Center for Research in the Economics of Development | CRED

SEMIPARAMETRIC FIXED-EFFECTS ESTIMATOR

FRANÇOIS LIBOIS AND VINCENZO VERARDI

WP 1201

DEPARTMENT OF ECONOMICS

WORKING PAPERS SERIES

Semiparametric Fixed-Effects Estimator*

François Libois† and Vincenzo Verardi‡

Abstract

This paper describes the Stata implementation of Baltagi and Li's (2002) series estimator of partially linear panel data models with fixed effects. After a brief description of the estimator itself, we describe the new command `xtsemipar`. We then simulate data to show that this estimator performs better than a fixed effect estimator if the relationship between two variables is unknown or quite complex.

Keywords: `xtsemipar`, semiparametric estimations

JEL Classification: C14, C21

1 Introduction

The objective of this note is to present our Stata implementation of Baltagi and Li's (2002) series estimation of partially linear panel data models.

The structure of the note is the following: section 2 describes Baltagi and Li's (2002) fixed effects semiparametric regression estimator is described. Section 3 presents the implemented Stata command (**xtsemipar**). Some simple simulations assessing the performance of the estimator are shown in Section 4. Section 5 concludes.

*We would like to thank our colleagues at CRED and ECARES and especially Wouter Gelade and Peter-Louis Heudtlass who helped improve the quality of the paper. The usual disclaimer applies.

†Corresponding author, CRED, Facultés Universitaires Notre Dame de la Paix de Namur. E-mail: flibois@fundp.ac.be.

‡CRED, Facultés Universitaires Notre Dame de la Paix de Namur; ECARES, CKE, Université Libre de Bruxelles. E-mail: vverardi@fundp.ac.be. Vincenzo Verardi is Associated Researcher of the FNRS and gratefully acknowledges their financial support.

2 Estimation method

2.1 Baltagi and Li's (2002) semiparametric fixed effects regression estimator

Consider a general panel data semiparametric model with distributed intercept of the type:

$$y_{it} = \mathbf{x}_{it}\theta + f(z_{it}) + \alpha_i + \varepsilon_{it}, \quad i = 1, ..., N; t = 1, ..., T \text{ where } T << N \tag{1}$$

To eliminate the fixed effects α_i, a common procedure, inter alia, is to difference (1) over time which leads to

$$y_{it} - y_{it-1} = (\mathbf{x}_{it} - \mathbf{x}_{it-1})\theta + [f(z_{it}) - f(z_{it-1})] + \varepsilon_{it} - \varepsilon_{it-1} \tag{2}$$

An evident problem here is to estimate consistently the unknown function of $z \equiv G(z_{it}, z_{it-1}) = [f(z_{it}) - f(z_{it-1})]$. What Baltagi and Li (2002) propose is to approximate $f(z)$ by series $p^k(z)$ (and therefore approximate $G(z_{it}, z_{it-1}) = [f(z_{it}) - f(z_{it-1})]$ by $p^k(z_{it}, z_{it-1}) = [p^k(z_{it}) - p^k(z_{it-1})]$) where $p^k(z)$ are the first k terms of a sequence of functions $(p_1(z), p_2(z), ...)$. They then demonstrate the $\sqrt{N}$ normality for the estimator of the parametric component (i.e., $\hat{\theta}$) and the consistency at the standard non-parametric rate of the estimated unknown function (i.e., $\hat{f}$). Equation (2) therefore boils down to

$$y_{it} - y_{it-1} = (\mathbf{x}_{it} - \mathbf{x}_{it-1})\theta + [p^k(z_{it}) - p^k(z_{it-1})]\gamma + \varepsilon_{it} - \varepsilon_{it-1} \tag{3}$$

which can be estimated consistently using ordinary least squares. Having estimated $\hat{\theta}$ and $\hat{\gamma}$, it is easy to fit the fixed effects $\hat{\alpha}_i$ and go back to (1) to estimate the error component residual

$$\hat{u}_{it} = y_{it} - \mathbf{x}_{it}\hat{\theta} - \hat{\alpha}_i = f(z_{it}) + \varepsilon_{it}. \tag{4}$$

The curve f can be fitted by regressing $\hat{u}_{it}$ on z_{it} using some standard non-parametric regression estimator.

A typical example of p^k series is spline which is a fractional polynomial with pieces defined by a sequence of knots $c_1 < c_2 < ... < c_k$ where they join smoothly.

The simplest case is a linear spline. For a spline of degree m the polynomials and their first $m-1$ derivatives agree at the knots, so that $m-1$ derivatives are continuous (see Royston and Sauerbrei, 2007 for further details)

A spline of degree m with k knots can be represented as a power series:

$$S(z) = \sum_{j=0}^{m} \theta_j z^j + \sum_{j=1}^{k} \lambda_j (z - c_j)_+^m \text{ where } (z - c_j)_+^m = \begin{cases} z - c_j \text{ if } z > c_j \\ 0 \text{ otherwise} \end{cases}$$

The problem here is that successive terms tend to be highly correlated. A probably better representation of splines is a linear combinations of a set of basic splines called (k^{th} degree) B-splines which are defined, for a set of $k+2$ consecutive knots $c_1 < c_2 < ... < c_{k+2}$ as

$$B(z, c_1 ... c_{k+2}) = (k+1) \sum_{j=1}^{k+2} \left[\prod_{1 \leq h \leq k+2, h \neq j} (c_h - c_j) \right]^{-1} (z - c_j)_+^k$$

B-splines are intrinsically a rescaling of each of the piecewise functions. The technicalities of this method are beyond the scope of this paper and we refer the reader to Newson (2001) for further details.

We implemented this estimator in Stata under the command **xtsemipar**. We describe the command here below.

3 The xtsemipar command

The **xtsemipar** command fits Baltagi and Li's double series fixed-effects estimator in the case of one single variable entering the model nonparametrically.

The general syntax for the command is:

`xtsemipar` *varlist* [`if`] [`in`] [`weight`], `nonpar`(*varname*) [`generate`([*string1*] *string2*) `degree`(#) `nograph` `spline` `bwidth`(#) `robust` `cluster`(*varname*) `ci` `level`(#)]

The first option, `nonpar`, is mandatory. It declares which variable enters the model nonparametrically. None of the remaining options are compulsory. The user has the opportunity to recover the error component residual - the left hand side of equation (4) - whose name can be chosen by specifying *string2*. This error component can then be used to draw any kind of nonparametric regression. Since the error component has already been partialled out from fixed effects and from the parametrically dependent variables, this amounts to estimating the net nonparametric relation between the dependent and the variable that enters the model nonparametrically. By default, `xtsemipar` reports one estimation of this net relationship. *string1* allows to reproduce the values of the fitted dependent variable. It is worth noting that the plot of residuals is re-centered around its mean. The remaining part of this section describe options that affect this fit.

A key option in the quality of the fit is `degree`. It determines the power of the B-splines that are used to estimate consistently the function resulting from the first difference of $f(z_{it})$ and $f(z_{it-1})$ functions. By default it is set to 4. If the `nograph` option is not specified, i.e. the user wants the graph of the nonparametric fit of the variable in `nonpar` to appear, `degree` will also determine the degree of the local weighted polynomial fit used in the epanechnikov kernel performed at the last stage fit. If `spline` is specified, this last nonparametric estimation will also be estimated by the B-spline method and `degree` is then the power of these splines. More details about B-spline can be found in Newson (2001). The `bwidth` option can only be used if `spline` is not specified. It gives the half-width of the smoothing window in the epanechnikov-kernel estimation. If left unspecified, a rule-of-thumb bandwidth estimator is calculated and used (see `lpoly` for more details).

The remaining options refer to the inference. The `robust` and `cluster` options correct the inference respectively for heteroskedasticity and for clustering of error terms. In the graph, confidence intervals can be displayed by a shaded area around the curve of fitted values by specifying the option `ci`. Confidence intervals are set to 95% by default,

however it is possible to modify them by setting a different confidence level through the `level` option. This affects the confidence intervals both in the nonparametric and in the parametric part of estimations.

4 Simulation

In this section we show, using some simple simulations, how `xtsemipar` behaves in finite samples. At the end of the section we illustrate how this command can be extended to tackle some endogeneity problems.

In brief, the simulation setup is a standard panel fixed-effects of 200 individuals over 5 time periods (1000 observations). For the design space, four variables x_1, x_2, x_3 and d are generated from a Normal distribution with mean $\mu = (0, 0, 0, 0)$ and variance-covariance matrix:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ d \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & d \end{array} \\ \begin{pmatrix} 1 & & & \\ 0.2 & 1 & & \\ 0.8 & 0.4 & 1 & \\ 0 & 0.3 & 0.6 & 1 \end{pmatrix} \end{array}$$

Variable d is categorized such that five individuals are identified by each category of d. In practice we generate these variables in a two-step procedure where x's have two components. The first one is fixed for each individual and is correlated with d. The second one is a random realization for each time period.

500 replications are carried-out and for each replication an error term e is drawn from a $N(0, 1)$. The dependent variable y is generated according to DGP: $y = x_1 + x_2 - x_3 - 2 * x_3^2 - 0.25 * x_3^3) + d + e$. As it is obvious from this estimation setting, multivariate regressions with individual fixed effects should be used if we want to estimate consistently the parameters. So, we regress y on the x's using three regression models.

Table 1: Comparison between `xtsemipar` and `xtreg`

	Bias x_1	Bias x_2	MSE x_1	MSE x_2
`xtsemipar` with nonparametric control for x_3	-0.0006	-0.0007	0.00536	0.00399
`xtreg` with linear control for x_3	-0.2641	0.03752	0.07383	0.00462
`xtreg` with second and third order polynomial control for x_3	-0.0023	-0.0009	0.00410	0.00321

1. `xtsemipar` considering that x_1 and x_2 enter the model linearly and x_3 non-parametrically.

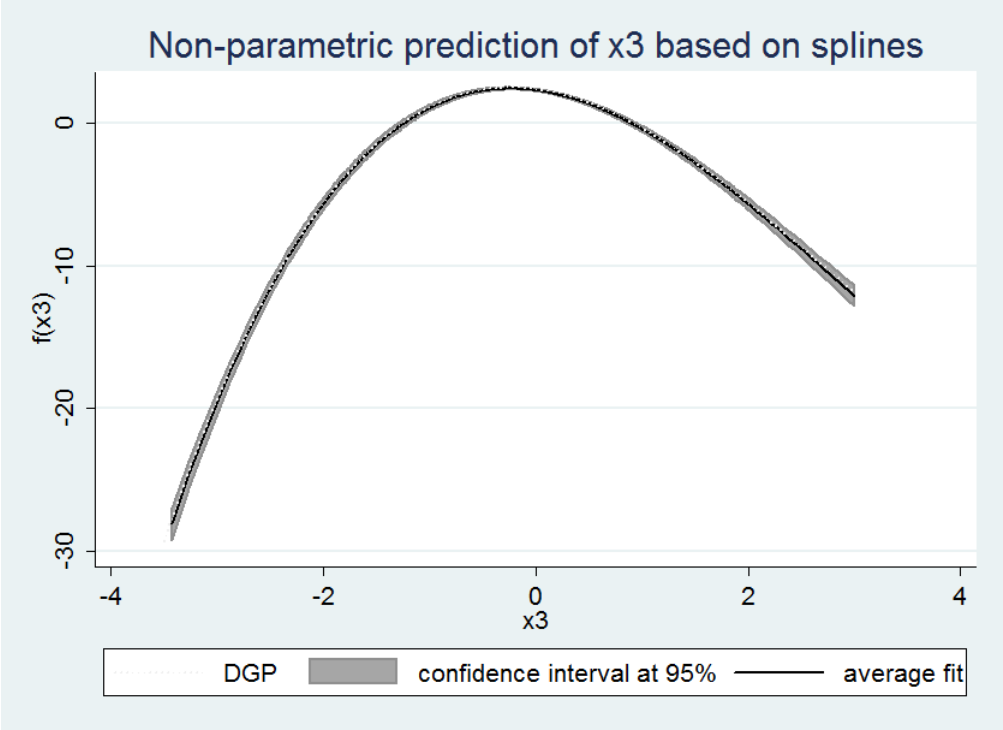
2. `xtreg` considering that x_1, x_2 and x_3 enter the model linearly.

3. `xtreg` considering that x_1 and x_2 enter the model linearly whereas x_3 enters the model parametrically with the correct polynomial form (i.e. x_3^2 and x_3^3).

Table 1 reports the bias and mean squared error (MSE) of coefficients associated with x_1 and x_2 for the three regression models. What we find is that Baltagi and Li's (2002) estimator performs much better than the usual fixed effect estimator with linear control for x_3, both in terms of bias and efficiency. As expected, the most efficient and unbiased estimator remains the fixed effect estimator with the appropriate polynomial specification. However this specification is generally unknown. Figure 1 displays the average non-parametric fit of x_3 (plain line) obtained in the simulation with the corresponding 95% band. The true DGP is represented by the dotted line.

If we want efficient and consistent estimates of parameters, estimations relying on the correct parametric specification are always better. Nevertheless, this correct form has to be known. It could be argued that a sufficiently flexible polynomial fit could be preferable than to a semi-parametric model. This is however not the case. Indeed let's consider the same simulation setting described above, but now the dependent variable y is created according to the new DGP $y = x_1 + x_2 + 3sin(2.5x_3) + d + e$. Figure 2 reports the average non-parametric fit of x_3 in a black solid line, with a 95% confidence band around it. The dotted grey line represents the true DGP which is quite close to the

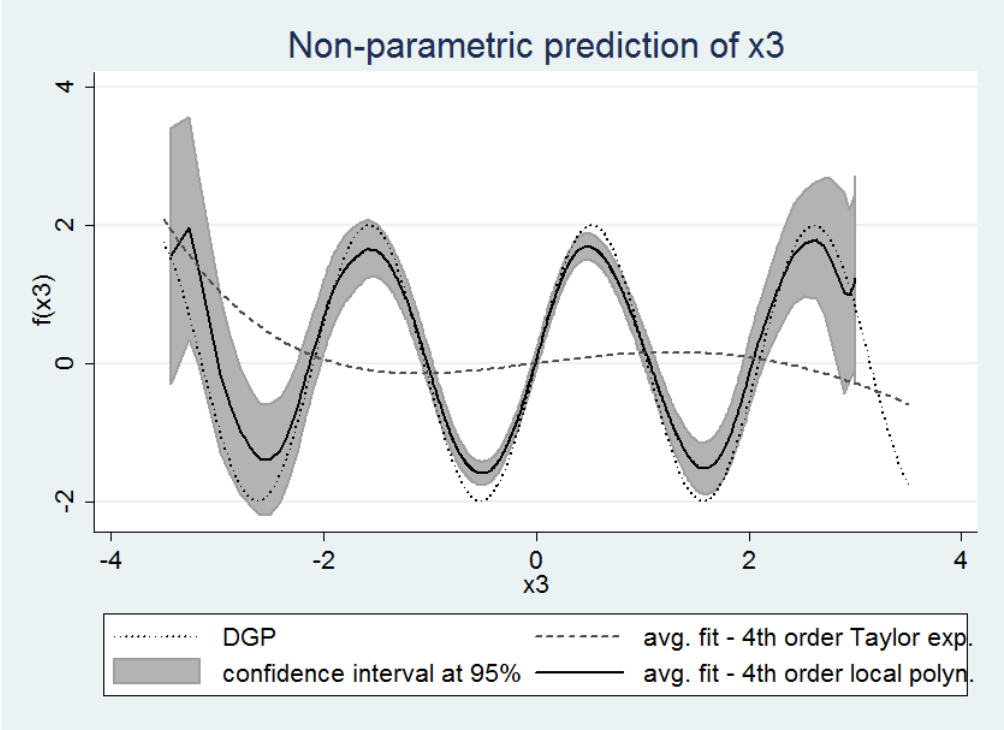
Figure 1: Average semi-parametric fit of x_3



average fit estimated by `xtsemipar` using a 4^{th} order kernel regression with a bandwith set to 0.33. The dashed grey line is the average 4th order polynomial fixed-effects parametric fit. As it is clear from this figure, `xtsemipar` provides a much better fit for this quite complex DGP. `xtsemipar` can also ease up the identification of the relevant parametric form and avoid some trial and error that often faces applied researchers.

In much of the empirical research in applied economics, measurement errors, omitted variable bias and simultaneity are common issues that can be solved through IV estimation. Baltagi and Li (2002) extend their results to address this kind of problems and establish the asymptotic properties for a partially linear panel data model with fixed effects and possible endogeneity of the regressors. In practice, our estimator can be used within a two-step procedure to obtain consistent estimates of the β's. In the first stage, the right-hand side endogenous variable has to be regressed (and fitted) by using (at least) one valid instrument. It should be noted that at this stage of the procedure, the non-parametrical variable enters linearly into the estimation procedure.

Figure 2: Average semi-parametric fit of x_3



In the second stage, the semi-parametric fixed effect panel data model can be used to estimate the relation between the dependent variable and the set of regressors. The non-parametrical variable now enters the model nonparametrically, exactly as explained before. If the instrument is valid, this procedure leads to consistent estimations.

Another problem can arise if the non-parametrical variable is subject to endogeneity problems. In this case, we suggest, as first step of the estimation procedure, to use a control functional approach as explained by Ahamada and Flachaire (2008). However we believe that the technicalities associated to this method go well beyond the scope of this note.

5 Conclusion

In econometrics, semiparametric regression estimators are becoming standard tools for applied researchers. In this paper, we present Baltagi and Li's (2002) series semipara-

metric fixed effects regression estimator. We then introduce the Stata codes we created to implement it in practice. Some simple simulations to illustrate the usefulness and the performance of the procedure are also shown.

References

[1] Baltagi, B. H. and Li, D. (2002). "Series Estimation of Partially Linear Panel Data Models with Fixed Effects". Annals of Economics and Finance 3: 103-116.

[2] Ahamada, I. and Flachaire, E. (2008). "Econométrie non paramétrique",Eds, Economica (Corpus Economie).

[3] Royston P., Sauerbrei W. (2007). "Multivariable modeling with cubic regression splines: a principled approach". Stata Journal 7: 45-70.

[4] Newson R., (2001). "B-splines and splines parameterized by their values at reference points on the x-axis," Stata Technical Bulletin, StataCorp LP, vol. 10(57)