Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors

Arthur Lewbel, Yingying Dong, and Thomas Tao Yang
Boston College, University of California Irvine, and Boston College

revised May 2012

Abstract

We discuss the relative advantages and disadvantages of four types of convenient estimators of binary choice models when regressors may be endogenous or mismeasured, or when errors are likely to be heteroskedastic. For example, such models arise when treatment is not randomly assigned and outcomes are binary. The estimators we compare are the two stage least squares linear probability model, maximum likelihood estimation, control function estimators, and special regressor methods. We specifically focus on models and associated estimators that are easy to implement. Also, for calculating choice probabilities and regressor marginal effects, we propose the average index function (AIF), which, unlike the average structural function (ASF), is always easy to estimate.

Keywords: Binary choice, Binomial Response, Endogeneity, Latent Variable Model, Measurement Error, Heteroskedasticity, Discrete Endogenous, Special Regressor, Control Functions, Linear Probability Model, Random Coefficients. JEL codes: C25, C26. The authors would like to thank Jeff Wooldridge, David Green, and anonymous referees for helpful discussions. Corresponding Author: Arthur Lewbel, Department 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

A common empirical problem is estimation of binary choice (binomial response) models when one or more regressors are endogenous or mismeasured, or when errors are heteroskedastic. For example, this problem arises in estimation of treatment effects when treatment is not randomly assigned and outcomes are binary. These models are often estimated using linear two stage least squares, despite having a binary dependent variable. This corresponds to estimation of a linear probability model (LPM) with instrumental variables.

In this paper we discuss the relative advantages and disadvantages of four different kinds of convenient estimators that can be applied to these types of models. These are linear probability model estimators, maximum likelihood estimation, control function based estimation, and special regressor methods. Each of these kinds of estimators has both advantages and drawbacks, with some of the latter being rarely recognized or acknowledged. Our discussion includes a previously unrecognized drawback of linear probability model based estimators, which is that they can provide a negative average treatment effect estimate even if the true treatment effect is nonnegative for every individual in the sample.

We specifically focus here on practical, easy to apply estimators, rather than on theoretical econometric generality. We therefore do not consider estimators that employ high dimensional nonparametric components, or entail complications like inequality constraints, estimation of bounds, difficult numerical searches, etc. In all of the models we consider, regressors X and the model error ε are assumed to appear in the standard linear form $X'\beta + \varepsilon$.

After comparing estimation methods, we propose an "Average Index Function" for calculating choice probabilities and regressor marginal effects. The average index function is an alternative to other choice probability measures like the average structural function and the propensity score, which has the advantage of always being easy to construct and estimate.

Linear probability model estimation, maximum likelihood, and control functions are commonly used estimators. We also discuss a fourth, less well known alternative: special regressor estimation. A companion paper to this one, Dong and Lewbel (2012), provides more details regarding special regressor estimation, including extremely simple (both mathematically and numerically) implementations of the method. One such simple special regressor estimator is summarized later in section 6. Even with multiple endogenous regressors, this estimator requires nothing more than a few linear regressions to implement, and so is on par with the linear probability model for ease of use. As we discuss later, the special regressor method differs substantially from others in both virtues and drawbacks. That feature, and the simplicity with which it can be implemented, suggests that special regressor methods may be particularly useful for providing robustness checks of results against more standard alternative estimators.

We also propose an alternative to Blundell and Powell's (2003, 2004) average structural function (ASF) measure for defining choice probabilities and marginal effects given estimates of a binary choice model with endogenous regressors and hetereoskedastic errors. A virtue of our alternative, which we call the average index function (AIF), is that it is often easier to calculate than the ASF, and, like the ASF, it is equivalent to the standard propensity score estimator of choice probabilities and marginal effects when the model errors ε are independent of the regressors X.

In the next section we set up our notation. Sections 3, 4, and 5 then summarize and compare the data and model requirements for linear probability model, maximum likelihood, and control function based estimators. We do not describe the implementation of these alternatives in detail, since these can be found in standard textbooks such as Greene (2008, chapter 23.7) and Wooldridge (2010 chapter 15.7). Next, section 6 covers special regressor estimators, and section 7 discusses other types of estimators. Section 8 is about using the models to estimate choice probabilities and the marginal effects of regressors, including the ASF and our proposed AIF. We then conclude in section 9.

2 Binary Choice Models

Let D be an observed dummy, binary dependent variable, which equals either zero or one. Let X be a vector of observed regressors, which in a treatment model would include a treatment indicator variable T. Let β be a vector of coefficients to be estimated and let ε be an unobserved error. Define $I(\cdot)$ to be the indicator function that equals one if its argument $\cdot$ is true and zero otherwise. The standard binary choice model that would be estimated by control function or maximum likelihood methods is $D = I\left(X'\beta + \varepsilon \geq 0\right)$, meaning that D is one when $X'\beta + \varepsilon$ is positive and zero otherwise. This type of specification is known as a threshold crossing model, with D switching from zero to one when the latent variable $X'\beta + \varepsilon$ crosses the threshold zero. Special cases of the threshold crossing model are the probit and logit models, in which ε has a normal or logistic distribution, respectively. The initial goal is to estimate β, but ultimately we are interested in choice probabilities and the marginal effects of X, looking at the probability that D equals one for any given value of X, and how that probability changes when X changes.

For any dummy variable D, the probability that D is one given X equals $E(D|X)$. Let $F_{-\varepsilon}$ denote the probability distribution function of $-\varepsilon$. In the model $D = I\left(X'\beta + \varepsilon \geq 0\right)$, if X is independent of ε then the probability that D equals one given X is $F_{-\varepsilon}\left(X'\beta\right)$, which is the probability that $-\varepsilon \leq X'\beta$. However, when regressors are endogenous (or when ε is heteroskedastic), then the probability that D equals one will depend on the conditional distribution of ε given X rather than on the marginal distribution of ε, but in this case the marginal distribution $F_{-\varepsilon}\left(X'\beta\right)$ is still often used as a summary measure of the choice probability, and is an example of what is called the average structural function. See, e.g., Blundell and Powell (2003, 2004).

Instead of a threshold crossing model, the linear probability model assumes that $D = X'\beta + \varepsilon$ where ε has mean zero. Note that β equals the marginal effect of each coefficient in the LPM, while β is only proportional to marginal effects in threshold crossing models. In the LPM with all exogenous regressors $E(D|X) = X'\beta$, and so in that model $X'\beta$ equals the probability that D equals one given X. This probability is linear in X, hence the name LPM. In the LPM with endogenous regressors, $X'\beta$ equals the average structural function.

Suppose now that some elements of X (including possibly treatment indicators) 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 may have an unknown distribution. Let X^e denote the vector of endogenous regressors, and let X^o be the vector of all the other regressors, which are exogenous. Let Z be a vector of observed exogenous covariates to be used as instruments. Typically, all of the elements of X^o would be included in Z. The threshold crossing model is then $D = I\left(X^{e\prime}\beta_e + X^{o\prime}\beta_o + \varepsilon \geq 0\right)$, while the linear probability model assumes $D = X^{e\prime}\beta_e + X^{o\prime}\beta_o + \varepsilon$.

To complete the model, let $G\left(X^e, Z, e\right) = 0$ describe the relationship between X^e and Z, where e is a vector of errors (which may not be known to the econometrician) and G is a vector valued function generally containing as many elements as the vector X^e. The problem of endogeneity is then that e and ε are not (after conditioning on Z if necessary) independent. We can think of $G\left(X^e, Z, e\right) = 0$ along with $D = I\left(X'\beta + \varepsilon \geq 0\right)$ or $D = X'\beta + \varepsilon$ as a block triangular system of equations. Alternatively, if a fully simultaneous system was specified in which X^e was some function of D, Z, and errors, then $G\left(X^e, Z, e\right) = 0$ will be what remains after substituting $D = I\left(X'\beta + \varepsilon \geq 0\right)$ or $D = X'\beta + \varepsilon$ in for D in that function. Note however, that a fully simultaneous system containing a binary variable D often causes problems of coherency and com-

pleteness (see, e.g., Lewbel 2007b and reference therein) that are outside the scope of this paper. The focus of this paper will be on estimation rather than identification, so we will assume that the β parameters are identified. Typically, identification can be assured by exclusion restrictions, that is, the existence of elements of Z that are not in X.

A crucial issue that arises in the comparison of estimators is the restrictions that are placed on X^e and on the model $G(X^e, Z, e)$ (that is, how X^e relates to Z) and what information the econometrician is assumed to know about G and e. A rough summary is that the LPM and special regressor impose the weakest conditions on X^e and Z, in that they allow X^e to be continuous, discrete, limited, etc., and, regarding Z, only require the linear two stage least squares assumptions that $E(Z\varepsilon) = 0$ and that $rank\left(Z'X^e\right)$ equal the number of elements of X^e (the LPM in addition imposes constraints on how one of the exogenous regressors affects G). Maximum likelihood also allows X^e to be continuous, discrete, limited, etc., but requires that the model G (and hence the relationship of X^e to Z) and the joint distribution of e and ε be fully parameterized and correctly specified. Control function estimators generally require X^e to be continuously distributed, ruling out discrete endogenous regressors, and also requires that G be correctly (though not necessarily parametrically) specified, in the sense that the error vector e is required to satisfy certain properties.

3 The Linear Probability Model

Consider first estimation based on the linear probability model. The LPM is estimated by linearly regressing D on X, by ordinary least squares if the regressors are all exogenous (i.e., if $X = X^o$), or by linear two stage least squares with instruments Z if some regressors are endogenous. The LPM therefore assumes that $D = X'\beta + \varepsilon$ and either that $E(X\varepsilon) = 0$ in the exogenous case, or $E(Z\varepsilon) = 0$ when some regressors are endogenous.

One flaw in the LPM is that the error ε in the LPM cannot be independent of any regressors, even exogenous ones (unless X itself consists of only a single binary regressor). This is because for each possible realization of the regressors X, the error ε must equal either $1 - X'\beta$ or $-X'\beta$, which are functions of all the elements of X. It's difficult to see any way of constructing a plausible behavioral economic model that can justify this forced dependence of ε on all elements of X while at the same time satisfying the required uncorrelatedness assumption that either $E(X\varepsilon) = 0$ or $E(Z\varepsilon) = 0$.

Another problem with the LPM is that fitted probabilities should look like a distribution function (typically S shaped), as in the threshold crossing model where the average structural function is $F_{-\varepsilon}\left(X'\beta\right)$. In contrast, the fitted probabilities in the LPM are estimates of $X'\beta$. This yields the most commonly recognized drawback of the LPM, that $X'\beta$, and hence fitted probabilities, can go below zero or above one.

Formally, the linear probability model requires that no regressor have infinite support, e.g., no element of X is permitted to have a normal distribution (or any distribution that extends to plus or minus infinity), because otherwise the fitted probability $X'\beta$ will be below zero or above one for some observable value of X. More generally, the LPM requires that any regressor that can take on a large range of values must have a very small coefficient, or again some probabilities implied by the model will take on impossible values.

The usual counter argument to the problem of probabilities going below zero or above one is to claim that the LPM is only intended to approximate true probabilities for a limited range of X

values, and that the LPM should provide good approximations to marginal effects. However, there at least three objections that can made to this defense of the LPM.

First, essentially all models are approximations, so if one is willing to consider the LPM, then one should be at least as generous in considering alternative estimators when their assumptions are also unlikely or impossible to hold perfectly. For example, taken purely as an approximation to an unknown true model, we don't know of any evidence that probit estimation ignoring endogeneity is generically any worse or better then the LPM estimated by two stage least squares, and at least the former always lies between zero and one. Similarly, the special regressor method we describe later imposes strong support assumptions on one regressor. It would be illogical to argue against the special regressor method as an approximation because it formally requires very strong support restrictions on one regressor, while supporting a method like the LPM that formally imposes strong support restrictions on every regressor.

The second objection to the claim that the LPM is justified as an approximation is to compare a straight line approximation to the S shape of most actual distribution functions. The straight line will depart greatly from the S shaped function long before the line actually crosses the zero and one boundaries. That is, the LPM is likely to deliver fitted probabilites that are too extreme (too close to zero and one) even for moderate ranges of $X'\beta$ values that don't actually go outside the zero to one range.

The third objection concerns the common claim that LPM based estimates provides good approximations to true marginal effects, including treatment effects, even if they do have problems fitting choice probabilities. For example, Angrist and Pischke (2009, p. 107) provide an empirical application in which the estimated marginal effect of a binary treatment indicator on a binary outcome is almost the same when estimated either by a probit or by a linear probability model estimator. They then conclude that, "while a nonlinear model may fit the CEF (conditional expectation function) for LDV's (limited dependent variable models) more closely than a linear model, when it comes to marginal effects, this probably matters little. This optimistic conclusion is not a theorem, but, as in the empirical example here, it seems to be fairly robustly true."

In contrast to this claim, consider the following example. Let T be a treatment indicator (equalling one for treated individuals and zero for untreated) and let R be another covariate, so X is the vector of elements T and R. Now suppose individual's outcomes D are given by the probit specification $D = I\left(1 + T + R + \varepsilon \geq 0\right)$ with normal errors ε that are independent of the regressors, so in this example there is no endogeneity to deal with. Assume that the errors ε have mean zero and very low variance, perhaps making the standard deviation of ε be 0.01, so in this model an individual's outcome can be predicted with a high degree accuracy just by seeing T and R. The treatment effect for an individual is defined as the difference in outcome between being treated and untreated, which is $I\left(2 + R + \varepsilon \geq 0\right) - I\left(1 + R + \varepsilon \geq 0\right) = I\left(-2 \leq R + \varepsilon \leq -1\right)$ for any given R and ε. In this model, by construction nobody can have a negative treatment effect, regardless of whatever value R or ε takes on.

Suppose we observe six individuals with the following covariate values: $R_1 = -1.8$, $R_2 = -0.9$, $R_3 = -0.92$, $R_4 = -2.1$, $R_5 = -1.92$, $R_6 = 10$, $T_1 = 0$, $T_2 = 0$, $T_3 = 0$, $T_4 = 1$, $T_5 = 1$, and $T_6 = 1$. With such small errors, this model will with very high probability generate the outcomes $D_1 = 0$, $D_2 = 1$, $D_3 = 1$, $D_4 = 0$, $D_5 = 1$, and $D_6 = 1$. In this data set half the people are treated, and the true treatment effect equals one for the fifth individual (who is treated in the sample) and zero for the others. The true average treatment effect for this sample is therefore $1/6$. However, if you use these data to estimate the linear probability model $D = \beta_0 + T\beta_1 + R\beta_2 + \varepsilon$

by ordinary least squares, the estimate of β_1 is -0.16, and the estimate of the marginal rate of substitution (MRS) β_1/β_2 is -3.2.[1,2] Not only does LPM estimation give the wrong sign here, but the estimated value is relatively large, i.e., the LPM β_1 estimate is about the same magnitude (with the wrong sign) as the true positive treatment effect of 1/6, and the MRS has the wrong sign and is over three times the size of the true MRS.

Examples like this one are possible because probit and other similar threshold crossing models make $E(D \mid X)$ highly nonlinear, and hence the derivatives of this function can be quite far from the derivatives of a linear approximation (i.e., the coefficients of the LPM), particularly when the regressors X may be correlated with each other [3]

This example shows that even in a trivial model (standard probit) with tiny errors, in which every single individual has an either positive or zero treatment effect, the LPM can still give the wrong sign of the treatment effect. Of course, this is a cooked example, but it illustrates the general point that linear regressions do not necessarily provide good estimates of marginal effects in highly nonlinear models like those of binary choice. While getting the wrong sign for every data point is an extreme example, more likely are applications where signs are correctly estimated but magnitudes may be far off.

The sole advantage of linear probability model estimators relative to other estimators is that all LPM requires for estimation is the assumption that $E(Z\varepsilon) = 0$ along with the standard linear model rank condition that $E\left(X'Z\right)$ have full rank. In particular, models for the endogenous regressors X^e (described earlier as $G(X^e, Z, e) = 0$) do not need to be specified, and elements of X^e can have any type of distribution, including discrete, continuous, continuous with mass points (like censored data), etc. The linear probability model also permits general forms of heteroskedasticity in the errors ε, e.g., the regressors can have random coefficients. The efficiency of LPM estimators cannot be directly compared to the efficiency of other estimators, because the specification of the LPM is incompatible with the threshold crossing model assumed by other estimators.

4 Maximum Likelihood

Maximum Likelihood assumes $D = I\left(X^{e\prime}\beta_e + X^{o\prime}\beta_o + \varepsilon \geq 0\right)$ and $G(X^e, Z, e, \theta) = 0$, where we have now added a parameter vector θ to the model for X^e. For maximum likelihood estimation, the function G is assumed to be known to the econometrician up to a parameter vector θ, and

[1]The standard error on β_1 in this LPM will be large because the sample size is tiny, but it would be simple to scale up this example with more observations to generate a negative β_1 estimate with any desired significance level.

[2]For comparison, if one estimated the coefficients β_0, β_1, and β_2 in the correctly specified probit model $D = I\left(\beta_0 + T\beta_1 + R\beta_2 + \varepsilon \geq 0\right)$ using this data set, then because of the small sample size the likelihood function will be maximized at a range of values for the coefficients. Every beta vector in this range will perfectly predict the observed outcomes D. But every beta vector in this range has β_1 and the MRS β_1/β_2 positive, so the sign of the treatment effect will be correctly estimated. In particular, this range has $0.12 \leq \beta_1/\beta_2 \leq 1.18$ and $0.10s \leq \beta_1 \leq 1.14s$ where s is an arbitrary positive scale normalization that applies to all of the coefficients. Standard probit chooses the scale s to make $s^2 = var(\varepsilon) = 1$, but with a perfect fit $var(\varepsilon)$ is not identified, and typical numerical implementations of the probit model will just arrive at one point in the set of optimizing values. For example, with this data the probit command in Stata chooses a value of s around a hundred and estimates $\beta_1/\beta_2 = 1.03$. Of course, probit is also consistent for this specification, unlike the LPM. The same math shows that, with this data, other threshold crossing model estimators such as logit or maximum score, will also give the proper sign of the treatment effect and the MRS.

[3]In this example, both the lowest and highest values of R are associated with $T = 1$.

the joint distribution of ε and e, conditional on Z, is also fully specified up to some vector of parameters. All these parameters are assumed to be jointly identified.

For example, the biprobit command in Stata estimates this maximum likelihood model when X^e is a single binary endogenous regressor and the model G is probit with e and ε jointly normal. Another example (which we will discuss more later) is the maximum likelihood option in Stata's ivprobit command, which estimates this model when X^e is a single continuously distributed regressor and the model G is linear with e and ε jointly normal.

Like linear probability model estimators, maximum likelihood permits endogenous regressors X^e to be continuous, discrete, limited, etc., as long as an appropriate parametric model G for each can be specified. Maximum likelihood also requires that the joint distribution of ε and e conditional on Z be fully parameterized and correctly specified. Maximum likelihood permits general forms of heteroskedasticity, but the associated error variances need to be fully and correctly parameterized. Maximum likelihood is generally more efficient than other estimators, because it makes stronger specification assumptions, both requiring and using more information than the alternatives.

In our binary choice framework, maximum likelihood also requires not just any set of instruments Z, but one specific complete set of instruments. This is because, when the model is correctly specified, dropping any element of Z will then cause misspecification of G. As a result, maximum likelihood usually becomes inconsistent if any element of Z is omitted[4]. This is in contrast to linear model two stage least squares, which only loses efficiency but not consistency when some elements of Z are dropped from the estimation (as is commonly done with variables that one is not certain are truly exogenous and hence might not be valid instruments).

In addition to the difficulty of correctly specifying the complete set of instruments Z, the vector valued function $G\left(X^e, Z, e, \theta\right)$, and the joint distribution of ε and the vector e, maximum likelihood also often suffers from numerical difficulties associated with the estimation of nuisance parameters. For example, the covariances between e and the latent ε might not be strongly identified, resulting in a likelihood function with ridges, multiple local maxima, etc. This is particularly likely when there are multiple endogenous regressors, or when one or more of the endogenous regressors is itself discrete, censored, or otherwise limited, making both e and ε latent.

5 Control Functions

There are a number of variants of control function methods, which can be found in standard textbooks such as Greene (2008) or Wooldridge (2010). The control function methodology traces 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 (2004) estimator with multiple nonparametric components.

Given a general model $D = M\left(X, \beta, \varepsilon\right)$, one way to describe the control function method is to assume first that the model $G\left(X^e, Z, e\right) = 0$ is parametrically or nonparametrically specified and so can be estimated, second that the model $G\left(X^e, Z, e\right)$ can be solved for the vector of errors

[4]We say "usually" here because sometimes misspecified likelihood functions can still yield consistent estimates, as in quasi-maximum likelihood estimation. Examples are completely linear models with normal errors, yielding estimates that are asymptotically equivalent to linear three stage least squares, which like two stage least squares only loses efficiency but not consistency when some instruments are dropped from the model.

e, and third that there exists a function h and a "clean" error U such that $\varepsilon = h(e, U)$ and U is independent of both X and e.

To implement the control function estimator, one would then first estimate the function $G(X^e, Z, e)$ and get fitted values of the errors e. Then plugging h into the D model gives $D = M[X, \beta, h(e, U)] = \widetilde{M}(X, e, \beta, U)$. Treating e as if it was a vector of additional regressors, the error term in the model $\widetilde{M}$ is U, which is independent of the regressors X and e. As a result, the model $\widetilde{M}$ no longer has an endogeneity problem, and so can be estimated in some standard way in place of the original model M.

For example, suppose we have the threshold crossing model $D = I\left(X^e\beta_e + X^{o\prime}\beta_o + \varepsilon \geq 0\right)$ where X^e is a scalar, suppose the function G is given by the linear model $X^e - Z'\alpha - e = 0$, and suppose ε and e are jointly normal with mean zero. A property of normals is that they can be linearly decomposed as $\varepsilon = e\lambda + U$ where U is independent of e and the constant λ depends on the covariance matrix of (e, ε). In this case, a control function estimator would consist of first linearly regressing X^e on Z. The residuals from that equation are then estimates of e. Plugging $\varepsilon = \lambda e + U$ into the original model gives $D = I\left(X^{e\prime}\beta_e + X^{o\prime}\beta_o + e\lambda + U \geq 0\right)$, which is just an ordinary probit model with independent normal error U after including e in the model as an additional regressor along with X^e and X^o (and estimating λ as another coefficient in addition to β_e and β_o). The model described in this paragraph is the model that is estimated by the ivprobit command in Stata.[5] Despite its name, ivprobit is actually a control function estimator, not an instrumental variables estimator, and this has important implications discussed below.

Control function estimators are more general than maximum likelihood because they can have the first stage G function be semiparametrically or nonparametrically identified and estimated, and they often do not require fully parameterizing the joint distribution of ε and e. However, two key requirements underlie control function estimation. First, we need to be able to solve the estimated first stage vector of G equations for their errors e, and second, we require that including the errors e in the model for D fixes the endogeneity problem. Another way of saying the second point is this: control functions assume the only way in which the endogenous regressors X^e relate to the model error ε (i.e., the entire source of any endogeneity problems) is through the first stage errors e.

One substantial limitation of control function methods for binary choice models is that they generally require the endogenous covariates X^e to be continuous, and so can typically not be used when the endogenous regressors X^e are discrete, censored, or otherwise noncontinuously distributed. This is because control function estimation requires the ability to solve for the error term e in $G(X^e, Z, e) = 0$. So for example if X^e is binary and G is a probit or similar threshold crosssing model making $G(X^e, Z, e) = X^e - I\left(Z'\gamma + e \geq 0\right) = 0$, then one would not be able to solve for the latent e given observations of X^e and Z.

For any X^e one could always define a model G by $G(X^e, Z, e) = E(X^e \mid Z) + e$ and estimate e as the residual from a nonparametric regression of X^e on Z. However, the fact that X^e is discrete would mean that e defined in this way is not independent of Z, which will generally cause violations of the required assumptions regarding U. [6]

[5] Since the errors are normal in this example, the model could have been estimated either by the two step procedure described here, or maximizing the likelihood function associated with the system of equations for D and X^e. Stata provides both options.

[6] This issue about control functions not working when endogenous regressors are discrete or limited is specific to binary choice and other nonlinear models. The same is true about the later comment that control functions become

For example, the ivprobit command in Stata will usually provide inconsistent estimates if any of the endogenous regressors in the model are binary, discrete, censored, or otherwise not continuously distributed. This is because the errors e in the first stage regression $X^e = Z'\alpha + e$ cannot be normal and independent of the regressors, as the ivprobit command requires. The ivprobit command applied to a model with discrete, censored, or otherwise limited endogenous regressors will provide estimates, but, as in the LPM, the assumptions required for these estimates to be consistent will generally not hold.[7]

This illustrates a fundamental difference between control function models and true instrumental variable estimators like linear two stage least squares. Instrumental variable estimators only require that instruments Z be correlated with regressors X^e and uncorrelated with the model errors. Two stage least squares does not impose any structural assumptions on the errors in the first stage regression. In contrast, control function estimators require that the first stage regression errors satisfy some strong properties, and so in that sense control functions, like maximum likelihood estimation, require that the first stage model be correctly specified.

This gives rise to another limitation of control function estimators, which is that they, like maximum likelihood, require not just any set of instruments Z, but the exact right set of instruments. This is because if e in the model $G\left(X^e, Z, e\right) = 0$ satisfies the control function assumptions, then dropping any element of Z will change e, thereby generally causing a violation of the assumptions required for control function consistency. So, for example, in linear two stage least squares models one can omit an instrument that is suspected of being invalid, and the only problem that results is a reduction in efficiency. In contrast, with both maximum likelihood and control function estimation, if the model is correctly specified and hence consistent with a given set of instruments, then the estimates will generally become inconsistent, not just inefficient, when any instrument is dropped.[8]

Control functions require fewer modeling assumptions than maximum likelihood, so control function estimates will generally be less efficient than maximum likelihood. However, some control function estimators are parametrically or semiparametrically efficient relative to their given information set.

6 Special Regressor Estimators

To illustrate the simplicity and basic form of special regressor methods, an example is provided below, but interested readers are encouraged to see Dong and Lewbel (2012) and references therein for a general overview of special regressor methods and alternative ways in which they can be implemented.

Special regressor methods assume that the model includes a single regressor, call it V, that has the following properties. First the special regressor V is exogenous (conditionally independent of the model error ε) and shows up in the model additively to the error. In all of the models

inconsistent when instruments are dropped. These problems do not arise in linear regression models, where control function estimators become numerically equivalent to linear instrumental variables estimators.

[7]The inconsistency of ivprobit estimates with noncontinuous X^e holds regardless of whether it is applied using the two step or maximum likelihood options in Stata. Either way, ivprobit assumes that $X^e = Z'\gamma + e$ with e independent of Z and/or normal, which generally be violated when X^e is discrete or limited.

[8]The comments in footnote 5 apply here also.

considered in this paper, every regressor has this property of appearing additively to the error. Second, the special regressor V is continuously distributed, and has a large support, so it can take on a wide range of values. For example, any normally distributed regressor would automatically satisfy this continuous with large support condition. No matter how many endogenous regressors are in the model, only one special regressor that satisfies these properties is needed.

A third condition that is not strictly necessary, but is desirable for efficiency (and can affect rates of convergence), is that V have a thick tailed distribution. So, other things equal, if more than one regressor in the model satisfies the required conditions to be special, in general the one with the thickest tails (e.g., having the largest variance or interquartile range) will typically be the best choice of special regressor.

The binary choice special regressor model first proposed in Lewbel (2000) has the threshold crossing form $D = I\left(X^{e\prime}\beta_e + X^{o\prime}\beta_o + V + \varepsilon \geq 0\right)$, or equivalently $D = I\left(X'\beta + V + \varepsilon \geq 0\right)$. This is exactly the same basic form for D as the maximum likelihood and control function models. The only difference is that for ease of notation we have separated the special regressor V from the other exogenous regressors X^o, and we have normalized the coefficient of V to equal one. Assuming the effect of V on D is positive, which can be tested, this is a completely free normalization in binary choice models. Standard models like logit or probit instead usually normalize the variance of the error ε, so e.g. probit chooses ε to have variance one instead of choosing V to have a coefficient of one, but this makes no difference for the calculation of choice probabilities, marginal effects or other applications of the estimates.

Given a special regressor V, the only other requirements for special regressor estimation are identical to those required for linear two stage least squares, that is, we require a set of instruments Z having the property that $E\left(Z'\varepsilon\right) = 0$ and $E\left(Z'X\right)$ has full rank. Unlike other exogenous regressors, the special regressor V should not be included in Z.

6.1 An Example of Simple Special Regressor Estimation

Since special regressor methods are less well known than the previously discussed alternatives, in this subsection we provide one specific example of a numerically simple special regressor estimator. See Dong and Lewbel (2012) and references therein for details on this and other special regressor estimators.

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. Assume

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0, \quad E(V) = 0, \\ V &= S'b + U, \quad E(U) = 0, \quad U \perp (S, \varepsilon), \quad U \sim f(U) \end{aligned}$$

where $\perp$ denotes statistical independence and $f(U)$ is a mean zero density function having support $supp(U)$ that contains $supp(-S'b - X'\beta - \varepsilon)$. Note that $U \perp (S, \varepsilon)$ implies that $V = S'b + U$ is conditionally independent of ε, conditioning on S.

Define T by $T = [D - I(V \geq 0)]/f(U)$. Given these assumption, it can be shown that $T = X'\beta + \widetilde{\varepsilon}$ where $E\left(Z\widetilde{\varepsilon}\right) = 0$, and therefore that $E(ZT) = E\left(ZX'\right)\beta$, which can be solved for β assuming that $E\left(ZX'\right)$ has rank equal to the number of elements of X.

Based on this result, we have the following simple estimator. Assume we have data observations D_i, X_i, Z_i, and V_i for $i = 1, ..., n$. 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.

A Simple Special Regressor Estimator

Step 1. Assume V_i has mean zero (if not, 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. 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 $\widehat{U}_i$. For $i = 1, ..., n$, define $\widehat{f}_i = 2/\left[\left(\widehat{U}_i^+ - \widehat{U}_i^-\right) n\right]$.

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

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.

The construction of T involves dividing by a density function, which is what can generate extreme outliers in step 4. These outliers can greatly increase standard errors and slow the rate of convergence of the estimator (see, e.g., Khan and Tamer 2010). It is therefore advisable to look for and remove outliers when using special regressor estimators. Removing outliers, which formally corresponds to asymptotic trimming, can improve the mean squared error of the estimator by trading off bias for variance.

An alternative to step 2 is to estimate $\widehat{f}_i$ using a kernel density estimator. This is more efficient than the above estimator, but requires choosing a kernel and bandwidth. See Lewbel and Schennach (2007) for details.

Most other estimators like probit do not normalize the coefficient of one of the regressors to equal one, so for example probit would assume that $D = I\left(X'\gamma + V\alpha + \varepsilon \geq 0\right)$ where ε has variance one. So to directly compare the special regressor model to something like this probit, one would either need to compare estimated marginal effects from the two models (using, e.g., the AIF discussed later), or compare the special regressor based coefficient vector β to the scaled coefficients γ / α.

6.2 Special Regressor Pros and Cons

The main drawbacks of the special regressor method are the restrictions it imposes regarding V. Formally, V (or more generally, the error term U in any model for V) must be conditionally independent of ε, conditioning on Z and X. Even if V is exogenously determined, this conditional independence assumption could be violated because of the way in which V might affect other, endogenous regressors (we'd like to thank Jeffrey Wooldridge for pointing this out).

To illustrate this issue, consider the estimator from the previous subsection and write the unknown implicit model for X^e in terms of errors e and all the other variables in our model as a vector valued equation $\widetilde{M}\left(X^e, Z, V, U, \varepsilon, e\right) = 0$, noting that X^o is included in Z. The function G defined in previous sections of this paper could be either a reduced form or a conditional mean of $\widetilde{M}$. At this level of generality X^e is endogenous either because it depends directly on ε, or because e may be correlated with ε. The problem is that the model of the previous subsection requires $U \perp (S, \varepsilon)$. Since S includes X^e , this means that U shouldn't affect X^e either directly,

or indirectly through V. Given the simple model of the previous section, it suffices to assume that $\widetilde{M}(X^e, Z, V, U, \varepsilon, e) = M(X^e, Z, V - U, \varepsilon, e) = 0$ for some function M. Then the model for X^e becomes $M\left(X^e, Z, S'b, \varepsilon, e\right) = 0$, and so U doesn't affect X^e. Roughly, what these equations say is that if we divide the equation for V into a fitted value and a residual U, then it suffices that X^e only depend on V through the fitted value, not the residual. One virtue of special regressor estimation is that this function M describing the model for X^e does not need to be known to the econometrician. It does not need to be specified, identified, or estimated.

Dong and Lewbel (2012) provide an empirical example of a special regressor model where D is a worker's decision to migrate from one state to another, endogenous regressors X^e are income and a homeownership dummy, and the special regressor V is minus age. Both theory and empirical evidence suggests that V does in fact appear additively in the model as required (e.g., the present value of future wage gains from moving decline linearly with age), and age is both continuous and exogenous. In this example the regression of V on S does not have any structural interpretation, but the above restriction on M would assume that it is not age itself that affects income and homeownership, but rather just the fitted value of age regressed on all the covariates and instruments. This is not strictly required for validity of the special regressor assumptions, but it's difficult to make all the necessary conditions hold otherwise.

Special regressor estimation could also be used when the threshold crossing model errors ε just have heteroskedasticity of unknown form, instead of endogeneity. In this case the difficulties of the previous section do not arise. See, e.g., Lewbel, Linton, and McFadden (2010) for general results of that type.

Another limitation on V is that it needs to be continuously distributed after conditioning on the other regressors. In the model of the previous section, this means that U be must be continuously distributed. This restriction implies that we cannot include a term like V^2 in the model for D as an additional regressor, because if we did, then V^2 would be included in S, and if we then regressed V on S, we would be regressing V on V^2, and this regression would not have a continuous residual U. However, we could instead replace V and V^2 in the D model with some known, nonlinear transformation of V if necessary.

Another issue with special regressor estimation, as noted by Khan and Tamer (2010) and Dong and Lewbel (2012), is if the conditional distribution of V does not have either thick tails, or strictly larger support than ε, or satisfy a condition Magnac and Maurin (2007) call tail symmetry, then the semiparametric convergence rate of special regressor estimators can be slow. Correspondingly, Lewbel (2007a) reports in simulations that finite sample biases may decline rather slowly with sample size when the sample variance of V is not large relative to that of $X'\beta + \varepsilon$.

Apart from these restrictions on the one regressor V, the special regressor estimation method has none of the previously discussed drawbacks of maximum likelihood, control function, and linear probability model estimators. In particular, unlike linear probability model estimates, special regressor based choice probability estimates always stay in the range of zero to one and are consistent with economically sensible threshold crossing models, including nesting logit and probit models. Unlike maximum likelihood and control functions, the special regressor is a true instrumental variables estimator, and so does not require specification of the model $G(X^e, Z, e)$ or estimation of the control function errors e. Similarly, the special regressor method does not impose assumptions regarding the joint distribution of the first stage error e and the structural model error ε.

Special regressor estimators can use any valid set of instruments, given only the standard linear

instrumental variables assumptions that $E(Z\varepsilon) = 0$ and $E\left(X'Z\right)$ have full rank. Unlike control functions and maximum likelihood, dropping some candidate instruments only affects efficiency, not consistency of the special regressor estimator. Unlike maximum likelihood, special regressor models can be estimated without numerical searches. Unlike control function methods, the special regressor method can be used when the distribution of endogenous regressors X^e is discrete or otherwise limited. Unlike maximum likelihood, the specification of the special regressor method is the same regardless of whether X^e is continuous, discrete, censored, truncated or otherwise limited. The special regressor estimator permits general, unknown forms of heteroskedasticity in the model errors, e.g., all of the regressors (including the endogenous ones) other than V can have random coefficients.

Regarding efficiency, the special regressor imposes far fewer assumption on the distribution of error terms (in particular on the errors e in the X^e equations) than control functions or maximum likelihood, and so will in general be less efficient than these alternatives, when these alternatives are consistent. Therefore, we expect special regressor estimators to have larger standard errors and less precise estimates than other methods, when the other methods are valid. However, as noted above, if we have a special regressor V, then the special regressor method will be valid under more general conditions regarding the endogenous regressors X^e and instruments Z than the other methods.

Magnac and Maurin (2007) and Jacho-Chávez (2009) show that the special regressor estimator of Lewbel (2000) is semiparametrically efficient (relative to its given information set), when the distribution of V is thick tailed and nonparametrically estimated. So, given the weaker assumptions of the special regressor method, it produces estimates that are as efficient as possible.

7 Other Estimators

The above comparison of estimators focused on general methods that can handle a vector of endogenous regressors X^e. Vytlacil and Yildiz (2007) instead consider a model in which X^e is a single binary endogenous regressor. Their estimator can be interpreted as a mixture of control function and special regressor estimation. Roughly, their model estimates β_o by control function methods, then treats $X^{o\prime}\beta_o$ as a special regressor for the purpose of estimating a single scalar coefficient β_e. In their context, the support of this special regressor only needs to be larger than the support of $X^e\beta_e$, instead of larger than the support of $X^e\beta_e + \varepsilon$ as ordinary special regressor identification requires.

Maximum score based estimators like Manski (1975, 1985) and Horowitz (1992) deal with heteroskedasticity in ε, obtaining identification by assuming that the median of ε given X is zero. Hong and Tamer (2003) propose an extension of maximum score assumptions to handle endogeneity, by assuming that the conditional median of ε given intruments Z is zero. However, Shaikh and Vytlacil (2008) show that obtaining point identification with these assumptions imposes severe restrictions on β_e and on the permitted conditional distribution of X^e given Z. Maximum score methods use the .5 quantile, and estimators could be based on other quantiles as well. In particular, Chesher (2009) and Hoderlein (2009) propose quantile extensions of control functions to allow for both endogeneity and heteroskedasticity in ε.

All of the estimators described here and in the previous sections have been discussed in the context of models containing a single linear index $X^{e\prime}\beta_e + X^{o\prime}\beta_o$ with a separable additive error ε.

Most of these estimators have extensions that allow for nonlinear indices and nonseparable errors, and there is a rapidly growing literature dealing with these more general classes of models. A couple of examples among many are Altonji and Matzkin (2005) and Imbens and Newey (2009).

Another feature of the models and estimators summarized here is that they provide point identification, that is, assumptions are made that are strong enough to identify the coefficients β and other related parameters of interest. There exists a very extensive literature that imposes weaker assumptions, and thereby only obtains bounds on parameters, or more generally identifies sets of values that parameters of interest may lie in. Examples include Manski (1988, 2007), Magnac and Maurin (2008), and Chesher (2010).

Still other large related literatures that are not covered here are dynamic binary choice models, binary choice with panel data, multinomial and ordered choice, binary choice of strategies within games, and reduced form average treatment effect or program evaluation models with binary outcomes. On this last point, note that when treatment is not randomly assigned, the treatment indicator will often be an endogenous regressor. The typical estimator for these models is essentially linear instrumental variables (using an instrument derived from some natural or constructed experiment), and hence many of the objections and problems associated with linear probability model estimation also apply to standard estimators of average treatment effects on binary outcomes.

8 Choice Probabilities, Marginal Effects, and the Average Index Function

In this section we return to writing the model as $D = I(X'\beta + \varepsilon \geq 0)$. If special regressor estimation is used for estimation (or if any other estimator is used that normalizes the coefficient of a regressor to equal one), all of the formulas given below still hold, letting one of the elements of X be V and letting the corresponding coefficient (element of β) of that coefficient be one. The results given here can be applied regardless of what method is used to estimate the binary choice model.

Given the model $D = I(X'\beta + \varepsilon \geq 0)$, if ε is independent of the regressors X then the propensity score or choice probability is defined as the conditional probability that $D = 1$ conditioning on the regressors X, which equals $E(D \mid X)$. If $\varepsilon \perp X$ then $E(D \mid X) = E\left(D \mid X'\beta\right) = F_{-\varepsilon}\left(X'\beta\right)$ where $F_{-\varepsilon}$ is the marginal distribution function of $-\varepsilon$.

When some regressors are endogenous, or more generally when ε is not independent of X (e.g., when ε is heteroskedastic), then the appropriate definition of a choice probability is less obvious. In that case all three of the above expressions, $E(D \mid X)$, $E\left(D \mid X'\beta\right)$, and $F_{-\varepsilon}\left(X'\beta\right)$, which are identical in the case of independent errors, may differ from each other.

With endogeneity or heteroskedasticity, $E(D \mid X)$ still equals the propensity score. It has the disadvantage of not using the information given specifically by the index $X'\beta$, although one could write this propensity score as $F_{-\varepsilon|X}\left(X'\beta \mid X\right)$, where $F_{-\varepsilon|X}$ is the conditional distribution of $-\varepsilon$ given X. Another disadvantage is that estimation of $E(D \mid X)$ either requires parameterizing the distribution function $F_{-\varepsilon|X}$, or estimating a high dimensional nonparametric regression. Also, for a propensity score one might want to condition on the instruments Z as well, since when X is endogenous, Z will contain observed covariates that have additional explanatory power for D.

Blundell and Powell (2003, 2004) suggest using the average structural function (ASF) to sum-

marize choice probabilities. For the binary response model the ASF is given by $F_{-\varepsilon}\left(X'\beta\right)$, where $F_{-\varepsilon}$ is still the marginal distribution function of $-\varepsilon$ even though ε is now no longer independent of X. An advantage of the ASF is that it is based on the estimated index structure $X'\beta$, and it equals the propensity score when errors are independent of regressors. The ASF can therefore be given the economic interpretation of what the propensity score would have been if the errors had not been endogenous, that is, a sort of counterfactual propensity score. A disadvantage of the ASF is that, in addition to β, the ASF requires estimation of the marginal distribution function $F_{-\varepsilon}$, which may be difficult to recover depending on the exact form of dependence of ε on X, and because ε is latent and so cannot be directly observed.

We propose using the measure $E\left(D \mid X'\beta\right)$, which we call the average index function (AIF), to summarize choice probabilities. Like the ASF, the AIF is based on the estimated index $X'\beta$, and like the ASF, the AIF equals the propensity score when errors are independent of regressors. However, an advantage of the AIF over both the propensity score and the ASF is that (when some regressors are endogenous or errors are heteroskedastic), the AIF is usually easier to estimate, as a one dimensional nonparametric regression of D on $X'\beta$. The AIF can also be given the interpretation of a counterfactual propensity score, specifically, it would equal the propensity score if the errors ε had depended on regressors only through the linear index $X'\beta$, instead of more generally on X. This is less likely to be an economically relevant counterfactual than the ASF case, which is a disadvantage of the AIF relative to the ASF.

The AIF can be written as $F_{-\varepsilon|X'\beta}\left(X'\beta \mid X'\beta\right)$, so all three measures can be written as a distribution function for $-\varepsilon$ evaluated at $X'\beta$. In this sense, the AIF may be interpreted as a middle ground between the propensity score and the ASF, since the propensity score conditions on all the covariates by using $F_{-\varepsilon|X}$, the ASF conditions on no covariates by using $F_{-\varepsilon}$ while the AIF is an in-between case that conditions on the index of covariates, $F_{-\varepsilon|X'\beta}$.

Define the function $M\left(X'\beta\right) = E\left(D \mid X'\beta\right)$, and let m denote the derivative of the function M. The marginal effects of changing regressors on choice probabilities as measured by the AIF are $\partial E\left(D \mid X'\beta\right)/\partial X = m\left(X'\beta\right)\beta$ so the average marginal effects just equal the average derivatives $E\left(m\left(X'\beta\right)\right)\beta$.

In the special case of the linear probability model, both the ASF and AIF will just equal the fitted values of the linear (two stage least squares) regression of D on X. Otherwise, given estimates $\widehat{\beta}$, the AIF choice probabilities could be estimated using a standard one dimensional kernel regression of D on $X'\widehat{\beta}$ (using, e.g., the lpoly module in Stata, with the 'at' option to evaluate the regression at the observed data points). In particular, let $M_i = M\left(X_i'\beta\right)$, let K denote a standard one dimensional kernel function (e.g., a symmetric density function like the standard normal) and let h denote a bandwidth. Then the estimated AIF for any observation i would be given by the kernel regression formula

$$\widehat{M}_i = \frac{\sum_{j=1}^{n} D_j K\left(\frac{\left(X_i'\widehat{\beta}\right)-\left(X_j'\widehat{\beta}\right)}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{\left(X_i'\widehat{\beta}\right)-\left(X_j'\widehat{\beta}\right)}{h}\right)} \quad \text{for } i = 1, ..., n \tag{1}$$

The bandwidth h could be obtained by cross validation, or more simply by Silverman's rule. One could evaluate equation (1) for each data point i, and then calculate the average or median AIF

using the sample average or sample median of $\widehat{M}_i$. Equation (1 is the exact same formula that is used to estimate the ordinary propensity score with nonparametric binary choice estimators when we do not have endogeneity or heteroskedasticity, e.g., $\widehat{M}_i$ equals the propensity score used in the Klein and Spady (1993) estimator.

Let $m_i = m\left(X_i'\beta\right)$ and let K' denote the derivative of the kernel function K. Based on the AIF, the vector of marginal effects of the regressors X (that is, the derivative of the probability of $D = 1$ with respect to X) at observation i would be estimated as $\widehat{m}_i\widehat{\beta}$ where the scalar $\widehat{m}_i$ (an estimator of m_i) is defined by

$$\widehat{m}_i = \frac{\frac{1}{h}\sum_{j=1}^{n}\left(D_j - \widehat{M}_j\right)K'\left(\frac{\left(X_i'\widehat{\beta}\right)-\left(X_j'\widehat{\beta}\right)}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{\left(X_i'\widehat{\beta}\right)-\left(X_j'\widehat{\beta}\right)}{h}\right)} \tag{2}$$

The estimated mean marginal effects of X on choice probabilities are then[9]

$$\overline{m}\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_i\widehat{\beta}. \tag{3}$$

9 Conclusions

We have summarized advantages and disadvantages of some simple, practical ways to deal with endogenous regressors and heteroskedasticity in empirical binary choice models. These are linear probability model estimators, control functions, maximum likelihood estimation, and special regressor methods. We also propose the average index function (AIF) as a simple method for summarizing choice probabilities and marginal effects in these types of models.

We have three main conclusions. First, linear probability model estimators have more disadvantages than are generally recognized. Second, in contexts where the average structural function (ASF) is difficult to calculate, the AIF provides a numerically simple alternative. Third, special regressor methods may be useful, at least in providing robustness checks of results against alternative, more standard estimators. This is because, while requiring strong restrictions on one regressor, they can be both trivial to implement numerically, and they are consistent under conditions that (apart from the one special regressor) are otherwise much weaker than the conditions required for common alternative estimators.

Other issues that arise in general endogenous regressor models involve weak instruments and tests for overidentification and exogeneity. A useful area for future work would be comparisons of how well each of the estimators provided here can deal with these issues.

[9] If special regressor estimation was used for coefficient estimation (or if any other estimator is used that normalizes the coefficient of a regressor V to equal one), then all of the preceeding formulas would need to replace $X'\beta$ with $X'\beta + V$, the marginal effect of V would just be $\widehat{m}_i$ and the mean marginal effect of V would be just $\overline{m}$.

References

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

[2] Angrist, J., and S. Pischke (2009), Mostly Harmless Econometrics: An Empiricists' Companion, Princeton University Press, Princeton, NJ.

[3] 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).

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

[5] 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.

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

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

[8] Dong, Y. and Lewbel, A. (2012), "A Simple Estimator for Binary Choice Models with Endogenous Regressors," unpublished working paper.

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

[10] 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.

[11] 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.

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

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

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

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

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

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

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

[19] Klein, R. W. and Spady, R. H. (1993), "An Efficient Semiparametric Estimator for Binary Response Models", Econometrica, 61, 387-421.

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

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

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

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

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

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

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

[27] Manski, C. F. (1975), “Maximum Score Estimation of the Stochastic Utility Model of Choice”, Journal of Econometrics, 3, 205-228.

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

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

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

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

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

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

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