Semiparametric Estimation of a Nonstationary Panel Data Transformation Model under Symmetry

Yahong Zhou[*]
Shanghai University of Finance and Economics

August 2006.

Abstract

This paper considers semiparametric estimation of a nonstationary transformation model with panel data. While nonstationarity is a common phenomenon in applied research, one of the drawbacks of most existing semiparametric procedures is the requirement of stationarity assumption. In this paper, a new semiparametric estimator is proposed under a symmetry condition, allowing for nonstationarity of the error term. Under some mild regularity conditions, the proposed estimator is consistent and asymptotically normal. A simulation study illustrates its usefulness.

Keywords: Panel Data; Transformation Model; Nonstationarity; Fixed-effects model; Maximum Rank Correlation Estimator.
JEL classification: C14, C23, C41.

1 Introduction

In this paper we consider semiparametric estimation of a nonstationary transformation model with panel data. The transformation model is in the following form:

$$h\left(y\right)=x^{0}\beta+\epsilon \tag{1}$$

where $h(.)$ is a strictly increasing function which is unknown, y is an observed dependent variable, x is an observed random vector, β is a vector of constant parameters that is

[*]This research was supported partly by research grants of SHUFE through 211-3-70. I would like to acknowledge useful discussion with professor Songnian Chen as well as seminar participants at HKUST and SHUFE for their valuable comments.

conformable with x, ϵ is an unobservable random variable. $h(.)$ is assumed to be strictly increasing to insure that (1) uniquely determines y as a function of x and ϵ. In applied econometrics, models of the form of (1) are used frequently for the analysis of duration data which include the proportional hazards model, the accelerated failure time model, and the Box-Cox regression model. A large literature has focused on this linear transformation model. But most researcher have concentrated on the cross-sectional version of the model

$$h\left(y_i\right) = x_i^{0}\beta + \epsilon_i \quad (i = 1, \cdots, n;) \tag{2}$$

Since h is unspecified (2), β is identifiable and estimable only up to scale. Several semiparametric estimators are available and can be used to estimate β (see, for example, Ichimura,1993; Horowitz and Hardle,1996; Han,1987; Powell et.al (1989)).

This paper considers the individual fixed-effects panel data form model

$$h\left(y_{it}\right) = x_{it}'\beta + \alpha_i + \epsilon_{it} \qquad (i = 1, \cdots, n; t = 1, \cdots, T) \tag{3}$$

where α is a fixed-effect, which may be correlated with other regressors in an arbitrary way, the aim of this paper is to estimate the homogeneous coefficient β. Similar to the cross-sectional case, β is only estimable up to scale. We only consider the case with two periods, and extension to general cases is straightforward. Under the assumption of stationarity that the marginal distributions of ε_{it} are the same over time for each individual, Abrevaya (1999) proposed the " leapfrog" estimator for β; he established that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal.

In applied economics, however, the stationarity is often violated, and nonstationarity is common in practice. So the "leapfrog" estimator is not applicable in these cases. Under a symmetry condition, we propose a semiparametric estimator for β, but allowing general forms of nonstationarity.

The remainder of the paper is organized as follows: the estimator is introduced in section 2. Section 3 introduces the large sample properties of the estimator. In Section 4, we report some simulation results. Section 5 concludes.

2 The Estimator

Recall the two-period panel data model subject to a monotone transformation

$$h\left(y_{it}\right) = x_{it}'\beta + \alpha_i + \epsilon_{it} \qquad (i = 1, \cdots, n; t = 1, 2) \tag{4}$$

We first briefly discuss the "leapfrog" estimator proposed by Abrevaya (1999) to motivate our estimator.

The basic idea behind Abrevaya's estimator is to compare dependent variables across different observational units indexed by i and j (with $i \neq j$) in the same time period,

$$h(y_{it}) - h(y_{jt}) = (x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j) + (\epsilon_{it} - \epsilon_{jt}). \tag{5}$$

If $(\epsilon_{i1}-\epsilon_{j1})$ and $(\epsilon_{i2}-\epsilon_{j2})$ have the same distribution (conditional on $(x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j)$), Abrevaya (1999) observed the following rank condition

$$\Delta x_i'\beta > \Delta x_j'\beta \Leftrightarrow \Pr(y_{i2} > y_{j2}|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j) > \Pr(y_{i1} > y_{j1}|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j) \tag{6}$$

He then further proposed the "leapfrog" estimator by maximizing

$$S_n(b) = \frac{1}{n(n-1)} \sum_{i \neq j} \text{sign}((\Delta x_i - \Delta x_j)'b)(1(y_{i2} > y_{j2}) - 1(y_{i1} > y_{j1}))$$

However, Abrevaya's estimator relies on the stationarity condition that $(\varepsilon_{i1}, \varepsilon_{j1})$ and $(\varepsilon_{i2}, \varepsilon_{j2})$ have the same conditional marginal distribution. In applied economics, the stationarity assumption is usually not satisfied. In order to accommodate possible nonstationarity, the idea for our proposed estimator is to compare dependent variables between different time period across different observational units; notice that

$$h\,(y_{i1}) - h(y_{j2}) = \alpha_i - \alpha_j + (x_{i1} - x_{j2})'\beta + \epsilon_{i1} - \epsilon_{j2} \tag{7}$$

and

$$h\,(y_{i2}) - h(y_{j1}) = \alpha_i - \alpha_j + (x_{i2} - x_{j1})'\beta + \epsilon_{i2} - \epsilon_{j1}$$

Thus

$$y_{i1} > y_{j2} \iff \alpha_i - \alpha_j + (x_{i1} - x_{j2})'\beta + \epsilon_{i1} - \epsilon_{j2} > 0.$$

and

$$y_{i2} > y_{j1} \iff \alpha_i - \alpha_j + (x_{i2} - x_{j1})'\beta + \epsilon_{i2} - \epsilon_{j1} > 0$$

Hence

$$\begin{aligned} \Pr(y_{i1} &> y_{j2}|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j) \\ &= \Pr(\epsilon_{j2} - \epsilon_{i1} < \alpha_i - \alpha_j + (x_{i1} - x_{j2})'\beta|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j). \end{aligned} \tag{8}$$

and

$$\begin{aligned} \Pr(y_{i2} &> y_{j1}|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j) \\ &= \Pr(\epsilon_{j1} - \epsilon_{i2} < \alpha_i - \alpha_j + (x_{i2} - x_{j1})'\beta|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j) \end{aligned} \tag{9}$$

Under symmetry restriction on the error terms, we can easily deduce that $\epsilon_{j2} - \epsilon_{i1}$ has the distribution of $\epsilon_{i2} - \epsilon_{j1}$, which, in turn, has the same distribution as $\varepsilon_{j1} - \varepsilon_{i2}$, all conditional on $x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j$. Consequently, with symmetry condition, if the common marginal density of ($\epsilon_{j2} - \epsilon_{i1}$) and $(\epsilon_{j1} - \epsilon_{i2})$ is positive everywhere along the real line. Considering (8) and (9), we have the following rank condition

$$\Pr(y_{i1} > y_{j2}|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j) > \Pr(y_{i2} > y_{j1}|x_{i1}, x_{j1}, x_{i2}, x_{j2,}\alpha_i, \alpha_j)$$

if and only if $(x_{i1} - x_{j2})'\beta > (x_{i2} - x_{j1})'\beta$. Based on this rank condition, we propose to estimate β by b_n, which maximizes

$$T_n(b) = \frac{1}{n^2} \sum_{i,j} \text{sign}((\Delta x_i + \Delta x_j)'b)(1\,(y_{i1} > y_{j2}) - 1\,(y_{i2} > y_{j1})).$$

over the parameter space B which will be defined below, where $\Delta x_i = x_{i1} - x_{i2}$ and $\Delta x_j = x_{j1} - x_{j2}$.

3 Large Sample Properties of the Estimator

In this section we establish the large sample properties of our estimator. First we make the following assumption.

Assumption 1. β belongs to a parameter space B, where B is a compact subset of R^k satisfying $B = \{b \in R^k : |b_1| = 1\}$, and β is an interior point of B.

This is normalization requirement since β is only identified up-to-scale. The requirement that B is a compact set is a common assumption in estimation literature.

Assumption 2. $\{(y_{i1}, y_{i2}, x'_{i1}, x'_{i2})\}_{i=1}^n$ is a random sample drawn from the population, satisfying the model in (3).

Assumption 3. Δx is a random vector satisfying: (i) The support of Δx is not contained in a proper linear subspace of R^k; (ii) The first component of it has everywhere positive Lebesgue density, conditional on the other components.

This assumption is common for identification of the parameters, like in estimating binary choice and index models (see, e.g., Manski (1987), Ichimura (1993), Klein and Spady (1993)).

Assumption 4: For all i, ϵ_{i1} and ϵ_{i2} have the symmetric distribution around 0, independent of $(x_{i1}, x_{i2}, \alpha_i)$. Furthermore, the marginal density of ϵ_{i1} and ϵ_{i2} is positive everywhere along the real line.

This assumption is crucial in allowing possible nonstationarity; for example, ε_{i1} and ε_{i2} can be normally distributed, but with different variances, as in Honoré (1992).

Following the notation of Sherman (1993), define $Z = (\Delta X, Y_1, Y_2)$, $T(b) = ET_n(b)$, then $z = (\Delta x, y_1, y_2)$ denotes an observation from the distribution P on the set $S \subseteq R^k \otimes R \otimes R$. Define $H(y_1, y_2, \Delta x'\beta)$ as follows:

$$H(y_1, y_2, v) = \Pr[(y_1 > Y_2, y_2 < Y_1)|\Delta X'\beta = v] - \Pr[(y_2 > Y_1, y_1 < Y_2)|\Delta X'\beta = v]$$

$$\tau(z, b) = E_Z(1(\Delta x + \Delta X)'b > 0)(1(y_1 > Y_2, y_2 < Y_1) - 1(y_2 > Y_1, y_1 < Y_2))$$

Assumption 5 let N denote a neighborhood of β and S denotes the support of the vector of ΔX.

1. For each z in S, all mixed third partial derivatives of $\tau(z, .)$ exist on N .
2. There is a function $M(z)$ satisfying $EM(z) < +\infty$ and $M(z) \geq |\partial^3\tau(z, \theta)/\partial\theta_i\partial\theta_j\partial\theta_k|$ where $(1, \theta)' = \beta$. for all θ in N, all triples (i, j, k), and all z in S.
3. $E|\partial\tau(Z, \beta)/\partial\theta_i|^2 < +\infty$ for all i.
4. $E|\partial^2\tau(Z, \beta)/\partial\theta_i\partial\theta_j| < +\infty$ for all pairs (i, j).
5. The matrix $E\partial^2\tau(Z, \beta)/\partial\theta\partial\theta'$ is negative definite.

This assumption is smoothness and boundness condition used in proving the theorem below, it can be justified by more primitive conditions on the distribution of variables in the model (see Lee, 1994; Sherman,1993; for some discussion on similar conditions).

Theorem If assumptions 1-5 hold, then b_n is consistent and asymptotic normal,

$$\sqrt{n}(b_n - \beta) \longrightarrow \overset{\mu}{} \begin{matrix} 0 \\ W \end{matrix} \overset{\P}{}$$

where $W \sim N(0, V^{-1}\Lambda V^{-1})$ with $\Lambda = E\frac{\partial\tau(Z,\beta)}{\partial\theta}\frac{\partial\tau(Z,\beta)}{\partial\theta^0}$ $V = \frac{1}{2}E\frac{\partial^2\tau(Z,\beta)}{\partial\theta\partial\theta^0}$.

The proof of this theorem can follow Sherman (1993) and Abrevaya (1999) exactly. For the simplicity of this paper, it is omitted here.

As in Section 7 of Sherman (1993), the estimation of covariance matrix in this theorem is necessary to the statistical inference for the parameter β. Numerical derivatives can be used to consistently estimate Λ and V. But some restrictions are needed to guarantee this consistency, one among them is that the bandwidth choice should satisfy $\varepsilon_n \to 0, n^{1/4}\varepsilon_n \to \infty.(\varepsilon_n$

denotes the bandwidth). Also in section 2 of Abrevaya (1999), a closed-form expression for Λ and V were mentioned, and it can be non-parametrically estimated. In our case, it is feasible but in more complicated form.

With more than two time periods, any pair of time periods can be used to construct an estimator proposed in section 2. which will yield $T(T-1)/2$ different estimators. Specifically, we can define our estimator as a solution that maximizes

$$\begin{aligned} T_n^T(b) &= \frac{1}{n^2}\sum_{t<s}\sum_{i,j}[1((x_{it}+x_{jt})'b > (x_{is}+x_{js})'b)(1\,(y_{it} > y_{js}) - 1\,(y_{is} > y_{jt})) - \\ &\quad 1((x_{it}+x_{jt})'b < (x_{is}+x_{js})'b)(1\,(y_{it} > y_{js}) - 1\,(y_{is} > y_{jt}))] \end{aligned}$$

4 A Simulation Study

In this section we present a small Monte Carlo study to illustrate the usefulness of our estimator. We will report the results for our estimator $\hat{\beta}_{ys}$ and Abrevaya's leapfrog estimator $\hat{\beta}_{lf}$. Throughout, we report the Mean, Bias, SD (standard deviation), and RMSE (root mean square error) of these two estimators based on 500 replications for each design with sample size equal to 200. The results are in the table below.

The simulation is based on the model

$$\begin{cases} y_{i1} = \alpha_i + x_{1i1} + x_{2i1} + \epsilon_{i1} \\ y_{i2} = \alpha_i + x_{1i2} + x_{2i2} + \epsilon_{i2} \end{cases}$$

where x_{1i1} and x_{1i2} are standard normally distributed, $x_{2i1} \sim 2{*}\text{uniform}(0,1).x_{2i2} \sim 0.5{*}\text{uniform}(0,1)$.the fixed effect $\alpha_i = 0.5 * (x_{2i1} + x_{2i2}) + \eta_i$, where η_i follows the standard normal distribution.

In the first design, the error terms ε_{i1} and ε_{i2} are drawn from the standard normal, independent of each other. In this case, the stationarity is satisfied, as expected, both $\hat{\beta}_{ys}$ and $\hat{\beta}_{lf}$ perform satisfactorily, and the Leapfrog estimator is even better than our estimator. In the second design, ϵ_{i1} is drawn from $N(0,2)$, whereas ϵ_{i2} is drawn from $N(0,1)$, independent of each other, thus stationarity is violated; as a result, $\hat{\beta}_{lf}$ has large biases, whereas $\hat{\beta}_{ys}$ performs well. In the three design, ϵ_{i1} is drawn from $N(0,3)$, whereas ϵ_{i2} is drawn from $N(0,1)$, independent of each other. In this case, the degree of nonstationarity increases, which leads to larger biases for $\hat{\beta}_{lf}$, while our estimator still performs satisfactorily.

Table

Design I	True value	Mean	Bias	SD	RMSE
$\hat{\beta}_{lf}$	1	1.009	0.009	0.353	0.354
$\hat{\beta}_{sy}$	1	1.013	0.013	0.293	0.294
Design II					
$\hat{\beta}_{lf}$	1	0.949	-0.051	0.413	0.417
$\hat{\beta}_{sy}$	1	1.024	0.024	0.361	0.362
Design III					
$\hat{\beta}_{lf}$	1	0.860	-0.140	0.401	0.424
$\hat{\beta}_{sy}$	1	1.047	0.047	0.391	0.409

5 Conclusion

In this paper , under symmetry restriction we have considered semiparametric estimation of the nonstationary panel data transformation model . Compared with Abrevaya's (1999) approach, the main advantage of our approach is to allow the unobservable disturbance terms for the same individual to be correlated. Our estimator is shown to be consistent and asymptotically normal. A Monte Carlo study shows the estimator performs satisfactorily in finite samples.

References

[1] Abrevaya, J., 1999. Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. Journal of Econometrics 93, 203-228.

[2] Han, A.K, 1987. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. Journal of Econometrics 35, 303-316.

[3] Horowitz, J.L., Hardle, W., 1996. Direct semiparametric estimation of single-index models with discrete covariates. Journal of the American Statistical Association 91, 1632-1640.

[4] Honore, B.E., 1992. Trimmed LAD and least squares estimation of truncated and censored regression model with fixed effects.Econometrica 60, 533-565.

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

[6] Lee, L.F., 1994. Semiparametric instrumental variable estimation of simultaneous equation sample selection models. Journal of Econometrics 63, 341-388.

[7] Manski, C.F., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica 55, 357-362.

[8] Powell, J.L., Stock, J.H., Stocker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 474-523.

[9] Sherman, R.P., 1993. The limiting distribution of the maximum rank correlation estimator. Econometrica 61, 123-137.