Binary Choice With Endogenous Or Mismeasured Regressors

Arthur Lewbel

Boston College

Binary Choice With Endogenous or Mismeasured Regressors - Outline

1. The model and normalizations

2. Comparison of Parametric and Semiparametric Estimators
   a. Linear Probability Model
   b. Maximum Likelihood
   c. Control Functions
   d. Special Regressor (at length)

3. Fitted choice probabilities and marginal effects

4. Simultaneity, Coherence, and Completeness

This lecture is based primarily on:

Lewbel, A., Y. Dong, and T. Yang, (2012) "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors," forthcoming, Canadian Journal of Economics.

Dong, Y. and A. Lewbel, (2012) "A Simple Estimator for Binary Choice Models with Endogenous Regressors," forthcoming, Econometrics Reviews.

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

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

The model

D = Binary (dummy) dependent variable, e.g. decision to migrate or to buy a car.
X = Vector of regressors, e.g., income, age, demographics, treatment indicator.
$\beta, \widetilde{\beta}$ = Vectors of coefficients.
ε = error term (unobserved heterogeneity).

Standard threshold crossing binary choice model (e.g., logit or probit):
$D = I\left(X'\beta + \varepsilon \geq 0\right)$

Exception: linear probability model: $D = X'\widetilde{\beta} + \varepsilon$

When convenient to separate out one regressor, will call that regressor V and write the model as $D = I\left(V\alpha + X'\beta + \varepsilon \geq 0\right)$

Binary Choice Topics NOT Covered Here

- Nonlinear or nonparametric indices in place of $X'\beta$.
- Nonadditive or nonseparable error ε.
- Dynamic choice, binary choice panels, multinomial choice, ordered choice.
- Bounds or set identified models.
- Binary choice of strategies within games.
- Reduced form treatment effect / program evaluation (though note these are often equivalent to linear probability models).

However, most of the estimators here can be extended to include most of the above topics.

Normalizations

With $D = I\left(X'\beta + \varepsilon \geq 0\right)$, β and ε moments are only identified up to location and scale.

Scale: Fit stays the same if replace β and ε with $\beta\lambda$ and $\varepsilon\lambda$ for any $\lambda > 0$.

Location: Let β_0 be the constant term in $X'\beta$. Fit stays the same if replace β_0 and ε with $\beta_0 + \kappa$ and $\varepsilon - \kappa$ for any $\kappa > 0$.

Probit chooses location and scale by setting $E\left(\varepsilon\right) = 0$, $var\left(\varepsilon\right) = 1$.

Semiparametric often chooses location and scale by setting $\beta_0 = 0$ and coefficient α of some regressor (call it V) equal to one.

Other estimators have different normalizations, e.g., average derivative estimation, maximum score.

Normalizations - continued

Choice of location and scale normalizations has no effect on estimated choice probabilities or marginal effects.

Sometimes choice of normalization has an economic interpretation. Suppose the model is $D = I\left(V\alpha + X'\beta + \varepsilon\right)$, where D is a purchase decision and V is $-\left(price\right)$. If normalize scale to make $\alpha = 1$, and location to make $E\left(\varepsilon\right) = 0$, then $x'\beta$ is average willingness to pay (reservation price) for consumer of type $X = x$.

When comparing estimators, write both in terms of marginal effects, or convert to same normalization. E.g. If a semiparametric normalizes $\alpha = 1$, then given a probit $\widehat{\beta}$ can calculate $\widehat{\beta}/\widehat{a}$ to compare.

Estimators for Models with Endogenous regressors

Partition X into two vectors:
X^0 = Exogenous regressors (uncorrelated with ε), e.g., age.
X^e = Endogenous regressors (correlated with ε), e.g., income.
Similar for coefficients β_e, β_o or $\widetilde{\beta}_e$, $\widetilde{\beta}_o$
Z = Vector of instruments (includes X^0), e.g. gov't defined benefits.

For Linear probability model $D = X'\widetilde{\beta} + \varepsilon = X^{e\prime}\widetilde{\beta}_e + X^{o\prime}\widetilde{\beta}_o + \varepsilon$

For Other estimators: $D = I\left(X'\beta + \varepsilon \geq 0\right) = I\left(X^{e\prime}\beta_e + X^{o\prime}\beta_o + \varepsilon \geq 0\right)$

Assume for now triangular system: $X^e = g\left(Z, e\right)$ or latent simultaneous $X^e = G^*\left(Z, e^*, X'\beta + \varepsilon\right)$ which has reduced form $X^e = g\left(Z, e\right)$.

But not $X^e = G\left(D, Z, e\right)$.

The Linear Probability Model - LPM

Assume $D = X'\widetilde{\beta} + \varepsilon$, $E(Z\varepsilon) = 0$, $\text{rank}(ZX') = \dim\left(\widetilde{\beta}\right)$
Estimator: linear 2SLS with instruments Z.

LPM has all the advantages of linear 2SLS:

- $X^e = g\left(Z, e\right)$ does not need to be specified.
- Computationally easy, no numerical searches required (nice for bootstrapping).
- Omitting some instruments inefficient, not inconsistent.
- X^e can include discrete, continuous, limited, etc.,
- Estimator is the same regardless of whether X^e has discrete or continuous elements.
- Allows general heteroskedasticity, e.g., random coefficients.

The Linear Probability Model - LPM - continued

$$D = X'\widetilde{\beta} + \varepsilon,\ E(Z\varepsilon) = 0,\ \text{rank}(ZX') = \dim\left(\widetilde{\beta}\right).$$

Disadvantages of linear probability models:

- Fitted distribution function $\widehat{F}_D$ is linear, not S shaped; an adequate approx only for a limited range of X.
- Often implausibly extreme $\widehat{F}_D$, easily negative or above one.
- Formally, no element of X can have ∞ support, e.g., no normal regressors.
- ε not independent of any regressor, even exogenous ones, because for any value of X, must have ε equal $1 - X'\widetilde{\beta}$ or $-X'\widetilde{\beta}$. What economic model justifies both this dependence and $E\left(X^0\varepsilon\right) = 0$?
- Does not nest probit, logit, etc., can't compare efficiency.

The Linear Probability Model - LPM - continued

The LPM can get signs wrong.
Suppose $D = I\left(1 + T + R + \varepsilon \geq 0\right)$, true coefficients are all 1.
$T =$ binary treatment indicator, $R =$ a regressor, ε is small.
True treatment effect is $I\left(2 + R + \varepsilon \geq 0\right) - I\left(1 + R + \varepsilon \geq 0\right) \geq 0$
equals 0 or 1 for everybody, is never negative in this model.

Suppose data are: $R_1 = -1.8$, $R_2 = -0.9$, $R_3 = -0.92$, $R_4 = -2.1$, $R_5 = -1.92$, $R_6 = 10$, $T_1 = 0$, $T_2 = 0$, $T_3 = 0$, $T_4 = 1$, $T_5 = 1$, $T_6 = 1$.
Given tiny errors, get outcomes $D_1 = 0$, $D_2 = 1$, $D_3 = 1$, $D_4 = 0$, $D_5 = 1$, $D_6 = 1$.

Half the sample are treated, average treatment effect ATE is $1/6$.
LPM $D = \beta_0 + T\beta_1 + R\beta_2 + \varepsilon$ by OLS has $\widehat{\beta}_1 = -0.16$, MRS $\widehat{\beta}_1 / \widehat{\beta}_2 = -3.2$.

LPM estimates a negative treatment effect, even though every single person actually has a zero or positive effect!

Maximum Likelihood Estimation - ML

Assume $D = I\left(X'\beta + \varepsilon \geq 0\right)$, and $X^e = G\left(Z, \theta, e\right)$
vector function G fully specified, parameterized, e.g., G could be linear if X^e is continuous, or a probit if X^e is binary.
Joint distribution of $(\varepsilon, e \mid Z)$ fully specified, parameterized.
Implementation: see standard textbooks like Greene (2008) or Wooldridge (2010).

Advantages of ML:

- Nests standard logit, probit, etc., as special cases
- X^e can include discrete, continuous, limited, etc.,
- Allows general heteroskedasticity, e.g., random coefficients.
- Asymptotically efficient (requires and uses the most info).

ML - continued

$D = I\left(X'\beta + \varepsilon \geq 0\right)$, and $X^e = G\left(Z, \theta, e\right)$

Disadvantages of ML:

- Need to parameterize everything: G and $F_{\varepsilon,e|Z}$
- Numerical problems are common (ridges, failure to converge, multiple solutions).
- Many nuisance parameters, sometimes poorly identified, e.g. correlations of latent ε with e. Particular problem when some X^e are also discrete.
- Need to know all the required instruments Z. Omitting any causes inconsistency from G misspecification. (Not sure something's exogenous? Too bad).

Control Function Estimation - CF

Assume $D = I\left(X'\beta + \varepsilon \geq 0\right)$,
$X^e = G\left(Z\right) + e$, or $X^e = G\left(Z, e\right)$ identified and invertible in e.
$\varepsilon = \lambda' e + U$ or $\varepsilon = H\left(U, e\right)$ with conditions, and $U \perp Z, e$.

e.g. Stata ivprobit has $G\left(Z, e\right)$ linear, and $\left(e, \varepsilon\right)$ jointly normal and independent of Z.

Simplest CF Estimator:

1. Estimate vector of functions G in the X^e models, get estimated errors $\widehat{e}$.

2. Estimate the D model including $\widehat{e}$ as additional regressors in addition to X. The errors U in this model are now clean.

Control Function Estimation continued

$D = I(X'\beta + \varepsilon \geq 0)$, $X^e = G(Z, e)$, $\varepsilon = H(U, e)$, $U \perp X, e$

Key requirements for CF estimators:

- Must be able to solve for errors e in X^e equations.
- Must have endogeneity caused only by ε related to e, so after conditioning on e have ε no longer depend on X^e.

Advantages of CF Estimation:

- Nests standard logit, probit, etc., as special cases.
- Requires less info than ML.
- Some versions are computationally easy not needing numerical searches (nice for bootstrapping)
- Less efficient than ML due to less info used, but CF sometimes semiparametrically efficient relative to the info used.

Disadvantages of Control Functions (not well known)

$D = I\left(X'\beta + \varepsilon \geq 0\right)$, $X^e = G\left(Z, e\right)$, $\varepsilon = H\left(U, e\right)$, $U \perp X, e$

- Only permits limited forms of heteroskedasticity.
- Need to correctly specify the vector of high dimensional functions $G\left(Z, e\right)$, including knowing Z. Omitting instruments or wrong G can cause inconsistency, because e needs to satisfy joint conditions with ε.
- CF is usually inconsistent when any endogenous regressor in X^e is discrete, censored, limited, or otherwise not continuously distributed!

Can't solve for a latent e in $G\left(Z, e\right)$. E.g. if $X^e = \max\left(0, Z'\gamma + e\right)$ then can't get $\widehat{e}$ for the censored at zero observations

An observable e is $e = X^e - E\left(X^e \mid Z\right)$, but for discontinuous X^e that e violates assumptions (except in freaky special cases). Example: $\varepsilon = \left[X^e - E\left(X^e \mid Z\right)\right]\lambda + U$ satisfies CF, but if X^e is discrete this forces ε to have a strange distribution that depends on all regressors. What behavioral model yields this ε?

Example of Control Function issues - IVPROBIT in STATA

ivprobit: $D = I\left(X'\beta + \varepsilon \geq 0\right)$, $X^e = Z'\gamma + e$, $(e, \varepsilon) \perp Z$, normal.

Despite the name, ivprobit is not an iv estimator. It's a control function.

ivprobit uses $D = I\left(X^e\beta_e + e\lambda + U \geq 0\right)$ where U is a homoskedastic normal, Estimate as a probit after plugging in first stage e.

1. Let $Z'\gamma = Z_1\gamma_1 + Z_2\gamma_2$. What if you leave out instrument Z_2? Get $X^e = Z_1\widetilde{\gamma}_1 + \widetilde{e}$, and $D = I\left(X^e\beta_e + \widetilde{e}\widetilde{\lambda} + \widetilde{U} \geq 0\right)$ where $\widetilde{U} = U + \delta Z_2$. So if Z_2 is not normal then $\widetilde{U}$ is not normal and generally heteroskedastic. Dropping an instrument causes the model to no longer be probit.

2. What if X^e is discrete or censored or limited? Then again in general e is not normal and not independent of Z, violating ivprobit assumptions.

Stata ivprobit has two estimation options: two step or MLE. Both are generally inconsistent when X^e is not continuous or when an instrument is left out.

Control Functions and Generalized Residuals

Can control functions ever be used when the endogenous regressor is discrete? Yes, but...

Given a probit estimate of the endogenous regressor, it is sometimes possible to construct a "generalized residual," e^g (see, e.g., Imbens and Wooldridge lecture notes). This e^g is constructed to be proportional to $E(\varepsilon \mid Z, e)$. An estimate $\widehat{e}^g$ of e^g can be included as a regressor in the model to fix the endogeneity problem, just as $\widehat{e}$ would have been used if the endogenous regressor were continuous.

Problem: In linear models this is unnecessary, since IV will work there with far fewer restrictions. But in nonlinear models, like our discrete choice model, constructing $\widehat{e}^g$ requires almost the same assumptions as ML, so in that case better to do ML which is efficient.

Special Regressor Estimation - Literature

Binary, ordered, and multinomial choice, censored regression, selection, and treatment models (Lewbel 1998, 2000, 2007a), truncated regression models (Khan and Lewbel 2007), binary panel models with fixed effects (Honore and Lewbel 2002), dynamic choice models (Heckman and Navarro 2007, Abbring and Heckman 2007), contingent valuation models (Lewbel, Linton, and McFadden 2008), market equilibrium models of multinomial choice (Berry and Haile 2009a, 2009b), models with (partly) nonseparable errors (Lewbel 2007b, Matzkin 2007, Briesch, Chintagunta, and Matzkin 2009).

Other empirical applications: Anton, Fernandez Sainz, and Rodriguez-Poo (2002), Cogneau and Maurin (2002), Goux and Maurin (2005), Stewart (2005), Lewbel and Schennach (2007), and Tiwari, Mohnen, Palm, and van der Loeff (2007).

Precursors: Matzkin (1992, 1994) and Lewbel (1997).

Recent theory: Magnac and Maurin (2007, 2008), Jacho-Chávez (2009), Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b).

Special Regressor (SR) Estimation - overview

To explain SR identification, consider the model where the only regressors are a constant term and single continuous exogenous regressor V:

$D = I(\alpha + V + \varepsilon \geq 0)$, $\varepsilon \perp V$, unknown α, unknown F_ε.

Let $F_{-\alpha-\varepsilon}()$ be CDF of $-\alpha - \varepsilon$.

$$E(D \mid V = v) = \Pr(D = 1 \mid V = v) = \Pr(-\alpha - \varepsilon \leq v) = F_{-\alpha-\varepsilon}(v).$$

Estimating $E(D \mid V = v)$ gives $F_{-\alpha-\varepsilon}(v)$. From knowing $F_{-\alpha-\varepsilon}(v)$ for all $v \in (-\infty, \infty)$ can calculate the mean of $\alpha + \varepsilon$, which is α, and knowing both α and $F_{-\alpha-\varepsilon}$ gives F_ε.

If ε is bounded, only need $v \in [\alpha + \min \varepsilon, \alpha + \min \varepsilon]$

If tails of ε are symmetric, even smaller range of v works, because the unestimated upper and lower tails of $F_{-\alpha-\varepsilon}$ can cancel each other out in calculating the mean (Magnac and Maurin 2007).

Special Regressor (SR) Estimation - overview continued

With $D = I(\alpha + V + \varepsilon \geq 0)$, could estimate $E(D \mid V = v)$ by a nonparametric regression, then obtain α by

$$\int_{\Omega_V} -v \frac{\partial E(D \mid V = v)}{\partial v} dv = -\int_{\Omega_V} v \frac{\partial F_{-\alpha-\varepsilon}(v)}{\partial v} dv$$

$$= -\int_{\Omega_V} v f_{-\alpha-\varepsilon}(v)\, dv = -E(-\alpha - \varepsilon) = \alpha$$

This shows identification, and provides a way one could estimate α. But there is an easier estimator.

Special Regressor (SR) Estimation - overview continued

Define $T = \dfrac{D - I(V \geq 0)}{f_v(V)}$ where $f_v()$ is pdf of V. Will show $E(T) = \alpha$

$$E(T) = E\left[E\left(\frac{D - I(V \geq 0)}{f_v(V)} | V = v\right)\right] \quad \text{by iterated expectations}$$

$$= E\left(\frac{(D \mid V = v) - I(v \geq 0)}{f_v(v)}\right) = \int_{\Omega_V} \frac{E(D \mid V = v) - I(v \geq 0)}{f_v(v)} f_v(v)\, dv$$

$$= \int_{\Omega_V} [E(D \mid V = v) - I(v \geq 0)]\, dv. \quad \text{Integrate by parts:}$$

$$= v[E(D \mid V = v) - I(v \geq 0)]|_{\Omega_V} - \int_{\Omega_V} v \frac{\partial E(D \mid V = v)}{\partial v} dv$$

$$= v[F_{-\alpha-\varepsilon}(v) - I(v \geq 0)]|_{\Omega_V} + \alpha = \alpha$$

So easy estimator is $\widehat{\alpha} = \overline{T}$, sample average of T.

Special Regressor (SR) Estimation - Tail Symmetry

Let $A = E(T) = -\int_{\Omega_V} v f_{-\alpha-\varepsilon}(v)\, dv$. We showed $A = \alpha$, assuming big support: $\Omega_V \subseteq supp(-\alpha - \varepsilon)$.

Suppose $\Omega_V = [\ell, u]$ and $supp(-\alpha - \varepsilon) = [\ell - L, u + U]$. If L or U is positive then the support of V isn't big enough. In this case

$$\alpha = \int_{\ell-L}^{u+U} v f_{-\alpha-\varepsilon}(v)\, dv = \delta + A$$

where

$$\delta = \int_{\ell-L}^{L} v f_{-\alpha-\varepsilon}(v)\, dv + \int_{u}^{u+U} v f_{-\alpha-\varepsilon}(v)\, dv$$

Tail symmetry is when $\delta = 0$, so $\alpha = E(T)$ and the special regressor works even though the support of V is not large enough (Magnac and Maurin 2007).

If support of V is reasonably large, and tails of ε are thin or close to symmetric, then δ will be small and so bias will tend to be small even if tail symmetry or support is violated.

Special Regressor Estimation - SR

Assume $D = I(V + X'\beta + \varepsilon \geq 0)$, $E(Z\varepsilon) = 0$, $\text{rank}(ZX') = \dim(\beta)$, $\varepsilon \perp V \mid Z, X^e$, $supp(-X'\beta - \varepsilon) \subseteq supp(V)$ (or tail symmetry), $V \mid Z, X^e$ continuously distributed.

One regressor V is "strongly" exogenous (conditionally independent of ε) and continuously distributed with a large support or tail symmetry.

All instruments Z, and all other regressors X (including all the endogenous regressors X^e) satisfy only the same conditions as linear 2SLS: $E(Z\varepsilon) = 0$ and $\text{rank}(ZX') = \dim(\beta)$.

SR Estimator:

1. Construct $T = [D - I(V \geq 0)] / f_v(V \mid Z, X^e)$.
   Can show that $T = X'\beta + \varepsilon^*$ where $E(Z\varepsilon^*) = 0$.
2. Do linear 2SLS regression of T on X using instruments Z to get β.

Special Regressor Estimation - SR

$D = I(V + X'\beta + \varepsilon \geq 0)$, $E(Z\varepsilon) = 0$, $\text{rank}(ZX') = \dim(\beta)$,
$\varepsilon \perp V \mid Z, X^e$, $supp(-X'\beta - \varepsilon) \subset supp(V)$, $V \mid Z, X^e$ continuous.

Disadvantages of SR estimation:

- To construct $T = [D - I(V \geq 0)] / f_v(V \mid Z, X^e)$ need an estimator of f_v, the conditional density function of V. Will later give parametric, semiparametric, and nonparametric examples.
- Need V to have thick tails relative to ε, else T estimator may behave badly (large outliers and/or slow convergence rate). Formally, need $var(V)$ infinite *or* $supp(-X'\beta - \varepsilon)$ finite *or* tail symmetry. Informally, biases tend to be small when $var(V)$ comparable to $var(-X'\beta - \varepsilon)$, or when ε is thin tailed and/or symmetric ε.
- Need V 'strongly' exogenous (will discuss later).
- V must appear linearly in the model (e.g., X^o can't include V^2, because $V \mid V^2$ is not continuous.

Special Regressor Estimation - SR

(The Sales Pitch): SR has almost all of the advantages and none of the disadvantages of all the other estimators.

Advantages of SR Estimation:

- $X^e = g(Z, e)$ does not need to be specified, only need $E(Z\varepsilon)$.
- Omitting some instruments is inefficient, not inconsistent.
- X^e can include discrete, continuous, limited, etc.,
- Estimator is the same regardless of whether X^e has discrete or continuous elements.
- Allows general heteroskedasticity, e.g., random coefficients.
- Computationally easy, no numerical searches required (nice for bootstrapping)
- Nests standard logit, probit etc.,
- Requires less info than ML and CF (except info on V).
- Less efficient than ML and CF due to less info used, but some versions are semiparametrically efficient relative to the info used.

Implementing the SR estimator

$D = I(V + X'\beta + \varepsilon \geq 0)$, $E(Z\varepsilon) = 0$, $\text{rank}(ZX') = \dim(\beta)$,
$\varepsilon \perp V \mid Z, X^e$, $supp(-X'\beta - \varepsilon) \subset supp(V)$, $V \mid Z, X^e$ continuous.

Original SR Estimator (Lewbel 2000):

1. Demean or otherwise center V at zero. By exogeneity assume conditional density $f_v(V \mid Z, X^e) = f_v(V \mid Z)$ and let $\widehat{f}_v(V \mid Z)$ be a nonparametric kernel estimator of $f_v(V \mid Z)$. Or just use kernel estimator of $\widehat{f}_v(V \mid Z, X^e)$.

2. For each observation i, Construct $\widehat{T}_i = I[D_i - I(V_i \geq 0)] / \widehat{f}_v(V_i \mid Z_i)$.

3. Do a linear two stage least squares regression of $\widehat{T}$ on X using instruments Z to get the estimated coefficients $\widehat{\beta}$.

Here $\widehat{f}_v(V \mid Z)$ is high dimensional. So will now consider some simpler parametric or semiparametric specifications for f_v.

Numerically Trivial SR estimator

Let $S = Z, X^e$, or just Z if V doesn't depend on X^e. Assume

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

$$supp(-S'b - X'\beta - \varepsilon) \subseteq supp(U) \text{ or tail symmetry}$$

$f(U)$ is a mean or median zero pdf. $T = [D - I(V \geq 0)] / f(U)$. Then $T = X'\beta + \widetilde{\varepsilon}$ where $E(Z\widetilde{\varepsilon}) = 0$.

V is modeled as linear in covariates S and an independent error U.

1. Demean or demedian V_i. Estimate $\widehat{b}$ by linear OLS $V_i = S_i'b + U_i$. Construct residuals $\widehat{U}_i = V_i - S_i'\widehat{b}$.
2. Estimate density f of $\widehat{U}_i$.
3. Construct data $\widehat{T}_i = D_i - I(V_i \geq 0) / \widehat{f}\left(\widehat{U}_i\right)$.
4. Do a linear 2SLS of $\widehat{T}$ on X using instruments Z to get the estimated coefficients $\widehat{\beta}$.

This simple SR estimator needs a one dimension density $\widehat{f}\left(\widehat{U}_i\right)$. Possible estimators include:

A parametric estimator, or standard one dimensional kernel density estimator (e.g., kdensity command in Stata), or:

Sorted data estimator (Lewbel and Schennach 2007):
1. Sort $\widehat{U}_i$, from lowest to highest.
2. For each $\widehat{U}_i$, 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 $\widehat{U}_i$.
3. Let

$$\widehat{f}\left(\widehat{U}_i\right) = \frac{2/n}{\widehat{U}_i^+ - \widehat{U}_i^-} \qquad \text{for } i = 1, ..., n$$

No searching, kernel, or bandwidth.

Special Regressor Estimation - Asymptotics

Simplest: Since estimators are numerically trivial and no numerical searches are required, bootstrapping is practical and easy (for iid data, draw all the variables with replacement and repeat all the estimation steps with each drawn data set).

Another option: If the density f is parameterized, can do all the steps either sequentially or jointly as a GMM estimator.

Example: Numerically trivial estimator with normal f is estimating desired parameters β with nuisance parameters b, σ^2 using moments

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

note: can replace $I(V \geq 0)$ with $F(V)$ for any distribution function F. This makes objective function smoother.

Interpreting Exogeneity of V Assumption:

In the simple estimator, exogeneity is $U \perp S, \varepsilon$ where $V = g(S) + U$ and S is all other regressors X and instruments Z. Simplest estimator has $g(S) = S'b$, but can be any g.

$U \perp \varepsilon$ is ordinary exogeneity, being independent of the model error term.

Having $V = g(S) + U$ divides V into a fitted part $g(S)$ and a residual U. This decomposition of V into $g(S)$ and U is just a regression, not a structural model.

U independent of S is stronger. This requires that, if any endogenous regressors in the model depend on V, they do so only by depending on the fitted part $g(S)$ and not on the residual U.

Special Regressor Estimation - Tails and Asymptotics

SR involves dividing by f, the density of V, which can cause the constructed variable T to have thick tails. Implications:

- Standard root n limiting distribution theory requires V to have infinite variance OR $X'\beta + \varepsilon$ have bounded support OR ε satisfies tail symmetry.
- Without root n, convergence rates can be slow, requiring the limiting distribution theory given by Khan and Tamer (2010), Khan and Nekipelov (2010) or Hill and Renault (2010).
- In practice, should check for outliers in the final 2SLS step, maybe trim the data (discarding outliers) to robustify.
- In practice, compare variance, spread of V versus $X'\widehat{\beta}$. Small sample biases tend to be small when spread of V is relatively large.

Special Regressor Estimation - An Example
Empirical estimates are in Dong and Lewbel (2011).

D = one if move (migrate), zero if not. V = age.
X^e = income before moving, homeownership dummy before moving.
X^o = demographic characteristics.
Z = government benefits, state median residential property tax rates.

$$D = I\left(V + X^{e\prime}\beta_e + X^{o\prime}\beta_o + \varepsilon \geq 0\right).$$

ML impractical, need much more info to fully model endogenous income and homeownership.

CF assumptions violated by homeownership binary and not having all instruments Z needed to get independent errors in models of X^e.

SR is ok. Age (or residual U in regression of Age on X^e and Z) is exogenous with large support. Empirically and by human capital theory, the returns to moving (present value of associated wage gains) decline linearly with age V.

A Few Other Estimators

Vytlacil and Yildiz (2007): If $D = I\left(X^{e\prime}\beta_e + X^{o\prime}\beta_o + \varepsilon \geq 0\right)$ and X^e is a single binary dummy, treat $X^{o\prime}\beta_o$ like the special regressor V. Then $D = I\left(X^e\beta_e + V + \varepsilon \geq 0\right)$. With instruments Z to hold $E\left(X^e \mid Z\right)$ fixed, shows only need $supp\left(V\right) \supseteq supp\left(-X^e\beta_e\right)$ instead of $supp\left(V\right) \supseteq supp\left(-X^e\beta_e - \varepsilon\right)$. If β_e can be anything, then still need infinite support for V.

Maximum Score (Manski 1975, 1985, Horowitz 1992) mainly deals with heteroskedasticity in ε, not endogeneity. Hong and Tamer (2003) extend to handle endogeneity assuming median$(\varepsilon \mid Z)$, but Shaikh and Vytlacil (2008) show point identification requires strong restrictions on β_e and on $X^e \mid Z$.

Quantile extensions of control functions can cover endogeneity and heteroskedasticity, e.g, Chesher (2009) and Hoderlein (2009).

Yet More Other Estimators

Nonseparable errors: Altonji and Matzkin (2005), Imbens and Newey (2009), many, others. ML, CF, and SR all have nonseparable error extensions.

Bounds and set identification: Manski (2007), Magnac and Maurin (2008), Chesher (2010), many, many other.

Reduced form treatment effects and program evaluation with binary outcomes and endogenous treatment. See Angrist and Pischke (2008) and many, many, many others. Note that the standard ATE model is equivalent to the linear probability model variant $D = g\left(X^0\right) + h\left(X^0\right) X^e + \varepsilon$ where X^e is an endogenous binary treatment indicator.

Fitted Choice Probabilities

With exogenous regressors X, the propensity score or choice probability is defined as $p(X) = \Pr(D = 1 \mid X) = E(D \mid X)$.

If $D = I(X'\beta + \varepsilon \geq 0)$ with $\varepsilon \perp X$, then

$$p(X) = E(D \mid X) = E\left(D \mid X'\beta\right) = F_{-\varepsilon}\left(X'\beta\right)$$

where $F_{-\varepsilon}$ is the marginal distribution function of $-\varepsilon$.

Given $\widehat{\beta}$, could estimate $\widehat{p}(X)$ by one dimensional nonparametric regression of D on $X'\widehat{\beta}$

Fitted Choice Probabilities - continued

For ε independent of X, had $E(D \mid X) = E(D \mid X'\beta) = F_{-\varepsilon}(X'\beta)$.

What if $D = I(X'\beta + \varepsilon \geq 0)$ but ε is *not* independent of X, due to heteroskedasticity or endogenous regressors? Then all three are different:

$E(D \mid X)$ is propensity score. This is a high dimensional nonparametric regression of D on X. It ignores β.

$F_{-\varepsilon}(X'\beta)$ is ASF (average structural function, Blundell and Powell 2000). ASF is the marginal distribution $F_{-\varepsilon}$ evaluated at $X'\beta$. The $F_{-\varepsilon}$ function can be difficult to identify and estimate because ε is latent.

$E(D \mid X'\beta)$ is AIF (average index function, Dong and Lewbel 2011). AIF equals $E(D \mid X'\beta)$. This is an easy to estimate one dimensional nonparametric regression of D on $X'\beta$.

Fitted Choice Probabilities - continued

Propensity score: $E(D \mid X) = F_{-\varepsilon|X}(X'\beta \mid X)$.

AIF: $E(D \mid X'\beta) = F_{-\varepsilon|X'\beta}(X'\beta \mid X'\beta)$.

ASF: $F_{-\varepsilon}(X'\beta)$

- AIF (conditioning on $X'\beta$) is a middle ground between propensity score (conditioning on all of X) and ASF (conditioning on nothing): $F_{-\varepsilon|X}$ vs $F_{-\varepsilon|X'\beta}$ vs $F_{-\varepsilon}$.
- Unlike propensity score, AIF uses β and doesn't require high dimensional estimation.
- Unlike ASF, AIF is always identified and easy to estimate.
- ASF, AIF and propensity score are all identical under exogeneity.

Fitted Choice Probabilities - Marginal Effects

With exogenous X: marginal effects are $m(X) = p'(X) = \dfrac{\partial E(D \mid X)}{\partial X}$

Let $f_{-\varepsilon}$ be marginal pdf of $-\varepsilon$. If $D = I(X'\beta + \varepsilon \geq 0)$ with $\varepsilon \perp X$, then

$$m(X) = \frac{\partial E(D \mid X)}{\partial X} = \frac{\partial E(D \mid X'\beta)}{\partial X'\beta}\beta = f_{-\varepsilon}(X'\beta)\beta$$

With endogenous X:

Propensity score marginal effects are $m(X) = p'(X) = \dfrac{\partial E(D \mid X)}{\partial X}$.

ASF marginal effects are $m(X) = \dfrac{\partial \text{ASF}(X'\beta)}{\partial X'\beta}\beta = f_{-\varepsilon}(X'\beta)\beta$.

AIF marginal effects are $m(X) = \dfrac{\partial \text{AIF}(X'\beta)}{\partial X'\beta}\beta = \dfrac{\partial E(D \mid X'\beta)}{\partial X'\beta}\beta$.

Given $\widehat{\beta}$, and ASF and AIF marginal effects just require one dimensional index derivative.

Simultaneity, Coherence, and Completeness

Consider a trivial simultaneous system:

$$D = I(Y + e_1 \geq 0), \qquad Y = \alpha D + e_2$$

$$\text{Then} \quad D = I(\alpha D + e_1 + e_2 \geq 0)$$

If $D = 0$ then $Y = e_2$ and $0 = I(e_1 + e_2 \geq 0)$ so $e_1 + e_2 < 0$

If $D = 1$ then $Y = \alpha + e_2$ and $1 = I(\alpha + e_1 + e_2 \geq 0)$, so $\alpha + e_1 + e_2 \geq 0$

Incomplete: both $D = 0$ and $D = 1$ are solutions if $-\alpha \leq e_1 + e_2 < 0$.
Incoherent: Neither $D = 0$ nor $D = 1$ hold if $0 \leq e_1 + e_2 < -\alpha$.

This model is incoherent or incomplete unless $e_1 + e_2$ is constrained not to lie between zero and $-a$.

Simultaneity, Coherence, and Completeness - continued

Incoherence: For parameter values, the model is not internally consistent.

A model must be coherent before one should even consider identification and estimation.

Incompleteness: For some values of parameters, the same values of all exogenous variables (both observed and unobserved) can yield multiple values of the endogenous variables.

Multiple equilibria in games is an example of incompleteness. Usually causes identification problems and requires special estimators.

Heckman (1978), Gourieroux, Laffont, and Monfort (1980), Blundell and Smith (1994), Dagenais (1997), Bresnahan and Reiss (1991), Tamer (2003), Aradillas-Lopez (2005), Lewbel (2007).

Theorems that follow are from Lewbel (2007 IER McFadden Festschrift).

Simultaneity, Coherence, and Completeness - continued

Let W be a vector of observed or unobserved exogenous variables. With $D \in \{0, 1\}$ consider the system

$$D = H_1(D, Y, W), \qquad Y = H_2(D, Y, W)$$

Theorem (Lewbel 2007): This system of equations is coherent and complete iff for some g:

$$H_1[1, g(1, W), W] = H_1[0, g(0, W), W], \qquad Y = g(D, W)$$

To prove, solve for reduced form equation $Y = g(D, W)$, substitute it into

H_1, and show incoherence or incompleteness whenever $H_1[D, g(D, W), W]$ is not the same for both values of D.
Note here Y is continuous or discrete or limited. Requirement that $H_1[D, g(D, W), W]$ not depend on D is very restrictive.

Simultaneity, Coherence, and Completeness - continued

Suppose the D model is a general threshold crossing model:

$$D = I\left[h(D, Y, W) + e_1 \geq 0\right], \qquad Y = H_2(D, Y, W)$$

here W can include e_1. Y is continuous or discrete or limited.

Theorem: This system is coherent and complete iff for some g, $Y = g(D, W)$ and either $s_0(W) = s_1(W)$, or $e_1 \notin$ interval $[-s_0(W), -s_1(W)]$ where $s_d(W) = h[d, g(d, W), W]$.

Proof: By previous Theorem coherency requires $I[s_0(w) + e_1 \geq 0] = I[s_1(w) + e_1 \geq 0]$, which holds if $s_0(w) = s_1(w)$ or by limiting e_1.

Shows must either restrict the error support or strongly restrict h.

Simultaneity, Coherence, and Completeness - continued

Theorem 2 models can be rewritten in either of the following ways: Generalize Blundell and Smith (1994). For some H (To preserve threshold crossing, make D depend on D):

$$D = I[H[Y + [g(0, W) - g(1, W)]D, W] + e_1 \geq 0], \qquad Y = g(D, W)$$

or generalize Heckman (1978) (triangular, but direction of dependence can depend on W:

$$D = I[\phi[(1 - t(W))Y, W] + e_1 \geq 0], \qquad Y = g[t(W)D, W]$$

where $t(W)$ is zero or one, which determines direction of causality.

An alternative is instead assume simultaneity of an unobserved continuous latent variable that determines D, e.g. $D^* = h(Y, W)$ and $Y = g(D^*, W)$ are fully simultaneous with $D = I(D^* \geq 0)$.

Binary Choice With Endogenous Regressors - Conclusions

- Linear probability models, Maximum Likelihood, and Control functions (including ivprobit) have more drawbacks and limitations than are usually recognized.
- Special Regressor estimators are a viable alternative (or at least they have completely different drawbacks and may be more generally applicable than has been recognized).
- In practice, best might be to try all estimators and check robustness of results. Can use marginal effects to normalize them the same when comparing.
- Average Index Functions can be used to construct estimated probabilities and comparable marginal effects across estimators, often simpler to calculate than Average Structural Functions.
- Fully simultaneous systems involving a binary regressor are restrictive: must check for coherence and completeness before considering identification and estimation.