Quantitative Economics 4 (2013), 125–148 1759-7331/20130125

Modeling structural equations with endogenous regressors and heterogeneity through derivative constraints

Tomás Rau
Instituto de Economía, Pontificia Universidad Católica de Chile

In this paper, I present a general modeling framework for nonparametric models with endogenous regressors and heterogeneity. I show that many existing models in the literature can be derived from a structural equation with unobserved heterogeneity by imposing constancy assumptions on the first and second derivatives. I consider a less restrictive model that imposes constancy assumptions on the second partial derivative of the structural equation. Assuming the existence of suitable instrumental variables, I provide identification results and show that the model can be estimated using a generalized control function approach. I consider an application to the estimation of the returns to education in Chile, exploiting variation across regions and cohorts in educational infrastructure and compulsory schooling laws. Using penalized spline functions to approximate the components of the average structural function, I find that the local average returns to schooling are highly nonlinear and typically underestimated by flexible models that ignore the endogeneity of schooling. I also find evidence of credential effects for high school and college graduates, and limited evidence of comparative advantage bias in the returns to certain levels of education.

Keywords. Nonparametric regression, endogenous regressors, control function, endogenous treatment, returns to schooling.

JEL classification. C14, C21, C31, J31.

Tomás Rau: trau@uc.cl

This article corresponds to a revised version of the first chapter of my doctoral dissertation, which circulated under the title “Semiparametric estimation of microeconometric models with endogenous regressors and sorting.” I am highly indebted to David Card and Jim Powell for their continuous guidance and support. I would like to thank Ken Chay, Frank Vella, and Juan Pablo Torres-Martínez for helpful discussions and suggestions. This paper has benefited from the comments of Matias Cattaneo, Bryan Graham, Michael Greenstone, Michael Jansson, David Lee, Rob McMillan, Enrico Moretti, Franco Peracchi, Marina Halac, Richard Crump, and seminar participants at UC Berkeley, Concordia University, Universidad de Chile, University of Minnesota, University of Texas-Austin, University of Washington-Seattle, University of Western Ontario, the 2006 North American Summer Meeting of the Econometric Society, and the Sixth Villa Mondragone Workshop in Economic Theory and Econometrics. I gratefully acknowledge financial support from the Institute of Business and Economic Research at the University of California–Berkeley and from the Millennium Science Initiative from the Ministry of Economy, Development and Tourism to Microdata Center, Project NS100041. All errors, however, are my own.

DOI: 10.3982/QE123

1. Introduction

Many relationships in economics and the social sciences can be written as structural equations that relate an outcome variable with observed and unobserved explanatory variables. Hedonic price schedules, earning equations, and production functions are some examples of such equations. These structural relationships may emerge from underlying theoretical models or from arbitrary assumptions of the researcher.

This paper considers the estimation of a nonparametric structural equation of the form

$$y = H(x, u), \tag{1}$$

where y is a scalar outcome variable, x is a $d_x \times 1$ vector, and u is a $d_u \times 1$ vector that reflects unobserved heterogeneity with $E[u|x] \neq 0$. I show that by imposing constraints in partial derivatives of $H(x, u)$, it is possible to obtain nonparametric additive/multiplicative models that can be easily identified by control functions.

Identification and estimation of the fully nonseparable model with endogenous regressors has been recently studied by Chesher (2003, 2005), Chernozhukov and Hansen (2006), Imbens and Newey (2009), and Torgovitsky (2011). Identification of such a structural equation relies on assumptions on marginal or joint independence of the instrument and the unobservable terms as pointed out by Chesher (2007), and large support assumptions as described by Florens, Heckman, Meghir, and Vytlacil (2008). These stochastic assumptions are stronger than those for separable models that typically rely on conditional mean independence such as Garen (1984) and Newey, Powell, and Vella (1999). Therefore, a trade-off between stochastic restrictions and shape restrictions in the functional form (such as linearity or additive separability) can be observed in the task of identifying parameters of interest from (1).

In this paper, I present a general modeling framework for econometric relationships with endogenous regressors and unobserved heterogeneity through shape restrictions in the structural equation. I show that the existing econometric models, such as linear models with endogenous regressors, correlated random coefficients (CRC) models, and nonparametric regressions with endogenous regressors can be derived from constancy constraints on partial derivatives of equation (1). These constraints strengthen the restrictions on the shape of $H(x, u)$ to some form of linearity or additive separability, but weaken the stochastic restrictions to conditional mean independence.

I develop a less restrictive microeconometric model that shares the essential characteristics of Garen (1984) and Newey, Powell, and Vella (1999). By imposing constancy constraints on the second partial derivative of $H(x, u)$, I obtain an additive model that includes two unknown functions of x and u and an interaction between a function of u and x. I provide formal identification results using only conditional mean independence assumptions instead of joint independence between the instruments and the error terms. The model can be easily estimated using a generalized control function approach.

I consider an application to the estimation of the returns to education in Chile, exploiting variation across regions and cohorts in educational infrastructure and compulsory schooling laws. Using penalized spline functions to approximate the components of the average structural function, I find that the average returns to schooling are highly nonlinear and typically underestimated by flexible models that ignore the endogeneity of schooling. I also find evidence of credential or sheepskin effects for high school and college graduates, and limited evidence of comparative advantage bias in the returns to certain levels of education.[1]

The paper is organized as follows. In Section 2, I propose a general modeling framework for a microeconometric relationship with endogenous regressors and unobserved heterogeneity. In Section 3, I analyze the identification strategy with control functions and provide formal identification results. In Section 4, I briefly described the estimation method to be implemented. Section 5 describes the data used and institutional changes that generate the instrument. Section 6 shows the empirical results and Section 7 concludes.

2. Modeling through derivative constraints

In microeconometrics, modeling heterogeneity can be viewed as imposing constancy constraints in the partial derivatives of $H(x, u)$. These constraints imply that the dependence of the pth order partial derivatives on the unobserved heterogeneity vanishes for some integer p. Certainly, the restrictions on partial derivatives of order p will impose further constraints on successive partial derivatives as well.

In general, most models in the empirical literature impose constancy constraints on the first partial derivative, which allows one to control only for the differences in the intercept of the structural function. A few models impose constancy constraints on the second derivative, which permits one to control for differences in the intercept and the slope. The model I propose is a natural extension of the previous models, analyzed by imposing a less restrictive derivative constraint.

Before stating the constraint, let me introduce the multi-index notation that will be useful to define partial derivatives. Let a d_x-tuple $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{d_x})$ of nonnegative integers ($\alpha \in \mathbb{N}_0^{d_x}$) be a d_x-dimensional multi-index such that the norm of α is defined by $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_{d_x}$ and x to the α (power) is defined by $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_{d_x}^{\alpha_{d_x}}$. Then, using this notation, a pth order derivative of $H(x, u)$ is defined as

$$\partial^\alpha H(x, u) = \frac{\partial^{|\alpha|} H(x, u)}{\partial x_1^{\alpha_1}\, \partial x_2^{\alpha_2} \cdots \partial x_{d_x}^{\alpha_{d_x}}}, \quad \alpha \in \mathbb{N}_0^{d_x}, |\alpha| = p.$$

Note that I define this derivative with respect to the regressor vector x.

Assumption 1 (DIFF). *The unknown function $H(x, u)$ is continuously differentiable such that a pth order derivative with respect to x is constant in u, so*

$$\partial^\alpha H(x, u) = h(x) \tag{2}$$

for some multi-index $\alpha \in \mathbb{N}_0^{d_x}$ with $|\alpha| = p$.

[1] The sheepskin effect is the increase in wage associated with obtaining a higher credential.

TABLE 1. Derivative constraints and examples.

Restrictions	Model	Model/Author(s)
$p=1$, $g(x)=x\beta$, $\phi_0(u)=u$, $d_u=1$	$H(x,u)=x\beta+u$	Linear, endogeneity
$p=2$, $g(x)=x\beta$, $\phi_i(u)=u$, $d_u=1$	$H(x,u)=x\beta+ux+u$	CRC, Garen (1984)
$p=1$, $\phi_0(u)=u$, $d_u=1$	$H(x,u)=g(x)+u$	Newey, Powell, and Vella (1999)
$p=2$, $d_x=1$	$H(x,u)=g(x)+\phi_0(u)+\phi_1(u)x$	Rau (2007)
$p=k+1$, $\phi_j(u)=h(u)$, $d_x=1$	$H(x,u)=g(x)+\sum_{j=0}^{k}h_j(u)x^j$	Florens et al. (2008)

Note that sequential integration in equation (2) implies that the characterization of the structural equation is

$$H(x,u)=g(x)+\sum_{|\alpha|=0}^{p-1}\phi_\alpha(u)x^\alpha, \tag{3}$$

with $\partial^\alpha g(x)=h(x)$. Note that when $d_x=1$, the solution is simply $H(x,u)=g(x)+\sum_{i=0}^{p-1}\phi_i(u)x^i$ and we can characterize the constants (relative to x) $\phi_i(u)$ in terms of $H(x_0,u)$ and its derivatives evaluated at an initial condition x_0. For instance, if $x_0=0$, we have that $\phi_0(u)=H(x_0,u)$ and $\phi_i(u)=\frac{1}{i!}\frac{\partial^i H(x,u)}{\partial x^i}|_{x=x_0}$ for $1\le i\le p-1$.

There are several interesting features about the model implied by Assumption 1. First, note that after each integration, a constant of integration (relative to x, which is a function of u) is generated. This generates a polynomial of order $p-1$ that incorporates interaction terms between powers of x and functions of u. This modeling framework nests several models analyzed in the literature recently. In Table 1, I present some models nested by the framework I propose.

The first row presents the linear model with endogenous regressors. The second row presents Garen's (1984) correlated random coefficient (CRC) model. This model incorporates an interaction term between the unobserved heterogeneity u and the endogenous choice variable x. Garen assumed a conditionally homoskedastic normal distribution for the endogenous regressor with linear conditional expectation and derived a consistent two-step estimator of the average treatment effect (ATE). This model has been estimated by Chay and Greenstone (2005) in the context of a hedonic model of housing prices and air quality.

The third row of Table 1 shows the nonparametric regression with endogenous regressors proposed by Newey, Powell, and Vella (1999). Assuming a first stage $x=\Pi(z)+v$ with $E[v|z]=0$, mean conditional independence $E(u|x,z)=E(u|v)$, and differentiability of the functions in the model, they showed that the average structural function (ASF) is identified and derived a consistent two-step nonparametric estimator of the ASF.

In the fourth row, we have Rau (2007), which combines some features of Garen (1984) and Newey, Powell, and Vella (1999) by assuming $p=2$. This leads to a model with an unknown nonparametric function $g(x)$ with endogenous regressors and two unknown terms: $\phi_0(u)$ and the interaction term $\phi_1(u)x$. There are several interesting features of this model. First, it allows interactions between the unobserved heterogeneity and the endogenous regressors, as in Garen (1984). This interaction term leads to heterogenous partial derivatives given that $E[\phi_1(u)|x] \neq 0$. I define this interaction as the *sorting effect* since it allows us to test if individuals sort into different levels of the endogenous regressors. Second, it allows us to recover the whole curve purged of endogeneity, that is, the ASF. Third, for estimation purposes, it is feasible in terms of computational intensity.

The last row shows the model of Florens et al. (2008) for a single continuous treatment. It considers interactions of functions of unobserved heterogeneity with powers of the endogenous treatment variable as in Rau (2007). They allowed that the unobserved heterogeneity affects higher order derivatives and provided identification results for the average treatment effect and the average treatment on the treated.

Consequently, I have provided a general modeling framework for microeconometric models with endogenous regressors and unobserved heterogeneity. In what follows, I focus on the model with $|\alpha| = p = 2$, $\alpha \in \mathbb{N}_0^{d_x}$, and $d_x \geq 1$. Note that Assumption 1 and these restrictions imply that the square matrix of second-order partial derivatives of $H(x, u)$ with respect to x does not depend on u. Exploiting this and assuming the existence of instruments, z, with a first stage $x = \Pi(z) + v$, I obtain the system

$$\begin{aligned} y &= g(x) + \phi_0(u) + \phi_1(u)x, \\ x &= \Pi(z) + v, \end{aligned} \tag{4}$$

where $\phi_0(u) = H(x_0, u) - \frac{\partial H(x_0,u)'}{\partial x} x_0$, $\phi_1(u) = \frac{\partial H(x_0,u)'}{\partial x}$, and $g(x) = \int_{x_0}^{x} \int_{x_0}^{t} h(s)\, ds\, dt$, $E[\phi_i(u)|x] \neq 0$. Also, I use the normalization $E[\phi_i(u)] = 0$ for $i \in \{0, 1\}$ and $E[v|z] = 0$. Note that $\phi_0(u)$ is scalar and $\phi_1(u)$ is a vector of dimension $1 \times d_x$. It is important to remark that the interaction term $\phi_1(u)x$ allows the slopes to differ among observationally equivalent individuals and is what I call the sorting effect as mentioned before. I provide formal identification results in the next section.

2.1 *Parameters of interest*

The *average structural function* (ASF) defined by Blundell and Powell (2003) is an object of central interest since it allows us to obtain an average response for a particular value of x, purged of endogeneity. It is the expected value of the function with respect to the marginal density of u; in this model, it is just $E_u[H(x, u)] = g(x)$.

Discrete variations in the ASF might be useful to compute treatment effects. The *average treatment effect* (ATE) of a change in x by Δx is defined as

$$\begin{aligned} \text{ATE} &= E\big[H(x + \Delta x, u) - H(x, u)\big] \\ &= g(x + \Delta x) - g(x). \end{aligned}$$

While it is an interesting object to estimate, in reality the treatments or policy interventions are applied to a subpopulation in which the selection rule might be correlated with unobserved heterogeneity, for instance, a program to complete high school education for dropouts.

In the case of self-selected individuals into the treatment, the distribution of the unobservable conditional on the selection rule will be the same in the treatment and control groups, but the interaction between the unobserved heterogeneity and the endogenous regressor will differ, so they will not cancel out. To take that bias into consideration, I define the *subpopulation average treatment effect* (SATE) of a change in x by Δx, for a subpopulation such that $x = \bar{x}$, in the manner

$$\begin{aligned}\text{SATE} &= E\bigl[H(x+\Delta x, u) - H(x, u)|x = \bar{x}\bigr] \\ &= g(\bar{x}+\Delta x) - g(\bar{x}) + E\bigl[\phi_1(u)|x = \bar{x}\bigr]\Delta x.\end{aligned}$$

This parameter is the discrete version of the *local average response* (LAR) in Altonji and Matzkin (2005) and has a causal interpretation. This parameter measures the effect of Δx extra units for the sample of individuals with $x = \bar{x}$, fixing the conditional distribution of the unobserved heterogeneity term $\phi_1(u)|x = \bar{x}$.

Also, one might be interested in the marginal effect of an infinitesimal change in the components of vector x on y. Given that there are many ways in which a vector can change, one has to choose a direction ν. Then the average treatment effect (ATE) is defined as

$$\begin{aligned}\lim_{h\to 0} E\biggl[\frac{H(x+h\nu, u) - H(x, u)}{h}\biggr] &= \lim_{h\to 0}\frac{g(x+h\nu) - g(x)}{h} \\ &= g_x(x)'\nu,\end{aligned}$$

where $g_x(x)$ is the gradient of $g(x)$ and $g_x(x)'\nu$ is the directional derivative along a given vector $\nu \in \mathbb{R}^{d_x}$. A particular case of interest is when $\nu = e_k$, where e_k is a $d_x \times 1$ standard basis vector with the kth element equal to 1 and the rest equal to zero. Then the directional derivative along vector e_k is the partial derivative of $g(x)$ with respect to x_k, so $g_x(x)'e_k = \partial g(x)/\partial x_k$.

Accordingly, the SATE parameter in the continuous cases is

$$\begin{aligned}\lim_{h\to 0} E\biggl[\frac{H(x+h\nu, u) - H(x, u)}{h}\bigg| x = \bar{x}\biggr] &= \lim_{h\to 0}\frac{g(\bar{x}+h\nu) - g(\bar{x})}{h} \\ &\quad + E\bigl[\phi_1(u)|x = \bar{x}\bigr]\nu \\ &= \bigl(g_x(\bar{x})' + E\bigl[\phi_1(u)|x = \bar{x}\bigr]\bigr)\nu.\end{aligned}$$

In case we are interested in moving in the direction of the basis vector e_k, the SATE is simply $\partial g(x)/\partial x_k + E[\phi_1(u)|x = \bar{x}]e_k$. The crucial term in these subpopulation treatment effects is $E[\phi_1(u)|x = \bar{x}]$, since it accounts for the sorting effect.

3. Identification strategy

Identification in this type of model has been analyzed by Newey, Powell, and Vella (1999), Rau (2007), Florens et al. (2008), and Hahn and Ridder (2011). To identify (4), I follow a control function approach.

Assumption 2 (CF). *Control function or mean-conditional independence is denoted*

$$E\big[\phi_0(u)|x,z\big]=E\big[\phi_0(u)|v,z\big]=E\big[\phi_0(u)|v\big]=\lambda(v),$$
$$E\big[\phi_1(u)|x,z\big]=E\big[\phi_1(u)|v,z\big]=E\big[\phi_1(u)|v\big]=\gamma(v),$$

where $\lambda(v)$ and $\gamma(v)$ are unknown control functions.

Then, under Assumption 2, my model stated in equation (4) becomes

$$E[y|x,z]=g(x)+\lambda(v)+\gamma(v)x. \tag{5}$$

The case when $E[\phi_1(u)|x,z]=\gamma(v)=0$ has been analyzed by Newey, Powell, and Vella (1999). This is a generalization of their model, since it includes the interaction term that controls for sorting that allows for heterogeneity in the slope.

As in Newey, Powell, and Vella (1999), identification in this model exploits the fact that conditional expectations are unique with probability 1. Then if we have two sets of functions satisfying equation (5), identification is equivalent to equality of conditional expectation, implying equality of these functions.

Theorem 1. *If $g(x)$, $\gamma(v)$, $\lambda(v)$, and $\Pi(z)$ are differentiable, the boundary of the support (z,v) has zero probability, and with probability* 1, $\operatorname{rank}(\Pi_z(z))=d_x$, *then $g(x)$ is identified.*

Proof. Let the two set of functions $\{g^1(x),\gamma^1(v),\lambda^1(v)\}$ and $\{g^2(x),\gamma^2(v),\lambda^2(v)\}$ be such that

$$E[y|x,v]=g^1(x)+\gamma^1(v)x+\lambda^1(v)=g^2(x)+\gamma^2(v)x+\lambda^2(v).$$

Then

$$\big(g^1(x)-g^2(x)\big)+x\big(\gamma^1(v)-\gamma^2(v)\big)+\big(\lambda^1(v)-\lambda^2(v)\big)=0.$$

Defining the functions $\tilde{g}(x)=g^1(x)-g^2(x)$, $\tilde{\gamma}(v)=\gamma^1(v)-\gamma^2(v)$, and $\tilde{\lambda}(v)=\lambda^1(v)-\lambda^2(v)$, then

$$0=\tilde{g}(x)+x\tilde{\gamma}(v)+\tilde{\lambda}(v), \tag{6}$$
$$0=\tilde{g}\big(\Pi(z)+v\big)+\big(\Pi(z)+v\big)\tilde{\gamma}(v)+\tilde{\lambda}(v), \tag{7}$$

identically in z and v. Differentiating (7) with respect to z and v gives

$$0=\Pi_z(z)\tilde{g}_x\big(\Pi(z)+v\big)+\Pi_z(z)\tilde{\gamma}(v), \tag{8}$$
$$0=\tilde{g}_x\big(\Pi(z)+v\big)+\tilde{\gamma}(v)+\big(\Pi(z)+v\big)\tilde{\gamma}_v(v)+\tilde{\lambda}_v(v), \tag{9}$$

where $\tilde{g}_x(x)=\partial g(x)/\partial x$ and $\Pi_z(z)=\partial\Pi(z)/\partial z$. If $\operatorname{rank}(\Pi_z(z))=d_x$, we have that

$$\tilde{g}_x\big(\Pi(z)+v\big)+\tilde{\gamma}(v)=0. \tag{10}$$

Now putting this result into (9) implies that

$$\big(\Pi(z)+v\big)\tilde{\gamma}_v(v)+\tilde{\lambda}_v(v)=0. \tag{11}$$

Differentiating (11) with respect to z, we have that $\Pi_z(z)\tilde{\gamma}_v(v)=0$, which implies that $\tilde{\gamma}_v(v)=0$ and $\tilde{\lambda}_v(v)=0$ by the rank condition $\operatorname{rank}(\Pi_z(z))=d_x$. Then $\tilde{\gamma}(v)$ and $\tilde{\lambda}(v)$ are constant. Given the normalization $E[\phi_i(u)]=0$ and the law of iterated expectations, we have that $E[\lambda(v)]=E[\gamma(v)]=0$. Hence $\tilde{\gamma}(v)=\tilde{\lambda}(v)=0$. Plugging this result into equation (6), we have that $\tilde{g}(x)=0$.

By differentiability of the functions and the boundary of the support of (z,v) having zero probability, equalities (8)–(11) hold identically on the interior of the support of (z,v). Then all the functions $\tilde{g}(x)$, $\tilde{\gamma}(v)$, and $\tilde{\lambda}(v)$ are zero with probability 1 implying identification of $g(x)$, $\lambda(v)$, and $\gamma(v)$. □

The identification result implies identification of the ASF when there is an interaction between a function of unobserved heterogeneity with the endogenous regressors. Certainly, by Theorem 1, the ATE is identified, but identification of $E[\phi_1(u)|x]$ is needed so as to identify the *sorting* effect and so the SATE. Using Assumption 2 and the law of iterated expectations, we have

$$\begin{aligned} E\big[\phi_1(u)|x\big] &= E\big[E\big[\phi_1(u)|x,z\big]|x\big] \\ &= E\big[\gamma(v)|x\big], \end{aligned} \tag{12}$$

which is identified since it is a conditional expectation of an identified function.

The identification result in Theorem 1 is a generalization of Newey, Powell, and Vella (1999) and rests on similar assumptions about differentiability of the functions and the rank condition $\operatorname{rank}(\Pi_z(z))=d_x$. As they pointed out, if $\Pi(z)$ were linear in z, then $\operatorname{rank}(\Pi_z(z))=d_x$ is the condition for identification of one equation of a linear simultaneous system, in terms of the reduced form coefficients. Thus, the rank condition in the general case is a nonlinear, nonparametric generalization of the rank condition.

The identification result of Florens et al. (2008) allows us to identify treatment effects that incorporate more interaction terms among unobserved heterogeneity and the endogenous regressors. However, they focus on the particular case of $d_x=1$. Additionally, the identifying assumptions they used are different than those used in Theorem 1. They imposed measurable separability for x and v, which is a type of rank condition.[2] Last, they assumed conditional independence between the instrument and the error term instead of conditional mean independence as assumed here.

4. Estimation

In this section, I briefly discuss the estimation strategy I follow in the empirical application discussed in Section 5. I estimate a single equation with a scalar endogenous regressor and I follow a two-step procedure similar to Newey, Powell, and Vella (1999).

[2] By definition, x and v are measurably separated if a function of x is equal to a function of v, almost surely, then they are constant, almost surely.

I propose the use of splines to approximate the functions $g(x)$, $\lambda(v)$, and $\gamma(v)$,

$$\begin{aligned} E[y|x,z] &= g(x) + \lambda(\hat{v}) + \gamma(\hat{v})x \\ &\cong \sum_{i=1}^{I} \alpha_i \rho_i(x) + \sum_{j=1}^{J} \beta_j \psi_j(\hat{v}) + \sum_{k=1}^{K} \gamma_k \chi(\hat{v})x, \end{aligned} \tag{13}$$

where $\hat{v}$ is the estimated residual from the first step and $\{\rho_i\}_{i=1}^{I}$, $\{\psi_j\}_{j=1}^{J}$, and $\{\chi_k\}_{k=1}^{K}$ are appropriate basis functions with J, L, and K increasing to infinity as the sample size increases. Note that this reduces to a least squares regression of y on the basis functions $\{\rho_i(x)\}_{i=1}^{I}$, $\{\psi_j(\hat{v})\}_{j=1}^{J}$, and $\{\chi_k(\hat{v})x\}_{k=1}^{K}$, and the estimators of $g(x)$ and $g_x(x)$ are given by

$$\hat{g}(x) = \sum_{i=1}^{I} \hat{\alpha}_i \rho_i(x), \qquad \hat{g}_x(x) = \sum_{i=1}^{I} \hat{\alpha}_i \rho_{x_i}(x), \tag{14}$$

where $\rho_{x_i}(\cdot)$ is the first derivative of $\rho_i(\cdot)$ and, imposing the normalization $E[\lambda(v)] = 0$, $E[\gamma(v)] = 0$. This is done by normalizing the basis functions to sum to 1 through the sample data and constraining the parameters of each function to sum to 0.

Newey, Powell, and Vella (1999) proposed the use of a cross-validation (CV) criterion to determine the optimal number of basis functions that act as smoothing parameters. I depart from their approach and follow the statistical literature of penalized regression splines (see Silverman (1985) and Wood (2004)).[3] In this approach, the basis functions are cubic splines and penalty matrices are incorporated into the objective function that is minimized. These penalty matrices penalize the "wiggliness" of the functions. Hence, the minimization problem is given by

$$\min_{\{\alpha_i, \beta_j, \gamma_k\}} \sum_{k=1}^{n} \bigl(y - g(x) - \lambda(\hat{v}) - \gamma(\hat{v})x\bigr)^2 + \sum_{i=0}^{2} \theta_i \mathcal{P}_i, \tag{15}$$

where θ_i are parameters that control the smoothness of the functions and $\mathcal{P}_i$ are the penalty matrices

$$\mathcal{P}_0 = \int \bigl[g''(x)\bigr]^2\,dx; \qquad \mathcal{P}_1 = \int \bigl[\lambda''(v)\bigr]^2\,dv; \qquad \mathcal{P}_2 = \int \bigl[\gamma''(v)\bigr]^2\,dv, \tag{16}$$

where $[g''(x)]^2$, $[\lambda''(v)]^2$, and $[\gamma''(v)]^2$ measure the square curvature of the functions, so these penalties measure global square curvature.

In penalized regression splines, the smoothness is controlled exclusively by the smoothing parameter. Since the smoothing parameter also controls for degrees of freedom, the approach here is to set the number of basis functions as large as possible in terms of computational manageability and/or in terms of finite support restrictions. Knots are placed uniformly across the support of the independent variable as well, but

[3] As a robustness check, I implement the unpenalized splines strategy of Newey, Powell, and Vella (1999), obtaining very similar results.

since the number of basis functions is large, the smoothness is not compromised by their positions. Given the continuous nature of the smoothing parameters, the cross-validation (CV) score can be minimized by Newton's method, which relieves some of the computational burden. In the empirical application, I will use a generalized cross-validation (GCV) criterion developed by Wood (2004).

5. Application to the returns to schooling

In this section, I implement the proposed method to the estimation of the returns to schooling in Chile. It is important to mention that even though Chesher (2003), Imbens and Newey (2009), Florens et al. (2008), and Torgovitsky (2011) motivated their work with the estimation of the returns to schooling, none of them actually applied their proposed method to real data.

The implementation of the model developed in Section 3 to the estimation of the returns to schooling is related to Card (1995). He assumed that the marginal return to schooling is a linear function of schooling (x) plus a heterogeneity component (b_i),

$$y'(x)/y(x) = b_i + k_1 x,$$

where b_i is an unobserved term that represents individual heterogeneity and k_1 is a constant parameter. The model proposed in Section 3 is well suited in this context since it can be recovered from a slightly different assumption about the marginal return, that is,

$$y'(x)/y(x) = g'(x) + \phi_1(u),$$

in which I have added nonlinearity to the marginal returns to schooling and replaced the heterogeneity term b_i with $\phi_1(u)$. Integrating with respect to x, it follows that

$$\ln y = \phi_0(u) + g(x) + \phi_1(u)x,$$

which is exactly Model 2—the one that controls for endogeneity and sorting.

This model is suitable for estimation of the returns to schooling for several reasons. First, it allows us to recover endogeneity-corrected *local*, instead of average, returns to schooling. Second, it introduces corrections for two types of bias: omitted variable bias (absolute advantage) and ability bias (comparative advantage). Third, it permits us to obtain causal parameters from the population and subpopulations, *deconfounding* the sorting effect and the ATE when estimating in subpopulations.

5.1 *Data*

The primary data source I use is the Chilean household survey CASEN (National Socioeconomic Characterization Survey). The CASEN survey is representative at the regional as well as national level. This survey is conducted by the Ministry of Social Development (MDS) and it has been conducted every 2 or 3 years since 1990. The last wave of the survey was conducted in 2011.

The main purpose of the CASEN survey is to describe the socioeconomic conditions in Chile and also to evaluate social policies. The interviews are conducted at household and individual levels. The information obtained for each member of the household includes a description of the person's income, educational characteristics, health services, participation in social programs, and assistance received.

In this paper, I work with the 1992 wave of the CASEN survey so as to identify the older cohorts of the period under study. The choice of the 1992 over the 1990 wave was mainly driven by sample size advantages of the former and to avoid contamination of the estimation of the returns to schooling given the political changes that took place in Chile in 1990.[4]

The sample used in the earnings equations includes males born between the 1921 and 1965 cohorts. I choose to work with males only to avoid selectivity bias, considering that the male participation rate was above 80% in 1992 and the female participation rate was only about 38% in the same year. The sample size was 26,011 observations (about 6000 more than the 1990 sample): 43% of the sample reside in the south, 41% in the center, and the rest in the north.

5.2 *School reforms and institutional changes in Chile 1930–1965*

Two important reforms were introduced in Chile in the late 1920s and early 1930s: the increase in the compulsory schooling attendance grade in 1929 and the new labor code approved in 1931. Compulsory attendance laws in Chile relate to compulsory minimum grade instead of age to drop out of school. The compulsory leaving grade was raised from 4th to 6th grade nationwide in 1929, but it is likely that the enforcement had begun after the recovery from the overwhelming effects of the Great Depression on the Chilean economy.[5] The new labor code included a contract law that prohibited 17-year-old or younger workers from working in underground jobs (mining) and on any kind of job involving excessive physical work. Given that most of the mining industry was concentrated in the north and it was almost the only economic activity in the region, this law, in conjunction with the increase in compulsory attendance, is likely to have created an asymmetric impact on school attainment by region.

Figure 1 shows the average school attainment for men born between 1920 and 1970 per geographic region.[6] The vertical lines correspond to the structural breaks in the series estimated by conventional time series analysis and are only reference points. The first thing to note is the behavior of school attainment in the north. While the south and center increase at a similar rate during the entire period, the north grows at a faster rate since the early 1930s, reaching the center in the mid 1950s. After that, the three attainment series seem to increase at a lower rate and behave similarly up to 1965. The series were calculated using the CASEN survey, were corroborated with census data, and include people currently living in that geographic area. Since internal migration in Chile

[4] In 1990, Chile returned to democracy after 17 years of dictatorship.

[5] A second increase in compulsory education, from 6 to 8 years of schooling, took place in 1965, but I will focus on the one that occurred in 1929.

[6] The geographic regions were grouped as follow: the north includes regions I–IV; the center includes regions V–VII and the metropolitan region; the south consists of regions VIII–XII.

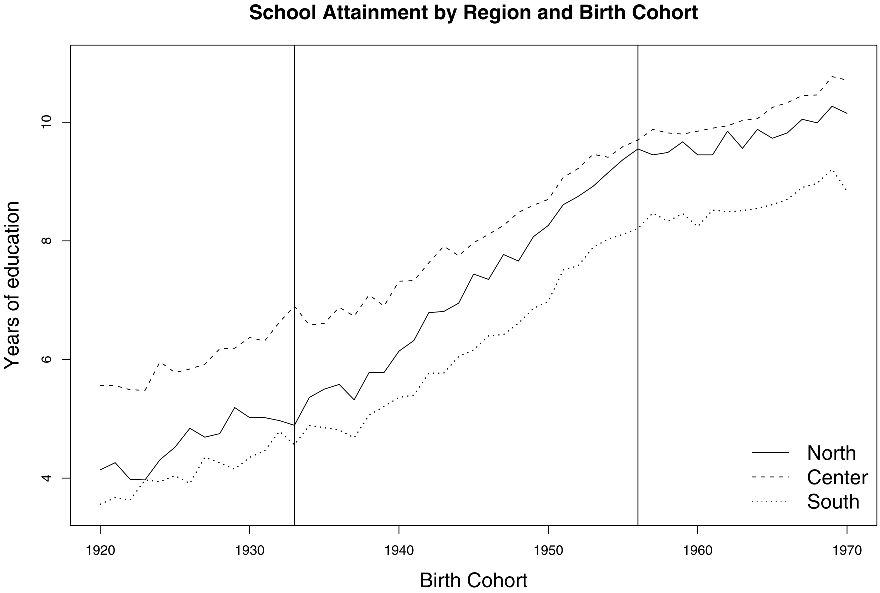


FIGURE 1. School attainment per birth cohort and geographical region for men, 1920–1970. The vertical lines correspond to the structural breaks in the series estimated by conventional time series analysis.

has been relatively low, I rely on these estimates as the average school attainment per cohort–region.[7]

One might expect to observe changes in the pattern of the series (if at all) for schooling attainment for individuals born earlier than 1930, since the increase in the compulsory leaving grade took place in 1929.

In my view, the political and economic crisis between 1924 and 1932 might have had an impact on children's labor participation (delaying their compliance with the new law) and in the enforcement of the new legislation as well.[8] A second thought concerns the impact of the further increase in the compulsory leaving grade (from 6th to 8th) in 1965: since the information displayed is for birth cohorts, one might expect changes in the pattern of the series circa 1953. Inspection suggests that there were no changes until 1956, and, indeed, a reduction of the growth rate of attainment in the three regions is observed. I do not have a neat interpretation of this fact, but it appears to be clear that the behavior of the series can be characterized, in terms of structural breaks, into three periods: from 1921 to 1933, between 1933 and 1956, and after 1956.

It is interesting to observe whether the behavior of school attainment relates to the behavior of dropout rates. Figure 2 shows the dropout rate from primary education of the three regions by cohort. These rates were calculated using the CASEN survey and

[7] Based on calculations using census data, about 0.72 percent of the northern population migrated to the center/south per year in the 1930–1960 period. Also, Soto and Torche (2004) showed that in the 1965–2000 period, only around 0.6 percent of the population moved between the 13 regions every year.

[8] In the 1924–1927 period, there were six presidents and two coups d'ètat, and in 1932, there were five presidents.

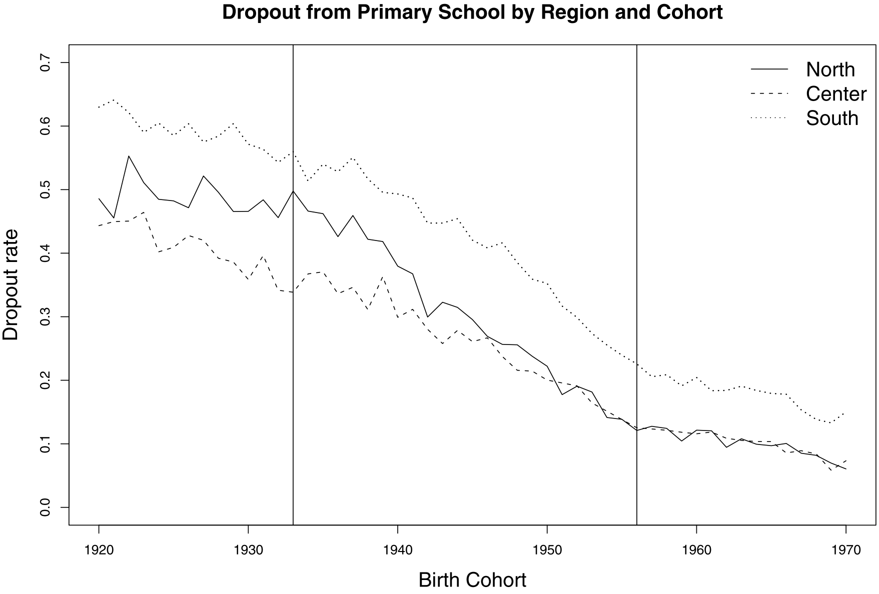


Figure 2. Dropout rate from primary education per birth cohort and geographical region for men, 1920–1970.

include people with 1–5 years of education divided by the number of people with some education in a particular cohort.[9] As can be observed, all the regions decline in parallel until 1936; then the north starts a fast decline until it reaches the center in the late 1950s. This asymmetric decline in the dropout rate agrees with the increase in school attainment in the north relative to the south and center. Now, it might be interesting to check what happened to the enrollment rate in secondary education since the change in schooling trend is so pronounced that one would expect that some people not only complied with the compulsory grade, but also continued their schooling to secondary education.

Figure 3 presents the enrollment rates for secondary education per region. These series were constructed from administrative data and correspond to the number of people enrolled in secondary education divided by the relevant population, where "relevant" refers to people in the 13–18 year range of age.[10] Observe that in the north, the enrollment rate for secondary education changes its behavior in 1940, starting a significant increase relative to the south until 1965. Secondary enrollment in 1940 includes individuals in the 1924–1928 birth cohort, which agrees with the increase in the compulsory leaving age in 1929.

The evidence provided suggests that interactions of region and cohort dummies can be used to generate a valid instrumental variable for schooling. The correlation between

[9] Until 1965, primary and secondary education took 6 years. In 1965, primary education was extended to 8 years, and so the compulsory leaving grade, and secondary education was reduced to 4 years.

[10] The relevant population was estimated using census data to project the population according to the age pyramid population growth rates.

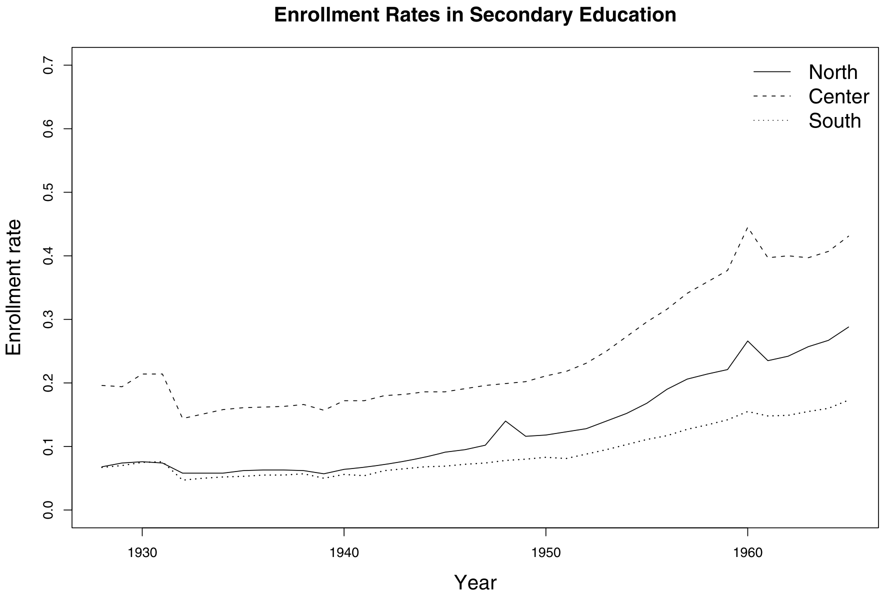


FIGURE 3. Enrollment rate in secondary education per year and geographical region for men, 1920–1970.

region–cohort trends and schooling appears to be quite strong by simple inspection and it will be corroborated in the first stage. I do not find a strong reason to believe that the mix of compulsory law and contract law might be correlated with unobservables such as ability or family background.[11] However, the pattern observed in the evolution of educational achievement in the three regions may have been correlated with economic activity and opportunities. In the Appendix, I provide two figures that show the evolution of exports and cabotage for the 1920–1970 period, and find no relationship with the pattern of educational outcomes.

6. Results

In this section, I provide the results for the nested models explained above. Table 2 presents the results of the first stage. As mentioned before, I exploit the asymmetric impact of the changes in compulsory schooling and labor laws across age and regions to generate a valid instrument for schooling. The F-statistic of the excluded instruments is equal to 14.9, which is larger than 10, a typical threshold value indicated as adequate.[12]

[11] One argument could be that parents with low taste for schooling might have migrated to the center to allow their children to work after they had achieved the compulsory leaving grade and the minimum legal age to work. Given that migrants were, in general, equally or more educated than central natives, it is hard to think that there were differences in taste for education between central natives and northern migrants. Also, the flows of internal migration from north to center and south were only about 0.72 percent of the population per year.

[12] The rule of thumb indicates that an F-statistic of the excluded instruments above 10 is more than adequate to rule out the presence of weak instruments (see Staigner and Stock (1997), Rothenberg (1984), and Stock and Yogo (2005)).

Table 2. First stage regression of years of education.

Variable	Coef.	Std. Err.	t-Stat.	p-Value
Age	1.881	0.643	2.92	0.0030
Age^2	−0.054	0.022	−2.5	0.0120
Age^3	0.001	0.000	1.91	0.0560
Age^4	−0.000	0.000	−1.36	0.1750
Center	−0.874	0.310	−2.82	0.0050
South	−1.389	0.308	−4.51	0.0000
Center × age	0.021	0.007	2.93	0.0030
South × age	−0.007	0.007	−0.96	0.3350
Constant	−11.695	6.966	−1.68	0.0930
R-squared	0.117			
$F(8, 26{,}002)$	430.320			
$F(2, 26{,}002)$	14.900		(excluded instruments[a])	

[a] Excluded instruments are age and region interactions.

Table 3. Models.[a]

Model	Equation
Uncorrected	$\ln y = g(x) + \xi$
Model 1	$\ln y = g(x) + \lambda(v) + \xi$
Model 2	$\ln y = g(x) + \lambda(v) + \gamma(v)x + \xi$

[a] All include a function of age and region dummies.

However, more formal procedures have been developed in recent years. Stock and Yogo (2005) showed that the F-statistic (in the case of one endogenous variable) is equivalent to the Cragg–Donald statistic and can be used to test for weak instruments. According to their criteria, I can bound the actual size of a Wald test with a nominal size of 5% to a maximum of 15% (the critical value is 11.59). The minimum bound for Wald tests with a nominal size of 5% they computed is 10%; hence my result can be considered to be a relatively strong instrument given its proximity to that value.

Table 3 summarizes the specification of the relevant models. First, I estimate an *uncorrected* model, which is a nonparametric regression of $\ln y$ on x that will provide the baseline estimates of the returns to schooling. Then I estimate Model 1, which is a nonparametric regression with endogenous regressor proposed by Newey, Powell, and Vella (1999). Last, I estimate Model 2, which is a nonparametric regression with endogenous regressor and sorting analyzed in Section 2.

To determine which model fits the data better, I perform an analysis of deviance (ANODEV) for the three nested models, that is, the uncorrected model, Model 1, and Model 2. In Table 4, the ANODEV analysis indicates that the null hypothesis of $\gamma(v)x = 0$ is rejected at the 5% level. Note that the *p-values* are approximate since they do not include the uncertainty added by the estimation of the smoothing parameters. Unless the model is unpenalized or the smoothing parameters are known, p-values tend to be somewhat low. According to Wood (2006), simulation results show that p-values might

Table 4. Deviance table for nested models.[a]

Model	Residual DOF	Residual Dev.	DOF	Deviance	F-Stat.	p-Value
Uncorrected	25,994.06	14,736.9				
Model 1	25,987.79	14,666.6	6.35	70.2	20.25	0.0000
Model 2	25,987.59	14,664.5	0.40	2.2	9.5	0.0144

[a] DOF = degrees of freedom.

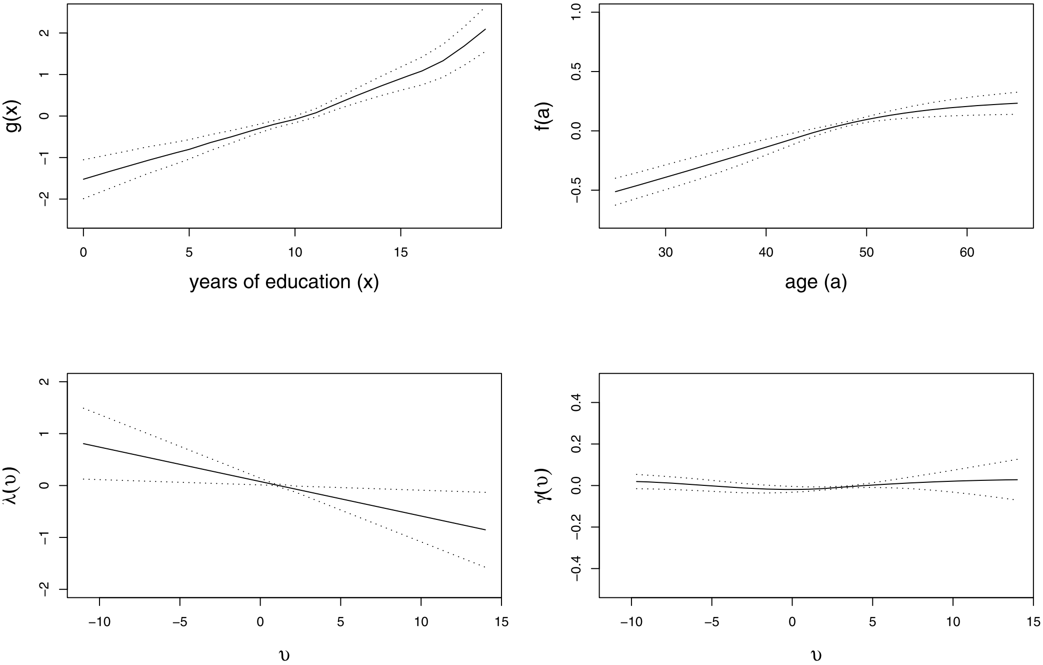


Figure 4. Estimated semiparametric functions and their 95% confidence intervals for Model 2.

be as little as half the correct value at the 5% level when the null is true. Given our p-value of 0.014, I am inclined to rely on the rejection of the null hypothesis discussed.

6.1 *Returns to schooling*

In this subsection, I analyze the effect of the two control functions—absolute and comparative advantages—on the log wage. Figure 4 graphs the four functions estimated in Model 2 and their 95% confidence intervals. The top-left panel shows the plot for $g(x)$, which is increasing in schooling and presents some curvature in the middle and upper parts that will translate into differences in local returns to schooling. The top-right panel shows the plot for the nonparametric function of age $f(a)$, which increases smoothly at a decreasing rate. The bottom-left panel shows the graph of $\lambda(v)$, which decreases almost linearly from 1 to -1, approximately. The bottom-right panel shows the behavior of $\gamma(v)$, which first decreases and then increases smoothly with a very small slope. This

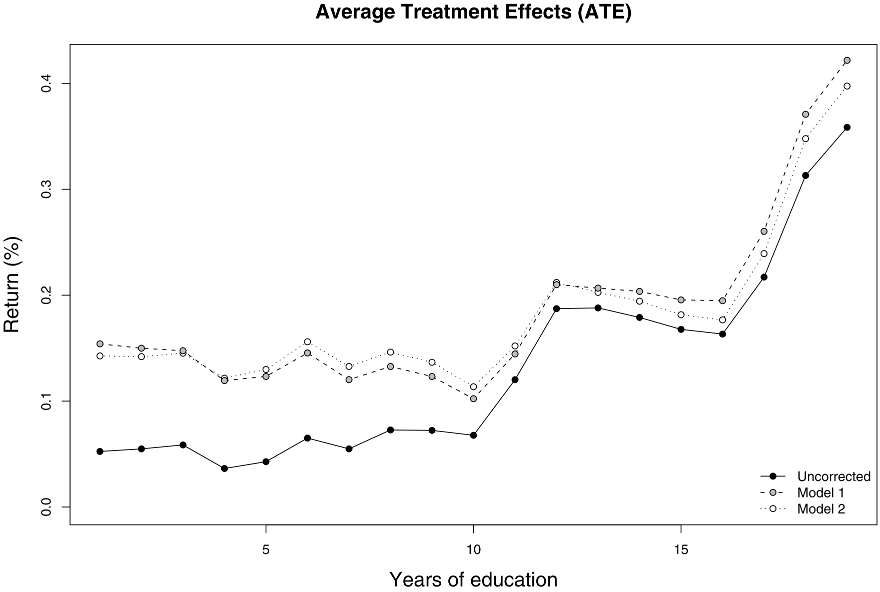


FIGURE 5. Average treatment effects for the three models: ATE $= g(x+1) - g(x)$.

suggests that the importance of the comparative advantage bias is low relative to the absolute advantage bias.

To see the effect of endogeneity in local returns to schooling, I present Figure 5, which summarizes the behavior of the local returns after correcting for endogeneity. Since $d_x = 1$ in this application, the average treatment effect is simply ATE $= E[H(x+1, u) - H(x, u)] = g(x+1) - g(x)$. It can be seen that endogeneity-corrected local returns to schooling are larger than those from the uncorrected model for individuals with up to a 10 years of education. The profile is flatter than for uncorrected measures, and there is some evidence of sheepskin effects for high school and college degree that I discuss later. After high school, the effect of endogeneity on local returns to schooling appears to vanish. This seems intuitive since the subpopulation more likely affected by the instrument are those in lower grades.

The pattern of bias-corrected local returns to schooling provides evidence that the change in compulsory and child labor laws affected a subpopulation of low-education individuals, decreasing their marginal cost of schooling by reducing their opportunity cost. Since this subpopulation had low schooling for reasons other than ability before the intervention, endogeneity-purged estimates of local returns to schooling are higher than the population local returns to schooling. This corresponds to the *discount rate* hypothesis in explaining why instrumental variable (IV) estimates of the returns to schooling are typically higher than ordinary least squares (OLS). People affected by the instrument had low education due to higher discount rates instead of lower ability.[13]

[13] This result of IV estimates being higher than OLS estimates of the returns to schooling have been called in the literature the OLS–IV puzzle, since if OLS estimates are upwardly biased due to the omission of ability,

Table 5. Returns to schooling (ATE), all models.

Years of Education	Uncorrected		Model 1		Model 2	
	ATE	SD[a]	ATE	SD[a]	ATE	SD[a]
1	0.052	0.038	0.154	0.023	0.143	0.036
2	0.055	0.031	0.150	0.014	0.142	0.031
3	0.059	0.030	0.148	0.016	0.145	0.030
4	0.036	0.029	0.119	0.014	0.122	0.030
5	0.043	0.028	0.123	0.012	0.130	0.029
6	0.065	0.031	0.145	0.013	0.156	0.029
7	0.055	0.028	0.120	0.013	0.133	0.030
8	0.073	0.029	0.133	0.015	0.146	0.031
9	0.072	0.031	0.123	0.014	0.137	0.032
10	0.068	0.029	0.102	0.017	0.114	0.033
11	0.120	0.033	0.145	0.019	0.152	0.033
12	0.187	0.035	0.210	0.022	0.212	0.037
13	0.188	0.036	0.207	0.024	0.203	0.041
14	0.179	0.032	0.204	0.021	0.194	0.042
15	0.168	0.035	0.196	0.024	0.182	0.043
16	0.163	0.035	0.195	0.023	0.177	0.046
17	0.217	0.037	0.260	0.026	0.239	0.051
18	0.313	0.051	0.371	0.038	0.348	0.062
19	0.359	0.063	0.422	0.052	0.397	0.075

[a] Wild bootstrapped standard deviations.

Complementing the results shown in Figure 5, in Table 5, I present the returns to schooling (ATE) for the three models with bootstrapped standard errors. I implement a *wild* bootstrap that provides valid inference in the case of penalized spline estimation as shown by Kauermann, Claeskens, and Opsomer (2009).[14]

It is worth noting the presence of sheepskin effects in high school and college degrees. In high school education, the returns to schooling jumps from about 15% to 21% from 11 to 12 years of education. Then it decreases to about 18% for the fourth year of college (16 years of education). For college degrees, the sheepskin effect is important too: the returns to schooling increases from 18% to 24% from 16 to 17 years of education and then from 24% to 35% from 17 to 18 years of education.[15]

The estimates for Models 1 and 2 do not differ statistically; however, there are significant differences between Model 2 and the uncorrected model up to 10th grade. In Figure 6, we can see the differences between the two models and the 95% confidence interval obtained by wild bootstrap. This provides strong evidence on the effects of endogeneity in local returns to schooling.

IV estimates should be lower than OLS, which is not the case in most empirical applications. For more details, see Belzil and Hansen (2005).

[14] The authors do not consider the case of generated regressors, which is a delicate issue of current research in nonparametric econometrics (Hahn and Ridder (2011) and Mammen, Rothe, and Schienle (2011)). Thus, the reported standard errors are likely to be underestimated.

[15] Most college careers last between 5 and 6 years in Chile.

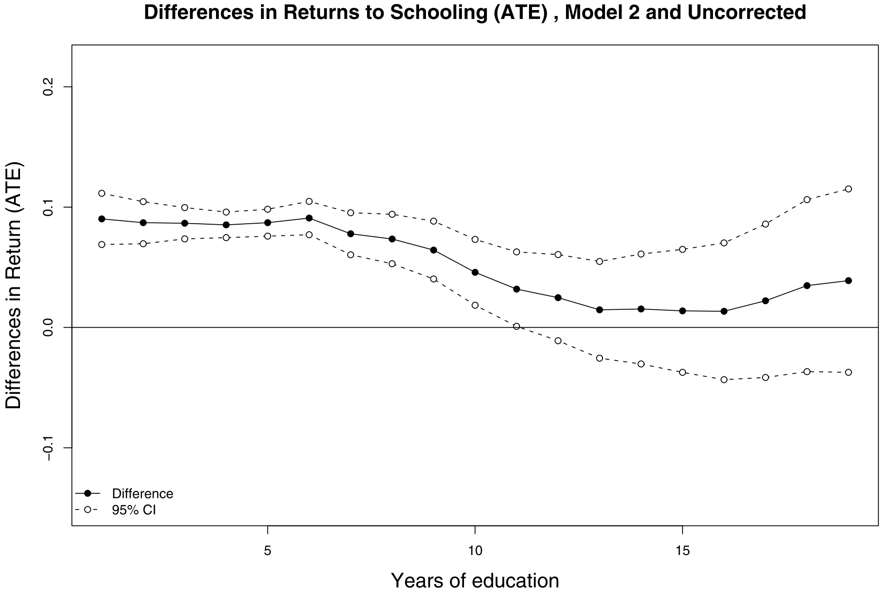


FIGURE 6. Differences in ATE comparing Model 2 and the uncorrected model. The solid line connects the point estimate difference and the dashed line connects the 95% confidence interval of the difference computed by wild bootstrap.

Last, we compare ATE and SATE from Model 2, since this is the only model in which both treatments are allowed to differ. In this setup, the SATE effect is simply SATE $= E[H(x+1,u) - H(x,u)|x] = g(x+1) - g(x) + E[\gamma(v)|x]$. In Figure 7, we can see both treatment effects: ATE and SATE. Model 2 allows us to obtain an estimate of the term $E[\gamma(v)|x]$, which is the sorting effect.

As mentioned before, population and local effects in Model 2 only differ by the sorting effect, that is, by the term $E[\gamma(v)|x]$. What is striking is the behavior of the sorting effect, which, until the last year of high school, is negative and then positive for college and graduate education. This seems intuitive since it is expected that more able individuals sort into upper grades. However, the effect seems rather small in comparison with the absolute comparative advantage addressed by the first control function. Figure 8 graphs the effect of sorting on subpopulation treatment effects. As mentioned before, it is negative until 12th grade and then positive thereafter.

7. CONCLUSIONS

In this paper, I develop a general modeling framework for microeconometric relationships with endogenous regressors and heterogeneity trough shape restrictions. By assuming constancy in partial derivatives of a structural equation $H(x,u)$, I show that many models in the related literature can be obtained.

I develop a nonparametric model to estimate nonlinear structural equations with endogenous regressors and heterogeneity. The model developed in this paper can be recovered from partial derivative equality constraints in the second derivative, which generates a model with additive separability in the first derivative. My model general-

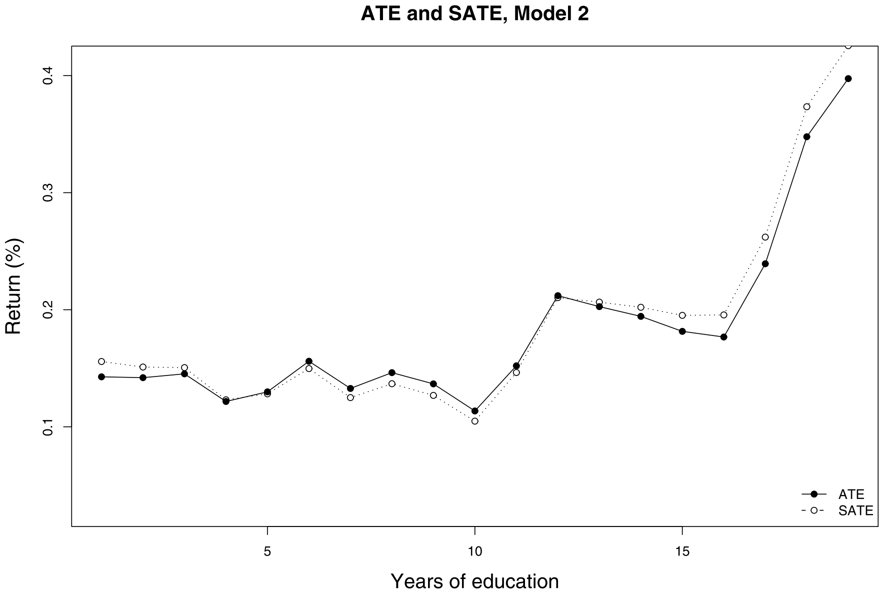


Figure 7. Average treatment effects from Model 2. ATE $= g(x+1) - g(x)$ and SATE $= g(x+1) - g(x) + E[\gamma(v)|x]$.

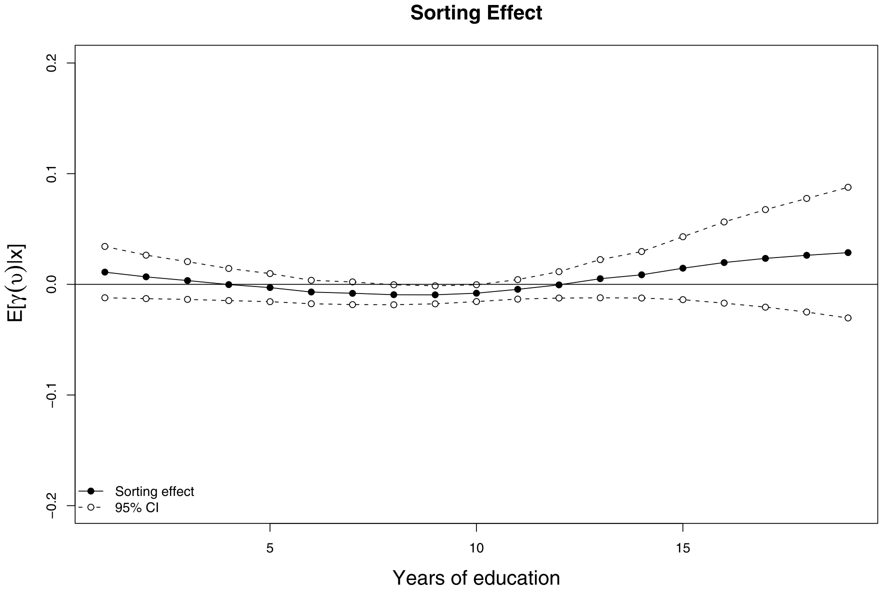


Figure 8. Effect of sorting on the treatment effects.

izes Garen (1984) and Newey, Powell, and Vella (1999), and shares the essential features of those models. It controls for endogeneity and sorting as in the correlated random coefficient model, and it permits nonlinearities in the endogenous regressor as in a nonparametric regression.

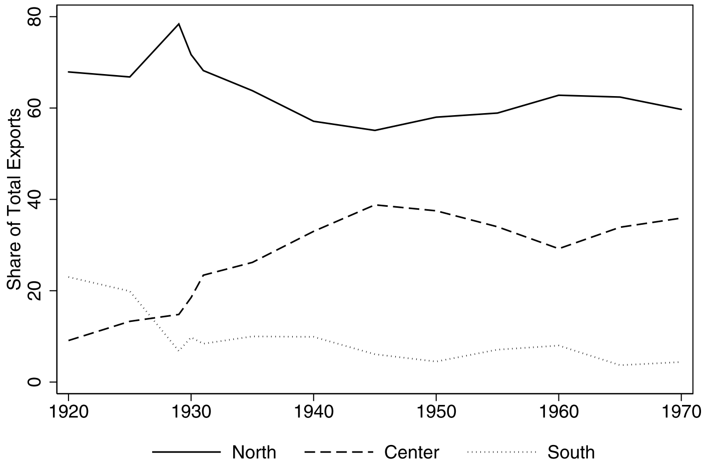
I show that this model allows one to obtain causal parameters from counterfactuals of "ideal" randomization in the population, as well as causal parameters from counterfactuals of randomization in subpopulations of self-selected individuals. I show that these subpopulation parameters can be decomposed into population parameters and a sorting effect. This may be very useful when evaluating policies in which the selection rule is correlated with the unobserved heterogeneity, such as high school completion programs.

The model developed here is identified with conditional mean independence assumptions and a generalized control function approach is considered. I provide formal identification results for the average structural function, the average treatment effect, and the subpopulation average treatment effect.

An application to the estimation of the returns to schooling is developed. I use Chilean data that exploit variation across regions and cohorts in educational infrastructure and compulsory attendance laws during the 1921–1965 period as a valid instrument for education. To estimate the model, I use penalized spline regression to estimate the additive components, where the smoothing parameters are chosen according to a generalized cross-validation criterion. The most interesting results are the asymmetric impact of endogeneity on the local returns to schooling and that I am able to identify the subpopulation affected by the instrument. It can be concluded that endogeneity causes a downward bias to the returns to schooling for individuals with up to 10 years of education.

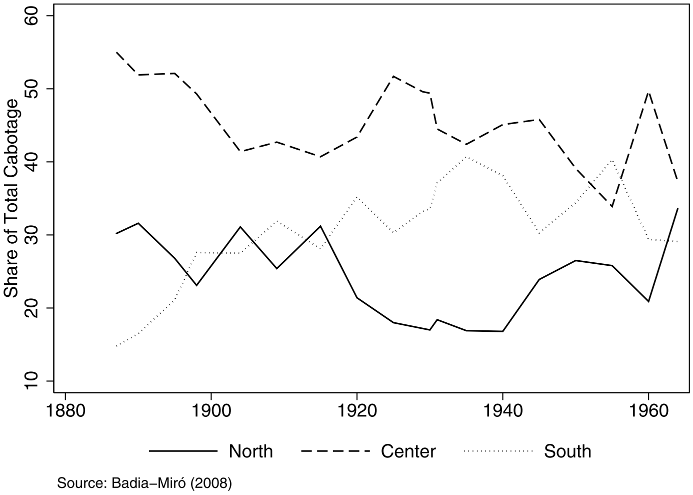
Appendix

Here I present data on exports and cabotage for the three regions between 1920 and 1970 taken from Badia-Miró (2008). As can be seen in Figures 9 and 10, there is no relationship of these patterns with those found in schooling in Figure 1.



Source: Badia–Miró (2008)

Figure 9. Exports by geographic region.



Source: Badia–Miró (2008)

Figure 10. Cabotage by geographic region.

References

Altonji, J. G. and R. L. Matzkin (2005), "Cross section and panel data estimators for nonseparable models with endogenous regressors." *Econometrica*, 73 (4), 1053–1102. [130]

Badia-Miró, M. (2008), *La Localización de la Actividad Económica en Chile, 1890–1973. Su Impacto de Largo Plazo.* Tesis Doctoral, Universidad de Barcelona. [145]

Belzil, C. and J. Hansen (2005), "A structural analysis of the correlated random coefficient wage regression model with an application to the OLS–IV puzzle." Discussion Paper 1585, Institute for the Study of Labor (IZA). [142]

Blundell, R. and J. L. Powell (2003), "Endogeneity in nonparametric and semiparametric regression models." In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. II (L. P. Hansen and S. J. Turnovsky, eds.), 312–358, Cambridge University Press, Cambridge. [129]

Card, D. (1995), "Earnings, schooling and ability revisited." *Research in Labor Economics*, 14, 23–48. [134]

Chay, K. and M. Greenstone (2005), "Does air quality matter? Evidence from the housing market." *Journal of Political Economy*, 113 (2), 376–424. [128]

Chernozhukov, V. and C. Hansen (2006), "Instrumental quantile regression inference for structural and treatment effect models." *Journal of Econometrics*, 132 (2), 491–525. [126]

Chesher, A. (2003), "Identification in nonseparable models." *Econometrica*, 71 (5), 1405–1441. [126, 134]

Chesher, A. (2005), "Nonparametric identification under discrete variation." *Econometrica*, 73 (5), 1525–1550. [126]

Chesher, A. (2007), "Identification of non-additive structural functions." In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Vol. 3 (R. Blundell, W. K. Newey, and T. Persson, eds.), 1–16, Cambridge University Press, New York. [126]

Florens, J. P., J. Heckman, C. Meghir, and E. Vytlacil (2008), "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects." *Econometrica*, 76 (5), 1191–1206. [126, 128, 129, 130, 132, 134]

Garen, J. (1984), "The returns to schooling: A selectivity bias approach with a continuous choice variable." *Econometrica*, 52 (5), 1199–1218. [126, 128, 144]

Hahn, J. and G. Ridder (2011), "Conditional moment restrictions and triangular simultaneous equations." *Review of Economics and Statistics*, 93 (2), 683–689. [130, 142]

Imbens, G. and W. K. Newey (2009), "Identification and estimation on triangular simultaneous equation models without additivity." *Econometrica*, 77 (5), 1481–1512. [126, 134]

Kauermann, G., G. Claeskens, and J. Opsomer (2009), "Bootstrapping for penalized spline regression." *Journal of Computational and Graphical Statistics*, 18, 126–146. [142]

Mammen, E., C. Rothe, and M. Schienle (2011), "Semiparametric estimation with generated covariates." Discussion Paper 6084, Institute for the Study of Labor (IZA). [142]

Newey, W. K., J. L. Powell, and F. Vella (1999), "Nonparametric estimation of triangular simultaneous equations models." *Econometrica*, 67 (3), 595–603. [126, 128, 130, 131, 132, 133, 139, 144]

Rau, T. (2007), *Essays on Applied Semiparametric Econometrics.* Ph.D. dissertation, University of California, Berkeley. [128, 129, 130]

Rothenberg, T. J. (1984), "Approximating the distribution of econometric estimators and test statistics." In *Handbook of Econometrics*, Vol. 2 (Z. Griliches and M. D. Intriligator, eds.), 881–935, North-Holland, Amsterdam. [138]

Silverman, B. W. (1985), "Some aspects of the spline smoothing approach to nonparametric regression curve fitting." *Journal of the Royal Statistical Society, Ser. B*, 47, 1–52. [133]

Soto, R. and A. Torche (2004), "Spatial inequality, migration and economic growth in Chile." *Cuadernos de Econonomía*, 41 (124), 401–424. [136]

Staigner, D. and J. H. Stock (1997), "Instrumental variables regression with weak instruments." *Econometrica*, 65 (3), 557–586. [138]

Stock, J. and M. Yogo (2005), "Testing for weak instruments in linear IV regression." In *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg* (D. W. K. Andrews and J. H. Stock, eds.), 80–108, Cambridge University Press, Cambridge. [138, 139]

Torgovitsky, A. (2011), "Identification of nonseparable models with general instruments." Working paper, Yale University. [126, 134]

Wood, S. N. (2004), "Stable and efficient multiple smoothing parameter estimation for generalized additive models." *Journal of the American Statistical Association*, 99 (467), 673–686. [133, 134]

Wood, S. N. (2006), *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, Boca Raton. [139]

Submitted December, 2010. Final version accepted October, 2012.