1.394	2.141	1.586	2.271	1.633
	MAD	0.605	0.488	0.547	0.513	0.550	0.462
n=5000	mean	1.837	1.460	2.215	1.650	2.254	1.656
	mc. st. dev	0.916	0.967	0.903	0.853	0.920	0.906
	RMSE	1.241	1.071	1.514	1.073	1.556	1.118
	median	1.822	1.471	2.195	1.641	2.197	1.617
	MAD	0.507	0.486	0.552	0.403	0.511	0.544

Table F.5: Simulation Results for DGP 1

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.4$	n=1000	mean	0.622	0.644	149.774	584.489
		mc. st. dev	3.195	3.200	11454.700	18081.155
		RMSE	3.218	3.220	11455.667	18090.568
		median	0.858	0.864	0.979	1.027
		MAD	1.581	1.539	1.280	1.246
	n=3000	mean	0.892	0.904	0.946	0.983
		mc. st. dev	1.439	1.436	1.604	1.521
		RMSE	1.444	1.439	1.605	1.521
		median	0.995	0.979	1.024	1.065
		MAD	0.953	0.955	0.778	0.774
	n=5000	mean	0.891	0.904	0.994	1.027
		mc. st. dev	1.104	1.101	1.014	0.980
		RMSE	1.110	1.105	1.014	0.980
		median	0.894	0.916	1.023	1.028
		MAD	0.728	0.714	0.585	0.564
$\rho = 0.5$	n=1000	mean	0.509	0.540	-6.616E+05	-1.324E+15
		mc. st. dev	3.336	3.309	2.069E+07	4.103E+16
		RMSE	3.371	3.340	2.070E+07	4.106E+16
		median	0.759	0.826	1.030	1.032
		MAD	1.614	1.590	1.248	1.182
	n=3000	mean	0.827	0.839	0.922	0.971
		mc. st. dev	1.463	1.464	1.497	1.427
		RMSE	1.473	1.472	1.499	1.428
		median	0.921	0.931	1.045	1.067
		MAD	0.933	0.930	0.727	0.712
	n=5000	mean	0.821	0.834	0.960	0.997
		mc. st. dev	1.131	1.128	1.026	0.979
		RMSE	1.145	1.141	1.027	0.979
		median	0.853	0.860	1.016	1.037
		MAD	0.737	0.734	0.582	0.559

See next page.

Table F.5 (cont'd)

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.6$	n=1000	mean	0.402	0.428	-9.205	-12.746
		mc. st. dev	3.387	3.364	173.719	301.270
		RMSE	3.440	3.412	174.019	301.584
		median	0.849	0.864	1.101	1.012
		MAD	1.620	1.614	1.244	1.213
	n=3000	mean	0.771	0.784	0.796	0.959
		mc. st. dev	1.503	1.500	1.888	1.465
		RMSE	1.520	1.516	1.899	1.466
		median	0.852	0.874	0.983	1.121
		MAD	0.958	0.959	0.722	0.689
	n=5000	mean	0.768	0.779	0.950	1.013
		mc. st. dev	1.149	1.147	1.009	0.945
		RMSE	1.172	1.168	1.011	0.945
		median	0.800	0.811	0.992	1.070
		MAD	0.766	0.768	0.541	0.550

Table F.6: Simulation Results for DGP 2

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.4$	n=1000	mean	0.791	0.503	-9.268	-3.302
		mc. st. dev	2.329	2.788	321.882	135.961
		RMSE	2.338	2.832	322.046	136.029
		median	1.062	0.823	1.085	1.121
		MAD	1.278	1.517	0.965	0.926
	n=3000	mean	0.918	0.619	1.085	1.130
		mc. st. dev	1.260	1.520	1.212	1.157
		RMSE	1.263	1.567	1.215	1.164
		median	0.930	0.653	1.016	1.041
		MAD	0.816	0.999	0.606	0.597
	n=5000	mean	0.857	0.553	0.998	1.015
		mc. st. dev	0.956	1.148	0.827	0.829
		RMSE	0.966	1.232	0.827	0.829
		median	0.898	0.580	0.974	0.987
		MAD	0.639	0.778	0.484	0.483
$\rho = 0.5$	n=1000	mean	0.523	0.100	0.422	0.578
		mc. st. dev	2.581	3.067	6.241	13.511
		RMSE	2.625	3.197	6.268	13.517
		median	0.824	0.449	0.975	1.203
		MAD	1.447	1.786	0.944	0.913
	n=3000	mean	0.905	0.512	1.145	1.177
		mc. st. dev	1.279	1.559	1.129	1.093
		RMSE	1.283	1.634	1.139	1.107
		median	1.013	0.621	1.122	1.148
		MAD	0.821	0.988	0.603	0.570
	n=5000	mean	0.810	0.411	1.038	1.048
		mc. st. dev	0.942	1.148	0.744	0.726
		RMSE	0.961	1.291	0.745	0.727
		median	0.847	0.466	1.029	1.035
		MAD	0.610	0.731	0.446	0.427

See next page.

Table F.6 (cont'd)

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.6$	n=1000	mean	0.540	0.035	-40.132	-2.731
		mc. st. dev	2.549	3.185	1267.348	77.542
		RMSE	2.590	3.328	1268.015	77.632
		median	0.884	0.423	1.069	1.064
		MAD	1.321	1.646	0.849	0.871
	n=3000	mean	0.822	0.345	1.137	1.163
		mc. st. dev	1.275	1.569	1.120	1.068
		RMSE	1.287	1.700	1.128	1.080
		median	0.904	0.475	1.153	1.173
		MAD	0.841	1.024	0.541	0.549
	n=5000	mean	0.739	0.253	1.069	1.075
		mc. st. dev	0.953	1.173	0.719	0.712
		RMSE	0.988	1.391	0.722	0.716
		median	0.750	0.269	1.080	1.077
		MAD	0.629	0.789	0.426	0.424

Table F.7: Simulation Results for DGP 3

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.4$	n=1000	mean	0.364	0.361	-1097.543	34.431
		mc. st. dev	6.723	6.285	34992.925	874.882
		RMSE	6.753	6.317	35010.165	875.520
		median	0.676	0.646	1.241	1.235
		MAD	2.681	2.697	1.836	1.826
	n=3000	mean	0.750	0.749	1.861	1.370
		mc. st. dev	3.039	3.059	14.839	3.718
		RMSE	3.049	3.070	14.864	3.736
		median	0.816	0.839	1.158	1.222
		MAD	1.599	1.622	0.989	0.953
	n=5000	mean	0.713	0.713	1.331	1.367
		mc. st. dev	2.198	2.219	1.959	1.693
		RMSE	2.217	2.237	1.987	1.733
		median	0.761	0.769	1.105	1.201
		MAD	1.239	1.260	0.769	0.772
$\rho = 0.5$	n=1000	mean	0.537	0.492	6.409E+05	2.076E+05
		mc. st. dev	6.434	6.505	1.969E+07	5.937E+06
		RMSE	6.451	6.525	1.970E+07	5.940E+06
		median	0.745	0.735	1.425	1.409
		MAD	2.611	2.651	1.571	1.356
	n=3000	mean	0.419	0.405	1.420	1.497
		mc. st. dev	3.132	3.145	2.629	2.370
		RMSE	3.185	3.201	2.663	2.422
		median	0.605	0.589	1.273	1.335
		MAD	1.719	1.704	0.888	0.852
	n=5000	mean	0.702	0.673	1.566	1.591
		mc. st. dev	2.253	2.304	2.160	1.563
		RMSE	2.273	2.327	2.233	1.671
		median	0.787	0.731	1.328	1.419
		MAD	1.297	1.338	0.746	0.789

See next page.

Table F.7 (cont'd)

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.6$	n=1000	mean	0.278	0.231	169.761	4308.951
		mc. st. dev	6.323	6.510	4842.930	131368.520
		RMSE	6.365	6.556	4845.870	131439.140
		median	0.744	0.715	1.466	1.605
		MAD	2.634	2.705	1.402	1.328
	n=3000	mean	0.575	0.541	2.485	1.755
		mc. st. dev	3.067	3.122	24.906	4.314
		RMSE	3.096	3.156	24.950	4.379
		median	0.768	0.729	1.506	1.526
		MAD	1.586	1.612	0.830	0.805
	n=5000	mean	0.638	0.599	1.703	1.630
		mc. st. dev	2.278	2.337	2.836	2.129
		RMSE	2.307	2.371	2.921	2.221
		median	0.673	0.610	1.473	1.416
		MAD	1.371	1.411	0.646	0.612

Table F.8: Simulation Results for DGP 4

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.4$	n=1000	mean	0.646	0.888	-18.308	-156.597
		mc. st. dev	3.689	3.854	205.225	1979.583
		RMSE	3.706	3.855	206.131	1985.846
		median	0.819	0.988	1.213	1.026
		MAD	2.099	2.069	1.764	2.137
	n=3000	mean	0.725	0.910	0.864	0.535
		mc. st. dev	1.881	1.834	1.998	3.068
		RMSE	1.901	1.836	2.003	3.103
		median	0.721	0.877	0.947	0.900
		MAD	1.217	1.187	1.058	1.232
	n=5000	mean	0.732	0.914	0.935	0.730
		mc. st. dev	1.389	1.354	1.439	1.937
		RMSE	1.415	1.356	1.440	1.955
		median	0.762	0.935	1.064	1.078
		MAD	0.917	0.841	0.770	1.004
$\rho = 0.5$	n=1000	mean	0.505	0.724	-8.922	-13.194
		mc. st. dev	3.577	3.512	93.284	107.502
		RMSE	3.611	3.523	93.810	108.435
		median	0.751	0.925	1.164	1.360
		MAD	2.017	1.979	1.468	1.609
	n=3000	mean	0.663	0.833	0.911	0.641
		mc. st. dev	1.847	1.806	1.713	2.531
		RMSE	1.878	1.814	1.715	2.557
		median	0.653	0.821	0.919	0.957
		MAD	1.232	1.230	0.903	1.059
	n=5000	mean	0.655	0.831	0.850	0.671
		mc. st. dev	1.394	1.366	1.393	1.792
		RMSE	1.436	1.376	1.401	1.822
		median	0.685	0.874	0.909	0.896
		MAD	0.895	0.834	0.835	1.113

See next page.

Table F.8 (cont'd)

			Linear Model		NFES	
			2SLS	LFES(Hkt)	2PQML	WNLS
$\rho = 0.6$	n=1000	mean	0.321	0.584	-7.682	-12.635
		mc. st. dev	4.308	3.733	64.476	117.247
		RMSE	4.361	3.756	65.058	118.038
		median	0.692	0.849	1.483	1.342
		MAD	2.039	1.954	1.463	1.403
	n=3000	mean	0.567	0.729	0.862	0.920
		mc. st. dev	1.822	1.794	1.816	1.853
		RMSE	1.873	1.814	1.821	1.854
		median	0.611	0.758	0.971	1.114
		MAD	1.215	1.164	0.912	0.869
	n=5000	mean	0.595	0.759	0.858	0.850
		mc. st. dev	1.380	1.353	1.480	1.547
		RMSE	1.438	1.374	1.487	1.554
		median	0.624	0.798	1.087	1.223
		MAD	0.885	0.852	0.822	0.690

Table F.9: 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 F.10: Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
children	2.267	2.222	0	13
ceb	2.441	2.406	0	13
mort	.173	.511	0	7
educ7	.555	.496	0	1
age	27.405	8.685	15	49
agesq	826.460	526.923	225	2401
evermarr	.476	.499	0	1
urban	.516	.499	0	1
electric	.140	.347	0	1
tv	.092	.290	0	1
radio	.701	.457	0	1
frsthalf	.540	.498	0	1

Table F.11: Regression Results: selection equation

		1 Regime	2 Regime
Variable	probit	1PQML	1PQML
age	-0.014	-0.013	-0.007
	(0.018)	(0.019)	(0.018)
agesq	-0.001**	-0.001**	-0.001***
	(0.000)	(0.000)	(0.000)
evermarr	-0.306***	-0.306***	-0.301***
	(0.049)	(0.049)	(0.050)
urban	0.257***	0.256***	0.258***
	(0.043)	(0.044)	(0.044)
electric	0.412***	0.413***	0.418***
	(0.079)	(0.079)	(0.073)
tv	0.828***	0.831***	0.811***
	(0.111)	(0.111)	(0.089)
radio	0.492***	0.491***	0.490 ***
	(0.045)	(0.046)	(0.048)
frsthalf	-0.215***	-0.211***	-0.231***
	(0.042)	(0.044)	(0.042)
constant	0.271***	0.269***	0.278***
	(0.032)	(0.032)	(0.031)
L-likelihood	-2371.668		

Note: The second and third columns report the results from single step estimation. All the figures in the parenthesis are bootstrap standard errors. *: significant at 10%, **: 5%, ***: 1%

Table F.12: Regression Results: dependent variable *children*

Variable	Linear Models					NET (1 Regime)		
	OLS	2SLS	LET(Hkt)	LFES(Hkt)		1PQML	2PQML	NLS
ATE						-0.103	-0.120	-0.681 *
						(0.363)	(0.324)	(0.394)
educ7	-0.398 ***	-1.185 *	-2.232 ***		-1.552	-0.046	-0.053	-0.303 *
	(0.046)	(0.691)	(0.432)		(0.979)	(0.158)	(0.142)	(0.172)
				R1	R0			
age	0.272 ***	0.262 ***	0.249 ***	0.251 ***	0.384 ***	0.340 ***	0.340 ***	0.265 ***
	(0.019)	(0.021)	(0.021)	(0.030)	(0.049)	(0.009)	(0.009)	(0.012)
agesq	-0.002 ***	-0.002 ***	-0.002 ***	-0.003	-0.003	-0.004 ***	-0.004 ***	-0.003 ***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
evermarr	0.694 ***	0.610 ***	0.499 ***	0.194 **	0.930 ***	0.326 ***	0.325 ***	0.291 ***
	(0.054)	(0.096)	(0.080)	(0.098)	(0.198)	(0.030)	(0.029)	(0.033)
urban	-0.246 ***	-0.178 **	-0.088	0.206 **	-0.478 ***	-0.101 ***	-0.101 ***	-0.085 ***
	(0.047)	(0.078)	(0.066)	(0.088)	(0.172)	(0.024)	(0.023)	(0.027)
electric	-0.337 ***	-0.233 **	-0.094	0.197	-0.512	-0.162 ***	-0.161 ***	-0.135 ***
	(0.074)	(0.114)	(0.098)	(0.126)	(0.347)	(0.044)	(0.042)	(0.051)
tv	-0.330 ***	-0.155	0.078	0.467 ***	-0.563	-0.203 ***	-0.201 ***	-0.108 *
	(0.085)	(0.182)	(0.124)	(0.142)	(0.698)	(0.061)	(0.058)	(0.070)
radio	0.027	0.153	0.322 ***	0.620 ***	-0.052	-0.015	-0.014	0.035
	(0.053)	(0.126)	(0.099)	(0.129)	(0.294)	(0.032)	(0.030)	(0.037)
constant	-3.540 ***	-2.880 ***	3.508 ***		1.878 **	-5.514 ***	-5.507 ***	-4.059 ***
	(0.035)	(0.385)	(0.246)		(0.943)	(0.085)	(0.077)	(0.240)
cov(epsilon, v)			1.108 ***	2.550 ***	-0.524	-0.060	-0.056	0.099
			(0.257)	(0.454)	(0.905)	(0.094)	(0.085)	(0.104)
L-likelihood						-1088.243	1283.413	8597.940
R^2	0.586	0.563	0.589		0.595			
sigma	1.431	1.471	1.427		1.417			

Table F.12 (cont'd)

Variable	NFES (2 Regime)					
	1PQML		2PQML		NLS	
ATE	-0.830 ***		-0.841 ***		-1.224 ***	
	(0.318)		(0.322)		(0.375)	
educ7						
	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(epsilon, 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		-1067.756		1303.732		8521.373
R^2						
sigma						

Note: All the covariates are demeaned. All the figures in the parenthesis are bootstrap standard errors.

Table F.13: Regression Results: dependent variable *ced*

Variable	Linear Models					NET (1 Regime)		
	OLS	2SLS	LET(Hkt)	LFES(Hkt)		1PQML	2PQML	NLS
ATE						-0.179	-0.189	-0.897 **
						(0.373)	(0.343)	(0.419)
educ7	-0.462 ***	-1.187	-2.496 ***		-2.118 *	-0.074	-0.078	-0.372 **
	(0.048)	(0.725)	(0.483)		(1.084)	(0.152)	(0.141)	(0.172)
				R1	R0			
age	0.269 ***	0.260 ***	0.243 ***	0.251 ***	0.388 ***	0.339 ***	0.339 ***	0.264 ***
	(0.021)	(0.023)	(0.023)	(0.033)	(0.053)	(0.009)	(0.009)	(0.012)
agesq	-0.002 ***	-0.002 ***	-0.002 ***	-0.003 ***	-0.003 ***	-0.004 ***	-0.004 ***	-0.003 ***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
evermarr	0.734 ***	0.657 ***	0.517 ***	0.174	0.936 ***	0.321 ***	0.321 ***	0.283 ***
	(0.058)	(0.101)	(0.087)	(0.107)	(0.221)	(0.029)	(0.029)	(0.034)
urban	-0.248 ***	-0.186 **	-0.073	0.233 **	-0.419 **	-0.093 ***	-0.093 ***	-0.071 ***
	(0.049)	(0.082)	(0.072)	(0.097)	(0.190)	(0.023)	(0.023)	(0.027)
electric	-0.389 ***	-0.293 **	-0.119	0.229 *	-0.507	-0.172 ***	-0.171 ***	-0.140 ***
	(0.078)	(0.119)	(0.105)	(0.138)	(0.374)	(0.043)	(0.042)	(0.050)
tv	-0.389 ***	-0.228	0.063	0.537 ***	-0.426	-0.215 ***	-0.214 ***	-0.106
	(0.089)	(0.190)	(0.136)	(0.153)	(0.769)	(0.079)	(0.057)	(0.070)
radio	0.001	0.118	0.329 ***	0.733 ***	-0.048	-0.023	-0.023	0.033
	(0.057)	(0.134)	(0.109)	(0.142)	(0.325)	(0.114)	(0.030)	(0.037)
constant	2.698 ***	3.101 ***	3.828 ***		2.308 **	0.518 ***	0.520 ***	0.783 ***
	(0.037)	(0.405)	(0.276)		(1.046)	(0.083)	(0.077)	(0.104)
cov(epsilon, v)			1.229 ***	2.944 ***	-0.296	-0.052	-0.049	0.131
			(0.287)	(0.493)	(1.003)	(0.091)	(0.084)	(0.103)
L-likelihood						-84.416	2287.245	-9545.770
R^2	0.605	0.588	0.608		0.615			
sigma	1.513	1.545	1.509		1.495			

Table F.13 (cont'd)

Variable	NFES (2 Regime)					
	1PQML		2PQML		NLS	
ATE		-0.974 *		-0.983 **		-1.488 ***
		(0.591)		(0.400)		(0.511)
educ7						
	R1	R0	R1	R0	R1	R0
age	0.405 ***	0.293 ***	0.404 ***	0.292 ***	0.318 ***	0.239 ***
	(0.018)	(0.015)	(0.018)	(0.014)	(0.021)	(0.018)
agesq	-0.005 ***	-0.003 ***	-0.005 ***	-0.003 ***	-0.004 ***	-0.003 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
evermarr	0.287 ***	0.291 ***	0.289 ***	0.289 ***	0.240 ***	0.273 ***
	(0.047)	(0.044)	(0.045)	(0.043)	(0.052)	(0.047)
urban	-0.016	-0.101 ***	-0.018	-0.099 ***	0.006	-0.081 **
	(0.041)	(0.035)	(0.040)	(0.033)	(0.045)	(0.037)
electric	-0.103 *	-0.139 *	-0.106 *	-0.136 *	-0.056	-0.128
	(0.060)	(0.080)	(0.057)	(0.076)	(0.067)	(0.085)
tv	-0.106	-0.069	-0.109	-0.061	0.033	0.007
	(0.083)	(0.158)	(0.081)	(0.151)	(0.096)	(0.174)
radio	0.053	0.001	0.051	0.004	0.165 **	0.025
	(0.063)	(0.051)	(0.061)	(0.046)	(0.073)	(0.053)
constant	0.164	0.645 ***	0.171	0.656 ***	0.098	0.809 ***
	(0.111)	(0.179)	(0.105)	(0.154)	(0.102)	(0.190)
cov(epsilon, v)	0.300	0.011	0.289 *	0.023	0.579 ***	0.091
	(0.183)	(0.179)	(0.171)	(0.159)	(0.212)	(0.185)
L-likelihood		-66.489		2305.122		-9480.499
R^2						
sigma						

Table F.14: Average Treatment Effects

		children	*ceb*	*mort*
2SLS	$\widehat{E}(\widehat{y}_1)$	1.741	1.914	0.173
	$\widehat{E}(\widehat{y}_0)$	2.926	3.101	0.175
	$\widehat{ATE}$	-1.185 *	-1.187	-0.002
		(0.691)	(0.725)	(0.220)
LET(Heckit)	$\widehat{E}(\widehat{y}_1)$	1.276	1.332	0.056
	$\widehat{E}(\widehat{y}_0)$	3.508	3.828	0.32
	$\widehat{ATE}$	-2.232 ***	-2.496 ***	-0.264 **
		(0.432)	(0.483)	(0.123)
LFES(Heckit)	$\widehat{E}(\widehat{y}_1)$	0.326	0.19	-0.136
	$\widehat{E}(\widehat{y}_0)$	1.878	2.308	0.43
	$\widehat{ATE}$	-1.552	-2.118 *	-0.566 *
		(0.979)	(1.084)	(0.291)
1PQML	$\widehat{E}(\widehat{y}_1)$	1.482	1.683	0.201
	$\widehat{E}(\widehat{y}_0)$	2.312	2.657	0.345
	$\widehat{ATE}$	-0.83 **	-0.974 **	-0.144
		(0.337)	(0.438)	(0.152)
2PQML	$\widehat{E}(\widehat{y}_1)$	1.499	1.697	0.198
	$\widehat{E}(\widehat{y}_0)$	2.34	2.68	0.34
	$\widehat{ATE}$	-0.841 ***	-0.983 ***	-0.142
		(0.310)	(0.373)	(0.123)
NLS	$\widehat{E}(\widehat{y}_1)$	1.264	1.402	0.138
	$\widehat{E}(\widehat{y}_0)$	2.488	2.89	0.402
	$\widehat{ATE}$	-1.224 ***	-1.488 ***	-0.264
		(0.357)	(0.454)	(0.166)

Note: All the figures in the parenthesis are bootstrap standard errors.
The three nonlinear estimators on the bottom are all from NFES model.

Appendix G

Figures in Chapter 2

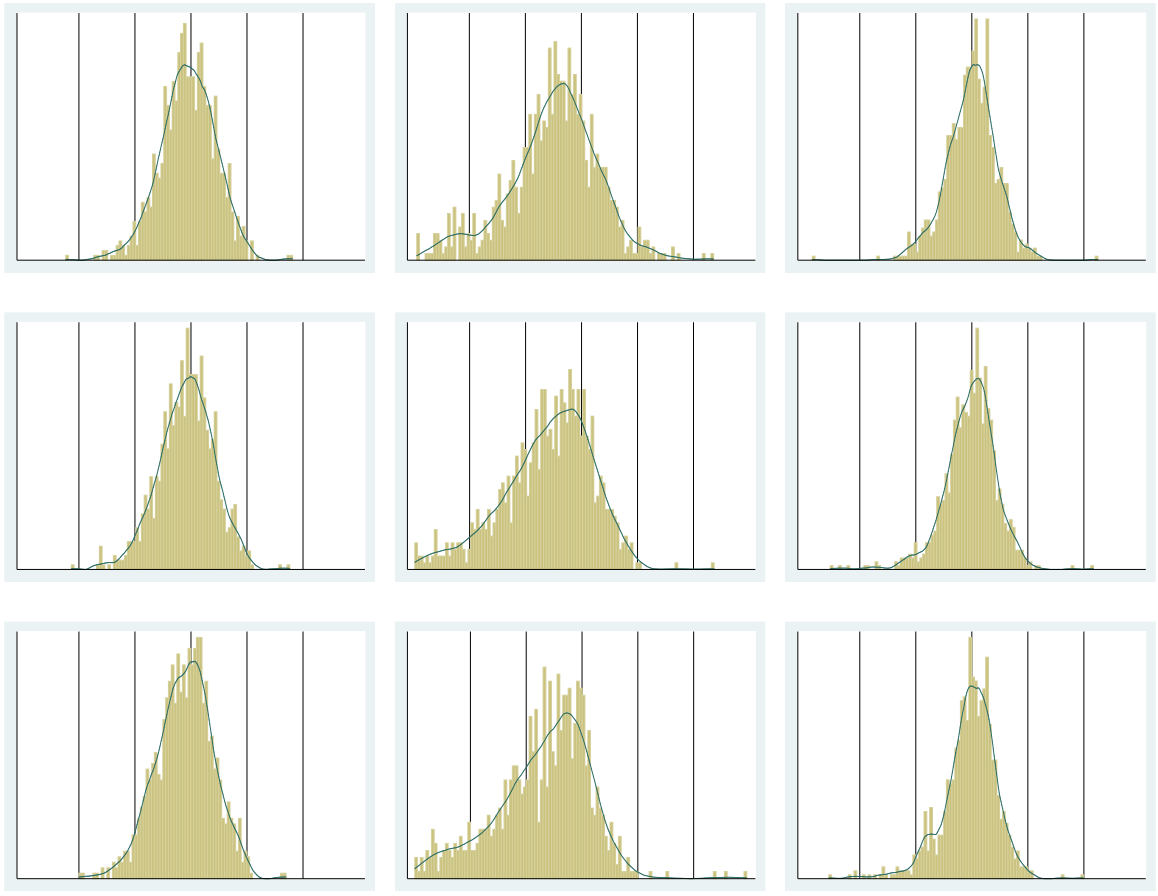

Figure G.1: Selected Monte Carlo Simulation Results for DGP 0.
Note: Each column represents the sampling distribution of 2SLS, NET (WNLS) and NFES (WNLS) estimator from left to right. Each row represents that of $\rho = .4, .5$ and .6 from top to bottom. The sample sizes are 5000 for all. Among these five grids, the middle one represents the population ATE, i.e. 1.

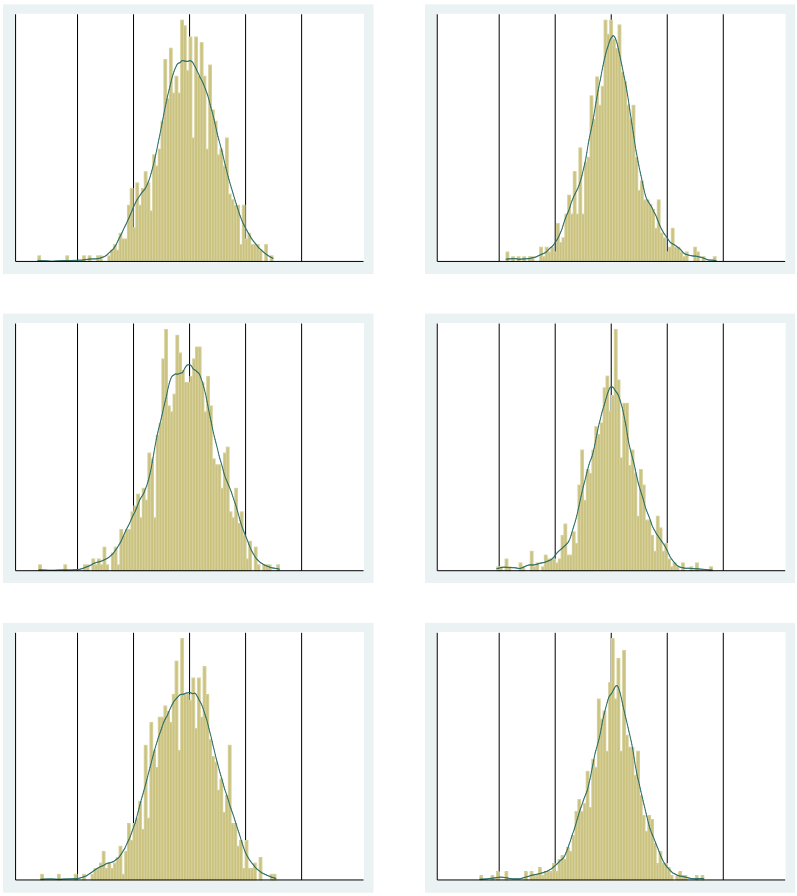

Figure G.2: Selected Monte Carlo Simulation Results for DGP 1.
Note: Each column represents the sampling distribution of 2SLS and NFES (WNLS) estimator from left to right. Each row represents that of $\rho = .4, .5$ and .6 from top to bottom. The sample sizes are 5000 for all. Among these five grids, the middle one represents the population ATE, i.e. 1.

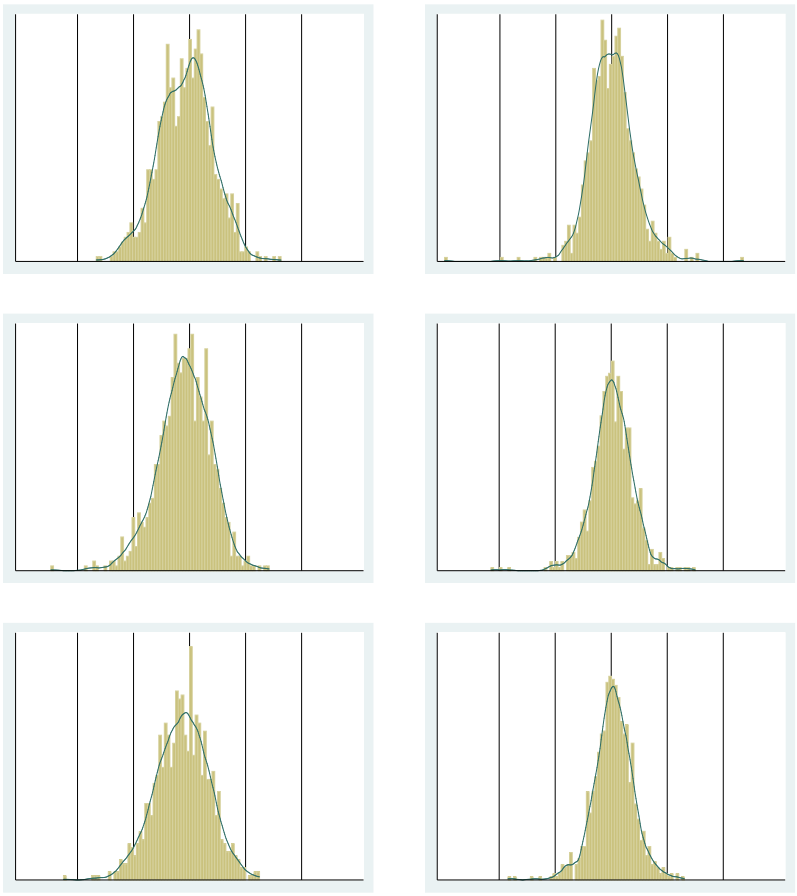

Figure G.3: Selected Monte Carlo Simulation Results for DGP 2.

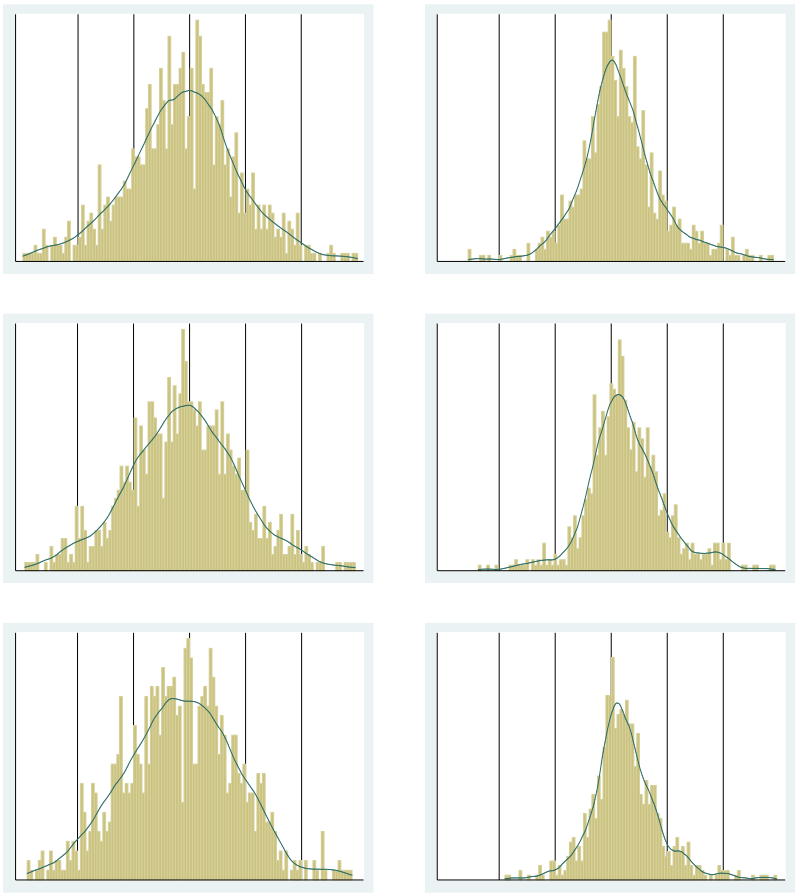

Figure G.4: Selected Monte Carlo Simulation Results for DGP 3.

Appendix H

Tables in Chapter 3

Table H.1: Simulation Results for $\rho = 0.3$

	DGP	I		II		III	
		linear	nonlinear	linear	nonlinear	linear	nonlinear
n=1000	mean	0.5640	2.192*	1.2664	1.584*	1.2505	5.366*
	mc. st. dev	1.4758	3.703*	1.7732	2.326*	2.4280	15.545*
	RMSE	1.5389	3.890*	1.7931	2.399*	2.4409	16.146*
	median	0.5930	1.031	1.1954	0.961	1.0098	1.041
	MAD	0.9266	1.023	1.0931	0.939	1.2917	1.322
	IR	3.5610	7.286	4.3419	5.947	5.0139	15.939
n=3000	mean	0.5781	1.361	1.3657	1.233	1.2210	2.430
	mc. st. dev	0.7665	1.613	0.9766	1.423	1.1970	5.627
	RMSE	0.8750	1.653	1.0429	1.442	1.2172	5.806
	median	0.5971	1.064	1.3409	0.913	1.1089	1.074
	MAD	0.5106	0.547	0.6304	0.488	0.7878	0.809
	IR	1.9470	2.398	2.4461	2.403	3.1001	5.169
n=5000	mean	0.5356	1.158	1.3692	1.035	1.2210	1.606
	mc. st. dev	0.6042	0.807	0.7265	0.766	0.9759	2.919
	RMSE	0.7620	0.822	0.8150	0.767	1.0006	2.981
	median	0.5220	1.032	1.3442	0.890	1.1648	1.046
	MAD	0.4223	0.394	0.5115	0.367	0.6017	0.589
	IR	1.5592	1.750	1.8633	1.705	2.3029	3.032

Note: The Monte Carlo standard deviation is substantially large for the sample size of 1000. The * indicates that the outlying values that are greater or smaller than the maximum and minimum values of for the sample size of 3000 were discarded from the generated data.

Table H.2: Simulation Results for $\rho = 0.5$

	DGP	I		II		III	
		linear	nonlinear	linear	nonlinear	linear	nonlinear
n=1000	mean	0.7653	1.358*	1.3439	1.634*	1.6385	3.457*
	mc. st. dev	1.4605	1.843*	1.7058	2.590*	2.3073	9.524*
	RMSE	1.4792	1.877*	1.7401	2.667*	2.3940	9.836*
	median	0.7267	0.984	1.3196	0.931	1.4693	1.045
	MAD	0.9558	0.927	1.0801	0.990	1.3670	1.200
	IR	3.6600	5.074	4.3459	6.482	5.5827	11.164
n=3000	mean	0.7548	1.147	1.3890	1.182	1.5936	1.937
	mc. st. dev	0.7953	1.044	0.9522	1.421	1.3847	4.738
	RMSE	0.8323	1.055	1.0286	1.432	1.5066	4.830
	median	0.7494	0.975	1.3690	0.872	1.4946	1.028
	MAD	0.5424	0.478	0.6060	0.505	0.8624	0.764
	IR	2.0694	2.085	2.4681	2.331	3.2039	3.935
n=5000	mean	0.7155	1.139	1.3440	1.024	1.7004	1.418
	mc. st. dev	0.6110	0.749	0.7306	0.788	1.0403	1.757
	RMSE	0.6740	0.762	0.8076	0.789	1.2541	1.806
	median	0.7179	1.031	1.3386	0.865	1.6007	1.006
	MAD	0.4144	0.413	0.5071	0.388	0.6847	0.574
	IR	1.5459	1.594	1.9085	1.657	2.5199	2.697

Table H.3: Regression Results for Linear and NFES models

Variables	OLS	2SLS	LFES		NFES (PQML)		NFES (NLS)	
ATE						-0.956***		-1.103***
						(0.278)		(0.315)
educ7	-0.419***	-1.152*		-1.168**				
	(0.049)	(0.596)		(0.528)				
			R1	R0	R1	R0	R1	R0
age	0.273***	0.263***	0.271***	0.431***	0.419***	0.300***	0.339***	0.250***
	(0.017)	(0.019)	(0.020)	(0.043)	(0.020)	(0.016)	(0.022)	(0.018)
agesq	-0.002***	-0.002***	-0.004***	-0.004***	-0.006***	-0.004***	-0.005***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
evermarr	0.685***	0.612***	0.135*	1.088***	0.212***	0.331***	0.174***	0.321***
	(0.052)	(0.080)	(0.078)	(0.138)	(0.047)	(0.043)	(0.055)	(0.046)
urban	-0.259***	-0.172**	0.187***	-0.596***	0.063	-0.152***	0.103*	-0.141***
	(0.046)	(0.084)	(0.064)	(0.120)	(0.051)	(0.040)	(0.059)	(0.044)
electric	-0.468***	-0.287*	0.117*	-0.781***	0.010	-0.242**	0.105	-0.244*
	(0.066)	(0.161)	(0.093)	(0.224)	(0.079)	(0.113)	(0.090)	(0.127)

See next page.

Table H.3 (cont'd)

Variables	OLS	2SLS	LFES		NFES (PQML)		NFES (NLS)	
constant	-3.514***	-2.815***	0.085	1.253***	-0.100	0.397**	-5.668***	-4.077***
	(0.245)	(0.620)	(0.268)	(0.478)	(0.097)	(0.154)	(0.327)	(0.469)
cov(ϵ, v)			2.202***	-1.484**	0.714	-0.204	0.963***	-0.213
			(0.304)	(0.605)	(1.398)	(0.192)	(0.295)	(0.218)
L-likelihood						1299.311		8534.8903
R^2	0.585	0.564		0.593				
σ	1.433	1.469		1.418				

Note: All the standard errors presented above are bootstrap standard errors. *, ** and *** indicates the significance at 10%, 5% and 1% levels respectively.

Table H.4: Regression Results for LTCRC and NTCRC models

Variable	LTCRC		NTCRC	
ATE				-1.020
				(1.633)
educ7		13.376***		
		(2.347)		
	R1	R0	R1	R0
age (1)	0.104***	0.343***	0.405***	0.106
	(0.030)	(0.040)	(0.040)	(0.088)
agesq	-0.003**	-	-	-0.001
	(0.001)	0.009***	0.005***	(0.001)
		(0.002)	(0.001)	
evermarr (2)	-0.047	1.377***	2.180***	0.690**
	(0.164)	(0.299)	(0.844)	(0.321)
urban (3)	0.132	-0.178	-0.265	-
	(0.155)	(0.287)	(0.382)	0.826***
				(0.319)
electric (4)	0.396**	0.686	-0.675**	-0.480
	(0.175)	(0.919)	(0.335)	(2.471)
age*evermarr			-0.097**	-0.009
			(0.041)	(0.008)
age*urban			0.013	0.012
			(0.025)	(0.008)
age*electric			0.009	-0.010
			(0.010)	(0.054)
evermarr*urban			0.060	0.132**
			(0.427)	(0.066)
evermarr*electric			0.479*	0.202
			(0.272)	(0.178)
urban*electric			0.295	0.088
			(0.287)	(0.248)
constant	0.999***	1.440***	-	-1.666
	(0.345)	(0.502)	7.010***	(1.842)
			(0.749)	

See next page.

Table H.4 (cont'd)

Variable	LTCRC		NTCRC	
	R1	R0	R1	R0
cov(ϵ, v)	1.208***	-0.677	2.438**	1.353
	(0.460)	(0.553)	(1.068)	(2.435)
cov(1)	0.022	0.135***	-0.108	-0.041
	(0.033)	(0.043)	(0.074)	(0.121)
cov(2)	0.675**	0.621**	0.290	6.954**
	(0.287)	(0.313)	(0.414)	(3.374)
cov(3)	-0.085	0.431	-0.312*	-0.225
	(0.263)	(0.279)	(0.182)	(2.127)
cov(4)	-0.869	0.953*	-0.094	-0.255
	(0.583)	(0.563)	(0.680)	(1.240)
L-likelihood				1341.602
R^2		0.6036		
σ		1.4026		

Note: The bootstrap error for NTCRC ATE estimator is calculated by excluding all the replication with ATE values over 13. cov(1) is the covariance between v and the coefficient of variable (1), i.e. *age*. Variable number for each basic covariate is indicated in the variable list. cov(2) and other numbers are defined similarly.

REFERENCES

REFERENCES

Altonji, J.G., Segal, L.M. (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics* 14, 353-366.

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

Anderson. T. W., Kunitomo, N., Sawa, T. (1982), "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator," *Econometrica* 50, 1009-1027.

Angrist, J.D. (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.D. (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.D., Evans, W.N. (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review* 88, 450-477.

Angrist, J.D., Imbens, G., Krueger, A.B. (1999), "Jackknife instrumental variables estimation," *Journal of Applied Econometrics* 14, 57-67.

Angrist, J.D., Krueger, A.B. (1991), "Does Compulsory School At- tendance Affect Schooling and Earnings?," *Quarterly Journal ofEconom- ics* 106, 979-1014.

Angrist J.D., 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.

Bekker, P.A. (1994), "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica* 62, 657-681.

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

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econoemtric Problems," *Econometrica* 69, 1127-1160

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.

Flores-Lagunes, A. (2007), "Finite Sample Evidence of IV Estimators under Weak Instruments," *Journal of Applied Econometrics* 22, 677-694.

Garen, J. (1984), "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica* 52, 1199-1218

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

Greene, W.H. (1998), "Gender Economics Courses in Liberal Arts Colleges: Further Results," *The Journal of Economic Education* 29, 291-300.

Hayashi, F. (2000), Econometrics. Princeton, NJ: Princeton University Press.

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.

Heckman, J.J., Vytlacil, E. (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model," *Journal of Human Resources* 33, 974-987

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.

Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics* 58, 71-120.

Imbens, G. (2005), "Semiparametric Estimation of Average Treatment Effects un-

der 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.

Johnson, N.L., Kotz, S. (1972), Distributions in Statistics. Continuous Multivariate Distributions. New York: Wiley.

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.

Klein, R. W., Spady, R.H. (1993), "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica* 61, 387?421.

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.

Maddala, G.S. (1986), "Limited-Dependent and Qualitative Variables in Econometrics", Cambridge: Cambridge University Press.

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,"

Newey, W.K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

Newey, W.K. (2009), "Two-step series estimation of sample selection models," *Econometrics Journal*, Supplement 1, 12, S217-S229.

Newey, W.K., Powell, J.L., Walker, J.R. (1990), "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review* 80, 324-328.

Olsen, R.J. (1980), "A Least-Squares Correction for Selectivity Bias," *Econometrica* 48: 1815-1820.

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.

Powell, J.L., Stock, J.H., Stoker, T.M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

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.

Staiger, D., Stock, J. (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65, 557-586.

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. (2008), "Parametric Regression and Health Policy Analysis: Estimation and Inference in the Presence of Endogeneity," Department of Economics, University of Florida.

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), http://stats.uis.unesco.org/unesco/TableViewer/tableView.aspx?ReportId =3674. (Retrieved on November 23, 2011)

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

Vella, F., Verbeek, M. (1999), "Estimating and Interpreting Models with Endogenous Treatment Effects," *Journal of Business and Economic Statistics* 17, 473-478.

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

Wooldridge, J. (1997), "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," *Economics Letters* 56, 129-133

Wooldridge, J. (2003), "Further Results on Instrumental Variables Estimation of

Average Treatment Effect in the Correlated Random Coefficient Model," *Economics Letters* 79, 185-191

Wooldridge, J. (2005), "Fixed Effects and Related Estimators for Correlated Random Coefficient and Treatment Effect Panel Data Models," *Review of Economics and Statistics* 87, 385-390

Wooldridge, J. (2007), "Control Function and Related Methods," What's New in Econometrics?, National Bureau of Economic Research

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

Wooldridge J. (2011), "Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables," Department of Economics, Michigan State University.

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.

Ziliak, J.P. (1997), "Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators," *Journal of Business and Economic Statistics* 15, 419-431.