Unobserved Heterogeneity in Regression Models: A Semiparametric Approach Based on Nonlinear Sieves[*]

Juliano Assunção[**]
Priscila Burity[***]
Marcelo C. Medeiros[****]

Abstract

This paper proposes a semiparametric approach to control for unobserved heterogeneity in linear regression models, based on an artificial neural network extremum estimator. We present a procedure to specify the model and use simulations to evaluate its finite sample properties in comparison to alternative methods. The simulations show that our approach is less sensitive to increases in the dimensionality and complexity of the problem. We also use the model to study convergence of per capita income across Brazilian municipalities.

Keywords: Semiparametric Models, Sieve Extremum Estimators, Neural Networks, Convergence, Unobserved Components.

JEL Codes: C31, C45, O47.

[*] Submitted in June 2014. Revised in February 2015.

[**] Department of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil. E-mail: juliano@econ.puc-rio.br

[***] Department of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil. E-mail: pburity@econ.puc-rio.br

[****] Department of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil. E-mail: mcm@econ.puc-rio.br

Brazilian Review of Econometrics
v. 35, n^o 1, pp. 47–63 May 2015

1. Introduction

Heterogeneity plays an important role in regression analysis. In particular, due to missing data, the econometrician may not be able to control for relevant factors in the estimation of regression coefficients, leading to inconsistent inference. Our paper provides an approach to address this problem, even with a cross section of data.

One way of confronting unobserved heterogeneity issues is to consider panel data methods that explore the potential similarity of observations across time periods, either in the variable-intercept models or in the variable-coefficient model (Hsiao, 1989). However, sometimes panel data are not available or, even worse, their use can lead to inadequate sources of variation in the estimation process. For example, in convergence studies, panel-data methods have advantages and disadvantages. Durlauf and Quah (1999) argue that "many economists regard growth analyses as relevant over long time spans. Averaging over the longest time horizon possible-as in cross-section regression work-comes with the belief that such averaging eliminates the business cycle effects that likely dominate per capita income fluctuations at higher frequencies. (...) Different time scales for analyzing the model are mutually appropriate only if the degree of misspecification in the model is independent of time scale. In growth work, one can plausibly argue that misspecification is greater at higher frequencies" (pp. 287). For the particular case of convergence studies, it is a challenge to reconcile the need to control for unobservables with the better source of variation provided by large time spans.

We propose a semiparametric framework based on a set of proxy variables to control for heterogeneity and unobserved effects in regression models. Contrary to Robinson (1988), who uses a kernel semiparametric correction, we consider a series (sieve) expansion of the unknown and possibly nonlinear term, as in Chen (2007). The use of sieve expansions has some advantages over kernel methods. First, multiple explanatory variables can easily be handled. Second, sieve expansions have better approximation capabilities than kernel methods, as several basis functions choices can be shown to be dense in a given functional space. For example, in this paper we advocate the use of artificial neural network sieves that can simultaneously approximate the unknown function and its derivatives; see Hornik et al. (1994) for further details. Although deriving the asymptotic properties for sieve extremum estimators is more complicated than the kernel case, results from Chen and Shen (1998), Chen and White (1998), and Chen (2007) can be applied in our context.

We conduct a Monte Carlo simulation to study the finite sample properties of our proposed estimator, comparing it with other alternatives available in the literature: ordinary least squares (OLS) with an omitted variable, OLS with linear specification and the kernel method proposed by Robinson (1988). The simulation exercises are organized in such a way that we consider a different number of regressors as all models with distinct levels of complexity. Our approach performs better than the alternatives. The differences become quite significant as the dimension and complexity of the problem increase.

In addition, we use our semiparametric model to investigate an old issue in economics: the convergence of per capita income. The origin of the debate around convergence goes back to David Hume's 1742 essay entitled "Of the Rise and Progress of the Arts and Sciences" (Elmslie, 1995). Whether the income levels of poorer economies grow faster than those of richer economies is not only an important and central question in the literature of development economics, but is also related to the issue of validating competing growth theories. In the recent empirical literature, a wide array of empirical results on the subject exists. These results were obtained using cross section, panel data, time series or distribution approaches to the investigation of convergence both within an economy and across economies (Islam, 2003) and, more recently, sectoral approaches to convergence (Rodrik, 2013).

Consistent with the literature, our results suggest that convergence is stronger when we account for unobserved differences across municipalities. The estimated parameters associated with unobserved heterogeneity are related to steady-state levels of per capita income, according to the neoclassical growth theory. In our application, the estimated parameters exhibit notable differences across Brazilian municipalities that do not necessarily coincide with aggregated administrative units such as states or macro-regions. Furthermore, it is possible to identify clusters of high-income steady-state municipalities in the central and southern parts of the country.

The paper is organized as follows. Section 2 describes the model and the estimation methodology. Section 3 presents the simulation exercises. Section 4 offers an application to the convergence of per capita income across Brazilian municipalities. Section 5 concludes the paper. All proofs are presented in the Appendix.

2. Semi-Parametric Fixed-Effects Regression Model

2.1 Model definition

Consider the following assumption concerning the data generating process (DGP).

Assumption 1 (Data Generating Process) *The observed sequence of real-valued dependent variable* $\{y_i\}_{i=1}^{N}$ *is generated as*

$$y_i = a_i + \boldsymbol{\beta}_0'\boldsymbol{x}_i + u_i, \quad i = 1, \ldots, N,$$

where y_i *is the dependent variable,* a_i *is the unobserved fixed-effect representing individual heterogeneities,* $\boldsymbol{x}_i \in \mathbb{R}^k$ *is a set of observed independent and identically distributed (i.i.d.) random variables, and* u_i *is an error term. The fixed-effects are* *possibly endogenous such that* $\mathbb{E}[a_i|\boldsymbol{x}_i] \neq 0$. *Furthermore, the sequence of i.i.d. random vectors* $\{\boldsymbol{x}_i\}$ *has a common joint distribution* $\mathcal{D}$ *on* Δ, *a measurable Euclidean space with measurable Radon-Nicodým density. Finally,* $\mathbb{E}\left[|\boldsymbol{x}_i|^\delta\right] < \infty$ *for* $\delta = 1, \ldots, 4$.

We propose a semiparametric approach to estimate the fixed-effects a_i, $i = 1, \ldots, N$, and the parameter of interest $\boldsymbol{\beta}_0$. The core idea relies on the following assumption.

Assumption 2. *The unobserved effects* a_i, $i = 1, \ldots, N$, *can be written as an unknown function of a set of observed, exogenous, i.i.d. proxy variables* $\boldsymbol{z}_i \in \mathbb{R}^q$, *distinct from* $\boldsymbol{x}_i$. *Therefore,* $a_i = \eta_0(\boldsymbol{z}_i) + \epsilon_i$, *where* $\eta_0(\cdot) : \mathbb{R}^q \longrightarrow \mathbb{R}$ *is an unknown function. Furthermore, the sequence of i.i.d. random vectors* $\{\boldsymbol{z}_i\}$ *has a common joint distribution* $\mathcal{D}_z$ *on* Δ_z, *a measurable compact space with measurable Radon-Nicodým density. Finally,* $\mathbb{E}\left[|\boldsymbol{z}_i|^\delta\right] < \infty$ *for* $\delta = 1, \ldots, 4$.

The model can be rewritten as

$$y_i = \eta_0(\boldsymbol{z}_i) + \boldsymbol{\beta}_0'\boldsymbol{x}_i + \epsilon_i + u_i. \tag{1}$$

The vector $\boldsymbol{\theta}_0 = (\eta_0, \boldsymbol{\beta}_0')'$ is defined as the parameter of interest, where η is the nonparametric nuisance parameter.[1]

The following assumption states two crucial conditions for the identification of (1).

Assumption 3 (Identification) *The following conditions hold:*

$$\mathbb{E}[y_i|\boldsymbol{x}_i, a_i, \boldsymbol{z}_i] = \mathbb{E}[y_i|\boldsymbol{x}_i, \boldsymbol{z}_i] \tag{2}$$

$$\mathbb{D}[a_i|\boldsymbol{x}_i, \boldsymbol{z}_i] = \mathbb{D}[a_i|\boldsymbol{z}_i], \tag{3}$$

where $\mathbb{D}[\cdot|\cdot]$ represents the conditional distribution.

Equation (2) implies that conditional on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, the expected value of y_i does not depend on a_i. This is true if $\boldsymbol{z}_i$ is sufficient as a proxy for a_i. If $a_i = \eta(\boldsymbol{z}_i) + \epsilon_i$ and $\mathbb{E}[\epsilon_i|\boldsymbol{x}_i] = 0$, (3) is trivially satisfied.

Finally, consider the following assumption about the error terms.

Assumption 4 (Errors). *Set* $\xi_i = u_i + \epsilon_i$, *and consider that the error sequence* $\{\xi_i\}_{i=1}^N$ *is formed by random variables drawn from an absolutely continuous (with respect to a Lebesgue measure on the real line), positive-everywhere distribution such that* $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}[|\xi_i|^\delta] < \infty$, $\delta = 1, \ldots, 4$, *and* $\mathbb{E}[\xi_i\xi_j] = 0$, $\forall i \neq j$.

[1] There are many ways to estimate $\eta(\cdot)$ parametrically. However, in this paper we solely consider the nonparametric alternative.

In addition, consider the following restrictions: (1) $\mathbb{E}\left[\xi_i|\boldsymbol{x}_i,\boldsymbol{z}_i\right]=0$ *and (2)* $\mathbb{E}\left[\xi_i^2|\boldsymbol{x}_i,\boldsymbol{z}_i\right]=\sigma_\xi^2(\boldsymbol{x}_i,\boldsymbol{z}_i)$*, where* $0<\sigma_\xi^2(\boldsymbol{x}_i,\boldsymbol{z}_i)<\infty$*,* $\forall\, i$*.*

Assumption 4 is standard, implying that spatial dependence can be successfully controlled once $\boldsymbol{z}_i$ is included in the regression. Moreover, the moment conditions in the assumption are important in deriving the asymptotic results.

2.2 Estimation method and asymptotic theory

The key idea of this paper is to jointly estimate both the parametric and nonparametric components of $\boldsymbol{\theta}$ by the sieve extremum estimation method.

Set $\rho(\boldsymbol{v}_i;\boldsymbol{\theta})\equiv\rho(\boldsymbol{v}_i;\boldsymbol{\beta},\eta(\cdot))=y_i-\boldsymbol{\beta}'\boldsymbol{x}_i-\eta(\boldsymbol{z}_i)$, where $\boldsymbol{v}_i=(y_i,\boldsymbol{z}_i',\boldsymbol{x}_i')'$, $i=1,\ldots,N$. Assume that the conditional expectation $\mathbb{E}[\rho(\boldsymbol{v}_i;\boldsymbol{\theta}_0)|\boldsymbol{x}_i,\boldsymbol{z}_i]=0$, where $\boldsymbol{\theta}_0\equiv(\boldsymbol{\beta}_0',\eta_0)'\in\boldsymbol{\Theta}$ is the true parameter vector. Furthermore, define $\sigma^2(\boldsymbol{v}_i)=\mathbb{E}\left[\rho(\boldsymbol{v}_i;\boldsymbol{\theta})^2|\boldsymbol{x}_i,\boldsymbol{z}_i\right]$. Let $\boldsymbol{\Theta}=\mathcal{B}\times\mathcal{H}$, where $\mathcal{H}$ is a space of continuous functions defined on a bounded set of $\mathbb{R}^q$ and $\mathcal{B}$ is a compact set in $\mathbb{R}^k$. Consider also a sequence of approximating parameter spaces (or sieves) represented as $\boldsymbol{\Theta}_N=\mathcal{B}\times\mathcal{H}_N$, where $\bigcup_N\mathcal{H}_N$ is dense in $\mathcal{H}$ in some desirable metric.

To obtain an efficient estimator of $\boldsymbol{\beta}_0$ we apply the following three-step procedure suggested by Ai and Chen (2003):

Step 1 Obtain an initial consistent sieve nonlinear least squares estimator $\widehat{\boldsymbol{\theta}}_N=(\widehat{\boldsymbol{\beta}}_N,\widehat{\eta}_N)$ by

$$\widehat{\boldsymbol{\theta}}_N=\arg\min_{(\boldsymbol{\beta},\eta)\in\mathcal{B}\times\mathcal{H}_N}\frac{1}{N}\sum_{i=1}^N\rho(\boldsymbol{v}_i;\boldsymbol{\theta})^2.$$

Step 2 Obtain a consistent estimator $\widehat{\sigma}^2(\boldsymbol{v}_i)$ of $\sigma_0^2(\boldsymbol{v}_i)\equiv\mathbb{E}\left[\rho(\boldsymbol{v}_i;\boldsymbol{\theta}_0)^2|\boldsymbol{x}_i,\boldsymbol{z}_i\right]$ using $\widehat{\boldsymbol{\theta}}_N=(\widehat{\boldsymbol{\beta}}_N,\widehat{\eta}_N)$.

Step 3 Obtain the optimally weighted estimator $\widetilde{\boldsymbol{\theta}}_N=(\widetilde{\boldsymbol{\beta}}_N,\widetilde{\eta}_N)$ by solving

$$\widetilde{\boldsymbol{\theta}}_N=\arg\min_{\boldsymbol{\theta}\in\mathcal{B}\times\mathcal{H}_N}\mathcal{Q}(\boldsymbol{\theta},M),$$

where $\mathcal{Q}(\boldsymbol{\theta},M)\equiv\frac{1}{N}\sum_{i=1}^N\frac{\rho(\boldsymbol{v}_i;\boldsymbol{\theta})^2}{\widehat{\sigma}^2(\boldsymbol{v}_i)}$.

There are a number of distinct sieve estimators. In this paper, we advocate the use of the Artificial Neural Network (ANN) sieve, defined as

$$\eta_0(\boldsymbol{z}_i)\in\mathcal{H}\equiv\left\{\alpha_0+\sum_{m=1}^{M_N}\alpha_m f(\boldsymbol{z}_i;\boldsymbol{\omega}_m,c_m)\right\},$$

where

$$f(\boldsymbol{z}_i;\boldsymbol{\omega}_m,c_m)=\frac{1}{1+e^{-(\boldsymbol{\omega}_m'\boldsymbol{z}_i-c_m)}},$$

and $|\alpha_0| < \infty$, $|\alpha_m| < \infty$, $|c_m| < \infty$, and $\|\boldsymbol{\omega}_m\| < \infty$, $m = 1, \ldots, M_N$. The class of ANN sieves is dense in $\mathcal{H}$.

Consider the following assumption.

Assumption 5 (Approximation Capability) *There exists a small number $\delta > 0$ such that $|a_i - \eta_0(\boldsymbol{z}_i)| < \delta$, $\forall\, i$, a.s.*

Assumption 5 states that the unobserved fixed-effects may be approximated in an arbitrarily accurate way by a function of $\boldsymbol{z}_i$, $i = 1, \ldots, N$. As the class of ANN sieves is dense in $\mathcal{H}_N$, and thus is a universal approximator of the unknown function $\eta_0(\boldsymbol{z}_i)$, the semi-parametric estimator $\widehat{\eta}$ can be used to control for unobserved characteristics, leading to an unbiased estimator of $\boldsymbol{\beta}$. Assumption 5 requires that the distribution of the approximation error has a compact support. Nevertheless, Assumption 5 can be dropped, but in this case the sieve approximation will not be a universal approximator anymore.

Define $\boldsymbol{D}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i - \mathbb{E}\left[\boldsymbol{x}_i|\boldsymbol{z}_i\right]$ and $\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{Z}) = \left[\boldsymbol{D}(\boldsymbol{x}_1, \boldsymbol{z}_1) \ldots \boldsymbol{D}(\boldsymbol{x}_N, \boldsymbol{z}_N)\right]'$. From the results in Ai and Chen (2003) and under Assumptions 1–5, it follows that

$$\sqrt{N}(\widetilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \xrightarrow{d} \mathsf{N}(0, \boldsymbol{G}_0^{-1}),$$

with

$$\boldsymbol{G}_0 = \mathbb{E}\left\{\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{Z})' \left[\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}, \boldsymbol{Z})\right]^{-1} \boldsymbol{D}(\boldsymbol{X}, \boldsymbol{Z})\right\}$$

and $\boldsymbol{\Sigma}(\boldsymbol{X}, \boldsymbol{Z}) = \text{diag}\left[\sigma_0^2(\boldsymbol{v}_1), \ldots, \sigma_0^2(\boldsymbol{v}_N)\right]$.

2.3 Model selection

In applications, the number of sieves is unknown and should be determined from the data. In the neural network literature, several approaches to determining the number of sieves have been proposed. A popular method is pruning, in which a model with a large number of hidden units is estimated first, and the size of the model is subsequently reduced by applying an appropriate technique such as cross-validation. Another technique used in this context is regularization. This procedure may be characterized as penalized maximum likelihood or least squares applied to the estimation of neural network models. Bayesian regularization, based on selecting a prior distribution for the parameters, is an example of this approach.

Another possibility, which is adopted in this paper, is to estimate R models, with $M_N = 1, ..., R$, for a sufficiently large R, and choose the optimal M_N^* based on the use of model selection criteria (MSC). In the simulation study we show that this procedure works reasonably well.

3. Simulations

3.1 Setup

In this section, we conduct a Monte Carlo simulation to check the finite sample properties of the estimator discussed in this paper and compare it with alternatives available in the literature. We simulate the following DGP:

$$y_i = a_i + x_i + u_i,$$

where $x_i \in \mathbb{R}$ and

$$a_i = f(\boldsymbol{z_i}; \zeta) = \sum_{j=1}^{\zeta} \left(\sum_{k=1}^{K} z_{ik} \right)^j + \varepsilon_i,$$

and $\boldsymbol{z}_i$ is a K-dimensional vector. The parameter ζ is a complexity index. $\{x_i, \boldsymbol{z}_i\}_{i=1}^N$ is generated from a normal distribution, such that:

$$(x_i, \boldsymbol{z'}_i)' \equiv (x_i, z_{1i}, z_{2i}, z_{3i})' \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, ...N,$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2.5 & \cdot & \cdot & \cdot \\ -0.3 & 1 & \cdot & \cdot \\ 1 & -.2 & 1.6 & \cdot \\ 1 & -.3 & -.1 & 1.3 \end{pmatrix}.$$

The disturbances u_i and ε_i are generated from independent standard normal distributions. Simulations with 400 repetitions are performed using different combinations of values for ζ, K and N: $\zeta \in \{2, 3, 4\}$, $K \in \{1, 2, 3\}$, $N \in \{200, 1000, 3000\}$. The sieve estimation procedure was applied to the data with the identity weighting matrix $\widehat{\boldsymbol{\Sigma}}_0(\boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{I}$. Three competing procedures are also evaluated: (1) an ordinary least squares model omitting a_i (OLS-OV), $y_i = x_i\beta + \xi_i$; an ordinary least squares model including as regressors $(x_i, \boldsymbol{z}_i')$ (OLS), $y_i = \beta x_i + \boldsymbol{\gamma}'\boldsymbol{z}_i + \xi_i$; and finally a version of Robinson's (1988) estimation method.

Note that we violated the assumption of compact support on purpose. The idea is to check how harmful this can be in finite samples.

In brief, our version of Robinson's (1988) estimation procedure consists of the following steps: (1) regress x_i on $\boldsymbol{z}_i$ using a nonparametric estimation method and collect the vector of residuals $\boldsymbol{U}_X$; (2) regress $\boldsymbol{Y}$ on $\boldsymbol{Z}$ using a nonparametric estimation method and collect the vector of residuals $\boldsymbol{U}_Y$; (3) obtain $\widetilde{\beta} = (\boldsymbol{U}'_X \boldsymbol{U}_X)^{-1} \boldsymbol{U}'_X \boldsymbol{U}_Y$. In steps (1) and (2), residual estimation is based on a Nadaraya-Watson kernel estimator with a Gaussian kernel.

In this exercises, the number of sieves M_N is chosen through the minimization of the Hannan-Quinn information criterion (HQC) in the range of 0 (linear case) to 15 sieves.

3.2 Results

The median, mean and standard error (SE) of each of the estimators of β across the 400 simulations are reported in Tables 1-3. In each table, we present the results for the competing estimation methods: OLS-OV, OLS, Robinson and sieves for different values of the complexity index ζ and the number of observations N.

When the number K of variables in the semi-parametric component is one (Table 1), Robinson's (1988) estimation method performs well. When the complexity index ζ equals 2, the mean of the estimated parameter ranges between 0.98 ($N = 200$) and 1.00 ($N = 3000$). When the complexity increases to four, Robinson's method performs more poorly. However the produced estimate is still not far from the true value of β, especially when the sample is large: for $N = 3000$ and ζ=4, the mean of Robinson's estimator is 0.98. The means of our sieve point estimations are all either 1.00 or 1.01 and the median of the number of sieves is between 2 and 4. The OLS and OLS-OV estimation methods perform poorly in all cases with different combinations of the number of observations and the complexity index.

Table 2 shows the simulation results for the case of two variables ($K = 2$). In this case, the performance of Robinson's estimation method worsens dramatically as the complexity index increases. As was already noted in Robinson (1988), the performance of this method is poor when the number of variables in the semi-parametric component is greater than one. On the other hand, our sieve estimator continues to perform well: in all cases with different combinations of the number of observations and the complexity index, its mean and median are between 1.00 and 1.01. The median number of sieves increases for a range between 3 and 8. The OLS and OLS-OV estimation methods perform even more poorly in the case of two variables than in the case of one variable.

The three-variable case results are displayed in Table 3 and are similar to the results for the one-variable and two-variable cases. The sieve estimator is very close to one, which is the real value of β, while the Robinson, OLS and OLS-OV estimation methods perform progressively more poorly as the complexity index increases.

4. Applications: Economic Growth and Convergence among Brazilian Municipalities

We illustrate our semi-parametric regression model by testing convergence among Brazilian municipalities between 1970 and 2000. Our starting point is the simplified convergence equation presented in Barro and Sala-i-Martin (1992):

$$\log\left(\frac{y_{i,t}}{y_{i,t-1}}\right) = a_i + \gamma \cdot \log y_{i,t-1} + \phi_i \cdot (t-1) + u_{i,t}, \tag{4}$$

Table 1
Simulation results for the case with one covariate (K=1): different combinations of number of observations N and complexity ζ

		N = 200			N = 1000			N = 3000		
		$\zeta = 2$	$\zeta = 3$	$\zeta = 4$	$\zeta = 2$	$\zeta = 3$	$\zeta = 4$	$\zeta = 2$	$\zeta = 3$	$\zeta = 4$
Method	Moment	β estimates (True value: $\beta = 1$)								
OLS-OV	Median	0.647	(0.020)	(1.817)	0.646	(0.061)	(1.897)	0.638	(0.089)	(2.014)
	Mean	0.644	(0.056)	(1.929)	0.645	(0.069)	(1.935)	0.638	(0.088)	(2.021)
	SE	0.214	0.595	1.872	0.096	0.269	0.849	0.055	0.156	0.502
OLS	Median	0.997	0.896	0.618	1.003	0.886	0.580	1.002	0.868	0.497
	Mean	0.999	0.888	0.543	1.003	0.882	0.557	1.000	0.875	0.507
	SE	0.091	0.281	1.145	0.040	0.132	0.546	0.025	0.082	0.340
Robinson	Median	0.983	0.929	0.746	0.998	0.966	0.882	0.998	0.983	0.935
	Mean	0.982	0.926	0.739	0.996	0.966	0.884	0.998	0.982	0.935
	SE	0.067	0.109	0.415	0.028	0.038	0.112	0.016	0.018	0.042
Sieves	Median	0.996	0.998	1.006	1.001	0.996	0.998	1.000	0.999	1.000
	Mean	0.996	1.000	1.004	1.002	0.996	0.999	1.001	0.998	1.000
	SE	0.066	0.067	0.065	0.028	0.032	0.028	0.016	0.016	0.018
Method	Moment	M_N estimates (Hannan-Quinn Information Criteria choice in a range between 0 and 15)								
Sieves	Median	2.000	3.000	3.000	2.000	3.000	3.000	2.000	3.000	4.000
	Mean	2.560	3.310	3.308	2.503	3.463	3.525	2.488	3.438	3.770
	SE	1.277	1.154	1.091	0.918	0.878	0.788	0.819	0.814	0.961

Table 2
Simulation results for the case with two covariates (K=2): different combinations of number of observations N and complexity ζ

		N = 200			N = 1000			N = 3000		
		$\zeta = 2$	$\zeta = 3$	$\zeta = 4$	$\zeta = 2$	$\zeta = 3$	$\zeta = 4$	$\zeta = 2$	$\zeta = 3$	$\zeta = 4$
Method	Moment	β estimates (True value: $\beta = 1$)								
OLS-OV	Median	4.205	28.093	196.863	4.422	29.335	210.768	4.381	29.371	212.203
	Mean	4.274	28.516	205.336	4.365	29.268	210.864	4.378	29.387	211.862
	SE	1.222	9.675	79.199	0.597	4.673	36.707	0.353	2.783	21.979
OLS	Median	2.739	21.234	167.701	2.736	21.694	177.650	2.728	21.657	177.608
	Mean	2.738	21.671	178.029	2.733	21.578	177.167	2.734	21.636	177.780
	SE	0.402	5.514	60.798	0.182	2.373	24.862	0.102	1.344	14.383
Robinson	Median	1.608	6.829	43.410	1.337	4.854	31.410	1.207	3.659	22.698
	Mean	1.621	7.022	46.146	1.337	4.849	31.242	1.209	3.687	22.907
	SE	0.276	3.190	28.910	0.078	0.946	8.953	0.035	0.446	4.412
Sieves	Median	1.004	1.001	1.006	0.999	1.003	1.000	1.001	1.000	1.001
	Mean	1.004	1.002	1.004	0.998	1.003	1.001	1.001	1.001	1.002
	SE	0.081	0.080	0.082	0.035	0.034	0.032	0.019	0.020	0.019
Method	Moment	M_N estimates (Hannan-Quinn Information Criteria choice in a range between 0 and 15)								
Sieves	Median	4.000	5.000	7.000	3.000	4.500	8.000	4.000	4.000	7.000
	Mean	5.165	5.903	7.723	3.778	5.290	8.438	4.095	4.680	7.250
	SE	3.535	3.530	3.699	2.242	2.574	2.775	2.083	2.461	2.291

Table 3
Simulation results for the case with three covariates (K=3): different combinations of number of observations N and complexity ζ

		N = 200			N = 1000			N = 3000		
		$\zeta = 2$	$\zeta = 3$	$\zeta = 4$	$\zeta = 2$	$\zeta = 3$	$\zeta = 4$	$\zeta = 2$	$\zeta = 3$	$\zeta = 4$
Method	Moment	β estimates (True value: $\beta = 1$)								
OLS-OV	Median	17.622	238.881	2856.905	17.121	236.053	2856.514	17.267	237.283	2877.016
	Mean	17.346	237.417	2867.028	17.212	236.390	2861.914	17.307	237.788	2880.014
	SE	4.110	49.613	588.705	1.728	21.167	254.812	1.070	13.092	156.859
OLS	Median	9.402	164.712	2239.662	9.448	164.943	2248.778	9.463	165.549	2259.428
	Mean	9.433	164.799	2245.658	9.459	165.431	2257.087	9.476	165.921	2265.451
	SE	0.764	18.785	324.801	0.369	9.038	154.257	0.209	5.107	88.062
Robinson	Median	9.923	137.307	1636.507	7.242	101.535	1217.770	5.695	80.278	966.420
	Mean	10.092	139.702	1660.477	7.261	101.738	1218.606	5.693	80.354	967.781
	SE	1.362	22.220	293.609	0.453	7.827	104.844	0.219	3.903	51.905
Sieves	Median	1.011	1.000	1.003	1.006	0.997	0.998	0.997	0.998	0.998
	Mean	1.008	0.996	1.006	1.004	0.998	0.999	0.998	0.998	0.999
	SE	0.111	0.115	0.119	0.046	0.047	0.048	0.027	0.026	0.024
Method	Moment	M_N estimates (Hannan-Quinn Information Criteria choice in a range between 0 and 15)								
Sieves	Median	5.000	5.000	7.500	4.000	3.000	6.000	4.000	3.000	6.000
	Mean	7.113	6.960	8.180	5.275	4.500	6.745	5.120	4.465	6.738
	SE	4.829	4.624	3.924	3.786	3.241	3.137	3.184	3.092	2.787

where $y_{i,t}$ is the per capita income of region i in period t, a_i is associated with the steady-state level of per capita income and the rate of technological progress, ϕ_i is a parameter related to the time trend determined by technological progress, and $u_{i,t}$ is the random term. Convergence corresponds to the parameter γ.

From a conceptual point of view, there are two alternative assumptions that determine the most important distinction of convergence concepts. First, we can assume that $a_i = a$ and $\phi_i = \phi$, i.e., that the basic parameters of preference and technology are the same for all economies represented in the sample. This is the case when $\gamma < 0$ represents *unconditional convergence* – a situation in which poor municipalities tend to grow unconditionally more quickly than rich ones. Alternatively, we can state a weaker assumption allowing for possible differences in the steady state across the economies considered. In terms of equation (4), a_i and ϕ_i are allowed to vary. In this case, $\gamma < 0$ means *conditional convergence* – controlling for differences in the steady-state per capita income, poor economies grow more quickly.

Here, we estimate (4) in a cross-section setup, where there is no identifiable time trend and we are not able to distinguish between ϕ_i and a_i. Thus, we estimate the following equation:

$$\log\left(\frac{y_{i,2000}}{y_{i,1970}}\right) = \alpha_i + \gamma \cdot \log y_{i,1970} + u_{i,2000}. \tag{5}$$

Our data comes from the Brazilian Demographic Censuses of 1970 and 2000. The geographical units were adapted to reflect the changes in the organization of the Brazilian territory during that period. In 1970, Brazil was comprised of 3,951 municipalities. By 2000,this number had grown to 5,507. Therefore, all the information collected in 2000 were aggregated to match the municipal structure of 1970. Our dependent variable is the average per capita income growth between 1970 and 2000 for each municipality. The independent variable is the logarithm of the per capital income level in 1970. These two variables are presented in Figures 1 and 2.

Figures 1 and 2 illustrate the large differences across municipalities, in terms of both growth rate and 1970 income level. Figure 2 shows a sharp contrast between the poor northeastern part of the country and the southeastern and southern regions. Figures 1 and 2 also show that the variations in both growth and income levels do not coincide with the administrative state frontiers. There are numerous variations within many of the Brazilian states.

We consider different formulations in the estimation of equation (5). For the case of unconditional convergence, where $\alpha_i = \alpha$, equation (5) can be estimated by OLS, providing consistent estimates for γ if $E\left(\log y_{i,1970} \cdot u_{i,2000}\right) = 0$. This result is reported in column (1) of Table 4 and in the scatter plot of Figure 3. The estimated γ coefficient is -0.004, significant at the 1% level. Thus, the poorer municipalities in 1970 experienced a (unconditionally) lower growth rate between

Figure 1
Map of the log(Brazilian per capita income growth 1970-2000)

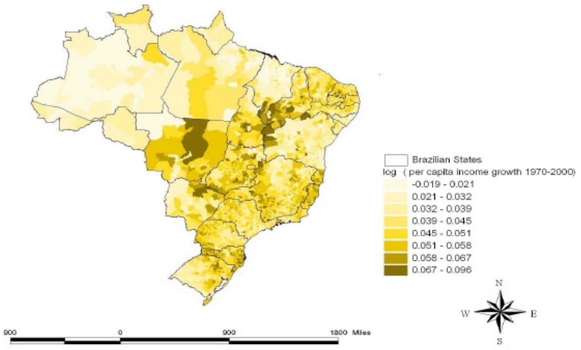


Figure 2
Map of the log(Brazilian per capita income level 1970)

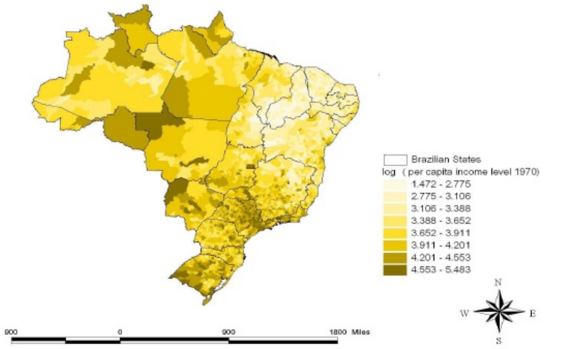


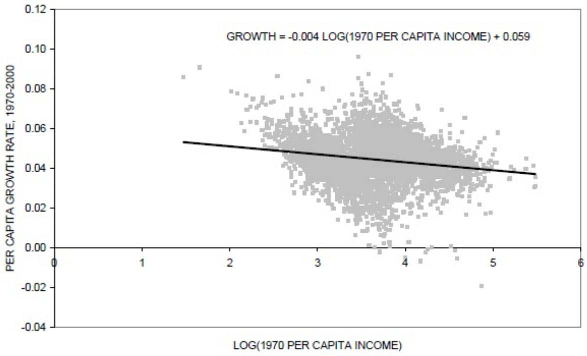
Table 4
Convergence Regressions – Brazilian Municipalities

Dependent variable: per capita growth rate, 1970-2000

	(1)	(2)	(3)	(4)
log(per capita income level 1970)	-0.004***	-0.014***	-0.014***	-0.017***
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.059***	0.104***	0.118***	-
	(0.001)	(0.002)	(0.009)	-
Number of Sieves	-	-	-	12
Macrorregion dummies (5 regions)	No	Yes	No	No
State dummies (27 states)	No	No	Yes	No
Method of estimation	OLS	OLS	OLS	Sieves GLS
Observations	3948	3948	3948	3948
R-squared	0.031	0.386	0.487	0.498

Note: Standard errors in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%

Figure 3
Growth rate from 1970 to 2000 vs. log(1970 per capita income level)



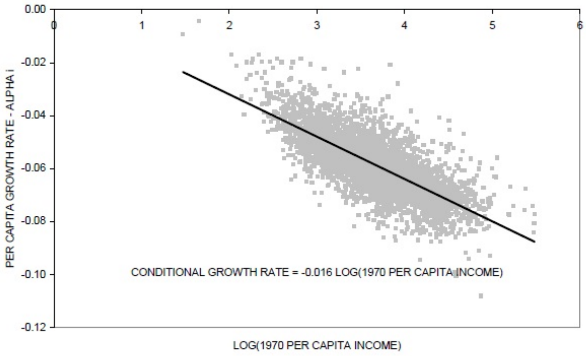
1970 and 2000.

Although we found significant evidence of unconditional convergence in Brazilian municipalities, Brazil is a large country that displays salient regional differences as shown in Figures 1 and 2. As a consequence, we might expect significant variation in the steady-state levels of per capita income across cities. Therefore, we consider the more flexible concept of conditional convergence.

However, the study of conditional convergence in this cross-section environment, is a more complex task. On the one hand, we use no additional data to approximate α_i. On the other hand, there is no degree of freedom to estimate α_i without additional statistical structure.

A natural strategy is to use aggregation through dummy variables to enable the

Figure 4
Conditional Growth rate from 1970 to 2000 vs. log(1970 per capita income level)



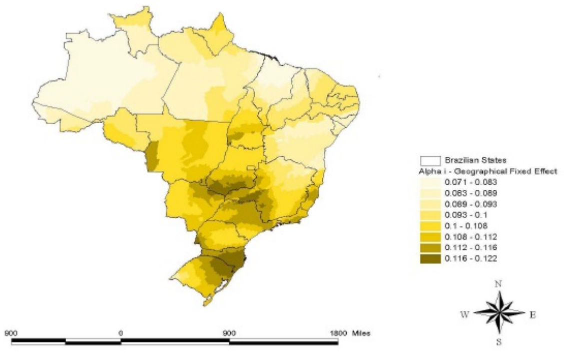
estimation of different steady-states, based on administrative divisions such as the five Brazilian macro-regions or the 27 states. The drawback of this approach is that administrative divisions of the territory do not necessarily represent differences in the steady-state levels of per capita income and technology. Figures 1 and 2 show that considerable variation exists within many states.

Columns (2) and (3) report the estimation of equation (5), controlling for dummy variables of the Brazilian macro-regions and states, respectively. In both cases, the estimated coefficient of convergence indicates a faster process of convergence, where $\widehat{\gamma} = -0.014$ at the 1% level.

Finally, we use the semi-parametric approach presented in the previous section to evaluate conditional convergence. Differences in preference and technological parameters are endogenously incorporated into the analysis through geographical similarities. The underlying assumption is that cities that are near each other face similar steady-states. In terms of the modeling, we estimate (5), considering that α_i is a semi-parametric function of the latitude and longitude coordinates, as in (1). The result is presented in the column (4) of Table 4, while the estimates for α_i are depicted in Figure 4.

The semi-parametric approach gives us a highly significant $\widehat{\gamma} = -0.017$, which is also significantly different from the point estimate -0.014 obtained from OLS using dummy variables. There are two possible interpretations of this result. First, one can argue that the semi-parametric approach makes it possible to incorporate unobserved heterogeneity across Brazilian municipalities in the estimation of the convergence parameter. This is a remarkable result, especially considering that the border Brazilian states are a highly non-linear function of the latitude and longitude coordinates. Second, from a statistical point of view, the parameter of convergence estimated by the OLS with dummy variables is missing significant unobserved heterogeneity that was incorporated in the semi-parametric approach.

Figure 5
Map of the Semi-Parametric Geographical Fixed Effects



Another aspect of this approach is the ability to estimate geographic fixed-effects at the municipality level. This allows us to investigate the evidence for unconditional convergence and conditional convergence in our sample. Figures 3 presents a plot growth versus initial per capita income for the OLS (without dummies), depicting the pattern of unconditional convergence. Figure 4, on the other hand, presents a plot of the growth rate adjusted by differences in the steady-state $\left(\log\left(\frac{y_{i,2000}}{y_{i,1970}}\right) - \alpha_i\right)$ versus the initial per capita income. A comparison of Figures 3 and 4 reveals that the evidence for conditional convergence is much clearer than that for unconditional convergence. Controlling for differences in the steady-state, poor cities grow more quickly than rich cities.

Figure 5 depicts the geographical fixed effects on a map. The estimated geographic fixed-effects vary from 0.071 to 0.122, with notable clusters of high-income municipalities in the central and southern parts of the country. There are significant geographic differences across municipalities that do not coincide with the geographic structure of the Brazilian states. This is a possible explanation for the significant differences between the OLS estimates and the values obtained using our semi-parametric method.

5. Conclusion

This paper proposes a semi-parametric approach to control for unobserved fixed effects in linear regression models. The approach is based on the artificial neural network sieve extremum estimator. We present a procedure to specify the model and use simulations to examine its finite sample properties.

The semi-parametric fixed-effect regression model is applied to the study of convergence across Brazilian municipalities for the period from 1970 to 2000. The estimated fixed effects account for differences in the steady-state levels of per capita income, allowing for a more evident convergence relation. The estimated coefficient of convergence is significantly different from the OLS counterparts, and the relationship between the (adjusted) growth rate and the per capita income level becomes closer to a straight and negatively sloped line. We find strong evidence of conditional convergence among Brazilian cities between 1970 and 2000.

References

Ai, C. & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unkown functions. *Econometrica*, 71:1795–1843.

Barro, R. & Sala-i-Martin, X. (1992). Convergence. *Journal of Political Economy*, 100:223–251.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In Heckman, J. & Leamer, E., editors, *Handbook of Econometrics*. Elsevier.

Chen, X. & Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 66:289–314.

Chen, X. & White, H. (1998). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 18:17–39.

Durlauf, S. & Quah, D. (1999). The new empirics of economic growth. In Woodford, J. B. T. . M., editor, *Handbook of Macroeconomics*. Elsevier.

Elmslie, B. (1995). Retrospectives: the convergence debate between David Hume and Josiah Tucker. *Journal of Economic Perspective*, 9:207–216.

Hornik, K., Stinchcombe, M., White, H., & Auer, P. (1994). Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, 6:1262–1274.

Hsiao, C. (1989). *Analysis of Panel Data*. Cambridge University Press.

Islam, N. (2003). What have we learnt from the convergence debate? *Journal of Economic Surveys*, 17:309–362.

Robinson, P. (1988). Root n consistent semiparametric regression. *Econometrica*, 56:931–954.

Rodrik, D. (2013). Unconditional Convergence in Manufacturing. *The Quarterly Journal of Economics*.