N° 17-855

October 2017

"Endogenous Variables in Binary Choice Models: Some Insights for Practitioners"

Christophe Bontemps and Céline Nauges

Endogenous Variables in Binary Choice Models: Some Insights for Practitioners

Christophe Bontemps

Toulouse School of Economics, INRA, University of Toulouse Capitole, Toulouse, France

Céline Nauges

Toulouse School of Economics, INRA, University of Toulouse Capitole, Toulouse, France

Abstract

The main purpose of this article is to offer practical insights to econometricians wanting to estimate binary choice models featuring a continuous endogenous regressor. We use simulated data to investigate the performance of Lewbel's *special regressor* method, an estimator that is relatively easy to implement and that relies on different identification conditions than more common control function and Maximum Likelihood estimators. Our findings confirm that the large support condition is crucial for the special regressor method to perform well and that one should be very cautious when implementing heteroscedasticity corrections and trimming since these could severely bias the final estimates.

Introduction

The main purpose of this article is to guide practitioners in their choice of a methodology for the estimation of a binary choice model with potentially endogenous explanatory variables. We are primarily interested in assessing the performance and robustness of an estimator that has been proposed recently (Lewbel et al., 2012; Dong and Lewbel, 2015). This estimator, which was originally described in Lewbel (2000), has been of greater interest lately after Dong and Lewbel (2015) provided a quite simple (step-by-step) estimation procedure. The availability of an estimation package in one of the most popular statistical software further increased its popularity.[1] Lewbel's estimator relies on a continuously distributed, strictly exogenous "special regressor" that has to satisfy a large support condition. When such a special regressor is available and satisfies all required assumptions, Lewbel's estimator may be a relevant alternative to more common control function and Maximum Likelihood (ML) estimators.[2]

Control functions and ML estimators have become natural candidates when estimating binary choice models with endogenous regressors since the early work of Rivers and Vuong (1988) and Blundell and Smith (1989). When the endogenous variable is continuously distributed, control functions can be used and are relatively easy to implement, while the ML method is traditionally used with a discrete endogenous variable (bivariate probit model).[3] Both the control function and ML approaches rely on assumptions that are very different from those required in Lewbel (2000) and Dong and Lewbel (2015). Control functions and ML approaches rely on the specification of a model for the endogenous regressor, written as a function of the set of exogenous regressors and instruments, and a random error term. These approaches also usually require the parametrization of the joint distribution of the errors in the structural model and the (reduced-form) model for the endogenous regressor.[4]

---

[1] An estimation package (*sspecialreg*) was made available in Stata (Baum, 2012).

[2] There also exists an extensive literature on estimators that rely on weaker assumptions and provide bounds on parameters rather than point identification (e.g., Magnac and Maurin, 2008; Chesher, 2010). These articles are not discussed here.

[3] For the estimation of a binary choice model featuring a discrete endogenous variable, see for example Lin and Wooldridge (2015) which proposes an estimator combining the quasi-limited information ML and control function approach.

[4] Some estimators that have been proposed recently are free from any distributional assumptions but are computationally quite complex; e.g., the two-step semi-parametric ML estimator proposed by Rothe (2009). In the first step, endogenous variables are regressed on a set of instruments. This auxiliary regression model can be either left unspecified (fully non-parametric model) or incorporate some parametric restrictions. Residuals

The different sets of identification conditions that are required for the two estimators (control function/ML and Lewbel's) to perform well provide a relevant robustness test for practitioners. If coefficient estimates obtained from the two methods are similar, then the analyst will have confidence in those estimates. On the contrary, if the estimates are very different, then it is an indication that some assumptions are violated.

Our purpose is to study various practical and computational issues that empirical researchers may face when implementing Dong and Lewbel's semi-parametric procedure. Using simulations we provide practitioners with additional insights into the pros and cons of this approach with respect to the (more common) control function/ML approaches and we offer recommendations to select variables that could qualify as special regressors.

We consider the case of a binary choice model featuring an endogenous continuous regressor and use simulated data to compare the performance of the control function/ML approach with the special regressor method. The simulation framework is the same as that used in Lewbel (2000). More precisely, we compare Dong and Lewbel's semi-parametric procedure with the ML estimation of a structural probit model estimated together with a reduced-form equation featuring the relationship between the endogenous variable and the instruments. These two estimators are comparable in terms of simplicity and ease of implementation. In the case of Dong and Lewbel's procedure, we discuss three specific issues that are commonly encountered in empirical work: the limited spread of the special regressor observed values that may cause failure of the large support condition; the presence of heteroscedasticity; and computational issues caused by the construction of a new (dependent) variable that involves dividing by a density.

Although the control function/ML approach has been used widely in empirical research featuring binary decision models, the special regressor model presented in Lewbel (2000) has gained in popularity in recent years.[5] Lewbel's procedure has been applied in empirical studies assessing individuals' willingness to pay or willingness to accept some level of risk in hypothetical situations where the bid or compensation amount was randomized (e.g.,

from the first-stage regression are then added non-parametrically to the structural model which, in turn, is estimated by semi-parametric ML.

[5] Earlier applications include Maurin (2002) and Goux and Maurin (2005). Both studied educational performance through a binary decision model describing the probability of being held back in school using data from France, with potentially endogenous variables in both cases.

Riddel, 2011; Kalisa et al., 2016). With randomization, the bid or compensation amount can be chosen by the econometrician such that it qualifies as a special regressor (distributed over a large support and uncorrelated with the endogenous regressor). Riddel (2011) applied Lewbel's approach to elicit willingness to accept the risk related to nuclear-waste transport on a specific US route, using the subjects' perceived mortality risk as one conditional, but potentially endogenous, factor. Survey respondents were asked if they would (or would not) accept the compensation offered to those living near the route and remaining at their current location. To account for the possible endogeneity of subjective risk assessment, Lewbel (2000)'s model was estimated using the compensation amount as the special regressor. Kalisa et al. (2016) used Lewbel's approach to estimate individuals' willingness to pay to reduce perceived mortality risk due to excess arsenic in their drinking water, using data from the US. The perception of mortality risk was elicited through a risk ladder and treated as endogenous. The authors studied individuals' decision to agree (or not agree) to pay an extra amount in their water bill to ensure that arsenic concentration levels complied with the federal standard; the bid offered to the subject was used as the special regressor.[6]

Our article is organized as follows: in the next section we review the theory and assumptions underpinning Lewbel's special regressor approach. In Section 2 we present the step-by-step methodology for estimating a special regressor model and we discuss three possible issues that practitioners may face: the failure of the large support condition; heteroscedasticity; and computational issues. In Section 3 the design of our simulation setting is explained and the simulation results are given. Section 4 offers practical recommendations for the choice of a suitable special regressor. Section 5 is the conclusion.

1. The special regressor approach: theory and assumptions

The description of the model and assumptions is mainly based on Dong and Lewbel (2015). The model of interest is:

$$y = \mathrm{I}\left(x'\beta + v + \varepsilon \geq 0\right) \tag{1}$$

[6] Other recent articles have estimated models using a special regressor chosen from the list of observable and exogenous covariates, which makes it more difficult to justify that all conditions about the special regressor are fulfilled (e.g.; Bontemps and Nauges, 2016; Zapata Diomedi and Nauges, 2016).

where y is the binary decision variable, x is a structural (continuous) explanatory variable that is potentially endogenous, v (identified as the *special regressor* in Lewbel's terminology) is an exogenous (and continuous) regressor with coefficient normalized to one without any loss of generality. The parameter of interest is β; ε has a zero mean distribution, and $\mathrm{I}(.)$ is the indicator function taking the value one if the latent variable $x'\beta+v+\varepsilon$ is positive and zero otherwise. We also define a set of instruments $\mathbf{z}$ which are assumed to be uncorrelated with the random shock: $E(\mathbf{z}'\varepsilon)=0$. Let $\mathbf{S}=(x,\mathbf{z})$ denote the vector of all regressors but v. The special regressor v has to satisfy the following assumptions:

(H1) $E(v)=0$; $v=g(u,\mathbf{S})$ with u continuously distributed and $u\sim f(\cdot)$ a mean zero density function, $E(u)=0$, $u\perp(\mathbf{S},\varepsilon)$, and $g(u,\mathbf{S})$ differentiable and strictly monotonically increasing in its first element.[7]

(H2) The support of $-x'\beta-\varepsilon$ is a subset of the support of v defined as $]v_L,v_H[$, with $v_L<0<v_H$.

Under (H1), (H2), and the assumption of uncorrelated instruments, the following moment condition holds:

$$E(\mathbf{z}'x).\beta=E(\mathbf{z}'\tilde{T}) \qquad (2)$$

where:

$\tilde{T}=\dfrac{y-\mathrm{I}(v>0)}{f[u]}$; Lewbel et al. (2012) and Dong and Lewbel (2015).

Assumption (H1) requires the special regressor v to be conditionally independent of the model error ε, conditioning on covariates $\mathbf{S}=(x,\mathbf{z})$.[8] It is the main difference with the

[7] Cf. Theorem 1 in the Appendix of Dong and Lewbel (2015).

(original) special regressor estimator developed in Lewbel (2000). In the latter the main assumption used for identification was that the conditional distribution of ε given $(x,\mathbf{z})$ is independent of the special regressor v: $F_\varepsilon(.|v,x,\mathbf{z}) = F_\varepsilon(.|x,\mathbf{z})$. This condition is referred as the *partial independence* assumption in Magnac and Maurin (2007).

The large support condition (H2) is best described by quoting Dong and Lewbel (2015, p. 102): "*the condition regarding the support of* v *is that the range of possible values of* $x'\beta+\varepsilon$ *lies in the range of possible values of* $-v$*, which implies that it is possible for* v *to be small enough or large enough to drive* y *to zero or one*."[9]

Magnac and Maurin (2007) have shown that the large support condition (H2) could be relaxed and replaced with an assumption about ε tail symmetry. More precisely, even if the large support condition (H2) is not satisfied, the moment condition (2) providing exact identification of β still holds if:

$$E\left(\mathbf{z}'y^*_{v_H}\mathrm{I}\left\{y^*_{v_H}>0\right\}\right) = E\left(\mathbf{z}'y^*_{v_L}\mathrm{I}\left\{y^*_{v_L}>0\right\}\right) \qquad (3)$$

where $y^*_{v_L} = \left(x\beta + v_L + \varepsilon\right)$ is the propensity for success for individuals with the smallest v and $y^*_{v_H} = -\left(x\beta + v_H + \varepsilon\right)$ is the propensity for failure for individuals with the largest v (cf. Proposition 5 in Magnac and Maurin, 2007). Note that individuals for whom $y^*_{v_L} = x\beta + v_L + \varepsilon > 0$ respond $y=1$ even when $-v$ is maximum (i.e. equal to $-v_L$) so, for any $v \geq v_L$, $x\beta + v + \varepsilon > 0$ and the probability of success $\mathrm{Prob}\left[y=1\right]$ is equal to one. Symmetrically, individuals for whom $y^*_{v_H} = -\left(x\beta + v_H + \varepsilon\right) > 0$ respond $y=0$ even when $-v$ is minimum (i.e., equal to $-v_H$) so, for any $v \leq v_H$, one gets $x\beta + v + \varepsilon < 0$ and the

[8] For condition (H1) to be satisfied, the special regressor v can only enter linearly in the structural model (1) and not through higher-order or interaction terms.

[9] This assumption is set without specifying any distribution for ε.

probability of success is equal to zero.[10] These cases in which the binary decision is certain are not informative for the identification of the parameter of interest. They are called subsets of "certain success" and "certain failure" in Magnac and Maurin (2003). The large support condition (H2) guarantees that these two subsets of uninformative conditions are empty.

When the endogenous variable is continuously distributed, control functions are a natural candidate for estimating model (1). The control function/ML approach requires the specification of a (reduced-form) model for the endogenous regressor, written as a function of the set of exogenous covariates and instruments. All exogenous regressors (including the one which could play the role of the special regressor in Lewbel's approach) are included in the list of instruments. The control function/ML approach also requires assumptions about the joint distribution of the errors in the structural model (1) and the (reduced-form) model for the endogenous regressor. For the ML estimator to be consistent, the error terms in the two models have to be independent and identically distributed multivariate normal. For more details on the control function approach, see Wooldridge (2010).

2. The special regressor estimator (in practice)

2.1. Step-by-step estimation procedure

Recent articles by Lewbel and co-authors have proposed a quite easy, step-by-step, methodology to implement the special regressor estimator in practice. In Dong and Lewbel (2015), v is assumed to be a linear function of $\mathbf{S}=(x,\mathbf{z})$ and an error u: $v=\mathbf{S}'\mathbf{b}+u$. The specification of a parametric model for the special regressor v makes the estimation procedure simpler compared to Lewbel (2000) which relied on the nonparametric estimation of a conditional density function.

Step 0: If not of mean zero, then v must be demeaned.

[10] If $-v$ is the price of a good and y is the purchasing decision, then individuals for whom $y^*_{v_L}=x\beta+\varepsilon>-v_L$ will always buy the good even when the price is at its maximum and individuals for whom $-y^*_{v_H}=x\beta+\varepsilon<-v_H$ will never buy the good even when the price is at its minimum.

Step 1: Compute, for each observation *i*, the residuals of the linear regression of v on $(x, \mathbf{z})$: $\hat{u}_i = v_i - x_i \hat{b}_x - \mathbf{z_i}'\hat{\mathbf{b}}_\mathbf{z}$, where $\hat{b}_x$ and $\hat{\mathbf{b}}_\mathbf{z}$ are ordinary least squares (OLS) estimated coefficients for variables x and $\mathbf{z}$, respectively.

Step 2: Estimate the density f of the residuals $\hat{u}_i$ computed in Step 1. The following non-parametric kernel estimator (with bandwidth *h*) can be used:

$$\hat{f}(u) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{\widehat{u_j} - u}{h} \right). \quad (4)$$

Step 3: For each observation i compute $\hat{T}_i$ as:

$$\hat{T}_i = \frac{y_i - \mathrm{I}(v_i \geq 0)}{\hat{f}_i(\widehat{u_i})}. \quad (5)$$

Step 4: An estimate of the parameter of interest, $\hat{\beta}$, is obtained by running a two-stage least squares (2SLS) regression of $\hat{T}$ on x using instruments $\mathbf{z}$.

Contrary to traditional two-stage instrumental variables methods, the set of instruments in the special regressor approach should not include the special regressor v. This requirement and the restriction that the special regressor can only enter the structural equation linearly and not through higher-order or interaction terms (cf. footnote 9) are seen as important limitations by some authors (Lin and Wooldridge, 2015).

2.2. Issues in the implementation of the special regressor estimator

We discuss three issues which practitioners are likely to encounter and which are going to be the focus of our simulation exercise: the failure of the condition on the large support; heteroscedasticity in *u* (Step 1); and computational issues in the construction of $\hat{T}$ (Step 4).

The large support condition: the large support condition (assumption H2) is a necessary condition for the moment condition (2) to hold and for the parameter of interest, β, to be

exactly identified. If this condition is not satisfied empirically, that is, if $-x'\widehat{\beta}-\hat{\varepsilon}$ takes values that are outside the observed spread of v, then some observed choices will carry no information. This may lead to an inconsistent estimation of β except if the condition about tail symmetry described in (3) holds. The difficulty for practitioners is that neither the large support condition nor the tail symmetry condition is testable prior to estimation. One can only verify, after estimation, whether the values of $-x'\hat{\beta}$ lie between the minimum and maximum observed values for v, but this does not guarantee that the large support condition holds for the true value of the parameters. In what follows we explore the performance of the two estimators (special regressor and probit model with instrumental variables) under different assumptions on the spread or standard deviation of v.[11]

Heteroscedasticity in u: Dong and Lewbel (2015) proposed an extension of the special regressor estimation procedure that allows for heteroscedastic errors *u* in the regression model of v on $\mathbf{S}=(x,\mathbf{z})$ in Step 1.[12] The revised step-by-step methodology is as follows:

Steps 0 and 1: same as before.

Step 2: Residuals $\hat{u}_i = v_i - x_i\hat{b}_x - \mathbf{z_i}'\hat{\mathbf{b}}_\mathbf{z}$ computed in Step 1 are raised to the square and regressed on $(x,\mathbf{z})$ plus all squares and cross-products of x and $\mathbf{z}$. Let $\mathbf{H}$ denote the vector including all regressors and $\hat{\boldsymbol{\mu}}_S$ be the corresponding vector of OLS coefficients. For each observation *i*, we compute the corrected residuals $\hat{\tilde{u}}_i = \left(\mathbf{H_i}'\hat{\boldsymbol{\mu}}_S\right)^{-1/2}\hat{u}_i$.

Step 3: Estimate the density f of the residuals $\hat{\tilde{u}}_i$ as in Step 2 before.

Step 4: For each observation i compute $\hat{T}_i$ as:

$$\hat{T}_i = \frac{\left[y_i - \mathrm{I}(v_i \geq 0)\right]\left[\left(\mathbf{H_i}'\hat{\boldsymbol{\mu}}_S\right)^{1/2}\right]}{\hat{f}_i\left(\hat{\tilde{u}}_i\right)}.$$

[11] Standard deviation is used as a measure of the spread of observations, as in Lewbel (2000).

[12] A White's test for heteroscedasticity on the regression of v on $\mathbf{S}=(x,\mathbf{z})$ should be undertaken in order to check whether heteroscedasticity is present (Dong and Lewbel, 2015).

Step 5: Same as Step 4 before.

Computational issues: Dong and Lewbel (2015) recommend the use of trimming or Winsorization to discard observations for which $\hat{T}$ takes on very large values (either negative or positive) because of very small values of the estimated density $\hat{f}(.)$. Very small values of the estimated density $\hat{f}(.)$ in (5) could result in *T* having infinite variance and slow the rate of convergence of the estimator.[13] Trimming and Winsorization are expected to improve the mean squared error performance of the estimator (Dong and Lewbel, 2015).[14]

In the next section, we explore how the large support condition (through assumptions on the standard deviation of v), heteroscedasticity correction, and trimming and Winsorization affect the performance of the special regressor estimator. Other practical issues, such as the choice of a kernel versus an ordered data estimator (Lewbel and Schennach, 2007) for the estimation of the density and the choice of the bandwidth with the kernel estimator are only briefly discussed.

3. Simulation design and results

We use as a basis the simulation setting of Lewbel (2000) and we consider the case of a continuous structural explanatory variable x. We start by considering a simple design where the covariate x is exogenous (Section 3.1) and then we study the case of an endogenous regressor (Section 3.2). We consider sample sizes of 100, 500 and 1,000, and we run 10,000 Monte-Carlo replications. All standard errors are computed using 399 standard bootstraps with replacement.

The structural binary-choice model is the following:

$$y = \begin{cases} 1 & \text{if } \beta_1 + \beta_2 x + v + \varepsilon \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

[13] For further discussions on density-weighted estimators, see Kahn and Tamer (2010).

[14] Winsorization implies setting tail values equal to some specified percentile of the data while trimming removes observations below and above the specified percentile (Ruppert, 2004). Correction for heteroscedasticity as well as trimming and Winsorization are included in the *sspecialreg* package in Stata (Baum, 2012).

with $\beta_1 = \beta_2 = 1$;

$v = \lambda(1+\gamma x)e_2$, the special regressor;

$x = e_1$, an exogenous regressor;

$z = x$, the instrument;

$\varepsilon = \rho e_1 + e_3$, the error term;

and $e_1 \sim U(0,1)$; $e_2 \sim N(0,1)$; $e_3 \sim N(0,1)$ are three independent random variables.

The parameter λ determines the standard deviation or spread of v, γ is a parameter introducing heteroscedasticity in the model, while ρ controls for the degree of endogeneity of x.

3.1. The case of an exogenous covariate

To start, we consider the simple case of an exogenous covariate x by setting $\rho = 0$ (labelled as "clean design" in Lewbel, 2000). In this setting we have $z = x$ so the 2SLS estimator run in Step 4 (Section 2.1) is equivalent to an OLS estimator. We compare the special regressor estimator[15] to a probit model with instrumental variables (called IV probit from now on) which, under the exogeneity assumption, reduces to a simple probit model. We study the benchmark case (Section 3.1.1) and follow by testing the robustness of the special regressor and probit estimator to varying support conditions (Section 3.1.2), the presence of heteroscedasticity (Section 3.1.3), and the application of trimming and Winsorization to remove observations with large values of $\hat{T}$ (Section 3.1.4).

3.1.1. Benchmark case

We run simulations in the benchmark case assuming no heteroscedasticity ($\gamma = 0$). Following Lewbel (2000) we set $\lambda = 2$ which should make the (theoretical) large support condition (H2) empirically satisfied in most cases: with $\lambda = 2$ the standard deviation of v is

[15] In the non-parametric estimation of the density (Step 2), we follow Dong and Lewbel (2015) and use an Epanechnikov kernel and a bandwidth equal to the Silverman's rule-of-thumb as defaults.

equal to 2 and hence is larger than the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$ (equal to $\sqrt{2}$ =1.42 when $\beta_1 = \beta_2 = 1$).

We report the mean, standard deviation, minimum and maximum of $\widehat{\beta_2}$ (its true value is 1), the proportion of replications in which $\widehat{\beta_2}$ falls within the 95% confidence interval, and the Root Mean Squared Error (RMSE); see Table 1.[16]

Table 1. Special Regressor (SR) and IV probit estimates; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Min of $\widehat{\beta_2}$	Max of $\widehat{\beta_2}$	Within 95% CI	RMSE
N=100						
SR	1.015	0.280	-0.116	2.715	0.957	0.255
IV Probit	1.121	0.361	0.266	6.698	0.987	0.283
N=500						
SR	1.011	0.127	0.574	1.660	0.964	0.261
IV Probit	1.019	0.120	0.650	1.623	0.954	0.290
N=1,000						
SR	1.009	0.088	0.680	1.456	0.968	0.261
IV Probit	1.009	0.083	0.695	1.364	0.949	0.291

Note: $\lambda = 2, \rho = 0, \gamma = 0$, 399 bootstraps.

With a sample size of 1,000, the two estimators produce virtually unbiased estimates but the RMSE is slightly larger for the IV probit estimator. The special regressor estimator is performing well even with a sample size of 100, while the coefficient estimated by the IV probit is slightly biased upwards (12% mean bias). The special regressor and IV probit estimates fall within the 95% confidence interval (around the value one) 95% (or more) of the time whatever the sample size.

[16] In all cases (special regressor and IV probit), the RMSE is computed as follows: $\sqrt{\sum_N (y - \hat{y})^2 / N}$ where $\hat{y}$ is the prediction of y and N is the sample size.

3.1.2. Large support condition: sensitivity to the spread of v

The condition about the support of v is essential for identification when using the special regressor approach, so we test the sensitivity of the special regressor estimator to the spread of v. In the benchmark case ($\lambda = 2$), the (theoretical) condition (H2) is likely to be empirically satisfied in most cases since the standard deviation of v is larger than the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$. We then consider a case where the spread of v is similar to the spread of $\beta_1 + \beta_2 x + \varepsilon$ by setting $\lambda = \sqrt{2}$, and two cases where the spread of v is smaller than the spread of $\beta_1 + \beta_2 x + \varepsilon$, by choosing $\lambda = 1$ and $\lambda = 0.7$. Based on the large support condition (H2), we expect that the larger the spread of v (hence the larger the parameter λ), the better the special regressor estimator will perform.

In Table 2 we report the mean and standard deviation of the special regressor estimator along with the frequency of non-informative conditions distinguishing between *perfect success* (y = 1 is certain whatever value v takes) and *perfect failure* (y = 0 is certain whatever value v takes); following the terminology of Magnac and Maurin (2003). A greater number of non-informative conditions should be observed for lower values of the parameter λ. The probit estimates are not shown here because they do not depend on condition (H2) and thus remain virtually unbiased whatever the standard deviation of v.

The number of non-informative conditions in columns 4 and 5 of Table 2 is computed based on the true values of the parameters β_1 and β_2, and the true ε. In practice, however, if the practitioner is willing to check on these conditions, they will have to be calculated using the estimated parameters $\left(\widehat{\beta_1}, \widehat{\beta_2}\right)$ because neither the true β_1 and β_2, nor the error ε, is observed.

Table 2. Sensitivity of the special regressor (SR) estimates to the spread of v; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Non-informative conditions (success) (β_1, β_2)	Non-informative conditions (failure) (β_1, β_2)
$\lambda = 2$: *Standard deviation of v* ***larger*** *than standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$					
N=100	1.015	0.280	0.957	0.005	0.000
N=500	1.011	0.127	0.964	0.000	0.000
N=1,000	1.009	0.088	0.968	0.000	0.000
$\lambda = \sqrt{2}$: *Standard deviation of v* ***comparable to*** *standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$					
N=100	0.975	0.278	0.948	0.046	0.001
N=500	0.987	0.146	0.947	0.011	0.000
N=1,000	0.990	0.104	0.950	0.006	0.000
$\lambda = 1$: *Standard deviation of v* ***smaller*** *than standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$					
N=100	0.879	0.278	0.860	0.160	0.006
N=500	0.928	0.184	0.812	0.083	0.002
N=1,000	0.942	0.155	0.793	0.062	0.001
$\lambda = 0.7$: *Standard deviation of v* ***much smaller*** *than standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$					
N=100	0.723	0.244	0.594	0.309	0.026
N=500	0.796	0.185	0.507	0.226	0.012
N=1,000	0.821	0.165	0.490	0.197	0.009

Note: λ varying, $\rho = 0, \gamma = 0$, 399 bootstraps.

In the benchmark case, i.e. when $\lambda = 2$ and the standard deviation of *v* is larger than the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$, the special regressor produces unbiased estimates of β_2 on average and there are almost no non-informative conditions. When the spread of v gets smaller, the number of non-informative conditions increases as does the mean bias. In the extreme case where the spread of v is about half the spread of $\beta_1 + \beta_2 x + \varepsilon$ (λ=0.7), the proportion of non-informative conditions varies from 20 to 30% depending on the sample size, and the mean bias of the special regressor varies from 18 to 28%.[17] These findings

[17] The non-informative conditions are almost all on "success", i.e. observations such that $y = 1$ whatever value v takes. The imbalance towards non-informative conditions of success (and the higher proportion of observations for which $y = 1$ in the sample) is due to the distribution of $\beta_1 + \beta_2 x + \varepsilon$ being centred on 1 while the distribution of v is centred on 0. We ran simulations under the assumption that the constant β_1 is equal to 0, which implies that the distributions of $\beta_1 + \beta_2 x + \varepsilon$ and v are both centred on 0 (and that the proportion of observations for which y=0 and y=1 is balanced in the sample). In Appendix A1 we report simple statistics on

confirm the importance of the large support condition when using the special regressor estimator. This theoretical condition is not directly testable but, in practice, one can verify that the standard deviation of the observed v is larger than the standard deviation of $\widehat{\beta}_1 + \widehat{\beta}_2 x$.

3.1.3. Sensitivity to the presence of heteroscedasticity

Under condition (H1) errors *u* in the regression model $v = xb_x - \mathbf{z}'\mathbf{b}_\mathbf{z} + u$ are independent of $\mathbf{S} = (x, \mathbf{z})$. If the homoscedasticity condition is not satisfied, the special regressor estimator will be biased. We control for heteroscedasticity with the parameter γ in $v = \lambda(1 + \gamma x)e_2$ and we run simulations for the following values: γ = 0 (homoscedasticity), γ = 0.5, and γ = 1, for two sample sizes (N=100 and N=500).[18] We set $\lambda = 2$ so that the standard deviation of v is larger than the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$ and the theoretical large support condition is likely to be satisfied in most cases.[19]

In Table 3 we report the mean and standard deviation of the special regressor estimate of β_2, along with the frequency of estimates falling within the 95% confidence interval, with and without the correction for heteroscedasticity.[20] We also show the IV probit estimates. Since the IV probit does not rely on the intermediate regression of v on $\mathbf{S} = (x, \mathbf{z})$, it should be robust to heteroscedasticity in errors *u*.

---

the special regressor estimator of β_2 along with the frequency of non-informative conditions for the four values of λ ($\lambda = 2$, $\sqrt{2}$, 1, and 0.7). Our results, compared to those reported in Table 2, show a smaller number of non-informative conditions for each level of λ and, consequently, a smaller mean bias (7% smaller when λ is equal to 1 and 12% smaller when λ is equal to 0.7) when the distributions of $\beta_1 + \beta_2 x + \varepsilon$ and v are centred on 0.

[18] With a sample size of N=100 and 10,000 replications, the White's test leads to the rejection of the null hypothesis of homoscedasticity (at the 5% level) in 75% of the cases when γ = 0.5 and in 83% of the cases when γ = 1.

[19] Note that the standard deviation of v now increases with γ.

[20] When correcting for heteroscedasticity, we regress residuals $\hat{u}_i = v_i - x_i\hat{b}_x$ raised to the square on x only ($x = \mathbf{z}$ in the clean design setting with no endogenous regressor) but the square of x could also be included as an additional regressor.

Table 3. Special regressor (SR) and IV probit estimates in the presence of heteroscedasticity; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI
	SR ***with*** heteroscedasticity correction			SR ***without*** heteroscedasticity correction			IV probit		
N=100									
$\gamma = 0$	1.015	0.286	0.956	1.015	0.280	0.957	1.121	0.361	0.987
$\gamma = 0.5$	0.979	0.314	0.947	0.937	0.355	0.912	1.093	0.319	0.982
$\gamma = 1$	0.807	0.307	0.842	0.840	0.392	0.858	1.115	0.366	0.982
N=500									
$\gamma = 0$	1.011	0.128	0.963	1.011	0.127	0.964	1.018	0.120	0.955
$\gamma = 0.5$	0.986	0.182	0.951	0.936	0.155	0.918	1.016	0.115	0.954
$\gamma = 1$	0.769	0.131	0.561	0.842	0.170	0.801	1.018	0.129	0.955

Note: $\lambda = 2, \rho = 0, \gamma$ varying, 399 bootstraps.

The presence of heteroscedasticity in *u* induces a mean bias in the special regressor estimator but this bias is not eliminated when the heteroscedasticity correction is implemented. When $\gamma = 1$, the mean bias is lower when the heteroscedasticity correction is not applied. For example when N=500 and $\gamma = 1$, the special regressor estimate of β_2 is 0.769 on average (23% mean bias) when the correction for heteroscedasticity is applied and the probability that the estimated coefficient falls within the 95% confidence interval around 1 is 0.56. When heteroscedasticity is not corrected for, the special regressor estimate is 0.842 on average (16% mean bias) and falls within the 95% confidence interval in 80% of the cases. As expected, the IV probit is robust to heteroscedasticity in *u*. When $\gamma = 1$, the IV probit outperforms the special regressor, even when the sample size is small (100). When $\gamma = 0.5$, the two estimators perform about the same.

3.1.4. Computational issues: the use of trimming and Winsorization

Dong and Lewbel (2015) recommend trimming/Winsorization in order to remove large values of $\left|\hat{T}\right|$, which correspond to observations with a small value for the estimated density $\hat{f}$ (see Step 3 and equation (5) where the estimated density appears at the denominator).

Small values indicate that the density could not be precisely estimated because of a small number of observations so we consider, as an alternative to trimming/Winsorizing on $\left|\hat{T}\right|$, the application of trimming/Winsorization in order to remove observations with very low values of the estimated density $\hat{f}$. We set $\lambda = 2$ so that the large support condition (H2) should be empirically satisfied in most cases and we assume no heteroscedasticity (γ = 0). Results are shown in Table 4.[21]

Under similar conditions on the spread of v ($\lambda = 2$) and without any trimming or Winsorization, Lewbel's procedure led to unbiased estimates of the coefficient of interest (Table 2). Results in Table 4 show that trimming observations based on large values of $\left|\hat{T}\right|$ induces a mean bias of 8-12% when the trim level is set at 2.5%, and a mean bias of 34% when the trim level is set at 5%. Increasing the sample size does not reduce the mean bias and lowers the probability of $\widehat{\beta_2}$ falling in the 95% confidence interval. In the extreme case of N=1,000 and the level of trimming is set at 5%, the probability of $\widehat{\beta_2}$ falling in the 95% confidence interval is close to zero. Winsorization induces a bias which is lower than the bias induced by trimming; the mean bias is 5-7% when the Winsorization level is set at 2.5%, and 9-11% when the level is set at 5%.

Our results also indicate that trimming observations with low values of the density $\hat{f}$ instead of trimming observations with large values of $\left|\hat{T}\right|$ does not create any bias (mean estimates are similar to the ones in the benchmark case, cf. Table 2). However, the standard deviation of $\widehat{\beta_2}$ is always larger when trimming based on the density $\hat{f}$ so the risk of getting extreme (very small or very large) $\widehat{\beta_2}$ is higher when trimming/Winsorizing based on $\hat{f}$, more so when the sample size is small.

[21] When Winsorizing at the 5% level, we replace observations with values of $\left|\hat{T}\right|$ above the 95th percentile by the 95th percentile of the distribution of $\left|\hat{T}\right|$.

Table 4. Special regressor estimator (SR) with trimming and Winsorization; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI
	Trimming/Winsor based on $\lvert\hat{T}\rvert$			Trimming/Winsor based on $\hat{f}$		
2.5% Trimming						
N=100	0.829	0.255	0.872	1.031	0.284	0.958
N=500	0.788	0.110	0.518	1.031	0.125	0.964
N=1,000	0.781	0.078	0.208	1.032	0.089	0.959
2.5% Winsorization						
N=100	0.953	0.259	0.942	1.015	0.278	0.956
N=500	0.933	0.114	0.910	1.011	0.126	0.964
N=1,000	0.929	0.079	0.859	1.008	0.087	0.967
5% Trimming						
N=100	0.667	0.250	0.668	1.050	0.280	0.961
N=500	0.662	0.110	0.125	1.053	0.126	0.953
N=1,000	0.658	0.077	0.007	1.051	0.088	0.947
5% Winsorization						
N=100	0.910	0.252	0.921	1.012	0.275	0.956
N=500	0.893	0.111	0.829	1.008	0.124	0.964
N=1,000	0.888	0.077	0.700	1.006	0.086	0.967

Note: $\lambda = 2, \rho = 0, \gamma = 0$, 399 bootstraps.

Our results show that, with an exogenous covariate, the application of trimming/Winsorization could lead to estimation bias. Discarding outliers that display low values of the density $\hat{f}$ (instead of large values of $\lvert\hat{T}\rvert$) may be preferable. A graphical illustration in Appendix A2 provides some intuition why trimming observations with low values of the density $\hat{f}$ could be preferable to trimming observations with large values of $\lvert\hat{T}\rvert$. Intuitively, the observations discarded when trimming on $\lvert\hat{T}\rvert$ are those for which $y_i \neq \mathrm{I}(v_i \geq 0)$, which provide useful information for the identification of β.

3.1.5. Estimation of the density

In all the cases discussed so far, the bandwidth *h* used for the kernel density estimator (Step 2) was chosen following Silverman's rule as suggested by Dong and Lewbel (2015). We tried both smaller and larger bandwidths and found the estimates of β_2 to be quite insensitive to the choice of bandwidth (see also Magnac and Maurin, 2003, for simulation results on the sensitivity of the special regressor estimator to window sizes).[22] We also ran simulations with the density in Step 2 estimated using the ordered data estimator of Lewbel and Schennach (2007).[23] The special regressor using the ordered data estimator of the density does not perform better, in general, than the (kernel-based) special regressor when the support condition (H2) is not satisfied, under the presence of heteroscedasticity, and when trimming/Winsorization is applied.[24]

3.2. The case of an endogenous covariate

This case is called "messy design" in Lewbel (2000). We assume $\rho \neq 0$ so that the regressor x is correlated with ε and is therefore endogenous. It is correlated with the instrument z through a new random variable e_4, defined as the mixture of an *N*(-0.3, 0.91) with probability 0.75 and an *N*(0.9, 0.19) with probability 0.25, so the simulation framework is now:

$v = \lambda(1+\gamma x)e_2 + e_4$, the special regressor;

$x = e_1 + e_4$, an endogenous regressor;

$z = e_4$, the instrument;

$\varepsilon = \rho e_1 + e_3$, the error term.

[22] Magnac and Maurin (2003) ran simulations to test additional identification assumptions when the large support condition fails. They tested the performance of Lewbel's estimator but the estimation methodology was not exactly the same as that used here. Magnac and Maurin (2003) estimated the density of the special regressor conditional upon *S* following Lewbel (2000).

[23] For each i define $\hat{U}_i^+$ as the nearest neighbour within $\left(\hat{U}_j\right)_{j\neq i}$ that is greater] than $\hat{U}_i$. For each i define $\hat{U}_i^-$ as the nearest neighbour within $\left(\hat{U}_j\right)_{j\neq i}$ that is smaller than $\hat{U}_i$ .The density $\hat{f}_i^0$ is defined as: $\hat{f}_i^o = (2/n)\Big/\left(\hat{U}_i^+ - \hat{U}_i^-\right)$.

[24] Estimates are not shown here but are available on request.

Note that, in the IV probit model, the special regressor v is used as an instrument.

In what follows we set the parameter ρ at 1. We do not present any simulation results for varying levels of ρ because the choice of ρ has an impact on the standard deviation of *v*. Hence we would not be able to separate the effect of the degree of endogeneity from the effect of the support condition.

3.2.1. Large support condition: sensitivity to the spread of v

We report in Table 5 the special regressor and IV probit estimates of the β_2 coefficient under various assumptions of the standard deviation of v. The IV probit estimates are virtually unbiased for sample sizes of 500 and 1,000. When the sample size is small (100), the estimate of β_2 can be severely biased upwards, which makes the mean difficult to interpret. For this reason we also report the median estimate of β_2.

As expected and as already observed in the case of an exogenous covariate, the large support condition has to be satisfied empirically for the special regressor estimator to be unbiased. In the most favourable case where the standard deviation of *v* is larger than the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$ ($\lambda = 3$), the mean bias of the special regressor is small; 8% with a sample size of 100 and less than 3% with a sample size of 1,000. When standard deviations are comparable, then a large number of observations is required for the special regressor estimator to be close on average to the true value: the mean bias is 10% with a sample size of 1,000 but 27% with a sample size of 100. Finally, when the standard deviation of *v* is smaller than the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$ ($\lambda = \sqrt{2}$), the special regressor estimator is severely biased downwards; the mean bias is 52% with a sample size of 100 and 30% with a sample size of 1,000.

Table 5. Sensitivity of the special regressor (SR) and IV probit estimates to the spread of v; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Median of $\widehat{\beta_2}$	Within 95% CI	Non-informative conditions (success) (β_1, β_2)	Non-informative conditions (failure) (β_1, β_2)
$\lambda = 2$: *Standard deviation of v* ***comparable to*** *standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$						
N=100						
SR	0.733	0.547	0.664	0.814	0.033	0.003
IV Probit	6.601	169.577	1.126	0.999		
N=500						
SR	0.858	0.358	0.798	0.799	0.006	0.000
IV Probit	1.030	0.165	1.023	0.965		
N=1,000						
SR	0.895	0.303	0.840	0.797	0.003	0.000
IV Probit	1.016	0.113	1.013	0.954		
$\lambda = \sqrt{2}$: *Standard deviation of v* ***smaller*** *than standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$						
N=100						
SR	0.476	0.504	0.406	0.563	0.096	0.015
IV Probit	10.254	498.881	1.118	0.998		
N=500						
SR	0.644	0.389	0.560	0.521	0.039	0.004
IV Probit	1.026	0.169	1.017	0.964		
N=1,000						
SR	0.699	0.346	0.619	0.506	0.026	0.002
IV Probit	1.014	0.115	1.011	0.957		
$\lambda = 3$: *Standard deviation of v* ***larger*** *than standard deviation of* $\beta_1 + \beta_2 x + \varepsilon$						
N=100						
SR	0.923	0.536	0.883	0.937	0.002	0.000
IV Probit	22.369	855.646	1.176	0.999		
N=500						
SR	0.965	0.251	0.951	0.942	0.000	0.000
IV Probit	1.035	0.174	1.023	0.969		
N=1,000						
SR	0.977	0.195	0.962	0.949	0.000	0.000
IV Probit	1.017	0.117	1.013	0.957		

Note: λ varying, $\rho = 1, \gamma = 0$, 399 bootstraps. In all cases the standard deviation of of $\beta_1 + \beta_2 x + \varepsilon$ is 2.449 while the standard deviation of *v* is 2.236 when $\lambda = 2$, 1.732 when $\lambda = \sqrt{2}$, and 3.162 when $\lambda = 3$.

3.2.2. Sensitivity to the presence of heteroscedasticity

We assume $\lambda = 2$ (that is, the standard deviation of *v* is comparable to the standard deviation of $\beta_1 + \beta_2 x + \varepsilon$), a degree of endogeneity $\rho = 1$, and various levels of heteroscedasticity (parameter γ).

Table 6. Special Regressor (SR) and IV probit estimates in the presence of heteroscedasticity; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI
	SR ***with*** heteroscedasticity correction			SR ***without*** heteroscedasticity correction			IV probit		
N=100									
$\gamma = 0$	0.711	0.507	0.813	0.733	0.547	0.814	6.601	169.577	0.999
$\gamma = 0.5$	0.648	0.522	0.782	1.076	0.857	0.886	21.064	1285.7	0.999
$\gamma = 1$	1.084	0.866	0.895	1.194	0.931	0.908	2.053	66.681	0.997
N=500									
$\gamma = 0$	0.850	0.344	0.800	0.858	0.358	0.799	1.030	0.165	0.965
$\gamma = 0.5$	0.733	0.270	0.718	1.332	0.540	0.968	1.023	0.160	0.963
$\gamma = 1$	1.336	0.568	0.913	1.393	0.448	0.918	1.021	0.149	0.958

Note: $\lambda = 2, \rho = 1, \gamma$ varying, 399 bootstraps.

As in the case of an exogenous covariate, applying the heteroscedasticity correction does not always reduce the mean bias. When $\gamma = 0.5$ and the sample size is 100, the mean bias is 8% without any heteroscedasticity correction while it is 35% when the heteroscedasticity correction is applied. When the sample size is 500, there is a 30% bias whether or not the correction for heteroscedasticity is applied (downward bias when the correction is applied and upward bias when it is not). When $\gamma = 1$, the mean bias is smaller when the heteroscedasticity correction is applied, whatever the sample size. As already observed, the mean IV probit estimates show a severe bias due to very large estimates in some runs for a sample size of 100. With a sample size of 500, the IV probit estimates are virtually unbiased.

3.2.3. Computational issues: the use of trimming and Winsorization

We assume $\lambda = 2$, a degree of endogeneity $\rho = 1$, and no heteroscedasticity ($\gamma = 0$). We apply trimming and Winsorization on $\left|\hat{T}\right|$ and $\hat{f}$ at the 2.5% and 5% levels. Results are shown in Table 7.

Table 7. Special Regressor estimator (SR) with trimming and Winsorization; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI
	Trimming/Winsor on $\hat{T}$			Trimming/Winsor on $\hat{f}$		
2.5% Trimming						
N=100	0.297	0.327	0.412	0.647	0.432	0.789
N=500	0.259	0.138	0.001	0.699	0.191	0.665
N=1,000	0.250	0.096	0.000	0.708	0.135	0.497
2.5% Winsorization						
N=100	0.508	0.352	0.661	0.687	0.448	0.812
N=500	0.521	0.152	0.155	0.766	0.210	0.772
N=1,000	0.527	0.108	0.016	0.785	0.151	0.717
5% Trimming						
N=100	0.049	0.288	0.104	0.545	0.381	0.734
N=500	0.045	0.125	0.000	0.599	0.170	0.418
N=1,000	0.048	0.089	0.000	0.611	0.122	0.179
5% Winsorization						
N=100	0.401	0.315	0.499	0.638	0.395	0.803
N=500	0.413	0.137	0.017	0.705	0.180	0.677
N=1,000	0.418	0.097	0.000	0.721	0.128	0.518

Note: $\lambda = 2, \gamma = 0, \rho = 1$, 399 bootstraps.

The mean bias of the special regressor estimator with a sample size of 100 is 27% when outliers are not discarded (Table 5). The mean bias increases when trimming or Winsorization is applied, more so with trimming; the mean bias increases to 60% with a 2.5% trimming, and to 95% with a 5% trimming. Winsorization also leads to an increase in the mean bias but the impact is less damaging; there is a 50% mean bias with a 2.5%

Winsorization and a 60% mean bias with a 5% Winsorization. The mean bias is similar whatever the sample size. As already observed in the case of an exogenous covariate, trimming/Winsorizing observations based on low values of the density $\widehat{f}$ would be preferable. However, trimming and Winsorization never eliminate the bias and do not really provide better estimates than those obtained when trimming or Winsorization is not applied.

4. Practical recommendations for the choice of the special regressor

The variable that is chosen as the special regressor has to be continuous, exogenous, and have a support large enough to encompass the (unobserved) support of $x'\beta+\varepsilon$. Intuitively, this means that one has to find an exogenous variable *v* so that the probability of the outcome of interest varies from 0 to 1 when *v* varies over its support. This assumption may hold for example when using age as a special regressor to study probabilities of having reached the primary school level in any industrialized country. If the support of the age variable is sufficiently large, one should be able to find individuals that are too young to attend primary school (and hence the probability of the outcome is 0) and individuals that are old enough and have already exited primary school (the probability of the outcome would be 1). Studies where the special regressor can be randomized by the analyst are typically good candidates for the application of Lewbel's method (e.g., Riddel, 2011, and Kalisa et al., 2016). In these cases, the support of the special regressor is "under the control" of the econometrician who can select the range of *v* so that the probability will be 0 and 1 for some observed values of *v*.

Simple tests on the suitability of *v* as a special regressor should also be performed. First, one should run a nonparametric regression of the binary outcome *y* on the special regressor *v* to check, at least visually, that the relationship is continuous and monotonic. Second, Dong and Lewbel's estimation procedure requires that the special regressor be excluded from the list of instruments. This assumption is easily testable and we recommend that practitioners who choose a special regressor from the set of observable variables test the significance of the coefficient of the special regressor in a reduced-form equation featuring the endogenous variable on the left hand side and instruments on the right hand side. For the special

regressor to be valid, its coefficient in this auxiliary regression should not be statistically different from zero. Third, squared or interaction terms involving the special regressor are not allowed in the structural model for Dong and Lewbel's estimation procedure to be valid. This assumption is also easily testable (Lin and Wooldridge, 2015).

5. Conclusion

Our simulation results show that the special regressor method as described in Dong and Lewbel (2015), which is free from any distributional assumptions and relatively easy to implement, may outperform classical ML approaches such as IV probit when a special regressor that satisfies all necessary conditions is available. When the sample size is small (around 100 observations), the IV probit model produces less precise estimates than the special regressor estimator and is slightly biased upwards in our simulation framework. In any case the comparison of estimates obtained using the special regressor method and control function/ML estimates provide a useful robustness check of the required identification conditions since both estimators rely on totally different assumptions.

For Lewbel's approach to perform well the choice of the special regressor *v* is crucial and some intuitive and practical recommendations for choosing a suitable *v* were provided in Section 4. Our simulations confirm the sensitivity of the special regressor estimates to violations of the large support condition and suggest a cautious use of the method if its spread is not under the control of the analyst. In our simulation setting, the special regressor estimates remained unbiased as long as the standard deviation of *v* was as large as the standard deviation of $x'\beta+\varepsilon$ when the unique regressor was assumed to be exogenous. However, with an endogenous regressor, the standard deviation of *v* had to be larger than the standard deviation of $x'\beta+\varepsilon$ for the special regressor to be unbiased. Thus, if there is no obvious reason for endogeneity, standard methods such as probit should be preferred since the failure of the large support condition will lead to biased estimates even without endogenous regressors.

We also found that correcting for heteroscedasticity in the construction of the estimator is likely to be detrimental to the performance of the special regressor estimator, whether or

not heteroscedasticity is present and with or without endogenous regressors. Another striking result regards the application of trimming and Winsorization, a procedure proposed by Dong and Lewbel (2015) and implemented in the *sspecialreg* Stata package (Baum, 2012). By construction $\hat{T}$ is likely to have large values, as it is defined as a ratio featuring the estimated density of the residual of an auxiliary regression as the denominator. In practice, we show that trimming or Winsorizing observations that have large values of $\hat{T}$ removes some informative values and introduces a bias in the estimates. Winsorization was found to be less damaging than trimming, and trimming/Winsorizing on the estimated density seemed to be preferable to trimming/Winsorizing on $\hat{T}$. However we were not able to find cases in which trimming or Winsorizing improved the estimates so we recommend using these with great caution.

It is important to keep in mind that our simulation results and recommendations are based on the chosen simulation framework, which was the one used in Lewbel (2000). Follow-up analyses could include testing the sensitivity of our results to the distributional assumptions made on the special regressor and the exogenous/endogenous regressor.

References

Baum, C.F., 2012. Sspecialreg: Stata Module to Estimate Binary Choice Model with Discrete Endogenous Regressor via Special Regressor Method.
Available at http://ideas.repec.org/c/boc/bocode/s457546.html.

Blundell, R.W., and Smith, R.J., 1989. Estimation in a class of simultaneous equation limited dependent variable models, Review of Economic Studies 56(1): 37–57.

Bontemps, C., and Nauges, C., 2016. The impact of perceptions in averting-decision models: An application of the special regressor method to drinking water choices, American Journal of Agricultural Economics 98(1): 297–313.

Chesher, A., 2010. Instrumental variable models for discrete outcomes, Econometrica 78(2): 575–601.

Dong, Y.I., and Lewbel, A., 2015. A Simple Estimator for Binary Choice Models with Endogenous Regressors, Econometric Reviews 34:1-2, 82–105.

Goux, D., and Maurin, E., 2005. The effect of overcrowded housing on children's performance at school, Journal of Public Economics 89: 797– 819.

Kalisa, T., Riddel M., and Shaw, W.D., 2016. Willingness to pay to avoid arsenic-related risks: a special regressor approach, Journal of Environmental Economics and Policy 5(2): 143–162.

Khan, S., and Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation, Econometrica 78(6): 2021–2042.

Lewbel, A., 2000. Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables, Journal of Econometrics 97(1):145–177.

Lewbel, A., Dong, Y., and Yang, T., 2012. Comparing features of convenient estimators for binary choice models with endogenous regressors, Canadian Journal of Economics 45(3): 809–829.

Lewbel, A., and Schennach, S., 2007. A Simple Ordered Data Estimator for Inverse Density Weighted Functions, Journal of Econometrics 136(1):189–211.

Lin, W., and Wooldridge, J.M., 2015. Estimating Binary Response Models with Endogenous Explanatory Variables, Combining Control Functions and Quasi-LIML. Working Paper, Michigan State University.

Magnac, T., and Maurin, E., 2003. Identification and information in monotone binary models. Working Paper, CREST, no. 2003–07, http://www.eco.uc3m.es/temp/MagnacMaurin06-03.pdf

Magnac, T., and Maurin, E., 2007. Identification and information in monotone binary models, Journal of Econometrics 139:76–104.

Magnac, T., and Maurin, E., 2008. Partial identification in monotone binary models: discrete regressors and interval data, Review of Economic Studies 75(3): 835–864.

Maurin, E., 2002. The impact of parental income on early schooling transitions. A re-examination using data over three generations, Journal of Public Economics 85: 301–332.

Riddel, M., 2011. Uncertainty and measurement error in welfare models for risk changes, Journal of Environmental Economics and Management 61: 341–354.

Rivers, D., and Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models, Journal of Econometrics 39(3): 347–366.

Rothe, C., 2009. Semiparametric estimation of binary response models with endogenous regressors, Journal of Econometrics 153(1): 51–64.

Ruppert, D., 2004. Trimming and Winsorization, in Encyclopedia of Statistical Sciences, John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0471667196.ess2768.pub2

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, 2nd ed. Cambridge, MA: MIT Press.

Zapata Diomedi, B., and Nauges, C., 2016. Pesticide handling practices: The case of coffee growers in Papua New Guinea, Australian Journal of Agricultural and Resource Economics 60(1): 112–129.

APPENDIX A1

Table A1. Special Regressor (SR) estimates under varying assumptions on the spread of v ; $\beta_1 = 0$; N=100; 10,000 replications

	Mean of $\widehat{\beta_2}$	Std dev of $\widehat{\beta_2}$	Within 95% CI	Non-informative conditions (success)	Non-informative conditions (failure)
$\lambda = 2$	1.032	0.265	0.953	0.004	0.004
$\lambda = \sqrt{2}$	1.006	0.244	0.950	0.008	0.008
$\lambda = 1$	0.952	0.247	0.917	0.043	0.043
$\lambda = 0.7$	0.838	0.233	0.745	0.117	0.117

Note: λ varying, $\rho = 0, \gamma = 0$, 399 bootstraps.

APPENDIX A2

An illustration is made using a sample of 1,000 observations randomly drawn from the model described in Section 3.1 with $\beta_1 = \beta_2 = 1$, $\rho = 0$ (the regressor x is exogenous), no heteroscedasticity ($\gamma = 0$), and a standard deviation for v large enough for the support condition (H2) to be empirically satisfied ($\lambda = 2$).The left graph shows (in dark grey) the observations that are removed when trimming observations with large values of $\left|\hat{T}\right|$ and the OLS fitted regression line (Step 4, regression of $\hat{T}$ on x) after trimming. Similarly, the right graph shows (in dark grey) the observations that are removed when trimming observations with low values of $\hat{f}$ and the OLS fitted regression line (Step 4, regression of $\hat{T}$ on x) after trimming. The estimated coefficient for β_2 is 0.627 when trimming observations based on $\hat{T}$ and 0.922 when trimming observations based on $\hat{f}$. The downward bias of β_2 when trimming observations based on $\hat{T}$ (left graph) is caused by the removal of large (positive and negative) values of $\hat{T}$. It is important to remember that, by construction, $\hat{T}_i = \dfrac{y_i - \mathrm{I}(v_i \geq 0)}{\hat{f}_i\left(\widehat{u}_i\right)}$ takes the value 0 in all cases where ($y_i = 1$ and $v_i \geq 0$) and ($y_i = 0$ and $v_i < 0$), which may represent a large share of the sample. Since identification of the parameter β_2 relies on information provided by observations for which $\hat{T}_i \neq 0$, removing observations for which $\hat{T}$ is strictly different from zero may induce a truncation bias. The observations that are removed are those for which y_i is different from $\mathrm{I}(v_i \geq 0)$. For example, such a case could be observed for individuals who are willing to purchase a good ($y_i = 1$) even when the price of the good is high: $\mathrm{I}(v_i \geq 0) = 0$. Such observations could be highly informative. Trimming observations based on low values of $\hat{f}$ led to the removal of observations for which $\hat{T}_i = 0$ and produced an estimate of the parameter of interest β_2 closer to its true value (β_2 = 1), but at the cost of an increased variance.

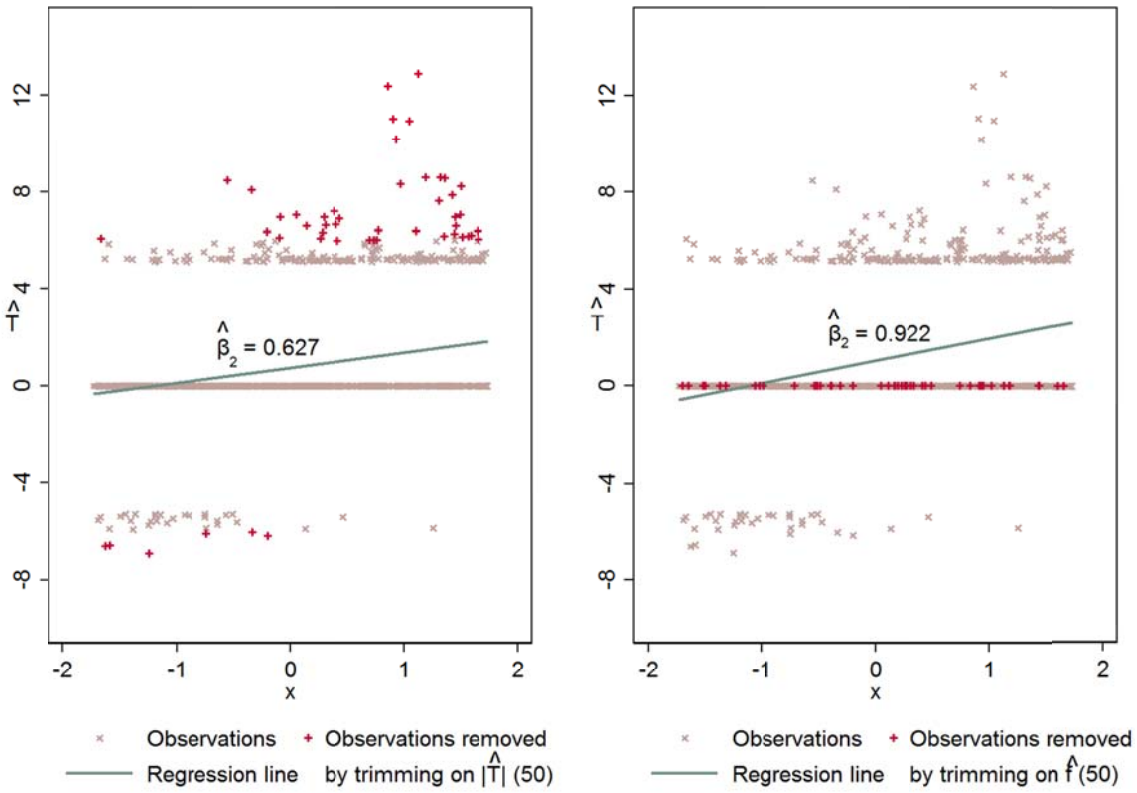

Figure A2. Trimming on $\hat{T}$ (left panel) versus trimming on $\hat{f}$ (right panel)
($\beta_1 = \beta_2 = 1;\ \lambda = 2;\ \text{trim level} = 5\%$; N=1,000 observations)