Binary Choice Models with Endogenous Regressors

Christopher F Baum, Yingying Dong, Arthur Lewbel, Tao Yang

Boston College/DIW Berlin, U.Cal–Irvine, Boston College, Boston College

Stata Conference 2012, San Diego

This presentation is based on the work of Lewbel, Dong & Yang, "Comparing features of Convenient Estimators for Binary Choice Models With Endogenous Regressors", a revised version of Boston College Economics Working Paper No. 789, forthcoming in the *Canadian Journal of Economics* and available from BC EC (www.bc.edu/economics), IDEAS (ideas.repec.org), and EconPapers (econpapers.repec.org). My contribution is the review and enhancement of the software developed in this research project.

Motivation

- Researchers often want to estimate a binomial response, or binary choice, model where one or more explanatory variables are endogenous or mismeasured.
- For instance: in policy analysis, the estimation of treatment effects when treatment is not randomly assigned.
- A linear 2SLS model, equivalent to a linear probability model with instrumental variables, is often employed, ignoring the binary outcome.

Motivation

- Researchers often want to estimate a binomial response, or binary choice, model where one or more explanatory variables are endogenous or mismeasured.
- For instance: in policy analysis, the estimation of treatment effects when treatment is not randomly assigned.
- A linear 2SLS model, equivalent to a linear probability model with instrumental variables, is often employed, ignoring the binary outcome.

Motivation

- Researchers often want to estimate a binomial response, or binary choice, model where one or more explanatory variables are endogenous or mismeasured.
- For instance: in policy analysis, the estimation of treatment effects when treatment is not randomly assigned.
- A linear 2SLS model, equivalent to a linear probability model with instrumental variables, is often employed, ignoring the binary outcome.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

In what follows, we focus on a particular disadvantage of the LPM, and propose a straightforward alternative based on ‘special regressor’ methods (Lewbel, *J. Metrics*, 2000; Dong and Lewbel, 2012, BC WP 604).

We also propose the *average index function* (AIF), an alternative to the average structural function (ASF; Blundell and Powell, *REStud*, 2004), for calculating marginal effects. It is easy to construct and estimate, as we will illustrate.

In what follows, we focus on a particular disadvantage of the LPM, and propose a straightforward alternative based on ‘special regressor’ methods (Lewbel, *J. Metrics*, 2000; Dong and Lewbel, 2012, BC WP 604).

We also propose the *average index function* (AIF), an alternative to the average structural function (ASF; Blundell and Powell, *REStud*, 2004), for calculating marginal effects. It is easy to construct and estimate, as we will illustrate.

Binary choice models

We define D as an observed binary variable: the outcome to be explained. Let X be a vector of observed regressors, and β a corresponding coefficient vector, with ε an unobserved error. In a treatment model, X would include a binary treatment indicator T. In general, X could be divided into X^e, possibly correlated with ε, and X^0, which are exogenous.

A binary choice or 'threshold crossing' model estimated by maximum likelihood is

$$D = I(X\beta + \varepsilon \geq 0)$$

where $I(\cdot)$ is the indicator function. This latent variable approach is that employed in a binomial probit or logit model, with Normal or logistic errors, respectively. Although estimation provides point and interval estimates of β, the choice probabilities and marginal effects are of interest: that is, $\Pr[D = 1|X]$ and $\partial \Pr[D = 1|X]/\partial X$.

Binary choice models

We define D as an observed binary variable: the outcome to be explained. Let X be a vector of observed regressors, and β a corresponding coefficient vector, with ε an unobserved error. In a treatment model, X would include a binary treatment indicator T. In general, X could be divided into X^e, possibly correlated with ε, and X^0, which are exogenous.

A binary choice or 'threshold crossing' model estimated by maximum likelihood is

$$D = I(X\beta + \varepsilon \geq 0)$$

where $I(\cdot)$ is the indicator function. This latent variable approach is that employed in a binomial probit or logit model, with Normal or logistic errors, respectively. Although estimation provides point and interval estimates of β, the choice probabilities and marginal effects are of interest: that is, $\Pr[D = 1|X]$ and $\partial \Pr[D = 1|X]/\partial X$.

Linear probability models

In contrast to the threshold crossing latent variable approach, a linear probability model (LPM) assumes that

$$D = X\beta + \varepsilon$$

so that the estimated coefficients $\hat{\beta}$ are themselves the marginal effects. With all exogenous regressors, $E(D|X) = \Pr[D = 1|X] = X\beta$.

If some elements of X (possibly including treatment indicators) are endogenous or mismeasured, they will be correlated with ε. In that case, an instrumental variables approach is called for, and we can estimate the LPM with 2SLS or IV-GMM, given an appropriate set of instruments Z.

Linear probability models

In contrast to the threshold crossing latent variable approach, a linear probability model (LPM) assumes that

$$D = X\beta + \varepsilon$$

so that the estimated coefficients $\hat{\beta}$ are themselves the marginal effects. With all exogenous regressors, $E(D|X) = \Pr[D = 1|X] = X\beta$.

If some elements of X (possibly including treatment indicators) are endogenous or mismeasured, they will be correlated with ε. In that case, an instrumental variables approach is called for, and we can estimate the LPM with 2SLS or IV-GMM, given an appropriate set of instruments Z.

As the LPM with exogenous explanatory variables is based on standard regression, the zero conditional mean assumption $E(\varepsilon|X) = 0$ applies. In the presence of endogeneity or measurement error, the corresponding assumption $E(\varepsilon|Z) = 0$ applies, with Z the set of instruments, including the exogenous elements of X.

An obvious flaw in the LPM: the error ε cannot be independent of *any* regressors, even exogenous regressors, unless X consists of a single binary regressor. This arises because for any given X, ε must equal either $1 - X\beta$ or $-X\beta$, which are functions of all elements of X.

As the LPM with exogenous explanatory variables is based on standard regression, the zero conditional mean assumption $E(\varepsilon|X) = 0$ applies. In the presence of endogeneity or measurement error, the corresponding assumption $E(\varepsilon|Z) = 0$ applies, with Z the set of instruments, including the exogenous elements of X.

An obvious flaw in the LPM: the error ε cannot be independent of *any* regressors, even exogenous regressors, unless X consists of a single binary regressor. This arises because for any given X, ε must equal either $1 - X\beta$ or $-X\beta$, which are functions of all elements of X.

The other, well recognized, flaw in the LPM is that its fitted values are not constrained to lie in the unit interval, so that predicted probabilities below zero or above one are commonly encountered. Any regressor that can take on a large range of values will inevitably cause the LPM's predictions to breach these bounds.

A common rejoinder to these critiques is that the LPM is only intended to approximate the true probability for a limited range of X values, and that its constant marginal effects are preferable to those of the binary probit or logit model, which are functions of the values of all elements of X.

The other, well recognized, flaw in the LPM is that its fitted values are not constrained to lie in the unit interval, so that predicted probabilities below zero or above one are commonly encountered. Any regressor that can take on a large range of values will inevitably cause the LPM's predictions to breach these bounds.

A common rejoinder to these critiques is that the LPM is only intended to approximate the true probability for a limited range of X values, and that its constant marginal effects are preferable to those of the binary probit or logit model, which are functions of the values of all elements of X.

Consider, however, the LPM with a single continuous regressor. The linear prediction is an approximation to the S-shape of any cumulative distribution function: for instance, that of the Normal for the probit model. The linear prediction departs greatly from the S-shaped CDF long before it nears the (0,1) limits. Thus, the LPM will produce predicted probabilities that are too extreme (closer to zero or one) even for moderate values of $X\hat{\beta}$ that stay 'in bounds'.

Some researchers claim that although predicted probabilities derived from the LPM are flawed, their main interest lies in the models' marginal effects, and argue that it makes little substantive difference to use a LPM, with its constant marginal effects, rather than the more complex marginal effects derived from a proper estimated CDF, such as that of the probit model.

Consider, however, the LPM with a single continuous regressor. The linear prediction is an approximation to the S-shape of any cumulative distribution function: for instance, that of the Normal for the probit model. The linear prediction departs greatly from the S-shaped CDF long before it nears the (0,1) limits. Thus, the LPM will produce predicted probabilities that are too extreme (closer to zero or one) even for moderate values of $X\hat{\beta}$ that stay ‘in bounds’.

Some researchers claim that although predicted probabilities derived from the LPM are flawed, their main interest lies in the models’ marginal effects, and argue that it makes little substantive difference to use a LPM, with its constant marginal effects, rather than the more complex marginal effects derived from a proper estimated CDF, such as that of the probit model.

Example 1

Jeff Wooldridge's widely used undergraduate text, *Introductory Econometrics: A Modern Approach* devotes a section of the chapter on regression with qualitative variables to the LPM. He points out two flaws: computation of the predicted probability and marginal effects—and goes on to state

> *"Even with these problems, the linear probability model is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample." (2009, p. 249)*

Wooldridge also discusses the heteroskedastic nature of the LPM's error, which is binomial by construction, but does not address the issue of the lack of independence that this implies.

Example 2

Josh Angrist and Steve Pischke's popular *Mostly Harmless Econometrics* give several empirical examples where the marginal effects of a dummy variable estimated by LPM and probit techniques are 'indistinguishable.' They conclude that

> *"...while a nonlinear model may fit the CEF (conditional expectation function) for LDVs (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." (2009, p. 107)*

Angrist and Pischke (AP) go on to invoke the principle of Occam's razor, arguing that

> *"...extra complexity comes into the inference step as well, since we need standard errors for marginal effects." (ibid.)*

This is surely a red herring for Stata users, as the `margins` command in Stata 11 or 12 computes those standard errors via the delta method. AP also discuss the difficulty of computing marginal effects for a binary regressor: again, not an issue for Stata 12 users, with the new `contrast` command.

An alarming example

The most compelling argument against the LPM, though, dismisses the notion that its use is merely a matter of taste and convenience. Lewbel, Dong and Yang (2012) provide a simple example in which the LPM cannot even recover the appropriate sign of the treatment effect. To illustrate that point, consider the data:

```
. l R Treated D, sep(0) noobs

  +---------------------+
  |     R   Treated   D |
  |---------------------|
  |  -1.8         0   0 |
  |   -.9         0   1 |
  |  -.92         0   1 |
  |  -2.1         1   0 |
  | -1.92         1   1 |
  |    10         1   1 |
  +---------------------+
```

In this contrived example, three of the observations are treated (`Treated=1`) and three are not. The outcome variable D is generated by the probit specification

$$D = I(1 + Treated + R + \varepsilon \geq 0)$$

with Normal errors, independent of the regressors. The treatment effect for an individual is the difference in outcome between being treated and untreated:

$$I((2 + R + \varepsilon) \geq 0) - I((1 + R + \varepsilon) \geq 0) = I(0 \leq (1 + R + \varepsilon) \leq 1)$$

for any given R, ε. By construction, no individual can have a negative treatment effect, regardless of their values of R, ε.

In this sample, the true treatment effect is 1 for the fifth individual (who is treated) and zero for the others, and the true average treatment effect (ATE) is 1/6. So let's estimate the ATE with a linear probability model:

```
. reg D Treated R, robust

Linear regression                                      Number of obs =       6
                                                       F(  2,     3) =    1.02
                                                       Prob > F      =  0.4604
                                                       R-squared     =  0.1704
                                                       Root MSE      =  .60723

------------------------------------------------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Treated |  -.1550841   .5844637    -0.27   0.808    -2.015108     1.70494
           R |   .0484638   .0419179     1.16   0.331    -.0849376    .1818651
       _cons |   .7251463   .3676811     1.97   0.143    -.4449791    1.895272
------------------------------------------------------------------------------
```

The estimated ATE is -0.16, and the estimated marginal rate of substitution (β_1/β_2), via `nlcom`, is -3.2. Both these quantities have the wrong sign, and the MRS is more than three times the true value.

In this sample, the true treatment effect is 1 for the fifth individual (who is treated) and zero for the others, and the true average treatment effect (ATE) is 1/6. So let's estimate the ATE with a linear probability model:

```
. reg D Treated R, robust

Linear regression                                      Number of obs =       6
                                                       F(  2,      3) =    1.02
                                                       Prob > F      =  0.4604
                                                       R-squared     =  0.1704
                                                       Root MSE      =  .60723

------------------------------------------------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Treated |  -.1550841   .5844637    -0.27   0.808    -2.015108     1.70494
           R |   .0484638   .0419179     1.16   0.331    -.0849376    .1818651
       _cons |   .7251463   .3676811     1.97   0.143    -.4449791    1.895272
------------------------------------------------------------------------------
```

The estimated ATE is -0.16, and the estimated marginal rate of substitution (β_1/β_2), via `nlcom`, is -3.2. Both these quantities have the wrong sign, and the MRS is more than three times the true value.

In this sample, the true treatment effect is 1 for the fifth individual (who is treated) and zero for the others, and the true average treatment effect (ATE) is 1/6. So let's estimate the ATE with a linear probability model:

```
. reg D Treated R, robust

Linear regression                                      Number of obs =       6
                                                       F(  2,      3) =    1.02
                                                       Prob > F      =  0.4604
                                                       R-squared     =  0.1704
                                                       Root MSE      =  .60723

------------------------------------------------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Treated |  -.1550841   .5844637    -0.27   0.808    -2.015108     1.70494
           R |   .0484638   .0419179     1.16   0.331    -.0849376    .1818651
       _cons |   .7251463   .3676811     1.97   0.143    -.4449791    1.895272
------------------------------------------------------------------------------
```

The estimated ATE is -0.16, and the estimated marginal rate of substitution (β_1/β_2), via `nlcom`, is -3.2. Both these quantities have the wrong sign, and the MRS is more than three times the true value.

Thus, even in a trivial model with a minuscule stochastic element, where every individual has either a zero or positive treatment effect, the LPM cannot even get the sign right. This is a contrived example, of course, but illustrative of the dangers of assuming that the LPM will do a reasonable job.

If a LPM estimated with OLS exhibits these problems, it is evident that a more elaborate model, such as a LPM estimated with 2SLS or IV-GMM, would be as clearly flawed. We turn, then, to more reliable alternatives.

Thus, even in a trivial model with a minuscule stochastic element, where every individual has either a zero or positive treatment effect, the LPM cannot even get the sign right. This is a contrived example, of course, but illustrative of the dangers of assuming that the LPM will do a reasonable job.

If a LPM estimated with OLS exhibits these problems, it is evident that a more elaborate model, such as a LPM estimated with 2SLS or IV-GMM, would be as clearly flawed. We turn, then, to more reliable alternatives.

Maximum likelihood estimators

A maximum likelihood estimator of a binary outcome with possibly endogenous regressors can be implemented for the model

$$\begin{aligned} D &= I(X^e\beta_e + X^0\beta_0 + \varepsilon \geq 0) \\ X^e &= G(Z, \theta, e) \end{aligned}$$

which for a single binary endogenous regressor, $G(\cdot)$ probit, and ε and e jointly Normal, is the model estimated by Stata's `biprobit` command.

Like the LPM, maximum likelihood allows endogenous regressors in X^e to be continuous, discrete, limited, etc. as long as a model for $G(\cdot)$ can be fully specified, along with the fully parameterized joint distribution of (ε, e).

Maximum likelihood estimators

A maximum likelihood estimator of a binary outcome with possibly endogenous regressors can be implemented for the model

$$\begin{aligned} D &= I(X^e\beta_e + X^0\beta_0 + \varepsilon \geq 0) \\ X^e &= G(Z, \theta, e) \end{aligned}$$

which for a single binary endogenous regressor, $G(\cdot)$ probit, and ε and e jointly Normal, is the model estimated by Stata's `biprobit` command.

Like the LPM, maximum likelihood allows endogenous regressors in X^e to be continuous, discrete, limited, etc. as long as a model for $G(\cdot)$ can be fully specified, along with the fully parameterized joint distribution of (ε, e).

Control function estimators

Control function estimators first estimate the model of endogenous regressors as a function of instruments, like the ‘first stage’ of 2SLS, then use the errors from this model as an additional regressor in the main model.

This approach is more general than maximum likelihood as the first stage function can be semiparametric or nonparametric, and the joint distribution of (ε, e) need not be fully parameterized.

Control function estimators

Control function estimators first estimate the model of endogenous regressors as a function of instruments, like the 'first stage' of 2SLS, then use the errors from this model as an additional regressor in the main model.

This approach is more general than maximum likelihood as the first stage function can be semiparametric or nonparametric, and the joint distribution of (ε, e) need not be fully parameterized.

To formalize the approach, consider a model $D = M(X, \beta, \varepsilon)$, and assume there are functions G, h and a well-behaved error U such that $X^e = G(Z, e), \varepsilon = h(e, U)$, and $U \perp (X, e)$.

We first estimate $G(\cdot)$: the endogenous regressors as functions of instruments Z, and derive fitted values of the errors e. Then we have

$$D = M(X, \beta, h(e, u)) = \widetilde{M}(X, e, \beta, U)$$

where the error term of the $\widetilde{M}$ model is U, which is suitably independent of (X, e). This model no longer has an endogeneity problem, and can be estimated via straightforward methods.

To formalize the approach, consider a model $D = M(X, \beta, \varepsilon)$, and assume there are functions G, h and a well-behaved error U such that $X^e = G(Z, e), \varepsilon = h(e, U)$, and $U \perp (X, e)$.

We first estimate $G(\cdot)$: the endogenous regressors as functions of instruments Z, and derive fitted values of the errors e. Then we have

$$D = M(X, \beta, h(e, u)) = \widetilde{M}(X, e, \beta, U)$$

where the error term of the $\widetilde{M}$ model is U, which is suitably independent of (X, e). This model no longer has an endogeneity problem, and can be estimated via straightforward methods.

Given the threshold crossing model

$$\begin{aligned} D &= I(X^e\beta_e + X^0\beta_0 + \varepsilon \geq 0) \\ X^e &= Z\alpha + e \end{aligned}$$

with (ε, e) jointly normal, we can first linearly regress X^e on Z, with residuals being estimates of e. This then yields an ordinary probit model

$$D = I(X^e\beta_e + X^0\beta_0 + \lambda e + U \geq 0)$$

which is the model estimated by Stata's `ivprobit` command. Despite its name, `ivprobit` is a control function estimator, not an IV estimator.

A substantial limitation of control function methods in this context is that they generally require the endogenous regressors X^e to be continuous, rather than binary, discrete, or censored. For instance, a binary endogenous regressor will violate the assumptions necessary to derive estimates of the ‘first stage’ error term e. The errors in the ‘first stage’ regression cannot be normally distributed and independent of the regressors. Thus, the `ivprobit` command should not be applied to binary endogenous regressors, as its documentation clearly states.

In this context, control function estimators—like maximum likelihood estimators—of binary outcome models require that the first stage model be correctly specified. This is an important limitation of these approaches. A 2SLS approach will lose efficiency if an appropriate instrument is not included, but a ML or control function estimator will generally become inconsistent.

A substantial limitation of control function methods in this context is that they generally require the endogenous regressors X^e to be continuous, rather than binary, discrete, or censored. For instance, a binary endogenous regressor will violate the assumptions necessary to derive estimates of the 'first stage' error term e. The errors in the 'first stage' regression cannot be normally distributed and independent of the regressors. Thus, the `ivprobit` command should not be applied to binary endogenous regressors, as its documentation clearly states.

In this context, control function estimators—like maximum likelihood estimators—of binary outcome models require that the first stage model be correctly specified. This is an important limitation of these approaches. A 2SLS approach will lose efficiency if an appropriate instrument is not included, but a ML or control function estimator will generally become inconsistent.

Special regressor estimators

Special regressor estimators were first proposed by Lewbel (*J. Metrics*, 2000). Their implementation are fully described in Dong and Lewbel (2012, BC WP 604). They assume that the model includes a particular regressor, V, with certain properties. It is exogenous (that is, $E(\varepsilon|V) = 0$) and appears as an additive term in the model. It is continuously distributed and has a large support. Any normally distributed regressor would satisfy this condition.

A third condition, preferable but not strictly necessary, is that V have a thick-tailed distribution. A regressor with greater kurtosis will be more useful as a special regressor.

The binary choice special regressor proposed by Lewbel (2000) has the 'threshold crossing' form

$$D = I(X^e\beta_e + X^0\beta_0 + V + \varepsilon \geq 0)$$

or, equivalently,

$$D = I(X\beta + V + \varepsilon \geq 0)$$

This is the same basic form for D as in the ML or control function (CF) approach. Note, however, that the special regressor V has been separated from the other exogenous regressors, and its coefficient normalized to unity: a harmless normalization.

The binary choice special regressor proposed by Lewbel (2000) has the 'threshold crossing' form

$$D = I(X^e \beta_e + X^0 \beta_0 + V + \varepsilon \geq 0)$$

or, equivalently,

$$D = I(X\beta + V + \varepsilon \geq 0)$$

This is the same basic form for D as in the ML or control function (CF) approach. Note, however, that the special regressor V has been separated from the other exogenous regressors, and its coefficient normalized to unity: a harmless normalization.

Given a special regressor V, the only other requirements are those applicable to linear 2SLS: to handle endogeneity, the set of instruments Z must satisfy $E(\varepsilon|Z) = 0$, and $E(Z'X)$ must have full rank.

The main drawback of this method is that the special regressor V must be conditionally independent of ε. Even if it is exogenous, it could fail to satisfy this assumption because of the way in which V might affect other endogenous regressors. Also, V must be continuously distributed after conditioning on the other regressors, so that a term like V^2 could not be included as an additional regressor.

Given a special regressor V, the only other requirements are those applicable to linear 2SLS: to handle endogeneity, the set of instruments Z must satisfy $E(\varepsilon|Z) = 0$, and $E(Z'X)$ must have full rank.

The main drawback of this method is that the special regressor V must be conditionally independent of ε. Even if it is exogenous, it could fail to satisfy this assumption because of the way in which V might affect other endogenous regressors. Also, V must be continuously distributed after conditioning on the other regressors, so that a term like V^2 could not be included as an additional regressor.

Apart from these restrictions on V, the special regressor (SR) method has none of the drawbacks of the three models discussed earlier:

- Unlike the LPM, the SR predictions stay ‘in bounds’ and is consistent with other threshold crossing models.
- Unlike ML and CF methods, the SR model does not require correct specification of the ‘first stage’ model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike ML, the SR method has a linear form, not requiring iterative search.
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML, there is a single estimation method, regardless of the characteristics of X^e.
- Unlike ML, the SR method permits unknown heteroskedasticity in the model errors.

Apart from these restrictions on V, the special regressor (SR) method has none of the drawbacks of the three models discussed earlier:

- Unlike the LPM, the SR predictions stay ‘in bounds’ and is consistent with other threshold crossing models.
- Unlike ML and CF methods, the SR model does not require correct specification of the ‘first stage’ model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike ML, the SR method has a linear form, not requiring iterative search.
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML, there is a single estimation method, regardless of the characteristics of X^e.
- Unlike ML, the SR method permits unknown heteroskedasticity in the model errors.

Apart from these restrictions on V, the special regressor (SR) method has none of the drawbacks of the three models discussed earlier:

- Unlike the LPM, the SR predictions stay ‘in bounds’ and is consistent with other threshold crossing models.
- Unlike ML and CF methods, the SR model does not require correct specification of the ‘first stage’ model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike ML, the SR method has a linear form, not requiring iterative search.
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML, there is a single estimation method, regardless of the characteristics of X^e.
- Unlike ML, the SR method permits unknown heteroskedasticity in the model errors.

Apart from these restrictions on V, the special regressor (SR) method has none of the drawbacks of the three models discussed earlier:

- Unlike the LPM, the SR predictions stay ‘in bounds’ and is consistent with other threshold crossing models.
- Unlike ML and CF methods, the SR model does not require correct specification of the ‘first stage’ model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike ML, the SR method has a linear form, not requiring iterative search.
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML, there is a single estimation method, regardless of the characteristics of X^e.
- Unlike ML, the SR method permits unknown heteroskedasticity in the model errors.

Apart from these restrictions on V, the special regressor (SR) method has none of the drawbacks of the three models discussed earlier:

- Unlike the LPM, the SR predictions stay 'in bounds' and is consistent with other threshold crossing models.
- Unlike ML and CF methods, the SR model does not require correct specification of the 'first stage' model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike ML, the SR method has a linear form, not requiring iterative search.
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML, there is a single estimation method, regardless of the characteristics of X^e.
- Unlike ML, the SR method permits unknown heteroskedasticity in the model errors.

The special regressor method imposes far fewer assumptions on the distribution of errors—particularly the errors e in the ‘first stage’ equations for X^e—than do CF or ML estimation methods. Therefore, SR estimators will be less efficient than these alternatives when the alternatives are consistent.

SR estimators may be expected to have larger standard errors and lower precision than other methods, *when those methods are valid*. However, if a special regressor V can be found, the SR method will be valid under much more general conditions than the ML and CF methods.

The special regressor method imposes far fewer assumptions on the distribution of errors—particularly the errors e in the ‘first stage’ equations for X^e—than do CF or ML estimation methods. Therefore, SR estimators will be less efficient than these alternatives when the alternatives are consistent.

SR estimators may be expected to have larger standard errors and lower precision than other methods, *when those methods are valid*. However, if a special regressor V can be found, the SR method will be valid under much more general conditions than the ML and CF methods.

The average index function (AIF)

Consider the original estimation problem

$$D = I(X\beta + \varepsilon \geq 0)$$

where with generality one of the elements of X may be a special regressor V, with coefficient one. If ε is independent of X, the *propensity score* or *choice probability* is

$\Pr[D = 1|X] = E(D|X) = E(D|X\beta) = F_{-\varepsilon}(X\beta) = \Pr(-\varepsilon \leq X\beta)$, with $F_{-\varepsilon}(\cdot)$ the probability distribution function of $-\varepsilon$. In the case of independent errors, these measures are identical.

When some regressors are endogenous, or generally when the assumption $X \perp \varepsilon$ is violated (e.g., by heteroskedasticity), these expressions may differ from one another.

The average index function (AIF)

Consider the original estimation problem

$$D = I(X\beta + \varepsilon \geq 0)$$

where with generality one of the elements of X may be a special regressor V, with coefficient one. If ε is independent of X, the *propensity score* or *choice probability* is $\Pr[D = 1|X] = E(D|X) = E(D|X\beta) = F_{-\varepsilon}(X\beta) = \Pr(-\varepsilon \leq X\beta)$, with $F_{-\varepsilon}(\cdot)$ the probability distribution function of $-\varepsilon$. In the case of independent errors, these measures are identical.

When some regressors are endogenous, or generally when the assumption $X \perp \varepsilon$ is violated (e.g., by heteroskedasticity), these expressions may differ from one another.

Blundell and Powell (*REStud*, 2004) propose using the average structural function (ASF) to summarize choice probabilities: $F_{-\varepsilon}(X\beta)$, even though ε is no longer independent of X. In this case, $F_{-\varepsilon|X}(X\beta|X)$ should be computed: a formidable task.

Lewbel, Dong and Yang (BC WP 789) propose using the measure $E(D|X\beta)$, which they call the *average index function* (AIF), to summarize choice probabilities.

Like the ASF, the AIF is based on the estimated index, and equals the propensity score when $\varepsilon \perp X$. However, when this assumption is violated (by endogeneity or heteroskedasticity), the AIF is usually easier to estimate, via a unidimensional nonparametric regression of D on $X\beta$.

Blundell and Powell (*REStud*, 2004) propose using the average structural function (ASF) to summarize choice probabilities: $F_{-\varepsilon}(X\beta)$, even though ε is no longer independent of X. In this case, $F_{-\varepsilon|X}(X\beta|X)$ should be computed: a formidable task.

Lewbel, Dong and Yang (BC WP 789) propose using the measure $E(D|X\beta)$, which they call the *average index function* (AIF), to summarize choice probabilities.

Like the ASF, the AIF is based on the estimated index, and equals the propensity score when $\varepsilon \perp X$. However, when this assumption is violated (by endogeneity or heteroskedasticity), the AIF is usually easier to estimate, via a unidimensional nonparametric regression of D on $X\beta$.

Blundell and Powell (*REStud*, 2004) propose using the average structural function (ASF) to summarize choice probabilities: $F_{-\varepsilon}(X\beta)$, even though ε is no longer independent of X. In this case, $F_{-\varepsilon|X}(X\beta|X)$ should be computed: a formidable task.

Lewbel, Dong and Yang (BC WP 789) propose using the measure $E(D|X\beta)$, which they call the *average index function* (AIF), to summarize choice probabilities.

Like the ASF, the AIF is based on the estimated index, and equals the propensity score when $\varepsilon \perp X$. However, when this assumption is violated (by endogeneity or heteroskedasticity), the AIF is usually easier to estimate, via a unidimensional nonparametric regression of D on $X\beta$.

The AIF can be considered a middle ground between the propensity score and the ASF, as the former conditions on all covariates using $F_{-\varepsilon|X}$; the ASF conditions on no covariates using $F_{-\varepsilon}$; and the AIF conditions on the *index* of covariates, $F_{-\varepsilon|X\beta}$.

Define the function $M(X\beta) = E(D|X\beta)$, with derivatives m. The marginal effects of the regressors on the 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, $E(m(X\beta + V))\beta$.

The AIF can be considered a middle ground between the propensity score and the ASF, as the former conditions on all covariates using $F_{-\varepsilon|X}$; the ASF conditions on no covariates using $F_{-\varepsilon}$; and the AIF conditions on the *index* of covariates, $F_{-\varepsilon|X\beta}$.

Define the function $M(X\beta) = E(D|X\beta)$, with derivatives m. The marginal effects of the regressors on the 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, $E(m(X\beta + V))\beta$.

For the LPM, the ASF and AIF both equal the fitted values of the linear 2SLS regression of D on X. For the other methods, the AIF choice probabilities can be estimated using a standard unidimensional kernel regression of D on $X\hat{\beta}$: for instance, using the `lpoly` command in Stata, with the `at()` option specifying the observed data points. This will produce the AIF for each observation i, $\widehat{M}_i$.

Employing the derivatives of the kernel function, the individual-level marginal effects $\widehat{m}_i$ may be calculated, and averaged to produce average marginal effects:

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

Estimates of the precision of these average marginal effects may be derived by bootstrapping.

For the LPM, the ASF and AIF both equal the fitted values of the linear 2SLS regression of D on X. For the other methods, the AIF choice probabilities can be estimated using a standard unidimensional kernel regression of D on $X\hat{\beta}$: for instance, using the `lpoly` command in Stata, with the `at()` option specifying the observed data points. This will produce the AIF for each observation i, $\widehat{M}_i$.

Employing the derivatives of the kernel function, the individual-level marginal effects $\widehat{m}_i$ may be calculated, and averaged to produce average marginal effects:

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

Estimates of the precision of these average marginal effects may be derived by bootstrapping.

For the LPM, the ASF and AIF both equal the fitted values of the linear 2SLS regression of D on X. For the other methods, the AIF choice probabilities can be estimated using a standard unidimensional kernel regression of D on $X\hat{\beta}$: for instance, using the `lpoly` command in Stata, with the `at()` option specifying the observed data points. This will produce the AIF for each observation i, $\widehat{M}_i$.

Employing the derivatives of the kernel function, the individual-level marginal effects $\widehat{m}_i$ may be calculated, and averaged to produce average marginal effects:

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

Estimates of the precision of these average marginal effects may be derived by bootstrapping.

The Stata implementation

My Stata command `ssimplereg`, which is still being refined, estimates the Lewbel and Dong simple special regression estimator of a binary outcome with one or more binary endogenous variables. It is an optimized version of earlier code developed for this estimator, and provides significant (8–10x) speed improvements over that code.

Two forms of the special regressor estimator are defined, depending on assumptions made about the distribution of the special regressor V. In the first form of the model, only the mean of V is assumed to be related to the other covariates. In the second, 'heteroskedastic' form, higher moments of V can also depend in arbitrary, unknown ways on the other covariates. In practice, the latter form may include squares and cross products of some of the covariates in the estimation process, similar to the auxiliary regression used in White's general test for heteroskedasticity.

The Stata implementation

My Stata command `ssimplereg`, which is still being refined, estimates the Lewbel and Dong simple special regression estimator of a binary outcome with one or more binary endogenous variables. It is an optimized version of earlier code developed for this estimator, and provides significant (8–10x) speed improvements over that code.

Two forms of the special regressor estimator are defined, depending on assumptions made about the distribution of the special regressor V. In the first form of the model, only the mean of V is assumed to be related to the other covariates. In the second, ‘heteroskedastic’ form, higher moments of V can also depend in arbitrary, unknown ways on the other covariates. In practice, the latter form may include squares and cross products of some of the covariates in the estimation process, similar to the auxiliary regression used in White’s general test for heteroskedasticity.

The `ssimplereg` Stata command also allows for two specifications of the density estimator used in the model: one based on a standard kernel density approach such as that implemented by `density` or Ben Jann's `kdens`, as well as the alternative 'sorted data density' approach proposed by Lewbel and Schennach (*J. Econometrics*, 2007). Implementation of the latter approach also benefited greatly, in terms of speed, by being rewritten in Mata, with Ben Jann's help gratefully acknowledged.

Just as in a probit or ivprobit model, the quantities of interest are not the estimated coefficients derived in the special regressor method, but rather the marginal effects. In the work of Lewbel et al., those are derived from the average index function (AIF) as described earlier. Point estimates of the AIF can be derived in a manner similar to that of average marginal effects in standard limited dependent variable models. For interval estimates, bootstrapped standard errors for the marginal effects are computed.

A bootstrap option was also added to ssimplereg so that the estimator can produce point and interval estimates of the relevant marginal effects in a single step, with the user's choice of the number of bootstrap samples to be drawn.

Just as in a probit or ivprobit model, the quantities of interest are not the estimated coefficients derived in the special regressor method, but rather the marginal effects. In the work of Lewbel et al., those are derived from the average index function (AIF) as described earlier. Point estimates of the AIF can be derived in a manner similar to that of average marginal effects in standard limited dependent variable models. For interval estimates, bootstrapped standard errors for the marginal effects are computed.

A bootstrap option was also added to ssimplereg so that the estimator can produce point and interval estimates of the relevant marginal effects in a single step, with the user's choice of the number of bootstrap samples to be drawn.

An empirical illustration

In this example of the special regressor method, taken from Dong and Lewbel (BC WP 604), the binary dependent variable is an indicator that individual *i* migrates from one US state to another. The objective is to estimate the probability of interstate migration.

The special regressor V_i in this context is age. Human capital theory suggests that it should appear linearly (or at least monotonically) in a threshold crossing model. Migration is in part driven by maximizing expected lifetime income, and the potential gain in lifetime earnings from a permanent change in labor income declines linearly with age. Evidence of empirical support for this relationship is provided by Dong (*Ec. Letters*, 2010). V_i is defined as the negative of age, demeaned, so that it should have a positive coefficient and a zero mean.

An empirical illustration

In this example of the special regressor method, taken from Dong and Lewbel (BC WP 604), the binary dependent variable is an indicator that individual i migrates from one US state to another. The objective is to estimate the probability of interstate migration.

The special regressor V_i in this context is age. Human capital theory suggests that it should appear linearly (or at least monotonically) in a threshold crossing model. Migration is in part driven by maximizing expected lifetime income, and the potential gain in lifetime earnings from a permanent change in labor income declines linearly with age. Evidence of empirical support for this relationship is provided by Dong (*Ec. Letters*, 2010). V_i is defined as the negative of age, demeaned, so that it should have a positive coefficient and a zero mean.

Pre-migration family income and home ownership are expected to be significant determinants of migration, and both should be considered endogenous. A maximum likelihood approach would require an elaborate dynamic specification in order to model the homeownership decision. Control function methods such as `ivprobit` are not appropriate as `homeowner` is a discrete variable.

The sample used includes male heads of household, 23–59 years of age, from the 1990 wave of the PSID who have completed education and are not retired, so as to exclude those moving to retirement communities. The observed $D = 1$ indicates migration during 1991–1993. In the sample of 4689 individuals, 807 were interstate migrants.

Exogenous regressors in the model include years of education, number of children, and indicators for white, disabled, and married individuals. The instruments Z also include the level of government benefits received in 1989–1990 and state median residential tax rates.

Pre-migration family income and home ownership are expected to be significant determinants of migration, and both should be considered endogenous. A maximum likelihood approach would require an elaborate dynamic specification in order to model the homeownership decision. Control function methods such as `ivprobit` are not appropriate as `homeowner` is a discrete variable.

The sample used includes male heads of household, 23–59 years of age, from the 1990 wave of the PSID who have completed education and are not retired, so as to exclude those moving to retirement communities. The observed $D = 1$ indicates migration during 1991–1993. In the sample of 4689 individuals, 807 were interstate migrants.

Exogenous regressors in the model include years of education, number of children, and indicators for white, disabled, and married individuals. The instruments Z also include the level of government benefits received in 1989–1990 and state median residential tax rates.

Pre-migration family income and home ownership are expected to be significant determinants of migration, and both should be considered endogenous. A maximum likelihood approach would require an elaborate dynamic specification in order to model the homeownership decision. Control function methods such as `ivprobit` are not appropriate as `homeowner` is a discrete variable.

The sample used includes male heads of household, 23–59 years of age, from the 1990 wave of the PSID who have completed education and are not retired, so as to exclude those moving to retirement communities. The observed $D = 1$ indicates migration during 1991–1993. In the sample of 4689 individuals, 807 were interstate migrants.

Exogenous regressors in the model include years of education, number of children, and indicators for white, disabled, and married individuals. The instruments Z also include the level of government benefits received in 1989–1990 and state median residential tax rates.

In the following table, we present four sets of estimates of the marginal effects computed by `ssimplereg`, utilizing the sorted data density estimator in columns 2 and 4 and allowing for heteroskedastic errors in columns 3 and 4.

For contrast, we present the results from an IV LPM (`ivregress 2sls`) in column 5, a standard `probit` (ignoring endogeneity) in column 6, and an `ivprobit` in the last column, ignoring its lack of applicability to the binary endogenous regressor `homeowner`.

Table: Marginal effects: binary outcome, binary endogenous regressor

	kdens	sortdens	kdens_hetero	sortdens_hetero	IV-LPM	probit	ivprobit
age	0.0146	0.0112	0.0071	0.0104	-0.0010	0.0019	-0.0005
	(0.003)***	(0.003)***	(0.003)*	(0.003)***	(0.002)	(0.001)**	(0.007)
log income	-0.0079	0.0024	0.0382	0.0176	0.0550	-0.0089	0.1406
	(0.028)	(0.027)	(0.024)	(0.026)	(0.080)	(0.007)	(0.286)
homeowner	0.0485	-0.0104	-0.0627	-0.0111	-0.3506	-0.0855	-1.0647
	(0.072)	(0.065)	(0.059)	(0.061)	(0.204)	(0.013)***	(0.708)
white	0.0095	0.0021	0.0021	0.0011	0.0086	-0.0099	0.0134
	(0.008)	(0.010)	(0.007)	(0.008)	(0.018)	(0.012)	(0.065)
disabled	0.1106	0.0730	0.0908	0.0916	0.0114	-0.0122	0.0104
	(0.036)**	(0.042)	(0.026)***	(0.037)*	(0.055)	(0.033)	(0.203)
education	-0.0043	-0.0023	-0.0038	-0.0036	0.0015	0.0004	0.0047
	(0.002)*	(0.003)	(0.002)*	(0.002)	(0.004)	(0.002)	(0.015)
married	0.0628	0.0437	0.0258	0.0303	0.0322	-0.0064	0.0749
	(0.020)**	(0.028)	(0.013)	(0.020)	(0.031)	(0.017)	(0.114)
nr. children	-0.0169	-0.0117	0.0006	-0.0021	0.0137	0.0097	0.0502
	(0.005)***	(0.005)*	(0.002)	(0.003)	(0.006)*	(0.005)*	(0.023)*

Note: bootstrapped standard errors in parentheses (100 replications)

The standard errors of these estimated marginal effects are computed from 100 bootstrap replications. The marginal effect of the ‘special regressor’ age of head is estimated as positive by the special regressor methods, but both the two-stage linear probability model and the ivprobit model yield negative (but insignificant) point estimates.

Household income and homeownership status do not seem to play significant roles in the migration decision. Among the special regression methods, the kernel data density estimator appears to yield the most significant results, with age of head, disabled status, years of education, marital status and number of children all playing a role in predicting the migration decision.

The standard errors of these estimated marginal effects are computed from 100 bootstrap replications. The marginal effect of the ‘special regressor’ age of head is estimated as positive by the special regressor methods, but both the two-stage linear probability model and the ivprobit model yield negative (but insignificant) point estimates.

Household income and homeownership status do not seem to play significant roles in the migration decision. Among the special regression methods, the kernel data density estimator appears to yield the most significant results, with age of head, disabled status, years of education, marital status and number of children all playing a role in predicting the migration decision.

Conclusions

We have discussed an alternative to the linear probability model for estimation of a binary outcome with one or more binary endogenous regressors. This alternative, Lewbel and Dong's 'simple special regressor' method, circumvents the drawbacks of the IV-LPM approach, and yields consistent estimates in this context in which `ivprobit` does not. Computation of marginal effects via the proposed average index function approach is straightforward, requiring only a single kernel density estimation and no iterative techniques. Bootstrapping is employed to derive interval estimates.

A Stata implementation of the simple special regressor method, `sspecialreg`, is being refined to take advantage of Stata's flexibility and Mata's potential for speed improvements. The routine will also be extended to the context of panel data. This routine will soon be made available to users via the SSC Archive.

Conclusions

We have discussed an alternative to the linear probability model for estimation of a binary outcome with one or more binary endogenous regressors. This alternative, Lewbel and Dong's 'simple special regressor' method, circumvents the drawbacks of the IV-LPM approach, and yields consistent estimates in this context in which `ivprobit` does not. Computation of marginal effects via the proposed average index function approach is straightforward, requiring only a single kernel density estimation and no iterative techniques. Bootstrapping is employed to derive interval estimates.

A Stata implementation of the simple special regressor method, `sspecialreg`, is being refined to take advantage of Stata's flexibility and Mata's potential for speed improvements. The routine will also be extended to the context of panel data. This routine will soon be made available to users via the SSC Archive.