A Simple Estimator for Binary Choice Models With Endogenous Regressors

Yingying Dong and Arthur Lewbel*

University of California Irvine and Boston College

Revised February 2012

Abstract

This paper provides a few variants of a simple estimator for binary choice models with endogenous or mismeasured regressors, or with heteroskedastic errors. Unlike control function methods, which are generally only valid when endogenous regressors are continuous, the estimators proposed here can be used with limited, censored, continuous, or discrete endogenous regressors, and they allow for latent errors having heteroskedasticity of unknown form, including random coefficients. The variants of special regressor based estimators we provide are numerically trivial to implement. We illustrate these methods with an empirical application estimating migration probabilities within the US.

JEL codes: C25, C26.

Keywords: Binary choice, Binomial response, Endogeneity, Measurement error, Heteroskedasticity, Discrete endogenous regressor, Censored regressor, Random coefficients, Identification, Latent variable model.

[*] We would like to thank Jeff Wooldridge for helpful comments. Corresponding Author: Arthur Lewbel, Dept of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, http://www2.bc.edu/~lewbel/

1 Introduction

This paper describes numerically very simple estimators that can be used to estimate binary choice (binomial response) models when some regressors are endogenous or mismeasured, and when latent errors can be heteroskedastic and correlated with regressors. These estimators have some significant advantages relative to leading alternatives such as maximum likelihood, control functions, and two stage least squares linear probability models. The model and associated estimators also allow for latent errors having heteroskedasticity of unknown form, including random coefficients on most of the regressors.

Consider a binary choice model $D = I\left(X'\beta + \varepsilon \geq 0\right)$, where D is an observed dummy variable that equals zero or one, X is a vector of observed regressors, β is a vector of coefficients to be estimated, ε is an unobserved error, and $I(\cdot)$ is the indicator function that equals one if its argument $\cdot$ is true and zero otherwise. The special case of a probit model has $\varepsilon \sim N(0, 1)$, while for logit ε has a logistic distribution. The initial goal is to estimate β, but ultimately we are interested in choice probabilities and the marginal effects of X, looking at how the probability that D equals one changes when X changes.

Suppose that some elements of X are endogenous or mismeasured, and so may be correlated with ε. In addition, the latent error term ε may be heteroskedastic (e.g., some regressors could have random coefficients) and has an unknown distribution. Let Z be a vector of instrumental variables that are uncorrelated with ε. There are three common methods for estimating such models: maximum likelihood, control functions, and linear probability models. We now briefly summarize each, noting that each method has some serious drawbacks, and we then discuss the relative advantage of this paper's alternative approach based on Lewbel's (2000) special regressor estimator. A more complete comparison of these estimators, including the special regressor method, is provided in Lewbel, Dong, and Yang (2012).

One method for estimation is maximum likelihood. Maximum likelihood estimation requires a complete parametric specification of how each endogenous regressor depends on Z and on errors. Let e denote the set of errors in the equations describing how each endogenous regressor depends on Z. For example, if the endogenous regressors are linear in Z, then e would be the errors in the first stage of a linear two stage least squares. Maximum likelihood requires that these first stage models be correctly parameterized, and it requires a correctly specified parametric functional form for the joint distribution of e and ε conditional upon Z. One drawback of maximum likelihood is the difficulty in correctly specifying all this information. A second problem is that the resulting joint likelihood function associated with binary choice and endogenous regressors will often have numerical difficulties associated with estimating nuisance parameters such as the covariances between e and ε.

A second type of estimator for binary choice with endogenous regressors is based on control functions. This methodology can be traced back at least to Heckman (1976) and Heckman and Robb (1985), and for binary choice with endogenous regressors can range in complexity from the simple ivprobit command in Stata (for a model like that of Rivers and Vuong (1988) and Blundell and Smith (1989)), to Blundell and Powell's (2003, 2004) estimators with multiple nonparametric components. Control function estimators for binary choice are typically consistent *only* when the endogenous regressors are continuously distributed, because one cannot otherwise estimate the latent error e, and so should not be used when the endogenous regressors are discrete or limited. Also, like maximum likelihood, control function estimators require models of the endogenous regressors as functions of Z and e to be correctly specified (though not necessarily parametric). In addition, control functions do not permit many types of heteroskedasticity, and can suffer from numerical problems similar to those of maximum likelihood.

A third approach to dealing with endogenous regressors is to estimate an instrumental variables linear probability model, that is, linearly regress D on X using two stage least squares with instruments Z. However, despite its simplicity and popularity, this linear probability model does not nest standard logit or probit models as special cases, is generally inconsistent with economic theory for binary choice, can get the signs and magnitudes of the effects of X on D incorrect, and can easily generate silly results such as fitted choice probabilites that are negative or greater than one. Additional drawbacks of the linear probability model (along with more details on all of the above estimator comparisons) are provided in Lewbel, Dong, and Yang (2012).

One reason for the popularity of the linear probability model, despite its serious flaws, is that for true linear regressions, two stage least squares has many desirable properties. In linear regression, two stage least squares does not require a correct specification, or indeed any specification, of models for the endogenous regressors. One might interpret the first stage of two stage least squares as a model of the endogenous regressors, but unlike maximum likelihood or control function based estimators, linear two stage least squares does not require the errors in the first stage regressions to satisfy any of the properties of a correctly specified model. Linear two stage least squares only requires that the instruments Z be correlated with regressors and uncorrelated with errors. Linear two stage least squares also allows for general forms of heteroskedasticity. Special regressor estimators possesses these desirable properties of linear two stage least squares, but without the drawbacks of the linear probability model.

This paper provides a simplified version of Lewbel's (2000) special regressor estimator. It overcomes all of the above listed drawbacks of linear probability models, control functions, and maximum likelihood. The special regressor based estimator consistently estimates β, nests logit and probit as special cases, allows for general and unknown forms of heteroskedasticity (including, e.g., random coefficients), does not require correctly specified models of the endogenous regressors, does not require endogenous regressors to be continuously distributed (e.g., permitting censored or discrete endogenous regressors), and does not suffer from computational convergence difficulties because it does not require numerical searches.

The price to be paid for these advantages is that the special regressor estimator requires one exogenous regressor to be conditionally independent of ε, appear additively to ε in the model, and be conditionally continuously distributed with a large support (though, as we discuss later, the support does not need to be as large as the first papers in this literature suggest). Call this special regressor V. Only one special regressor is required, no matter how many endogenous regressors appear in the model.

Let S denote the vector consisting of all the instruments and all the regressors other than V. A difficulty in implementing Lewbel's (2000) special regressor estimator is that it requires an estimate of the density of V, conditional upon S. In this paper we propose simple semiparametric specifications of this density, thereby yielding special regressor estimators that are numerically very easy and practical to implement.

Using a sample of individuals in the labor force, we empirically illustrate the special regressor method by applying our estimator to a model of migration. Specifically, we model the probability of moving from one state to another within the United States. Our special regressor V is an individual's age, which is arguably exogenous and continuous. The model contains both a discrete (binary) and a continuous endogenous regressor, namely, home ownership and family income.

For this model, linear probability is generically inconsistent as noted above, while maximum likelihood would require fully specifying a joint model of migration, home ownership, and income. Control function methods would also require modeling these variables, and will not in general be feasible here because homeownership is discrete. In contrast to the difficulty of maximum likelihood and the inconsistency of control function and linear probability estimates, we show that our simplified special regressor based estimator is numerically trivial to implement and provides reasonable estimates for this model.

1.1 Normalization of the Binary Choice Model

Let V be some conveniently chosen exogenous regressor that is known to have a positive coefficient, and now let X be the vector of all the other regressors in the model. We now write the binary choice model as

$$D = I\left(X'\beta + V + \varepsilon \geq 0\right) \tag{1}$$

where the variance of ε is some unknown constant σ_{ε}^{2}, and β is a vector of coefficients to be estimated.

Models like probit often normalize the variance of the error ε to be one, but it is observationally equivalent to instead normalize the positive coefficient of a regressor to equal one. Estimation of choice probabilities (propensity scores) and of marginal effects are unaffected by this choice of normalization. For special regressor estimators (and many other semiparametric estimators), equation (1) is more convenient than normalizing the variance of ε to one.

Normalizing a regressor coefficient rather than the error variance often makes better economic sense as well. For example, if D is the decision of a consumer to purchase a good and V is the negative logged price of the good, then having demand curves slope downward determines the sign of the coefficient of V, and in this scaling equation (1) $X'\beta + \varepsilon$ is the log of the consumer's reservation price (that is, their willingness to pay) for the good.

If unknown a priori, the sign of the coefficient of V can be determined as the sign of the estimated average derivative $E[\partial E(D \mid V, X)/\partial V]$, or weighted average derivative such as Powell, Stock, and Stoker (1989). The sign of this estimator converges faster than rate root n, so a first stage estimation of the sign won't affect the later distribution theory. Even simpler is to just graph the nonparametric regression of D on V and X, and see if the estimated function is upward or downward sloping in V.

1.2 The Special Regressor Method - Literature Review

The special regressor method is characterized by three assumptions. First, it requires additivity between the special regressor V and the model error ε (or some function of ε). In standard binary choice models where the latent variable, $X'\beta + V + \varepsilon$, is linear in regressors and an error term, all regressors in the model satisfy this assumption. Second, it requires the special regressor V to be conditionally independent of the model error ε, conditioning on other covariates. If the distribution of ε is independent of the exogenous regressors (e.g., is homoskedastic), then any exogenous regressor will satisfy this assumption. Third, the special regressor needs to be continuously distributed with a large support, though this last condition can sometimes be relaxed (see Magnac and Maurin 2007, 2008).

The special regressor method has been employed in a wide variety of limited dependent variable models including binary, ordered, and multinomial choice as well as censored regression, selection and treatment models (Lewbel 1998, 2000, 2007a), truncated regression models (Khan and Lewbel 2007), binary panel models with fixed effects (Honore and Lewbel 2002, Ai and Gan 2010), dynamic choice models (Heckman and Navarro 2007, Abbring and Heckman 2007), contingent valuation models (Lewbel, Linton, and McFadden 2011), market equilibrium models of multinomial discrete choice (Berry and Haile 2009a, 2009b), models of games, including entry games and matching games (Lewbel and Tang 2011, Khan and Nekipelov 2011, Fox and Yang 2012), and a variety of models with (partly) nonseparable errors (Lewbel 2007b, Matzkin 2007, Briesch, Chintagunta, and Matzkin 2009).

Additional empirical applications of special regressor methods include Anton, Fernandez Sainz, and Rodriguez-Poo (2002), Cogneau and Maurin (2002), Goux and Maurin (2005), Stewart (2005), Avelino

(2006), Pistolesi (2006), Lewbel and Schennach (2007), and Tiwari, Mohnen, Palm, and van der Loeff (2007). Earlier results that can be reinterpreted as special cases of special regressor based identification methods include Matzkin (1992, 1994) and Lewbel (1997). Vytlacil and Yildiz (2007) describe their estimator as a control function, but their identification of the endogenous regressor coefficient essentially treats the remainder of the latent index as a special regressor. Recent econometric theory involving special regressor models includes Jacho-Chávez (2009), Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b). The methods we propose in this paper to simplify special regressor estimation in binary choice models could be readily applied to many of the other applications of the method cited here, such as ordered choice and selection models.

To illustrate how a special regressor works to identify limited dependent variable models, consider the simple model $D = I(\alpha + V + \varepsilon \geq 0)$ where ε has an unknown mean zero distribution, V is independent of ε, and we want to estimate the constant α. Let $F_{-\alpha-\varepsilon}()$ and $f_{-\alpha-\varepsilon}()$ denote the unknown probability distribution function and probability density function (respectively) of $-\alpha - \varepsilon$. Suppose this distribution has support given by the interval $[L, U]$.

In this model, $E(D \mid V) = \Pr(-\alpha - \varepsilon \leq V) = F_{-\alpha-\varepsilon}(V)$, so by estimating the conditional mean of D given V, we estimate the distribution of $-\alpha - \varepsilon$, evaluated at V. Once we know this distribution, we can calculate its mean, which is $-\alpha$. In particular, by the definition of an expectation, $\alpha = -E(-\alpha - \varepsilon) = -\int_L^U V f_{-\alpha-\varepsilon}(V)\, dV = -\int_L^U V \left[\partial F_{-\alpha-\varepsilon}(V)/\partial V\right] dV = -\int_L^U V \left[\partial E(D \mid V)/\partial V\right] dV$, which shows one way in which α could be recovered from an estimate of $E(D \mid V)$. This illustrates the main idea that having an additive, independent regressor V can be used to identify the model. The actual special regressor estimator will applies an integration by parts argument to this integral to obtain a simpler expression.

Note that this construction requires V to take on all values in the interval $[L, U]$, since we need to evaluate $E(D \mid V)$ for all those values of V. This is the sense in which special regressor estimation requires a large support. However, suppose that V has a smaller support, say the interval $[\ell, u]$ where $L \leq \ell < 0$ and $U \geq u > 0$. Then we will only be able to estimate $\widetilde{\alpha} = -\int_\ell^u V \left[\partial E(D \mid V)/\partial V\right] dV$, which is in general not equal to α. But suppose the following equality between upper and lower tails of

$f_{-\alpha-\varepsilon}$ holds: $\int_L^\ell V f_{-\alpha-\varepsilon}(V)\,dV = \int_u^U V f_{-\alpha-\varepsilon}(V)\,dV$. Then $\alpha = \widetilde{\alpha}$ and the special regressor method estimator still works even though the support of V is not large enough relative to the support of $\alpha + \varepsilon$. This is tail symmetry, which is described in more detail in Magnac and Maurin (2007). Even when tail symmetry does not hold exactly, the size of bias term $\alpha - \widetilde{\alpha}$ will equal the magnitude of the difference between these two tail integrals, and so the the bias resulting from applying special regressor methods when the support of V is too small will generally be small if the density of ε either has thin tails or is close to symmetric in the tails.

In the more general model $D = I(X'\beta + V + \varepsilon \geq 0)$ with instruments Z, the conditional expectation $E(D \mid V, X, Z)$ will equal the conditional distribution of $X'\beta + \varepsilon$ conditioning on X and Z, evaluated at V, and this can be used to identify β (and the distribution of ε). Lewbel (2000) proposes a shortcut for directly estimating β that avoids the step of estimating $E(D \mid V, X, Z)$, however, this shortcut requires the conditional density of V given X and Z. This is the step that we simplify in this paper.

2 Special Regressor Binary Choice

Assume that $D = I(X'\beta + V + \varepsilon \geq 0)$ and that some or all of the elements of X are endogenous. Assume Z has the standard properties of instruments in linear regression models, i.e., that $E\left(ZX'\right)$ has rank equal to the number of elements of X, and that $E(Z\varepsilon) = 0$. As usual, Z would include all elements of X that are exogenous, including a constant. The special regressor V is not included in Z. The distribution of ε can be unknown, and be heteroskedastic, with second and higher moments depending on X and Z in unknown ways. If the model suffers only from heteroskedasticity and not endogeneity, then let Z equal X.

Unlike linear models, having $E\left(ZX'\right)$ full rank and $E(Z\varepsilon) = 0$ is not sufficient to identify β in the binary choice model. But by adding assumptions regarding the special regressor V, Theorem 1 in the Appendix shows how to construct a variable T having the property that $T = X'\beta + \widetilde{\varepsilon}$ and $E(Z\widetilde{\varepsilon}) = 0$. We will then be able to identify and estimate β by a linear two stage least squares regression of T on X using instruments Z, i.e., $\widetilde{Z} = Z'E\left(ZZ'\right)^{-1} E\left(ZX'\right)$ and $\beta = E\left(\widetilde{Z}X'\right)^{-1} E\left(\widetilde{Z}T\right)$.

Define S to be the union of all the elements of X and Z, so S is the vector of all the instruments and all of the regressors except for the special regressor V. The additional information that will be required regarding V is a model of the form $V = g\left(U, S\right)$ where U is an error term.

2.1 Simplest Estimator

To make estimation based on Theorem 1 in the appendix simple, a convenient parametric model is chosen here for g. This is given in Corollary 1. We then propose some generalizations that impose less restrictive assumptions on the special regressor while still being numerically simple to implement.

COROLLARY 1: Assume $D = I(X'\beta + V + \varepsilon \geq 0)$, $E(Z\varepsilon) = 0$, $E\left(V\right) = 0$, $V = S'b + U$, $E\left(U\right) = 0$, $U \perp \left(S, \varepsilon\right)$, and $U \sim f\left(U\right)$, where $f(U)$ denotes a mean zero density function having support $supp(U)$ that contains $supp(-S'b - X'\beta - \varepsilon)$. Define T by

$$T = \left[D - I(V \geq 0)\right]/f\left(U\right) \tag{2}$$

Then $T = X'\beta + \widetilde{\varepsilon}$ where $E\left(Z\widetilde{\varepsilon}\right) = 0$.

Corollary 1 assumes that the function $V = g\left(U, S\right) = S'b + U$ is linear in covariates S and an error U. By Theorem 1 in the Appendix, other regular parametric models for g could be assumed instead. This particular model is chosen for its simplicity. If U is normal, or has any other distribution over the whole real line, then the support assumption in Corollary 1 will be satisfied. Alternatively, as noted earlier, this large support assumption could be replaced by the ε tail symmetry of Magnac and Maurin (2007).

A limitation of the special regressor method is that we cannot include a term like V^2 in the model along with V (though we could replace V with some known transformation of V if necessary). This is because, if the model contained both V and V^2, then V^2 would be an element of S, and the errors U would not be continuous and independent in the model $V = g\left(U, S\right)$.

Other than the assumed model for the special regressor V, nothing is required for estimation using Corollary 1 other than what would be needed for a linear two stage least squares regression, specifically, that $E(ZX')$ have full rank and $E(Z\varepsilon) = 0$.

Based on Corollary 1 we have the following simple estimator. Assume we have data observations D_i, X_i, Z_i, and V_i. Recall that S_i is the vector consisting of all the elements of X_i and Z_i. Also note that X_i and Z_i should include a constant term.

ESTIMATOR 1

Step 1. Assume V_i has mean zero (if not, then demean it first). Let $\widehat{b}$ be the estimated coefficients of S in an ordinary least squares linear regression of V on S. For each observation i, construct data $\widehat{U}_i = V_i - S_i'\widehat{b}$, which are the residuals from this regression.

Step 2. For each i let $\widehat{f}_i$ be the nonparametric density estimator given later by equations (4) or (5). Alternatively, estimate a parametric f using $\widehat{U}_i$. For example, if f is normal or otherwise parameterized by its variance as $f\left(U \mid \sigma^2\right)$, then let $\widehat{\sigma}^2 = \sum_{i=1}^n \widehat{U}_i^2/n$ and for each observation i define $\widehat{f}_i = f\left(\widehat{U}_i \mid \widehat{\sigma}^2\right)$, where f is a standard normal (or other) density function.

Step 3. For each observation i construct data $\widehat{T}_i$ defined as $\widehat{T}_i = [D_i - I(V_i \geq 0)] / \widehat{f}_i$.

Step 4. Let $\widehat{\beta}$ be the estimated coefficients of X in an ordinary linear two stage least squares regression of $\widehat{T}$ on X, using instruments Z. It may be necessary to discard outliers in this step. Given $\widehat{\beta}$, choice probabilities and marginal effects can be obtained as in Lewbel, Dong, and Yang (2012).

Estimator 1 differs from Lewbel (2000) mainly in that it assumes a parametric or semiparametric model for V, while Lewbel (2000) uses a nonparametric conditional density estimator for V. However, Lewbel (2000) is not strictly more general than Estimator A, since Theorem 1 and Corollary 1 allow V to depend on all of the elements of X, including the endogenous regressors, while Lewbel (2000) assumes the conditional density of V does not depend on endogenous regressors.

This estimator is numerically trivial, requiring no numerical searches and no estimation steps more complicated than linear regressions. It is therefore also fast and easy to obtain standard errors, test statistics, or confidence intervals by an ordinary bootstrap if desired, drawing observations D_i, X_i, Z_i, and V_i with replacement.

In Estimator 1, nothing constrains the regression of V on S in the first step to be linear. For example, if necessary this first step regression could include squared and cross terms of S. The later steps of the estimator are estimator are unchanged by this generalization.

Also, Estimator 1 can be easily modified to allow for more general parametric specifications of f. In particular, suppose f is any regular continuous density function parameterized by a vector θ, which we may denote as $f(U \mid \theta)$. Then in step 2 we could estimate $\widehat{\theta}$ by maximizing $\sum_{i=1}^{n} \ln f\left(\widehat{U}_i \mid \theta\right)$, and then let $\widehat{f}_i = f\left(\widehat{U}_i \mid \widehat{\theta}\right)$ for each observation i. This step is then just a maximum likelihood estimator for θ.

As noted in step 4, it may be desirable to look for and discard outliers, that is, observations i for which $\widehat{T}_i$ takes on values that are very large in magnitude. This is discussed further in section 2.3.

2.2 Allowing for Heteroskedasticity in V

All the estimators in this paper allow the model errors ε to be heteroskedastic, e.g., X having random coefficients does not violate the assumptions of Theorem 1 or Corollary 1. More generally, the model errors ε can have second and higher moments that depend on the regressors in arbitrary, unknown ways. However, the model in the previous section assumes that $V = S'b + U$ where U is independent of the other covariates S, and so assume that the errors U in the V model are homoskedastic. In this section we provide a more general model for V that allows higher moments of V to depend on S, yielding a slightly more complicated, but still numerically trivial, estimator.

Let $\widetilde{S}$ denote the vector that consists of all the elements of S, and all the squares and cross products of all the elements of S.

COROLLARY 2: Assume $D = I(X'\beta + V + \varepsilon \geq 0)$, $E(Z\varepsilon) = 0$, $E(V) = 0$, $U \perp (S, \varepsilon)$, $V = S'b + \left(\widetilde{S}'c\right)^{1/2} U$, $E(U) = 0$, $var(U) = 1$, and $U \sim f(U)$ where $f(U)$ is a density function that has mean zero, variance one, and support $supp(U)$ that contains $supp\left[(-S'b - X'\beta - \varepsilon)\left(\widetilde{S}'c\right)^{-1/2}\right]$. Define T by

$$T = [D - I(V \geq 0)]\left[\left(\widetilde{S}'c\right)^{1/2}\right]/f(U) \tag{3}$$

Then $T = X'\beta + \widetilde{\varepsilon}$ where $E(Z\widetilde{\varepsilon}) = 0$.

Corollary 2 introduces a multiplicative heteroskedastic term, so the errors in the V regression are now $\left(\widetilde{S}'c\right)^{1/2}U$ instead of just U. Corollary 2 follows from Theorem 1 with $g\left(U,S\right)=S'b+(\widetilde{S}'c)^{1/2}U$, which puts the term $\widetilde{S}'c$ into equation (3). As with Corollary 1, the large support assumption for U could alternatively be replaced by the ε tail symmetry assumptions of Magnac and Maurin (2007). The estimator corresponding to Corollary 2 is an immediate generalization of Estimator 1, as follows.

ESTIMATOR 2

Step 1. Assume V_i has mean zero (if not, then demean it first). Let $\widehat{b}$ be the estimated coefficients of S in an ordinary least squares linear regression of V on S. For each observation i, construct data $\widehat{W}_i=V_i-S_i'\widehat{b}$, which are the residuals of this regression.

Step 2. Let $\widehat{c}$ be the estimated coefficients of $\widetilde{S}$ in an ordinary least squares linear regression of $\widehat{W}^2$ on $\widetilde{S}$. For each observation i, construct data $\widehat{U}_i=\left(\widetilde{S}_i'\widehat{c}\right)^{-1/2}\widehat{W}_i$.

Step 3. For each i let $\widehat{f}_i$ be a nonparametric density estimator given later by equation (4) or (5), alternatively, define $\widehat{f}_i=f\left(\widehat{U}_i\right)$ where f is a normal or any other distribution that has mean zero and variance one.

Step 4. For each observation i construct data $\widehat{T}_i$ defined as $\widehat{T}_i=\left[D_i-I\left(V_i\geq 0\right)\right]\left[\left(\widetilde{S}_i'\widehat{c}\right)^{1/2}\right]/\widehat{f}_i$.

Step 5. Same as step 4 of Estimator 1.

In Estimator 2, step 2 comes from $W=(\widetilde{S}'c)^{1/2}U$ with $U\perp S$, so $E\left(W^2\mid S\right)=\widetilde{S}'cE\left(U^2\right)=\widetilde{S}'c$. The step 2 regression of $\widehat{W}^2$ on $\widetilde{S}'$ is the same as the regression that would be used for applying White's (1980) test for heteroskedasticity in the step 1 regression of V on S. An easy way to test if Estimator 2 is required instead of Estimator 1 is to perform White's test for heteroskedasticity on the step 1 regression of V on S. The presence of heteroskedasticity in this regression would then call for Estimator 2.

As with Estimator 1, there is nothing that requires the functions $S'b$ and $\widetilde{S}'c$ to be linear. One could if desired specify these as higher order polynomial regressions, or for example one could estimate $e^{\widetilde{S}'c}$ or $\left(\widetilde{S}'c\right)^2$ in place of $\widetilde{S}'c$ everywhere in the above. This could be useful since Estimator 2 will not work if $\widetilde{S}_i'\widehat{c}$ is not positive for every observation i, since variances must be positive.

2.3 Asymptotics and Efficiency

Let $M(V)$ be any mean zero distribution on the support of V chosen by the econometrician. By Theorem 1, the term $I(V \geq 0)$ in the estimators can be replaced with $M(V)$. In particular, choosing M to be a simple differentiable function like $M(V) = I(-1 \leq V \leq 1)(V+1)/2$ (corresponding to a uniform distribution on -1 to 1) can simplify the calculation of limiting distributions and possibly improve the finite sample performance of the estimators.

To obtain standard error estimates without bootstrapping, and possibly to improve efficiency, the parameters in Estimators 1 and 2 can be estimated simultaneously instead of sequentially using GMM. Specifically, assuming f is parameterized by its variance, the steps comprising Estimator 1 correspond to the following moment conditions:

$$E\left[S\left(V - S'b\right)\right] = 0, \quad E\left[\left(V - S'b\right)^2 - \sigma^2\right] = 0, \quad E\left[Z\left(\frac{D - I(V \geq 0)}{f\left(V - S'b \mid \sigma^2\right)} - X'\beta\right)\right] = 0$$

These moments correspond respectively to the regression model of V, the estimator of the variance of U, and the transformed instrumental variables special regressor estimator.

If the density of f is parameterized more generally as $f(U \mid \theta)$ for some parameter vector θ, then the moment $E\left[\left(V - S'b\right)^2 - \sigma^2\right] = 0$ could be replaced by $E\left[\partial \ln f\left(V - S'b, \theta\right)/\partial\theta\right] = 0$, which is the vector of score functions associated with maximum likelihood estimation of θ.

The moments corresponding to Estimator 2 are

$$E\left[S\left(V - S'b\right)\right] = 0, \quad E\left[\widetilde{S}\left(\left(V - S'b\right)^2 - \widetilde{S}'c\right)\right] = 0, \quad E\left[Z\left(\frac{D - I(V \geq 0)}{f\left(V - S'b\right)}\left(\widetilde{S}'c\right)^{1/2} - X'\beta\right)\right] = 0$$

In estimator 2, U has mean zero and variance one by construction. If f, the density of U, is parameterized by parameters θ in addition to its mean and variance, then one could add the score function $E\left[\partial \ln f\left(V - S'b, \theta\right)/\partial\theta\right] = 0$ for estimating θ to this set of moments.

Applying ordinary two step GMM to either of these sets of moments provides estimates of the desired parameters β along with nuisance parameters b and σ^2 (or b and c for Estimator 2) that efficiently combine these estimation steps in the usual way for GMM, and also delivers asymptotic standard errors (possibly replacing $I(V \geq 0)$ with $M(V)$ as above). Alternatively, given the simplicity of the estimators, one could easily obtain standard errors, confidence intervals, or test statistics via bootstrapping.

We do not provide formal limiting distribution theory assumptions here, since the estimator is just GMM. However, a potential concern is that the definition of T involves dividing by a density. This could result in T having infinite variance, violating standard GMM limiting distribution theory. As shown by Khan and Tamer (2010), this generally leads to slower than root n convergence rates, unless the tails of U are very thick, or the distribution of ε is bounded, or Magnac and Maurin (2007) type tail symmetry conditions hold. If these conditions do not hold, then it would be necessary to apply the thick tailed GMM asymptotics of Hill and Renault (2010), or the asymptotics for irregularly identified models as described in Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b).

A practical implication of this construction of T is that one should watch out for outliers in the final step regression of $\widehat{T}$ on X. In particular, in some contexts it may be desirable to trim the data (that is, remove observations i where $\widehat{T_i}$ is extremely large in magnitude) before running the last step regression. Formally, this corresponds to asymptotic trimming, which trades off bias and variance to improve the mean squared error performance of the estimator.

Another implication is that the larger the variance (or other measures of spread such as interquartile range) of U or V, the better the estimator is likely to perform in practice. This should be borne in mind when choosing V. Lewbel (2000, 2007a) found in simulations that special regressor estimation tended to perform well in moderate size samples when the variance of V is as big or bigger than the variance of $X'\beta + \varepsilon$, and otherwise that finite sample biases decline rather slowly as the sample size is increased.

2.4 Other Density Estimators

Estimators 1 and 2 require an estimate of f_i, the density of U at each observation i. One possible nonparametric estimator is the standard one dimensional nonparametric kernel density estimator. This is

$$\widehat{f_i} = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{\widehat{U_i} - \widehat{U_j}}{h}\right) \quad \text{for } i = 1, ..., n \tag{4}$$

where the kernel function K is a symmetric density function like a standard normal density, and h is a bandwidth. Even with this nonparametric component, $\widehat{\beta}$ can be root n consistent and asymptotically normal, based on standard sets of regularity conditions, such as Newey and McFadden (1994), for two-step semiparametric estimation. The estimator for $\widehat{\beta}$ will still not require any numerical searches (except possibly a one dimensional search for the choice of bandwidth h), so bootstrapping would be entirely practical for estimating confidence intervals, tests, or standard errors, based on, e.g., Chen, Linton, and Van Keilegom (2003).

Instead of choosing a kernel and bandwidth, one could also use the sorted data density estimator of Lewbel and Schennach (2007), which is designed for estimating averages weighted by the inverse of a density, as is the case here. Given n observations of $\widehat{U}_i$, sort these observations from lowest to highest. For each observation $\widehat{U}_i$, let $\widehat{U}_i^+$ be the value of $\widehat{U}$ that, in the sorted data, comes immediately after $\widehat{U}_i$ (after removing any ties) and similarly let $\widehat{U}_i^-$ be the value that comes immediately before $\widehat{U}_i$. So $\widehat{U}_i^+$ is the smallest value of $\widehat{U}$ that is greater than $\widehat{U}_i$, and $\widehat{U}_i^-$ is the largest value of $\widehat{U}$ that is smaller than $\widehat{U}_i$[1],

Then the estimator is

$$\widehat{f}_i = \frac{2/n}{\widehat{U}_i^+ - \widehat{U}_i^-} \quad \text{for } i = 1, ..., n \tag{5}$$

Equation (5) is not a consistent estimator of f_i (its probability limit is random, not constant), but given regularity, averages of the form $\frac{1}{n}\sum_{i=1}^{n} \Pi_i / \widehat{f}_i$ will converge at rate root n, and our estimators entail averages of this form, e.g., Estimator 1 has $\Pi_i = Z_i \left(D_i - I\left(V_i \geq 0\right)\right)$. Asymptotic variance formulas are provided in Lewbel and Schennach (2007) and (in more generality) Jacho-Chávez (2009).

In addition to avoiding specification error in f, there is another advantage of using a nonparametric estimator for $\widehat{f}_i$. Results in Magnac and Maurin (2004) and Jacho-Chávez (2009) show that estimation of β will generally be more efficient using a nonparametric estimator of f than by using the true density, analogous to the result of Hirano, Imbens and Ridder (2003) that weighting by a nonparametrically estimated propensity score is more efficient than weighting by the true score in treatment effect estimation.

[1] For the endpoints, if no $\widehat{U}$ in the data is larger than $\widehat{U}_i$, then let $\widehat{U}_i^+ = \widehat{U}_i$, and if no $\widehat{U}$ in the data is smaller than $\widehat{U}_i$, then let $\widehat{U}_i^- = \widehat{U}_i$.

2.5 Combining Regressors

It may sometimes be possible to increase efficiency, or increase the relative support of the special regressor, by combining some exogenous covariates to construct a V. For example, suppose the model is $D = I(X'\beta + V_1 + V_2\alpha + \varepsilon \geq 0)$, where both V_1 and $V_1 + V_2\alpha$ satisfy the special regressor assumptions, i.e., V_1 is a special regressor and V_2 is exogenous and independent of ε (here V_2 could be continuous, discrete, limited, or could be a second special regressor). Then we can write down all the moments associated with estimator 1 or estimator 2 in section 2.3 treating V as V_1 and including αV_2 in $X'\beta$. These moments will identify α and β. We can also write down all the moments associated with estimator 1 or 2 based on defining V as $V_1 + V_2\alpha$. Then, to increase estimation efficiency, GMM could be applied to both sets of moments (those corresponding to either definition of V) simultaneously.

Alternatively, we could first obtain an estimate of $\widehat{\alpha}$ by estimating $E\left(D \mid V_1 + V_2\alpha, X, Z\right)$ using a conditional linear index model estimator such as Ichimura and Lee (2006), or just by weighted average derivative estimation of $\alpha = E\left[\partial E\left(D \mid V_1, V_2, X, Z\right)/\partial V_2\right] / \left[\partial E\left(D \mid V_1, V_2, X, Z\right)/\partial V_1\right]$ if both V_1 and V_2 are continuously distributed. Then, given a consistent $\widehat{\alpha}$ by one of these methods, we could construct $\widehat{V} = V_1 + V_2\widehat{\alpha}$ and then apply Estimator 1 or Estimator 2 using $\widehat{V}$ in place of V.

3 Empirical Illustration

In this section we illustrate our simple estimators (coded in Stata, available upon request) with an empirical application. Let D_i be the probability an individual i migrates from one state to another in the United States. Let age be the special regressor V_i, because it is exogenously determined, and human capital theory suggests it should appear linearly (or at least monotonically) in a threshold crossing model of the utility of migration. This is because workers migrate in part to maximize their expected lifetime income, and by construction the gains in expected lifetime earnings from a permanent change in wages decline linearly with age. Figure 1 provides empirical evidence for this relationship, showing a fitted kernel regression of D_i on age in our data, using a quartic kernel and bandwidth chosen by cross validation. We also depict the same nonparametric regression cutting the bandwidth in half, to verify that this near linearity is not an artifact of possible oversmoothing. Others have reported similar empirical evidence (See, e.g., Dong 2010 and the references therein) in accordance with the above human capital motivation for migration.

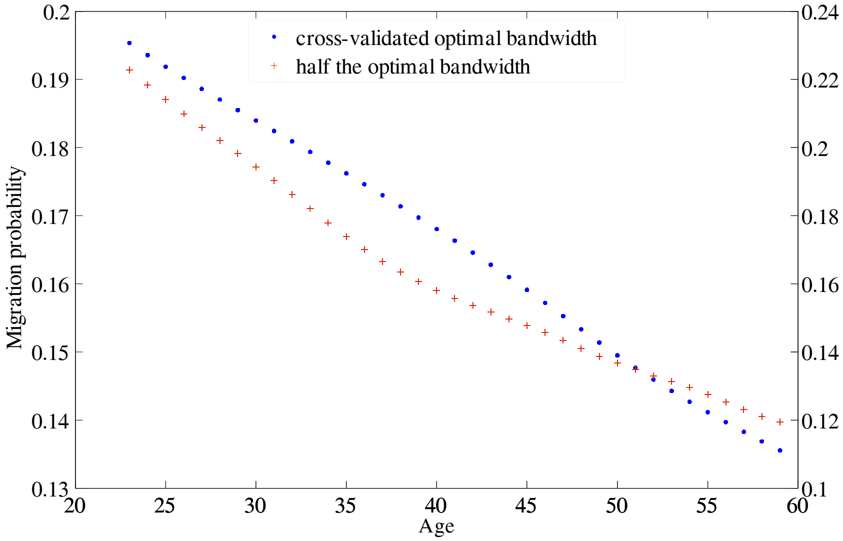

Figure 1: Nonparametric age profile of migration probabilities

Pre-migration income and home ownership greatly affect the decision of whether to move or not. Both are endogenous regressors in our binary choice model. Maximum likelihood would require an elaborate dynamic specification and an extensive amount of current and past information about individuals to completely model their homeownership decision and the determination of their wages and other income jointly with their migration decisions. See, e.g., Kennen and Walker (2011) for an example of a dynamic structural income based model of migration. Control function methods are also not appropriate for this application, because home ownership is discrete, and control functions are generally inconsistent when used with discrete endogenous regressors (see, e.g., Lewbel, Dong, and Yang 2012).

Our sample is 23 to 59 year old male household heads from the 1990 wave of the PSID, who have completed education and who were not retired at the time of their interview. This is intended to largely exclude people who are moving to retirement locations. The final sample has 4,689 observations, consisting of 807 migrants and 3882 nonmigrants. We let $D = 1$ if an individual changes his state of residence during 1991 - 1993, and 0 otherwise. We define the special regressor V to be the negative of age, minus its mean (ensuring it has a positive coefficient and mean zero).

Our endogenous regressors are log(income), defined as the logarithm of family income averaged over 1989 and 1990, and homeowner, a dummy indicating whether one owns a home in 1990. The remaining regressors comprising X, which we take as exogenous, are education (in years), number of children, and dummy indicators for white, disabled, and married. Our instruments Z consist of the exogenous regressors, along with government defined benefits received in 1989 and 1990, i.e., the value of food stamps and other welfare benefits such as Aid to Families with Dependent Children (AFDC), and state median residential property tax rates, computed from the 1990 U.S. Census of Population and Housing and matched to the original PSID data. Government benefits have been used by others as instruments for household income in wage and labor supply equations, based on their being determined by government formulas rather than by unobserved attributes like ability or drive. Similarly, property tax rates affect homeownership costs and hence the decision of whether to own or rent, while being exogenously set by government rules.

Although age is exogenously determined, that does not guarantee that age satisfies the required special regressor exogeneity assumptions, because age could affect the endogenous regressors in ways that cause a violation of conditional independence (we'd like to thank Jeffrey Wooldridge for pointing this out). This concern may be partially mitigated by our inclusion of the endogenous regressors in S, and hence in the model for V, since our estimator (unlike the original Lewbel 2000 version) only requires U, not V, to satisfy conditional independence.

The special regressor formally requires $-V$ to have the same or larger support than $X'\beta + \varepsilon$. As discussed earlier, in practice finite sample biases will tend to be small when measures of the empirical spread of V (standard deviation or interquantile ranges) are comparable to, or larger than, those of $X'\widehat{\beta}$. In our application, the standard deviation of $X'\widehat{\beta}$ (using $\widehat{\beta}$ from estimator 1) is either 16.3 or 12.4 depending on the choice of estimator for $\widehat{f}$ (kernel or sorted density, respectively). These are comparable in magnitude to the standard deviation of V, which is 9.0. Moreover, much of this difference in spread is due to a small fraction of outliers in $X'\widehat{\beta}$. Quantile measures of spread are similar, e.g., the difference between the 5th and 95th quantile of V is 30.0, while that of $X'\widehat{\beta}$ is 44.5 or 36.6.

Table 2 presents the estimated marginal effects of covariates by our special regressor estimators. For comparison, results from standard probit and ivprobit are also presented. Marginal effects are calculated from coefficient β estimates using formulas given in Lewbel, Dong, and Yang (2012). We report marginal effects because they have more direct economic relevance than β, and because they are directly comparable across specifications, including probit.

Table 2: The estimated migration equation - marginal effects

	Dependent variable: migration (0/1)					
	Estimator 1-(a)	Estimator 1-(b)	Estimator 2-(a)	Estimator 2-(b)	ivprobit	probit
age	0.003	0.004	0.002	0.002	-0.0008	0.002
	(0.001)**	(0.001)***	(0.001)**	(0.0007)***	(0.001)	(0.0007)***
log(income)	-0.013	-0.012	-0.026	-0.037	0.065	-0.009
	(0.013)	(0.015)	(0.014)*	(0.014)**	(0.035)*	(0.007)
homeowner	-0.055	-0.050	-0.043	-0.026	-0.330	-0.086
	(0.031)*	(0.033)	(0.030)	(0.033)	(0.058)***	(0.013)***
white	0.017	0.003	-0.004	-0.003	0.006	-0.010
	(0.012)	(0.012)	(0.007)	(0.006)	(0.014)	(0.012)
disabled	-0.165	-0.134	-0.187	-0.205	0.018	-0.012
	(0.073)**	(0.066)	(0.041)***	(0.041)***	(0.040)	(0.033)
education	0.005	0.006	0.003	0.004	0.0002	0.0004
	(0.002)**	(0.003)	(0.001)***	(0.001)***	(0.003)	(0.002)
married	-0.004	0.018	0.050	0.046	0.020	-0.006
	(0.011)	(0.018)	(0.015)***	(0.015)***	(0.025)	(0.017)
# of children	0.018	0.019	-0.006	-0.006	0.013	0.010
	(0.006)***	(0.007)***	(0.003)**	(0.003)**	(0.005)***	(0.005)**

Note: Bootstrapped standard errors are in the parentheses; *significant at the 10% level; **significant at the 5% level; ***significant at the 1% level.

Columns 1 and 2 of Table 2 are based on Estimator 1, (which assumes U is homoskedastic), using (a) an ordinary kernel density estimator and (b) the sorted data estimator, respectively. The kernel density estimator is given by equation (4), with a standard Epanechnikov kernel function K (though the results are not sensitive to the choice of kernel function) and bandwidth parameter h given by Silverman's rule. The sorted data estimator is given by equation (5). We used nonparametric density estimators for $\widehat{f}$ because estimates of $\widehat{U}$ from both Estimators 1 and 2 were somewhat skewed and kurtotic (with normality rejected by Jarque-Bera tests) and for the efficiency advantages discussed in section 2.4.

Columns 3 and 4 of Table 2 are from Estimator 2, using (a) kernel and (b) sorted data density estimators, respectively. White's (1980) test on the regression of V on S shows significant heteroskedasticity, indicating the more general Estimator 2 is necessary in this case. Recall that $\widetilde{S}$ in the heteroskedasticity term $\widetilde{S}'c$ was defined to be the vector of all elements of S and all of their squares and cross products. The total number of terms in $\widetilde{S}$ is rather high, so for parsimony we only included the squares and cross terms of the most relevant regressors of S in the construction of $\widetilde{S}'c$ (equivalent to setting the coefficients of other elements of $\widetilde{S}$ equal to zero). Note that all of this discussion regarding heteroskedasticity refers only to the equation for the special regressor V; all our estimators permit the model error ε to have variance and higher moments that depend on S in arbitrary ways.

For comparison with Estimators 1 and 2, Column 5 of Table 2 uses the ivprobit estimator from Stata. Let e_1 and e_2 respectively denote the errors in linear regressions of log(income) and the homeowner dummy on the instruments Z. The ivprobit estimator assumes that e_1, e_2, and the latent binary choice model error ε, are jointly distributed as homoskedastic trivariate normal. A drawback of ivprobit is that this assumption cannot hold for a discrete endogenous variable like our homeowner dummy, because the errors e_2 in a linear probability model (which is what the linear regression of homeownership on Z is) cannot be homoskedastic, and are generally nonnormal. As a result, ivprobit estimates, like other control function estimators, are generally inconsistent when used with discrete endogenous regressors. In contrast, our proposed Estimators 1 and 2 do not make any assumptions regarding properties of the errors e_1 and e_2. Also, ivprobit, unlike special regressor estimators, does not permit the model error ε to be heteroskedastic. See Lewbel, Dong, and Yang (2012) for more details on these points. Finally, column 6 in Table 2 reports ordinary probit estimates, which ignores any regressor endogeneity and possible heteroskedasticity in ε, and is provided here as a baseline benchmark.

The estimated marginal effect of negative age V is modest but statistically significant, and is similar across all specifications except ivprobit. Unlike the other specifications, ivprobit gives V the wrong sign, inconsistent with the human capital argument that potential wage gains from moving become smaller as one ages. Log income has a marginally significant coefficient in the heteroskedasticity corrected (Estimator 2 based) models. One would expect income to have a significant effect on migration. The relatively large standard errors on this variable may be due to weakness in the government defined benefits instrument, which for many people is zero. Unlike all the other estimators, the ivprobit estimates have a counterintuitive positive and statistically significant sign for log income. Probit and Estimator 1 give negative income effects, though small in magnitude compared to Estimator 2.

The endogenous homeowner dummy has a negative sign in all the estimators, consistent with the fact that fixed costs of moving are higher if one is a homeowner. The estimated magnitude of this effect is largest for ordinary probit, smallest for ivprobit, and roughly halfway between these two extremes in the special regressor estimators. Intuitively, people who buy a home should be those who prefer (or expect) to not move, so homeownership should be negatively correlated with any unobserved preference for migration. Ordinary probit fails to account for this endogeneity of homeownership on migration and so yields an overestimate of the negative impact of homeownership on migration probabilities, while ivprobit is inconsistent when endogenous regressors are discrete, which may be causing ivprobit to overcompensate for this endogeneity. Finally, being disabled significantly reduces the probability of migration in all the models except for ivprobit, and education has a small positive effect on migration across the board.

4 Conclusions

Commonly used methods to deal with heteroskedasticity and endogenous regressors in binary choice models are linear probability models, control functions, and maximum likelihood. Each of these types of estimators have some drawbacks. Unlike these other estimators (each of which only has some of the following attractive features), the special regressor based estimators we provide here possess all of the following attributes: They provide consistent estimates of the model coefficients β, they nest logit and probit as special cases, they allow for general and unknown forms of heteroskedasticity (including, e.g., random coefficients), they do not require correctly specified models of the endogenous regressors, they do not require endogenous regressors to be continuously distributed, and they do not require numerical searches. What special regressor estimators do require are ordinary instruments, and just one exogenous regressor (no matter how many regressors are endogenous) to be conditionally independent of the latent error ε and be conditionally continuously distributed with a large support.

In this paper, we provide variants of the special regressor model that are numerically almost as trivial to implement as linear probability models. We illustrate how our special regressor estimators can be implemented in practice, and apply our estimators to estimating migration probabilities in the presence of both discrete and continuous endogenous regressors. We compare our estimators with standard probit and ivprobit in this empirical application.

Special regressor methods can be applied in a variety of settings in addition to binary choice, as listed in our literature review. The same models for V and f that are proposed here could be used to simplify these other applications as well.

5 References

Abbring, J. H. and J. J. Heckman, (2007) "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in: J.J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics, edition 1, volume 6B, chapter 72 Elsevier.

Ai, C. and L. Gan, (2010) "An alternative root- consistent estimator for panel data binary choice models" Journal of Econometrics, 157, 93-100

Avelino, R. R. G. (2006), "Estimation of Dynamic Discrete Choice Models with Flexible Correlation in the Unobservables with an Application to Migration within Brazil," unpublished manuscript, University of Chicago.

Ait-Sahalia, Y., P. J. Bickel, and T. M. Stoker (2001), "Goodness-of-fit tests for kernel regression with an application to option implied volatilities," Journal of Econometrics, 105, 363-412.

Altonji, J. G. and R. L. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Econometrica, 73, 1053-1102.

Anton, A. A., A. Fernandez Sainz, and J. Rodriguez-Poo, (2002), "Semiparametric Estimation of a Duration Model," Oxford Bulletin of Economics and Statistics, 63, 517-533.

Berry, S. T., and P. A. Haile (2009a), "Identification in Differentiated Products Markets Using Market Level Data," Unpublished Manuscript.

Berry, S. T., and P. A. Haile (2009b), "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," Unpublished Manuscript.

Blundell R. and J. L. Powell (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models," in Dewatripont, M., L.P. Hansen, and S.J. Turnovsky, eds., Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Vol. II (Cambridge University Press).

Blundell, R. W. and J. L. Powell, (2004), "Endogeneity in Semiparametric Binary Response Models," Review of Economic Studies, 71, 655-679.

Blundell, R. W., and Smith, R. J. (1989), "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models", Review of Economic Studies, 56, 37-58.

Briesch, R., P. Chintagunta, and R.L. Matzkin (2009) "Nonparametric Discrete Choice Models with Unobserved Heterogeneity," Journal of Business and Economic Statistics, forthcoming.

Chesher, A. (2009), "Excess heterogeneity, endogeneity and index restrictions," Journal of Econometrics, 152, 37-45.

Chesher, A. (2010), "Instrumental Variable Models for Discrete Outcomes," Econometrica, 78, 575-601.

Cogneau, D. and E. Maurin (2002), "Parental Income and School Attendance in a Low-Income Country: A Semiparametric Analysis," Unpublished Manuscript.

Dong, Y. (2010), "Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration," forthcoming, Economics Letters

Fox, J. and C. Yang (2012), "Unobserved Heterogeneity in Matching Games," unpublished manuscript.

Goux, D. and E. Maurin (2005), "The effect of overcrowded housing on children’s performance at school, Journal of Public Economics, 89, 797-819.

Greene, W. H. (2008), Econometric Analysis, 6th edition, Prentice Hall.

Heckman, J. J., (1976) “Simultaneous Equation Models with both Continuous and Discrete Endogenous Variables With and Without Structural Shift in the Equations,” in Steven Goldfeld and Richard Quandt (Eds.), Studies in Nonlinear Estimation, Ballinger.

Heckman, J. J., and R. Robb, (1985) “Alternative Methods for Estimating the Impact of Interventions,” in James J. Heckman and Burton Singer (Eds.), Longitudinal Analysis of Labor Market Data, Cambridge:Cambridge University Press.

Heckman, J. J. (1990), "Varieties of selection bias, American Economic Review 80, 313–318.

Heckman, J. J. and Navarro, S. (2007), "Dynamic discrete choice and dynamic treatment effects," Journal of Econometrics, 136, 341-396.

Hill, J. B. and E. Renault (2010), "Generalized Method of Moments with Tail Trimming," unpublished manuscript.

Hirano, K., G. W. Imbens and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Econometrica, 71, 1161-1189.

Hoderlein, S. (2009) "Endogenous semiparametric binary choice models with heteroscedasticity," CeMMAP working papers CWP34/09.

Hong H. and E. Tamer (2003), "Endogenous binary choice model with median restrictions," Economics Letters 80, 219–225.

Horowitz, J. L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," Econometrica, 60, 505-532.

Honore, B. and A. Lewbel, (2002) "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," Econometrica, 70, 2053-2063.

Ichimura, H., and S. Lee (2006): "Characterization of the Asymptotic Distribution of Semiparametric M-estimators," CeMMAP working papers, CWP15/06.

Imbens, G. W. and Newey, W. K. (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," Econometrica, 77, 1481–1512.

Jacho-Chávez, D. T., (2009), "Efficiency Bounds For Semiparametric Estimation Of Inverse Conditional-Density-Weighted Functions," Econometric Theory, 25, 847-855.

Kennen, J. and J. R. Walker (2011), "The Effect of Income on Individual Migration Decisions," Econometrica, 79, 211-251.

Khan, S. and A. Lewbel (2007) "Weighted and Two Stage Least Squares Estimation of Semiparametric Truncated Regression Models," Econometric Theory, 23, 309-347.

Khan, S. and E. Tamer (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Econometrica, 78, 2021–2042.

Khan, S. and D. Nekipelov (2010a), "Semiparametric Efficiency in Irregularly Identified Models," unpublished working paper.

Khan, S. and D. Nekipelov (2010b), "Information Bounds for Discrete Triangular Systems," unpublished working paper.

Khan, S. and D. Nekipelov (2011), "Information Structure and Statistical Information in Discrete Response Models," unpublished working paper.

Lewbel, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," Econometric Theory, 13, 32-51.

Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors," Econometrica, 66, 105–121.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," Journal of Econometrics, 97, 145-177.

Lewbel, A. (2007a), "Endogenous Selection or Treatment Model Estimation," Journal of Econometrics, 141, 777-806.

Lewbel, A. (2007b), "Modeling Heterogeneity," in Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress (Econometric Society Monographs), Richard Blundell, Whitney K. Newey, and Torsten Persson, editors, Cambridge: Cambridge University Press, Vol. III, Chapter 5, 111-121.

Lewbel, A. (2007c), "Coherence and Completeness of Structural Models Containing a Dummy Endogenous Variable," International Economic Review, 48, 1379-1392.

Lewbel, A., Dong, Y., and T. Yang (2012), "Why and How to Avoid the Linear Probability Model, and a Simple Alternative," unpublished manuscript, Boston College.

Lewbel, A. and S. Schennach (2007), "A Simple Ordered Data Estimator for Inverse Density Weighted Functions," Journal of Econometrics, 186, 189-211.

Lewbel, A. and X. Tang (2011), "Identification and Estimation of Games with Incomplete Information using Excluded Regressors," unpublished manuscript.

Lewbel, A., O. Linton, and D. McFadden (2011), "Estimating Features of a Distribution From Binomial Data," Journal of Econometrics, 162, 170-188.

Magnac, T. and E. Maurin (2007), "Identification and Information in Monotone Binary Models," Journal of Econometrics, 139, 76-104.

Magnac, T. and E. Maurin (2008), "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data, Review of Economic Studies, 75, 835-864.

Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", Journal of Econometrics, 3, 205-228.

Manski, C. F. (1985), "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," Journal of Econometrics, 27, 313-333.

Manski, C. F. (1988), "Identification of Binary Response Models," Journal of the American Statistical Association, 83, 729-738.

Manski, C. F. (2007), "Partial Identification of Counterfactual Choice Probabilities," International Economic Review, 48, 1393–1410.

Matzkin, R.L. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and The Binary Choice Models," Econometrica, 60, 239-270.

Matzkin, R.L. (1994) "Restrictions of Economic Theory in Nonparametric Methods," in Handbook of Econometrics, Vol. IV, R.F. Engel and D.L. McFadden, eds, Amsterdam: Elsevier, Ch. 42, 2524-2554.

Matzkin, R. (2007), "Heterogeneous Choice," in Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress (Econometric Society Monographs), Richard Blundell, Whitney K. Newey, and Torsten Persson, editors, Cambridge: Cambridge University Press, Vol. III, Chapter 4, 75-110.

Pistolesi, N. (2006), "The performance at school of young Americans, with individual and family endowments," unpublished manuscript.

Powell, J. L., J. H. Stock, and T. M. Stoker, (1989) "Semiparametric Estimation of Index Coefficients," Econometrica, 57, 1403-1430.

Rivers, D., and Q. H. Vuong (1988), "Limited information estimators and exogeneity tests for simultaneous probit models," Journal of Econometrics 39, 347–66.

Shaikh, A. and E. Vytlacil (2008), "Endogenous binary choice models with median restrictions: A comment," Economics Letters, 23-28.

Stewart, M. B. (2005), "A comparison of semiparametric estimators for the ordered response model," Computational Statistics and Data Analysis, 49, 555-573.

Tiwari, A. K., P. Mohnen, F. C. Palm, S. S. van der Loeff, (2007), "Financial Constraint and R&D Investment: Evidence from CIS," United Nations University, Maastricht Economic and social Research and training centre on Innovation and Technology (UNU-MERIT) Working Paper 011.

Vytlacil, E. and N. Yildiz (2007), "Dummy Endogenous Variables in Weakly Separable Models," Econometrica, 75, 757-779.

White, H. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 48, 817-838.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data, 2nd edition, MIT press.

6 Appendix

LEMMA 1: Assume the distribution function of V given S is continuous and strictly monotonically increasing. Then there exists a function g and a random variable U such that $V = g(U, S)$ where $U \perp S$.

PROOF OF LEMMA 1: Define $U = F_{V|S}(V \mid S)$ where $F_{V|S}$ is the conditional distribution function V given S. Define g to be the inverse of the function $F_{V|S}$, so g is defined by $V = g\left[F_{V|S}(V \mid S), S\right]$. Then by construction $V = g(U, S)$ where $U \perp S$. In this construction U will have a uniform distribution but one could more generally let $U = \widetilde{f}\left(F_{V|S}(V \mid S)\right)$ for any strictly monotonic $\widetilde{f}$ to give U some other distribution, such as a normal.

Lemma 1 is not new, e.g., it is used in Vytlacil and Yildiz (2007) and Matzkin (2007). It is useful here because Theorem 1 below assumes existence of g and U with U independent of S, and the lemma shows that this assumption is made without loss of generality. The variable U can be interpreted as the error term in a model g for the variable V. In the main text we propose some simple functional forms for g.

Theorem 1 below generalizes Lewbel (2000), showing how to construct a variable T having the property that $E\left(ZT\right)=E\left(ZX'\right)\beta$, so a linear two stage least squares regression of T on X using instruments Z yields the desired coefficients β. Note this also proves point identification of β.

THEOREM 1: Assume $D=I(X'\beta+V+\varepsilon\geq 0)$, $E(Z\varepsilon)=0$, and $V=g\left(U,S\right)$. Assume $supp(X'\beta+\varepsilon)\subseteq supp(-V)$, $E\left(V\right)=0$, g is differentiable and strictly monotonically increasing in its first element, $U\perp(S,\varepsilon)$, and U is continuously distributed. Let $f(U)$ be the probability density function of U. Let $M(V)$ be any mean zero distribution function on $supp(V)$ such that $M(v_0)=0$ and $M(v_1)=1$ for some points v_0 and v_1 that are in the interior of $supp(V)$.

Define T by

$$T=\frac{D-M(V)}{f(U)}\frac{\partial g(U,S)}{\partial U} \tag{6}$$

Then $T=X'\beta+\widetilde{\varepsilon}$ where $E\left(Z\widetilde{\varepsilon}\right)=0$.

PROOF OF THEOREM 1: Define $D^*=X'\beta+\varepsilon$ so $D=I(D^*+V\geq 0)$. We first prove the theorem taking $M\left(V\right)=I(V\geq 0)$. By the definition of conditional expectation

$$\begin{aligned} E(T \quad | \quad S,\varepsilon) &= \int_{supp(U|S,\varepsilon)}\frac{I(D^*+g(U,S)\geq 0)-I(g(U,S)\geq 0)}{f(U)}\frac{\partial g(U,S)}{\partial U}f(U\mid S,\varepsilon)dU \\ &= \int_{supp(U|R)}\left[I(D^*+g(U,S)\geq 0)-I(g(U,S)\geq 0)\right]\frac{\partial g(U,S)}{\partial U}dU \\ &= \int_{supp(V|R)}\left[I(D^*+V\geq 0)-I(V\geq 0)\right]dV \end{aligned}$$

where the second equality follows from $U\perp(S,\varepsilon)$ which means $f(U)=f(U\mid S,\varepsilon)$, and the third equality uses a change of variables from U to V. If $D^*\geq 0$ then

$$E(T\mid S,\varepsilon)=\int_{supp(V|R)}I(-D^*\leq V\leq 0)dV=\int_{-D^*}^{0}1dV=D^*$$

and if $D^*\leq 0$ then

$$E(T\mid S,\varepsilon)=\int_{supp(V|R)}-I(0\leq V\leq -D^*)dV=-\int_{0}^{-D^*}1dV=D^*$$

This proves that $E(T\mid S,\varepsilon)=X'\beta+\varepsilon$. Defining $\widetilde{\varepsilon}=T-X'\beta$ we have

$$\begin{aligned} E(Z\widetilde{\varepsilon}) &= E[Z(T-X'\beta)]=E[E(Z(T-X'\beta)\mid S,\varepsilon)] \\ &= E[Z(E(T\mid S,\varepsilon)-X'\beta)]=E(Z\varepsilon)=0. \end{aligned}$$

To show the theorem holds for other choices of $M(V)$, replace $D-M(V)$ in equation (6) with $[D - I(V \geq 0)]+$ $[I(V \geq 0) - M(V)]$. Then $E(T \mid S, \varepsilon)$ equals the sum of the term given above and $\int_{supp(V|R)} [I(V \geq 0) - M(V)]\, dV$
Applying an integration by parts to this term gives

$$[I(V \geq 0) - M(V)]\, V|_{supp(V|R)} - \int_{supp(V|R)} -\frac{\partial M(V)}{\partial V} V dV$$

The first term here is zero because $M(V)$ is distribution function that equals zero and one strictly inside the support of V, and the second term is zero because $M(V)$ is a mean zero distribution function. So $E(T \mid S, \varepsilon)$ is unchanged by replacing $I(V \geq 0)$ with $M(V)$.

One way in which Theorem 1 generalizes previous results is that Lewbel (2000) used $I\,(V \geq 0)$ in place of $M(V)$. and we will usually let $M(V) = I\,(V \geq 0)$. The usefulness of this extension, first proposed by Lewbel and Tang (2012), is that taking $M(V)$ to be a differentiable function can simplify some limiting distribution theory.

Recall that S consists of all the elements of X and Z. As long as V given S is continuously distributed, the assumption that $V = g\,(U, S)$ with U independent of S holds without loss of generality. This is because, as shown by Lemma 1, it is always possible to construct a function g and an error term U that satisfies this independence assumption. Differentiability of g and continuity of the U distribution both correspond to smoothness of the distribution function of V. Having $E\,(V) = 0$ is not really necessary, but it simplifies T. Setting the median of V to zero would have the same effect. In practice one could simply recenter V (by demeaning or subtracting off the median) before using it in the model to make this hold. Note that X and Z will generally include a constant term, so recentering V will have no impact on the model.

Having $E(Z\varepsilon) = 0$ and rank of $E\left(ZX'\right)$ equal to the number of elements of β are just the minimal conditions that would be required for two stage least squares estimation of a linear model with endogenous regressors, so we maintain those minimal conditions in our nonlinear binary choice model.

The requirement that U be independent of ε is nothing more than an exogeneity assumption regarding the special regressor V. It says that after one has conditioned on other covariates, the remaining variation in V is unrelated to the binary choice model error ε.

Finally, 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 D to zero or one. In the case where the support of $X'\beta + \varepsilon$ is not bounded, this becomes an identification at infinity argument as in Heckman (1990), though it should be noted that consistent estimation of any moment, even a mean, requires observing data over their entire support, and Khan and Tamer (2010) point out that similar requirements apply to standard average treatment effect estimators.

The required support assumption is not in general testable prior to estimation, because it depends on β. After estimation of $\widehat{\beta}$ one can check whether the values of $X'\widehat{\beta}$ in the data lie in the range of observed values of $-V$, but even then, the true supports of the regressors and the support of the latent ε are not known in general. So one may worry about the support condition holding in empirical applications, to which there are a few responses.

First, in theory the support condition is easily satisfied, since e.g., it holds if V contains an additive component like an error term that is normal, t-distributed, or has any other full real line support distribution.

Second, as described earlier, the large support assumption can be relaxed and replaced with a tail symmetry assumption. See Magnac and Maurin (2007) for details. The construction of T and the conclusion of Theorem 1 is unchanged when the support of V is not as large as that of $X'\beta + \varepsilon$, provided that this tail symmetry condition holds. Moreover, even if tail symmetry does not hold exactly, as described earlier the asymptotic bias in estimation resulting from a violation of the large support assumption will generally be small if the tails of the distribution of ε are either thin or close to symmetric.

Third, Lewbel (2000, 2007a) shows that for special regressor based estimators, the finite sample bias in estimation of $\widehat{\beta}$ also tend to be small when the variance or interquantile ranges of V are comparable to or larger than the variance or interquantile ranges of $X'\beta + \varepsilon$. This makes intuitive sense, since in real data what matters is not the hypothetical extreme values that might possibly be seen, but rather the spread of values actually observed in the majority of the sample. Thus in practice one may check measures of the relative spread of the distributions of V versus $X'\widehat{\beta}$ to get a sense of whether the observed variation in V is likely to be large enough to provide reasonably accurate estimates.