Universität St.Gallen

A Note on Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables

Markus Frölich

Mai 2006 Discussion Paper no. 2006-11

Department of Economics University of St. Gallen

Editor:	Prof. Jörg Baumberger University of St. Gallen Department of Economics Bodanstr. 1 CH-9000 St. Gallen	
	Phone	+41 71 224 22 41
	Fax	+41 71 224 28 85
	Email	joerg.baumberger@unisg.ch
Publisher:	Department of Economics University of St. Gallen Bodanstrasse 8 CH-9000 St. Gallen	
	Phone	+41 71 224 23 25
	Fax	+41 71 224 22 98
Electronic Publication:	http://www.vwa.unisg.ch	

A Note on Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables[1]

Markus Frölich

Author's address: Dr. Markus Frölich
SIAW
Bodanstr. 8
9000 St.Gallen
Email markus.froelich@unisg.ch
Website www.siaw.unisg.ch/froelich

[1] The author is also affiliated with IZA (Bonn) and IFAU (Uppsala). I am grateful for discussions and comments to Samuel Berlinski, Michael Lechner, Patrick Puhani, Barbara Sianesi and Jeff Smith. This research was supported by the Marie Curie Individual Fellowship MEIF-CT-2004-006873.

Abstract

This note argues that nonparametric regression not only relaxes functional form assumptions vis-a-vis parametric regression, but that it also permits endogenous control variables. To control for selection bias or to make an exclusion restriction in instrumental variables regression valid, additional control variables are often added to a regression. If any of these control variables is endogenous, OLS or 2SLS would be inconsistent and would require further instrumental variables. Nonparametric approaches are still consistent, though. A few examples are examined and it is found that the asymptotic bias of OLS can indeed be very large.

Keywords

Endogeneity, nonparametric regression, instrumental variables

JEL Classification

C13, C14

1 Introduction

The recent literature on treatment evaluation pointed out two practical advantages of nonparametric matching methods over parametric regression such as OLS or 2SLS: Nonparametric methods relax the linearity assumption, and they assist in highlighting differences in the supports of the observable variables.[1] Nonparametric methods, however, have another advantage that is less often acknowledged but is of high practical relevance: nonparametric regression permits *endogenous control variables.*

In many empirical applications, additional "control" variables are added in a regression that are not of interest in themselves but are included to control for selection bias or omitted variable bias.[2] However, if any of these control variables is correlated with the error term,[3] generally all OLS estimates are inconsistent, and instrumental variables need to be found for the endogenous control variables. Similarly, with instrumental variable regression where it is also often necessary to include control variables to make the exclusion restriction valid. Again, if any of these control variables themselves are correlated with the error term in the main equation, the 2SLS estimates are inconsistent. Additional instrumental variables would be necessary.

Nonparametric regression approaches, however, would still be consistent and would thereby avoid the need for (additional) instrumental variables, which in many applications are difficult to find. The reason for this is that nonparametric regression compares only across individuals who have the same values for the control variables and differ only in the treatment variable, whereas parametric regression combines all observations in a single global regression. This situation is discussed in more detail in the following section and in section 3 a few examples are given to show that the asymptotic bias of parametric regression can indeed be very large.

[1] See in particular the discussion on propensity score matching in Black and Smith (2004).

[2] It is now common practice in many applied journals to report only the estimated coefficients on the variables of interest and to list only the names of all the additional control regressors in a footnote.

[3] Lechner (2006) analyzes endogenous control variables when the control variable is causally affected by the treatment variable. Here, the focus is on endogeneity in the conventional sense: a variable that is correlated with the error term.

2 Endogenous control variables

Consider an educational production function with true data generating process:

$$Y_{i,end} = \alpha + \beta \cdot Teacher_i + \gamma Y_{i,begin} + \delta X_i + U_i$$

where $Y_{i,end}$ is the educational achievement (e.g. exam results) of student i at the end of the school year, $Y_{i,begin}$ is achievement at the beginning of the school year,[4] X_i are some other student characteristics and $Teacher_i$ are some characteristics of the teacher of student i, e.g. salary, qualification, union status, participation in some teacher motivation scheme or material support.[5] This specification is often also called the value-added approach since achievement at the beginning of the school year is included, see e.g. Hanushek (1986).

In the following, we are interested only in the effect of the teacher characteristics on student achievement and treat the other regressors as control variables. In other words, we are interested in the *treatment effect* β of $Teacher_i$ on $Y_{i,end}$, but are not interested in α, γ or δ.[6] Under which conditions can we estimate β consistently by OLS?

Let $Teacher_i$ be abbreviated by T_i in the following. A simple regression of $Y_{i,end}$ on T_i and a constant would yield biased and inconsistent estimates of β due to omitted variable bias.[7] Therefore, we need to include also $Y_{i,begin}$ and X_i in the regression, i.e. to regress $Y_{i,end}$ on T_i, $Y_{i,begin}$ and X_i and a constant. The estimate of β would be unbiased if $E[U|Y_{begin}, X, T]$ is zero a.s. But, if *any* of these control variables is endogenous, generally the estimate of β would be biased.

In this particular example, it is likely that $Y_{i,begin}$ is endogenous in that the unobservable U_i may contain or reflect innate ability. (The other X may also be endogenous, but we focus here on $Y_{i,begin}$ for simplicity.) Innate ability probably affects exam results at the beginning of the school year $Y_{i,begin}$ and also at the end of the school year $Y_{i,end}$. With $cov(U, Y_{begin}) \neq 0$, the estimate of β will be biased and inconsistent.[8] The usual approach to counter this situation

[4] Or at the end of last year.

[5] See e.g. Angrist and Lavy (2002), Dearden, Ferri, and Meghir (2002), Glewwe, Kremer, Moulin, and Zitzewitz (2004), Hoxby (1996), Lavy (2002).

[6] To focus on the main issues, we consider the simplest linear model where there is no treatment effect heterogeneity, i.e. the average treatment effect is the same as the average treatment effect on the treated.

[7] Unless T_i is uncorrelated with $\gamma Y_{i,begin} + \delta X_i + U_i$.

[8] Placing $Y_{i,begin}$ on the left hand side, i.e. regressing $Y_{i,end} - \gamma Y_{i,begin}$ on $Teacher_i$, X_i and a constant also does not yield consistent estimates either, unless γ is known (e.g. to be one).

is to search for a valid instrument for $Y_{i,begin}$, which is often difficult to find.

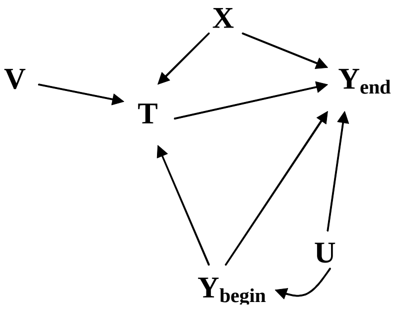
Nonparametric regression, on the other hand, will still produce consistent estimates of the impact of teacher characteristics on achievement if innate ability does not directly affect teacher characteristics. More precisely, $cov(U, Y_{begin}) \neq 0$ is permitted as long as

$$E[U|Y_{begin}, X, T] = E[U|Y_{begin}, X]. \tag{1}$$

It suffices if U has an equal mean for all values of T conditional on the control variables, but this mean can vary with the control variables.

This condition is satisfied, for example, if teachers are allocated within a school on the basis of past exam results but not on the basis of innate ability. To be specific, consider a school with two types of teachers: Teachers with only a standard pedagogical degree and teachers with additional training in advanced Math and Science teaching. Suppose that the best performing children on last year's exam are assigned to the teachers with the additional training, whereas all other students are assigned to the other teachers. Clearly, innate ability is correlated with last year's exam results and thus with teacher characteristics, but *conditional* on exam results, innate ability and teacher characteristics are uncorrelated. OLS would be inconsistent, but nonparametric regression is consistent.

This situation can more intuitively be explained by the following directed acyclic graph (DAG) that visualizes the encoded causal assumptions (Pearl 2000):



We are interested in the causal effect of T on Y_{end}, where we need to control for confounding variables. Y_{end} is assumed to be a (linear) function of teacher characteristics T, student characteristics X, achievement at the beginning of the school Y_{begin} and some unobserved characteristics U. The characteristics of student i's teacher T_i are a function of X, Y_{begin} and some unobservables V. The crucial assumption here is that U affects Y_{begin} but does not affect T directly.[9] In this situation, nonparametric regression is consistent but OLS is not. If U were not affecting Y_{begin}, OLS would also be consistent. On the other hand, if U also affects T directly, neither OLS nor nonparametric regression would be consistent.

In the above situation, the impact of teacher characteristics on student learning could be estimated nonparametrically by a matching estimator, which also permits nonlinear education production functions. Consider T_i binary, i.e. $T_i \in \{0,1\}$. Under the conditional independence assumption (1), the treatment effect β can be estimated as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_1\left(X_i, Y_{i,begin}\right) - \hat{m}_0\left(X_i, Y_{i,begin}\right)\ \right\}$$

where $\hat{m}_t(x,y)$ is a nonparametric regression estimator of $m_t(x,y) = E[Y_{end}|T = t, X = x, Y_{begin} = y]$. This estimator of β is consistent provided that $\hat{m}_t$ is consistent (i.e. bandwidth values converging to zero with increasing sample size) and that the supports of X and Y_{begin} are identical in the $T = 0$ and $T = 1$ population.[10] For further details on matching estimators see e.g. Heckman, Ichimura, and Todd (1998), Heckman, Ichimura, Smith, and Todd (1998), Lechner (2002), Frölich (2004), Imbens (2004), Smith and Todd (2005).

A similar example can also be found in the evaluation of active labour market programmes where one is interested in the effect of participating in a programme, e.g. receiving job search training ($T_i = 1$) or not ($T_i = 0$), on subsequent individual labour market outcomes Y. Past values of Y are important control variables as they are often strong predictors of programme participation (Heckman and Smith 1999), but they are also likely to be endogenous in that unobserved character traits have influenced past earnings and employment status.

A similar situation also arises with parametric and nonparametric *instrumental variables* estimation, where again nonparametric estimation permits endogenous control variables. This has the important practical implication that often one instrumental variable suffices, whereas for 2SLS estimation additional instrumental variables are required to instrument for the endogenous control variables.

[9] In other words, U and V are independent. Another assumption is also that T does not affect Y_{begin}. If there was a causal link from T to Y_{begin}, we would be measuring only a partial effect of T on Y_{end}, i.e. only that part of the effect that is not channeled via Y_{begin}.

[10] With treatment effect homogeneity this latter assumption is not necessary since an overlapping subset would suffice for identification.

Consider as an example the distance to college as an instrument for college attendance as in Card (1995). We are interested in the effect of attending or not attending college, i.e. $T_i \in \{0,1\}$, on earnings Y_i and intend to use distance to college Z_i as an instrument. Let the true earnings relationship be

$$Y_i = \alpha + \beta T_i + \gamma X_i + U_i,$$

where X_i are some individual characteristics, and where U_i and T_i may be correlated. A simple linear IV regression of Y on T with Z as instrument but *without* any additional "control" variables X is unlikely to give consistent estimates since family background characteristics such as parental education, profession and earnings are probably affecting Z as well as U. In other words, the choice of residence made by the parents is unlikely to have been completely random, and those families, who decided to live close to a college, may have different characteristics than those, who decided to live far away, and these characteristics may have a direct impact on their children's wages and returns to college. In addition, cities with a university may also have other facilities that might improve their earnings capacity (e.g. size of the city might matter). The exclusion restriction may thus only be valid after including several background variables X.

Linear IV estimation requires then that $cov(U, Z) = 0$ and that $cov(U, X) = 0$.[11] The latter assumption thus again requires the control variables X to be uncorrelated with U, an assumption which is not needed for nonparametric IV estimation where it suffices that

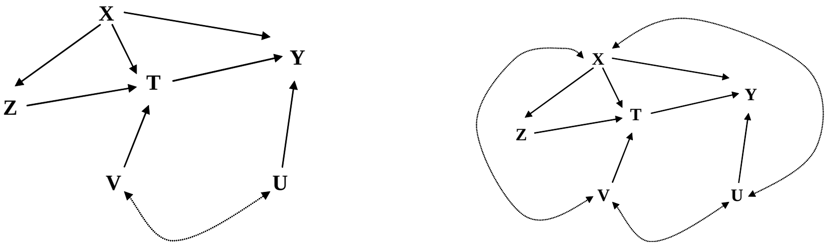
$$E\left[U|Z, X\right] = E\left[U|X\right],$$

which does not need to be zero. Hence, endogenous control variables are permitted with nonparametric IV estimation.

This difference is sketched in the following two graphs: Z, T and Y are functions of some family background characteristics X and some unobservables U and V. (U and V are correlated, thereby generating the endogeneity of T.) In the left graph, X is exogenous in that it is not correlated with the unobservables U and V, whereas it is endogenous in the right graph since X and U are related. As an example, let X be parental education, profession, job characteristics etc. Let U be innate ability of the child, and consider an additional variable: The innate ability of the parents. Innate ability of the parents will usually have affected X and will also be related

[11] Additionally, a rank restriction is needed, but the focus here is on the exclusion restriction.

(genetically) to the ability of their children, thus generating a correlation between X and U. This will bias 2SLS estimation. But, it still permits consistent estimation by nonparametric IV estimation as long as the innate ability of the parents has not directly influenced residence choice Z, but only indirectly via their observed characteristics X such as education, profession, job characteristics etc. For binary T, such a nonparametric $\sqrt{n}$-consistent IV estimator is developed in Frölich (2006). For continuous T, the estimator in Imbens and Newey (2003) can be adapted to endogenous control variables. Alternative nonparametric IV estimators are examined in Chesher (2003, 2005) and Chernozhukov and Hansen (2005).[12]



This difference between parametric and nonparametric approaches is of substantial practical relevance since plausibility of the exclusion restriction often requires conditioning on additional control variables. With 2SLS estimation, if any of these control variables is endogenous, we must find *additional* instruments for the control variables. This is not the case with nonparametric IV estimation.

3 Bias due to endogenous control variables

In this section, a few examples are given to show that bias due to endogenous control variables can be substantial, even in the situation where the treatment variable T is *binary*. (This is particularly interesting since a large number of alternative nonparametric matching estimators for binary T have been developed in the recent years and are readily available). Consider a data generating process:

$$Y_i = \alpha + \beta T_i + \gamma X_i + U_i,$$

where (α, β, γ) are 0.1 and $E[U] = 0$.

The OLS estimator of β is the second element of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ and the corresponding bias

[12] If treatment effects are heterogeneous, only average effects for certain subpopulations may be identified.

is thus

$$e_2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'U,$$

where e_2 is a row vector of zeros with second element one and $\mathbf{X}$ is the $N \times 3$ data matrix consisting of a constant, T_i and X_i, stacked for all N observations. The asymptotic bias of $\hat{\beta}$ can be shown to be

$$\underset{N\to\infty}{plim} \hat{\beta} - \beta = e_2 \begin{bmatrix} 1 & E\left[T\right] & E\left[X\right] \\ E\left[T\right] & E\left[T^2\right] & E\left[TX\right] \\ E\left[X\right] & E\left[TX\right] & E\left[X^2\right] \end{bmatrix}^{-1} \begin{bmatrix} E\left[U\right] \\ E\left[TU\right] \\ E\left[XU\right] \end{bmatrix} = \frac{\sigma_x^2\sigma_{ut} - \sigma_{xt}\sigma_{ux}}{\sigma_x^2\sigma_t^2 - \sigma_{xt}^2},$$

where $\sigma_x^2 = Var(X)$ and $\sigma_{ux} = Cov(U,X)$ and σ_t^2, σ_{ut} and σ_{xt} defined analogously.

If the variable of interest T_i is a linear function of X_i:

$$T_i = \delta X_i + V_i$$

where V_i is an error term independent of all other covariates, it follows that $\sigma_{xt} = \delta\sigma_x^2$ and that $\sigma_{ut} = \delta\sigma_{ux}$ such that

$$\underset{N\to\infty}{plim} \hat{\beta} - \beta = 0.$$

Hence, for T being linear in X the OLS estimate $\hat{\beta}$ is consistent.

However, for T *not being linear* in X, the estimate $\hat{\beta}$ will generally be inconsistent. Consider four examples. First, let V and W be two independent *discrete* random variables that take the values $\{0, 1, 2, 3, 4\}$ with equal probability. The variable T is generated as

$$T = X^2 + V \tag{2}$$

and the variables U and X as

$$\begin{aligned} U &= W^2 \\ X &= W. \end{aligned}$$

Clearly, the error term U is correlated with X as well as with T. However, U and T are not correlated after conditioning on X. For this example, the relative asymptotic bias is obtained after some tedious calculations as:

$$\underset{N\to\infty}{plim} \frac{\hat{\beta} - \beta}{\beta} = \frac{2 \cdot \frac{174}{5} - 8 \cdot 8}{2 \cdot \left(2 + \frac{174}{5}\right) - 8^2} / \beta = 583.\bar{3}\ \%.$$

Hence, the asymptotic bias is almost six times larger than the true value.

Now let V and W be independent standard *normal* variables. In this situation, X and U are dependent but they are no longer correlated. Nevertheless, $\hat{\beta}$ is still asymptotically biased since the relationship between X and T introduces a correlation between T and U, which leads to the relative asymptotic bias:

$$\underset{N\to\infty}{plim}\frac{\hat{\beta}-\beta}{\beta} = \frac{1\cdot 2-0}{1\cdot(1+2)-0}/\beta = 666.\bar{6}\ \%.$$

In the third example, let T be a *binary* variable

$$T = 1\left(X+V>2\right). \tag{3}$$

The relative asymptotic bias is calculated to be

$$\underset{N\to\infty}{plim}\frac{\hat{\beta}-\beta}{\beta} = -609.756\ \%$$

when V and W are *discrete* random variables drawn from $\{0,1,2,3,4\}$ with equal probability.

In the fourth example, for V and W being independent standard *normal* variables the asymptotic bias is:[13]

$$\underset{N\to\infty}{plim}\frac{\hat{\beta}-\beta}{\beta} = \frac{\frac{1}{\sqrt{2}}\phi\left(\sqrt{2}\right)}{\Phi\left(\sqrt{2}\right)\Phi(-\sqrt{2})-\frac{1}{2}\phi^{2}\left(\sqrt{2}\right)} = 1682.117.$$

These examples show that the asymptotic bias of OLS can be very large, whereas nonparametric methods would yield consistent estimates.

4 Conclusions

In this note it has been argued that nonparametric regression has not only the advantage of relaxing functional form assumptions but that it also permits endogenous control variables. Endogenous control variables may very often be of concern in applied empirical work, in a regression context as well as in instrumental variable estimation. A few examples have been discussed and it has been shown that the bias due endogenous control variables can be sizeable in linear regression. Although the bias can be smaller in other examples, it is worth emphasizing that, at least for binary treatment variables, alternative nonparametric "matching" estimators are readily available, which permit endogenous control variables.

[13] All calculations are available from the author.

References

Angrist, J., and V. Lavy (2002): "New Evidence on Classroom Computers and Pupil Learning," *Economic Journal*, 112, 735–765.

Black, D., and J. Smith (2004): "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching," *Journal of Econometrics*, 121, 99–124.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405–1441.

——— (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525–1550.

Dearden, L., J. Ferri, and C. Meghir (2002): "The Effect of School Quality on Educational Attainment and Wages," *The Review of Economics and Statistics*, 84, 1–20.

Frölich, M. (2004): "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *The Review of Economics and Statistics*, 86, 77–90.

——— (2006): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *forthcoming in Journal of Econometrics*.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2004): "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics*, 74, 251–268.

Hanushek, E. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141–1177.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J., and J. Smith (1999): "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies," *Economic Journal*, 109, 313–348.

Hoxby, C. (1996): "How Teachers' Unions Affect Education Production," *Quarterly Journal of Economics*, 111, 671–718.

Imbens, G. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G., and W. Newey (2003): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," presented at the EC2 conference London December 2003.

Lavy, V. (2002): "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy*, 110, 1286–1317.

Lechner, M. (2002): "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society, Series A*, 165, 59–82.

——— (2006): "A note on endogenous control variables in evaluation studies," *University of St. Gallen Economics Discussion Paper Series*, 2006.

Pearl, J. (2000): *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.

Smith, J., and P. Todd (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 125, 305–353.