

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 February 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, Measurement Error, Heteroskedasticity, discrete endogenous, censored, random coefficients, Identification, Latent Variable Model. JEL codes: C25, C26. The authors would like to thank Jeff Wooldridge 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 the linear probability

model, 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 the linear probability model, which is that it can estimate a negative average treatment effect even if the true treatment effect is nonnegative for every individual in the sample.

Given coefficient estimates from any of the estimators we examine, we propose the average index function (AIF) for calculating choice probabilities and regressor marginal effects. The AIF is an alternative to the average structural function (ASF) and the propensity score that has the advantage of always being easy to construct and estimate.

We specifically focus here on practical, easy to apply versions of these estimators, rather than on theoretical econometric generality. So we will for the most part 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 $X'\beta + \varepsilon$.

Linear probability models (LPM), 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 the Appendix. 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 heteroskedastic 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 the linear probability model, maximum likelihood, and control function 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, and in an appendix we summarize one simple implementation of the special regressor method.

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 $D = I(X'\beta + \varepsilon \geq 0)$, 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. The special case of a probit model has ε normal, 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 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 equals one given X equals $E(D|X)$. Let $F_{-\varepsilon}$ denote the probability distribution function of $-\varepsilon$. Ignoring endogeneity for the moment, in the model $D = I(X'\beta + \varepsilon \geq 0)$ the probability that D equals one given X is $F_{-\varepsilon}(X'\beta)$, which is the probability that $-\varepsilon \leq X'\beta$. When regressors are endogenous, the probability that $D = 1$ 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 $F_{-\varepsilon}(X'\beta)$ 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 (LPM) 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 $D = I(X^e\beta_e + X^o\beta_o + \varepsilon \geq 0)$, while the linear probability model assumes $D = X^e\beta_e + X^o\beta_o + \varepsilon$. For simplicity, let the system of equations for D and X^e be triangular, so X^e is endogenous in that it may depend on or correlate with ε , but is not itself directly a function of D . Equivalently, X^e and the latent index $X'\beta + \varepsilon$ may be simultaneously determined. The alternative full simultaneity that would permit X^e to be a function of D generally causes problems of coherency and completeness (see, e.g., Lewbel 2007b) that are outside the scope of this paper.

3 The Linear Probability Model

Consider first the linear probability model (LPM). This 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. So the LPM 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 is 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 $E(Z\varepsilon) = 0$ (noting that Z contains the exogenous regressors X^o and must correlate with the endogenous regressors X^e).

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 $F_{-\varepsilon}(X'\beta)$. In contrast, the fitted probabilities in the LPM are $X'\beta$. 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, because otherwise the 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 these problems is the claim that the LPM is only intended to approximate the true probability 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 be 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 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 than 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 will deliver fitted probabilities that tend to be 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 the LPM provides good approximate estimates of marginal effects, including treatment effects, even if it does 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. 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 from Lewbel and Yang (2012). 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 $D = I(1 + T + R + \varepsilon \geq 0)$ 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 $I(2 + R + \varepsilon \geq 0) - I(1 + R + \varepsilon \geq 0) = I(0 \leq 1 + R + \varepsilon \leq 1)$ 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 this 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 the LPM give the wrong sign, 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.

This example shows that even in a trivial model (standard probit) with tiny errors, in which

¹The 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.

²For comparison, if one estimated the coefficients β_0 , β_1 , and β_2 in the correctly specified probit model $D = I(\beta_0 + T\beta_1 + R\beta_2 + \varepsilon \geq 0)$ using this data, 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. The same logic 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.

Standard probit chooses the scale to make $s^2 = \text{var}(\varepsilon) = 1$, but with a perfect fit $\text{var}(\varepsilon)$ 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.

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 nonlinear models like those of binary choice. See Lewbel and Yang (2012) for other examples that include random assignment of treatment.

The sole advantage of the linear probability model relative to other estimators is that all it requires for estimation is the assumption that $E(Z\varepsilon) = 0$ along with the standard linear model rank condition that $E(X'Z)$ have full rank. In particular, models for the endogenous regressors X^e 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, e.g., the regressors can have random coefficients. The efficiency of linear probability model estimation cannot be directly compared to the efficiency of other estimators, because the specification of the linear probability model is incompatible with these other models.

4 Maximum Likelihood

Maximum Likelihood assumes $D = I(X^{e'}\beta_e + X^{o'}\beta_o + \varepsilon \geq 0)$ and $X^e = G(Z, \theta, e)$ where G is a vector of fully specified, parameterized models for each endogenous regressor in the vector X^e , and e is a vector of error terms for the models in G . 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 the linear probability model, 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 X^o and 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. 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 (which is commonly done for variables that one is not certain are truly exogenous and hence might not be valid instruments).

In addition to the difficulty in correctly specifying the functions $G(Z, \theta, e)$, the joint distribution of ε and the vector e , and the complete set of instruments Z , 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 model $D = M(X, \beta, \varepsilon)$, one way to describe the control function method is to assume there exists functions G and h and a "clean" error U such that $X^e = G(Z, e)$, $\varepsilon = h(e, U)$ and U is independent of both X and e . One would then first estimate the function G (this is the model of endogenous regressors as functions of instruments), and get fitted values of the errors e . Then plugging h into the original D model gives $D = M[X, \beta, h(e, U)] = \tilde{M}(X, e, \beta, U)$. Treating e as if it was another regressor, the error term in the model \tilde{M} is U , which is independent of regressors X and e . As a result, the model \tilde{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 $D = I(X^e \beta_e + X^o \beta_o + \varepsilon \geq 0)$, the function G is given by the linear model $X^e = Z'a + e$, and 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, the control function estimator could consist of first linearly regressing X^e on Z . The residuals from that equation are then estimates of e . Plugging $\varepsilon = e\lambda + U$ into the original model gives $D = I(X^e \beta_e + X^o \beta_o + e\lambda + U \geq 0)$, 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 . The model described in this paragraph is exactly the model that is estimated by the ivprobit command in Stata.³ Despite its name, ivprobit is actually a control function estimator, not an instrumental variables estimator, and this has important implications discussed below.

Generically, control function estimation can be described as first estimating the model of endogenous regressors as a function of instruments (like the first stage of two stage least squares), then taking the errors from those first stage models, and including them in the main model as additional regressors. 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 first stage equations for their errors e , and second that including the errors e in the model fixes the endogeneity problem. Another way of saying this second point is this: control functions

³Since 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.

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 they require the ability to solve for the error term e in $G(Z, e)$. So for example if X^e is binary and $G(Z, e) = I(Z'\gamma + e \geq 0)$, then one would not be able to solve for the latent e given observations of X^e and Z . In this case one could consider using an LPM for X^e by linearly regressing X^e on Z (as in the first stage of linear two stage least squares) and taking e to be the resulting residuals, or more generally one could define G as $G(Z, e) = E(X^e | Z) + e$ 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 either of these ways is not independent of Z , which will generally cause violations of the required assumptions regarding U .⁴

For example, the ivprobit command in Stata will generally 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 run and deliver estimates, but, as in the LPM, the assumptions required for these estimates to be consistent will generally not hold.⁵

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 $X^e = G(Z, e)$ 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

⁴This 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 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.

⁵The 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 cannot hold when X^e is discrete or limited. The documentation for the ivprobit command now correctly notes that it is intended for probit models with continuous endogenous regressors.

dropped.⁶

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 in the appendix of this paper. Interested readers are strongly urged to see Dong and Lewbel (2012) and references therein for a general overview of special regressor methods and various simple 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 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 $D = I(X^e \beta_e + X^o \beta_o + V + \varepsilon \geq 0)$, or equivalently $D = I(X' \beta + V + \varepsilon \geq 0)$. 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. This is a completely free normalization in binary choice models (standard models like logit or probit instead normalize the variance of the error ε , so e.g. probit chooses ε to have variance one), but this makes no difference for the calculation of choice probabilities, marginal effects or other applications of the estimates. The only impact of this normalization is that if one wishes to compare the β estimates from special regressor models to those in other threshold crossing models, it will be necessary to divide all of the β estimates in those other models by the estimated coefficient of V in those other models.

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 $E(Z' \varepsilon) = 0$ and $E(Z' X)$ has full rank.

The main drawback of the special regressor method is the restrictions it imposes on V . Formally, V (or more generally, the error term in any model for V as in Dong and Lewbel 2012)

⁶The comments in footnote 4 apply here also.

must be conditionally independent of ε . 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). Also V needs to be continuously distributed after conditioning on the other regressors in the model. This means that we cannot include a term like V^2 in the model for D as an additional regressor (though we could replace V with some known transform of V such as V^3 if necessary). Also, as noted by Khan and Tamer (2010) and Dong and Lewbel (2012), if the V distribution 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 functions, and linear probability models. In particular, unlike the linear probability model, the special regressor always stays in the range of zero to one and is 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 correct specification of the model $X^e = G(Z, e)$ (that is, the first stage of two stage least squares, or estimation of the control function error 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 $E(X'Z)$ 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 much more general conditions 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.

It should also be noted that because special regressor estimators involve dividing by a density function, they can sometimes generate extreme outliers that can greatly increase standard errors (and slow the rate of convergence). 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 means squared error of the estimator by trading off bias for variance and (again see Dong and Lewbel 2012).

7 Other Estimators

The above comparison of estimators focused on general methods that can handle a vector of endogenous regressors X^e . Vytlačil 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 $X^o\beta_o$ 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 instruments Z is zero. However, Shaikh and Vytlačil (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 $X^e\beta_e + X^o\beta_o$ 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 some of the objections and problems associated with the linear probability model 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 $E(D | X)$. If $\varepsilon \perp X$ then $E(D | X) = E(D | X'\beta) = F_{-\varepsilon}(X'\beta)$ 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, $E(D | X)$, $E(D | X'\beta)$, and $F_{-\varepsilon}(X'\beta)$, which are identical in the case of independent errors, may differ from each other.

With endogeneity or heteroskedasticity, $E(D | X)$ 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 $F_{-\varepsilon|X}(X'\beta | X)$, where $F_{-\varepsilon|X}$ is the conditional distribution of $-\varepsilon$ given X . Another disadvantage is that estimation of $E(D | X)$ 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 summarize choice probabilities. For the binary response model the ASF is given by $F_{-\varepsilon}(X'\beta)$, 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 $E(D | X'\beta)$, 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 $F_{-\varepsilon|X'\beta}(X'\beta | X'\beta)$, 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 $M(X'\beta) = E(D | X'\beta)$, 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 $\partial E(D | X'\beta) / \partial X = m(X'\beta) \beta$ so the average marginal effects just equal the average derivatives $E(m(X'\beta + V)) \beta$.

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 $\hat{\beta}$, the AIF choice probabilities could be estimated using a standard one dimensional kernel regression of D on $X'\hat{\beta}$ (using, e.g., the `lpoly` module in Stata, with the 'at' option to evaluate the regression at the observed data points). In particular, the AIF for any observation i would be given by the kernel regression formula

$$\hat{M}_i = \frac{\sum_{j=1}^n D_j K\left(\frac{(X'_i\hat{\beta}) - (X'_j\hat{\beta})}{h}\right)}{\sum_{j=1}^n K\left(\frac{(X'_i\hat{\beta}) - (X'_j\hat{\beta})}{h}\right)} \quad \text{for } i = 1, \dots, n \quad (1)$$

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 \hat{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., \hat{M}_i equals the propensity score used in the Klein and Spady (1993) estimator.

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 $\hat{m}_i \hat{\beta}$ where the scalar \hat{m}_i is defined by

$$\hat{m}_i = \frac{\frac{1}{h} \sum_{j=1}^n (D_j - \hat{M}_i) K'\left(\frac{(X'_i\hat{\beta}) - (X'_j\hat{\beta})}{h}\right)}{\sum_{j=1}^n K\left(\frac{(X'_i\hat{\beta}) - (X'_j\hat{\beta})}{h}\right)} \quad (2)$$

The estimated mean marginal effects of X on choice probabilities are then⁷

$$\bar{m}\hat{\beta} = \frac{1}{n} \sum_{i=1}^n \hat{m}_i \hat{\beta}. \quad (3)$$

⁷If 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 preceding formulas would need to replace $X'\beta$ with $X'\beta + V$, the marginal effect of V would just be \hat{m}_i and the mean marginal effect of V would be just \bar{m} .

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 models, control functions, maximum likelihood, 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, the linear probability model has more disadvantages than is 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 regressor) are otherwise weaker than the conditions required for common alternative estimators.

References

- [1] Altonji, J. G. and R. L. Matzkin (2005), "Cross Section and Panel Data Estimators for Non-separable 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), "Simple Estimators 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. and S. Schennach (2007), "A Simple Ordered Data Estimator For Inverse Density Weighted Functions," *Journal of Econometrics*, 186, 189-211.
- [24] Lewbel, A. and Yang, T. T. (2012), "Another Problem With the Linear Probability Model: Wrong Sign for Treatment Effects," unpublished working paper.
- [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.

10 Appendix: Example of Simple Special Regressor Estimation

In this appendix we describe an example of a numerically simple special regressor estimator. See Dong and Lewbel (2012) and references therein for details. 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 $E(ZX')$ 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 . 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 $f(U)$ denotes a mean zero density function having support $\text{supp}(U)$ that contains $\text{supp}(-S'b - X'\beta - \varepsilon)$. Define T by $T = [D - I(V \geq 0)] / f(U)$. Then it can be shown that $T = X'\beta + \tilde{\varepsilon}$ where $E(Z\tilde{\varepsilon}) = 0$.

Based on this result, 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.

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 $i = 1, \dots, n$, define $\widehat{f}_i = 2 / [(\widehat{U}_i^+ - \widehat{U}_i^-) n]$.

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.

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 will not normalize the coefficient of one of the regressors to equal one, so for example probit would assume that $D = I(X'\gamma + V\alpha + \varepsilon \geq 0)$ 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 as discussed earlier (using, e.g., the AIF), or compare the special regressor based coefficient vector β to the scaled coefficients γ / α .