The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 2, pp. 329–336

Semiparametric fixed-effects estimator

François Libois
University of Namur
Centre for Research in the Economics of Development (CRED)
Namur, Belgium
francois.libois@fundp.ac.be

Vincenzo Verardi
University of Namur
Centre for Research in the Economics of Development (CRED)
Namur, Belgium
and
Université Libre de Bruxelles
European Center for Advanced Research in Economics and Statistics (ECARES)
and Center for Knowledge Economics (CKE)
Brussels, Belgium
vincenzo.verardi@fundp.ac.be

Abstract. In this article, we describe the Stata implementation of Baltagi and Li's (2002, *Annals of Economics and Finance* 3: 103–116) 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-effects estimator if the relationship between two variables is unknown or quite complex.

Keywords: st0296, xtsemipar, semiparametric estimations, panel data, fixed effects

1 Introduction

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

The structure of the article is as follows. Section 2 describes Baltagi and Li's (2002) fixed-effects semiparametric regression estimator. Section 3 presents the implemented Stata command (`xtsemipar`). Some simple simulations assessing the performance of the estimator are shown in section 4. Section 5 provides a conclusion.

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

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

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

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

An evident problem here is to consistently estimate 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 $[p_1(z), p_2(z), \dots]$. They then demonstrate the $\sqrt{N}$ normality for the estimator of the parametric component (that is, $\widehat{\theta}$) and the consistency at the standard nonparametric rate of the estimated unknown function [that is, $\widehat{f}(.)$]. Equation (2) therefore boils down to

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

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

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

The curve f can be fit by regressing $\widehat{u}_{it}$ on z_{it} by using some standard nonparametric regression estimator.

A typical example of p^k series is spline, which is a fractional polynomial with pieces defined by a sequence of knots $c_1 < c_2 < \cdots < c_k$, 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 $m-1$ 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:

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

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

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

B-splines are intrinsically a rescaling of each of the piecewise functions. The technicalities of this method are beyond the scope of this article, and we refer the reader to Newson (2000b) for further details.

We implemented this estimator in Stata under the command `xtsemipar`, which we describe below.

3 The xtsemipar command

The `xtsemipar` command fits Baltagi and Li's double series fixed-effects estimator in the case of a single variable entering the model nonparametrically. Running the `xtsemipar` command requires the prior installation of the `bspline` package developed by Newson (2000a).

The general syntax for the command is

`xtsemipar` *varlist* [*if*] [*in*] [*weight*] `, nonpar(`*varname*`)` [`generate(`[*string1*] *string2*`)` `degree(`*#*`)` `knots1(`*numlist*`)` `nograph` `spline` `knots2(`*numlist*`)` `bwidth(`*#*`)` `robust` `cluster(`*varname*`)` `ci` `level(`*#*`)`]

The first option, `nonpar()`, is required. It declares that the 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 (4)—whose name can be chosen by specifying *string2*. This error component can then be used to draw any kind of nonparametric regression. Because 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* makes it possible to reproduce the values of the fitted dependent variable. Note that the plot of residuals is recentered around its mean. The remaining part of this section describes 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 consistently estimate the function resulting from the first difference of the $f(z_{it})$ and $f(z_{it-1})$ functions. The default is `degree(4)`. If the `nograph` option is not specified—that is, 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. `knots1()` and `knots2()` are both rarely used. They define a list of knots where the different pieces of the splines agree. If left unspecified, the number and location of the knots will be chosen optimally, which is the most common practice. `knots1()` refers to the B-spline estimation in (3). `knots2()` can only be used if the `spline` option is specified and refers to the last stage fit. More details about B-spline can be found in Newson (2000b). 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 [R] **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, by 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 fixed-effects panel of 200 individuals over five time periods (1,000 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

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

Variable d is categorized in such a way that five individuals are identified by each category of d. In practice, we generate these variables in a two-step procedure where the 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.

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

1. **xtsemipar**, considering that x_1 and x_2 enter the model linearly and x_3 enters nonparametrically.
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 (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-effects estimator with linear control for x_3, in terms of both bias and efficiency. As expected, the most efficient and unbiased estimator remains the fixed-effects estimator with the appropriate polynomial specification. However, this specification is generally unknown. Figure 1 displays the average nonparametric fit of x_3 (plain line) obtained in the simulation with the corresponding 95% band. The true DGP is represented by the dotted line.

Table 1. Comparison between **xtsemipar** and **xtreg**

	Bias x_1	Bias x_2	MSEx_1	MSEx_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 2nd- and 3rd-order polynomial control for x_3	−0.0023	−0.0009	0.00410	0.00321

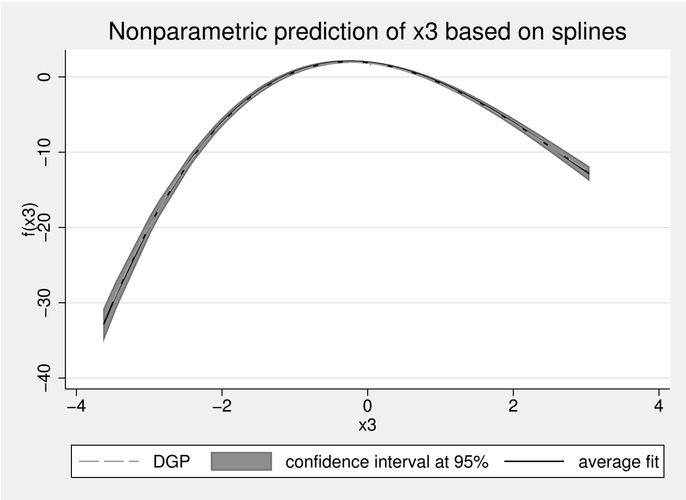

Figure 1. Average semiparametric fit of x_3

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 would be preferable to a semiparametric model. However, this is not the case. Indeed, let us consider the same simulation setting described above, but with the dependent variable y created according to the new DGP $y = x_1 + x_2 + 3\sin(2.5x_3) + d + e$. Figure 2 reports the average nonparametric fit of x_3 in a black solid line, with a 95% confidence band around it. The dotted gray line represents the true DGP, which is quite close to the average fit estimated by `xtsemipar` using a fourth-order kernel regression with a bandwidth set to 0.2. The dashed gray line is the average fourth-order polynomial fixed-effects parametric fit. As is clear from this figure, `xtsemipar` provides a much better fit for this quite complex DGP. `xtsemipar` can also help identify the relevant parametric form and help applied researchers avoid some trial and error.

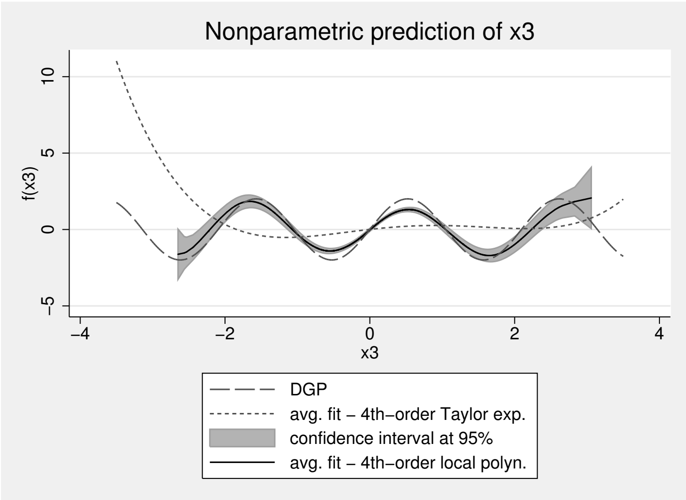

Figure 2. Average semiparametric fit of x_3

In much of the empirical research in applied economics, measurement errors, omitted variable bias, and simultaneity are common issues that can be solved through instrumental-variables estimation. Baltagi and Li (2002) extend their results to address these kinds 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 βs. In the first stage, the right-hand side endogenous variable has to be regressed (and fit) by using (at least) one valid instrument. At this stage, the nonparametric variable linearly enters into the estimation procedure. In the second stage, the semiparametric fixed-effects panel-data model can be used to estimate the relation between the dependent variable and the set of regressors. The nonparametric 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 nonparametric variable is subject to endogeneity problems. In this case, we suggest, as the first step of the estimation procedure, using a control functional approach as explained by Ahamada and Flachaire (2008). However, we believe that the technicalities associated with this method go well beyond the scope of this article.

5 Conclusion

In econometrics, semiparametric regression estimators are becoming standard tools for applied researchers. In this article, we presented Baltagi and Li's (2002) series semiparametric fixed-effects regression estimator. We then introduced the Stata program we created to put it into practice. Some simple simulations to illustrate the usefulness and the performance of the procedure were also shown.

6 Acknowledgments

We would like to thank Rodolphe Desbordes, Patrick Foissac, our colleagues at CRED and ECARES, and especially Wouter Gelade and Peter-Louis Heudtlass, who helped improve the quality of the article. The usual disclaimer applies. François Libois wishes to thank the ERC grant SSD 230290 for financial support. Vincenzo Verardi is an associate researcher at the FNRS and gratefully acknowledges their financial support.

7 References

Ahamada, I., and E. Flachaire. 2008. *Econométrie Non Paramétrique*. Paris: Économica.

Baltagi, B. H., and D. Li. 2002. Series estimation of partially linear panel data models with fixed effects. *Annals of Economics and Finance* 3: 103–116.

Newson, R. 2000a. bspline: Stata modules to compute B-splines parameterized by their values at reference points. Statistical Software Components S411701, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s411701.html.

———. 2000b. sg151: B-splines and splines parameterized by their values at reference points on the x-axis. *Stata Technical Bulletin* 57: 20–27. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 221–230. College Station, TX: Stata Press.

Royston, P., and W. Sauerbrei. 2007. Multivariable modeling with cubic regression splines: A principled approach. *Stata Journal* 7: 45–70.

About the authors

François Libois is a researcher and teaching assistant in economics at the University of Namur in the Centre for Research in the Economics of Development (CRED). His main research interests are new institutional economics with a special focus on development and environmental issues.

Vincenzo Verardi is a research fellow of the Belgian National Science Foundation (FNRS). He is a professor at the University of Namur and at the Université Libre de Bruxelles. His research interests include applied econometrics and development economics.