Microeconometrics
Blundell Lecture 1
Overview and Binary Response Models

Richard Blundell
http://www.ucl.ac.uk/~uctp39a/

University College London

February-March 2016

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

- social experiments methods

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

- social experiments methods
- natural experiment methods

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

- social experiments methods
- natural experiment methods
- matching methods

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

- social experiments methods
- natural experiment methods
- matching methods
- instrumental methods

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

- social experiments methods
- natural experiment methods
- matching methods
- instrumental methods
- regression discontinuity and regression kink methods

Overview

Subtitle: Models, Sampling Designs and Non/Semiparametric Estimation

1. **discrete data: binary response**
2. **censored and truncated data : cenoring models**
3. **endogenously selected samples: selectivity model**
4. **experimental and quasi-experimental data: evaluation methods**

- social experiments methods
- natural experiment methods
- matching methods
- instrumental methods
- regression discontinuity and regression kink methods
- control function methods

Binary Response Models

- Let $y_i = 1$ if an action is taken (e.g. a person is employed)
  $y_i = 0$ otherwise
  for an individual or a firm $i = 1, 2, ..., N$. We will wish to model the probability that $y_i = 1$ given a kx1 vector of explanatory characteristics $x_i' = (x_{1i}, x_{2i}, ..., x_{ki})$. Write this conditional probability as:

$$\Pr[y_i = 1 | x_i] = F(x_i' \beta)$$

Binary Response Models

- Let $y_i = 1$ if an action is taken (e.g. a person is employed)
  $y_i = 0$ otherwise
  for an individual or a firm $i = 1, 2, ..., N$. We will wish to model the probability that $y_i = 1$ given a kx1 vector of explanatory characteristics $x_i' = (x_{1i}, x_{2i}, ..., x_{ki})$. Write this conditional probability as:
  $$\Pr[y_i = 1 | x_i] = F(x_i'\beta)$$
- This is a single linear index specification. Semi-parametric if F is unknown. We need to recover F and β to provide a complete guide to behaviour.

Binary Response Models

- We often write the **response probability** as

$$\begin{aligned} p(x) &= \Pr(y = 1|x) \\ &= \Pr(y = 1|x_1, x_2, ..., x_k) \end{aligned}$$

for various values of x.

- We often write the **response probability** as

$$\begin{aligned} p(x) &= \Pr(y = 1|x) \\ &= \Pr(y = 1|x_1, x_2, ..., x_k) \end{aligned}$$

for various values of x.

- **Bernoulli (zero-one) Random Variables**
if $\Pr(y = 1|x) = p(x)$
then
$\Pr(y = 0|x) = 1 - p(x)$

$$\begin{aligned} E(y|x) &= p(x) \\ &= 1.p(x) + 0.(1 - p(x) \end{aligned}$$

$Var(y|x) = p(x)(1 - p(x))$

The Linear Probability Model

$$\begin{aligned}\Pr(y &= 1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \\ &= \beta' x\end{aligned}$$

Unless x is severely restricted, the LPM cannot be a coherent model of the response probability $P(y = 1|x)$, as this could lie outside zero-one.
Note:

$$\begin{aligned}E(y|x) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \\ Var(y|x) &= \beta' x(1 - x'\beta)\end{aligned}$$

which implies that the OLS estimator is unbiased but inefficient. The inefficiency due to the heteroskedasticity.
Homework: Develop a two-step estimator.

Binary Response Models

Typically express binary response models as a **latent variable** model:

$$y_i^* = x_i'\beta + u_i$$

where u is some continuously distributed random variable distributed independently of x, where we typically normalise the variance of u.

- The observation rule for y is given by $y = 1(y^* > 0)$.

$$\begin{aligned} \Pr[y_i^* &\geq 0|x_i] \iff \Pr[u_i \geq -x_i'\beta] \\ &= 1 - \Pr[u_i \leq -x_i'\beta] \\ &= 1 - G(-x_i'\beta) \end{aligned}$$

where G is the cdf of u_i.

Binary Response Models

In the symmetric distribution case (Probit and Logit)

$$\Pr[y_i^* \geq 0 | x_i] = G(x_i'\beta)$$

where G is some (monotone increasing) cdf. (Make sure you can prove this).

- This specification is the linear single index model.

- Show that for the linear utility and a normal *unobserved* heterogeneity implies the single index Probit model

Binary Response Models

Random sample of observations on y_i and x_i $i = 1, 2, ...N$.

$$\Pr[y_i = 1|x_i] = F(x_i'\beta)$$

where F is some (monotone increasing) cdf. This is the linear single index model.

Questions?

- How do we find β given a choice of $F(.)$ and a sample of observations on y_i and x_i?
- How do we check that the choice of $F(.)$ is correct?
- Do we have to choose a parametric form for $F(.)$?
- Do we need a random sample - or can we estimate with good properties from (endogenously) stratified samples?
- What if the data is not binary - ordered, count, multiple discrete choices?

ML Estimation of the Binary Choice Model

Assume we have N independent observations on y_i and x_i.
The probability density of y_i **conditional** on x_i is given by:

$F(x_i'\beta)$ if $y_i = 1$,

and

$1 - F(x_i'\beta)$ if $y_i = 0$.

Therefore the density of any y_i can be written:

$f(y_i|x_i'\beta) = F(x_i'\beta)^{y_i}(1 - F(x_i'\beta))^{1-y_i}$.

The joint probability of this particular sequence of data is given by the product of these associated probabilities (under independence). Therefore the joint distribution of the particular sequence we observe in a sample of N observations is simply:

$$f(y_1, y_2, ..., y_N) = \prod_{i=1}^{N} F\left(x_i'\beta\right)^{y_i}\left(1 - F\left(x_i'\beta\right)\right)^{1-y_i}$$

This depends on a particular β and is also the *'likelihood'* of the sequence $y_1, y_2 ..., y_N$,

$$\mathcal{L}(\beta; y_1, y_2, ..., y_N) = \prod_{i=1}^{N} F\left(x_i'\beta\right)^{y_i}\left(1 - F\left(x_i'\beta\right)\right)^{1-y_i}$$

ML Estimation of the Binary Choice Model

If the model is correctly specified then the MLE $\widehat{\beta}_N$ will be - consistent, efficient and asymptotically normal.
$\log \mathcal{L}$ is an easier expression:

$$\log \mathcal{L}(\beta; y_1, y_2, ..., y_N) = \sum_{i=1}^{N} [y_i \log F\left(x_i'\beta\right) + (1 - y_i) \log(1 - F\left(x_i'\beta\right)]$$

- The derivative of $\log \mathcal{L}$ with respect to β is given by:

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \beta} &= \sum_{i=1}^{N} [y_i \frac{f\left(x_i'\beta\right)}{F\left(x_i'\beta\right)} x_i + (1 - y_i) \frac{f\left(x_i'\beta\right)}{1 - F\left(x_i'\beta\right)} x_i] \\ &= \sum_{i=1}^{N} \frac{y_i - F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left(1 - F\left(x_i'\beta\right)\right)} .f\left(x_i'\beta\right) .x_i \end{aligned}$$

ML Estimation of the Binary Choice Model

The **MLE** $\widehat{\beta}_N$ refers to any root of the likelihood equation $\frac{\partial \log \mathcal{L}}{\partial \beta}|_N = 0$ that corresponds to a local maximum.

If $\log \mathcal{L}$ is a **concave function** of β, as in the Probit and Logit cases (Exercise: prove for the Probit using $\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial \beta \partial \beta'}$), then this is unique. Otherwise there exists a consistent root.

- We will consider the properties of the average log likelihood $\frac{1}{N} \log \mathcal{L}$, and assume that is converges to the 'true' log likelihood and that this is maximised at the true value of β, given by β_0.

- Notice that $\frac{\partial \log \mathcal{L}}{\partial \beta}$ is nonlinear in β. In general, no explicit solution can be found. We have to use 'iterative' procedures to find the maximum.

ML Estimation of the Binary Choice Model

Iterative Algorithms:

Choose an initial $\beta^{(0)}$.

Gradient method:
$\beta^{(1)} = \beta^{(0)} + \frac{\partial \log \mathcal{L}}{\partial \beta}|_{\beta^{(0)}}$
Convergence is slow

Deflected Gradient method:
$\beta^{(1)} = \beta^{(0)} + H^{(0)} \frac{\partial \log \mathcal{L}}{\partial \beta}|_{\beta^{(0)}}$

$H^{(0)} = \left(-\frac{1}{N} \frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial \beta \partial \beta'}|_{\beta^{(0)}} \right)^{-1}$ Newton

$H^{(0)} = \left(-E \frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial \beta \partial \beta'}|_{\beta^{(0)}} \right)^{-1}$ Scoring Method

$H^{(0)} = \left(E \left[\frac{\partial \ln \mathcal{L}_N(\beta)}{\partial \beta} \frac{\partial \ln \mathcal{L}_N(\beta)}{\partial \beta'} \right] |_{\beta^{(0)}} \right)^{-1}$ BHHH Method

ML Estimation of the Binary Choice Model

Theorem 1. (Consistency). If
(i) the true parameter value β_0 is an interior point of parameter space.
(ii) $\ln \mathcal{L}_N(\beta)$ is continuous.
(iii) there exists a neighbourhood of β_0 such that $\frac{1}{N} \ln \mathcal{L}_N(\beta)$ converges to a constant limit $\ln \mathcal{L}(\beta)$ and that $\ln \mathcal{L}(\beta)$ has a local maximum at β_0.
Then the MLE $\widehat{\beta}_N$ is consistent, or there exists a consistent root.

- Note:
  1. requires the correct specification of $\ln \mathcal{L}_N(\beta)$, in particular the $\Pr[y_i = 1|x_i]$.
  2. Contrast with MLE in the linear model.

ML Estimation of the Binary Choice Model

Theorem 2. (Asymptotic Normality). If

(i) $\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial\beta\partial\beta'}$ exists and is continuous

(ii) $\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial\beta\partial\beta'}$ evaluated at $\widehat{\beta}_N$ converges.

(iii) $\frac{1}{\sqrt{N}}\frac{\partial \ln \mathcal{L}_N(\beta)}{\partial\beta'} \sim^d N(0, H)$

then $\sqrt{N}(\widehat{\beta}_N - \beta_N) \sim^d N(0, H^{-1})$.

where

$$H = \lim_{N\to\infty}\left[-E\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial\beta\partial\beta'}|_{\beta_0}\right].$$

- Note:

$$-E\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial\beta\partial\beta'}|_{\beta_0} = E\frac{1}{N}\frac{\partial \ln \mathcal{L}_N(\beta)}{\partial\beta}\frac{\partial \ln \mathcal{L}_N(\beta)}{\partial\beta'}|_{\beta_0}$$

and $\left[-E\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial\beta\partial\beta'}|_{\beta_o}\right]^{-1}$ is the Cramer-Rao lower bound.

ML Estimation of the Binary Choice Model

Note that for **Probit (and Logit) estimators**

$$
\begin{aligned}
-E\frac{\partial^2 \ln \mathcal{L}_N(\beta)}{\partial\beta\partial\beta'} &= \sum_{i=1}^{N} \frac{[\phi(x_i'\beta)]^2}{\Phi(x_i'\beta)\left[1-\Phi(x_i'\beta)\right]} x_i x_i' \\
&= \sum_{i=1}^{N} d_i x_i x_i' \\
&= X'DX
\end{aligned}
$$

So that the $var(\widehat{\beta}_N)$ can be approximated by

$$
(X'DX)^{-1}
$$

- This expression has a similar form to that in the heteroscedastic GLS model.

In the case of the Probit there is another useful algorithm:

$$y_i^* = x_i'\beta + u_i \text{ with } u_i \sim N(0,1) \text{ and } y_i = 1(y_i^* > 0)$$

now note that

$$\begin{aligned} E(y_i^*|y_i &= 1) = x_i'\beta + E(u_i|x_i'\beta + u_i \geq 0) \\ &= x_i'\beta + E(u_i|u_i \geq -x_i'\beta) \\ &= x_i'\beta + \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} \end{aligned}$$

similarly

$$E(y_i^*|y_i = 0) = x_i'\beta - \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)}$$

If we now define

$$m_i = E(y_i^*|y_i)$$

then the derivative of the log likelihood can be written

$$\frac{\partial \log \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N} x_i(m_i - x_i'\beta)$$

set this to zero (to solve for β)

$$\sum_{i=1}^{N} x_i m_i = \sum_{i=1}^{N} x_i x_i' \beta$$

as in the OLS normal equations. We do not observe y_i^* but m_i is the best guess given the information we have.

Solving for β we have

$$\widehat{\beta} = \left(\sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i m_i .$$

Notice m_i depends on β.

This forms an **EM (or Fair) algorithm**:

- 1. Choose $\beta(0)$
- 2. Form $m_i(0)$ and compute $\beta(1)$, etc.
- This converges, but slower than deflected gradient methods.

Let $\Pr(y|x'\beta)$ be the population conditional probability of y given x.

Let $f(x)$ be the true marginal distribution of x.

Let $\pi(y|x'\beta)$ be the sample conditional probability.

- **Case 1: Random Sampling**
$\pi(y,x) = \pi(y|x'\beta)\pi(x)$
but $\pi(x) = f(x)$ and $\pi(y|x'\beta) = \Pr(y|x'\beta)$.

- **Case 2: Exogenous Stratification**
$\pi(y,x) = \Pr(y|x'\beta)\pi(x)$
Although $\pi(x) \neq f(x)$ the sample still replicates the **conditional probability** of interest in the population which is the only term that contains β in the log likelihood.

- **Case 3: Choice Based Sampling (Manski and Lerman)**

Suppose Q is the population proportion that make choice $y = 1$. Let P represent the sample fraction.

Then we can adjust the likelihood contribution by:

$$\frac{Q}{P}F(x_i'\beta).$$

If we know Q then the adjusted MLE is consistent for choice-based samples.

- (i) Semiparametric

$$E(y_i|x_i) = F\left(x_i'\beta\right)$$

retain finite parameter vector β in the linear index but relax the parametric form for F.

- (ii) Nonparametric

$$E(y_i|x_i) = F\left(g(x_i)\right)$$

both F and g are nonparametric. As you would expect, typically (i) has been followed in research.
What is the parameter of interest? β alone?

Notice that the function $F^*(a + bx_i'\beta)$ cannot be separately identified from $F(x_i'\beta)$. Therefore β is only identified up to location and scale.

To motivate, imagine $x_i'\beta \equiv z_i$ was known but $F(.)$ was not.
Seems obvious: run a general nonparametric (kernel say) regression of y on z.

- (i) How do we find β?
- (ii) How do we guarantee monotonic increasing F?

Semiparametric Estimation of β (single index models)
*** Iterated Least Squares and Quasi-Likelihood Estimation** (Ichimura and Klein/Spady)
Note that

$$E(y_i|x_i) = F\left(x_i'\beta\right)$$

so that

$$y_i = F\left(x_i'\beta\right) + \varepsilon_i \text{ with } E(\varepsilon_i|x_i) = 0.$$

A semiparametric least squares estimator can be derived. Choose β to minimise

$$S(\beta) = \frac{1}{N}\sum \pi(x_i)(y_i - F(x_i'\beta))^2$$

replacing F with a kernel regression F_h at each step with bandwidth h, simply a function of the scaler $x_i'\beta$ for some given value of β. $\pi(x_i)$ is a trimming function that downweights observations near the boundary of the support of $x_i'\beta$.

- Typically F_h is estimated using a leave-one-out kernel.

- Typically F_h is estimated using a leave-one-out kernel.
- Ichimura (1993) shows that this estimator of β up to scale is $\sqrt{N}-$consistent and asymptotically normal.

- Typically F_h is estimated using a leave-one-out kernel.
- Ichimura (1993) shows that this estimator of β up to scale is $\sqrt{N}-$consistent and asymptotically normal.
- We have to assume F is differentiable and requires at least one continuous regressor with a non-zero coefficient.

- Typically F_h is estimated using a leave-one-out kernel.
- Ichimura (1993) shows that this estimator of β up to scale is $\sqrt{N}-$consistent and asymptotically normal.
- We have to assume F is differentiable and requires at least one continuous regressor with a non-zero coefficient.
- ▶ Extends naturally to some other semi-parametric least squares cases.

- Typically F_h is estimated using a leave-one-out kernel.
- Ichimura (1993) shows that this estimator of β up to scale is $\sqrt{N}-$consistent and asymptotically normal.
- We have to assume F is differentiable and requires at least one continuous regressor with a non-zero coefficient.
- ▶ Extends naturally to some other semi-parametric least squares cases.
- ▶ It is also common to weight the elements in this regression to allow for heteroskedasticity.

Note that the average log-likelihood can be written:

$$\frac{1}{N}\log L_N(\beta) = \frac{1}{N}\sum \pi(x_i)\{y_i \ln F(x_i'\beta) + (1-y_i)y_i \ln(1-F(x_i'\beta))$$

So maximise $\log L_N(\beta)$, replacing $F(.)$ by kernel type non-parametric regression of y on $z_i = x_i'\beta$ at each step.

- Klein and Spady (1993) show asymptotic normality and that the outer-product of the gradients of the quasi-loglikelihood is a consistent estimator of the variance-covariance matrix.

Maximum Score Estimation (Manski)

Suppose F is unknown
Assume: the conditional median of u given x is zero (note that this is weaker than independence between u and x)
$\Longrightarrow$

$$\Pr[y_i = 1|x_i] > (\leq)\frac{1}{2} \text{ if } x_i'\beta > (\leq)0$$

- **Maximum Score Algorithm:**

score 1 if $y_i = 1$ and $x_i'\beta > 0$, or $y_i = 0$ and $x_i'\beta \leq 0$.
score 0 otherwise.
Choose $\widehat{\beta}$ that maximises the score, subject to some normalisation on β.

Note that the scoring algorithm can be written: choose β to maximise

$$S_N(\beta) = \frac{1}{N}\sum_{i=1}^{N}[2.1(y_i = 1) - 1]1(x_i'\beta \geq 0).$$

The complexity of the estimator is due to the discontinuity of the function $S_N(\beta)$.
Horowitz (1992) suggests a smoothed MSE:

$$S_N^*(\beta) = \frac{1}{N}\sum_{i=1}^{N}[2.1(y_i = 1) - 1]K(\frac{x_i'\beta}{h})$$

where K is some continuous kernel function with bandwidth h.

- No longer discontinuous. Therefore can prove $\sqrt{N}$ convergence and asymptotic distribution properties.

Consider the following (triangular) model

$$
\begin{aligned}
y_{1i}^* &= x_{1i}'\beta + \gamma y_{2i} + u_{1i} \qquad (1)\\
y_{2i} &= z_i'\pi_2 + v_{2i} \qquad (2)
\end{aligned}
$$

where $y_{1i} = 1(y_{1i}^* > 0)$. $z_i' = (x_{1i}', x_{2i}')$. The x_{2i}' are the excluded 'instruments' from the equation for y_1. The first equation is a the 'structural' equation of interest and the second equation is the 'reduced form' for y_2.

- y_2 is endogenous if u_1 and v_2 are correlated. If y_1 was **fully observed** we could use IV (or 2SLS).

Use the following othogonal decomposition for u_1

$$u_{1i} = \rho v_{2i} + \epsilon_{1i}$$

where $E(\epsilon_{1i}|v_{2i}) = 0$.

- *Note* that y_2 is uncorrelated with u_{1i} **conditional** on v_2. The variable v_2 is sometimes known as a **control function.**

- Under the assumption that u_1 and v_2 are jointly normally distributed, u_2 and ϵ are uncorrelated by definition and ϵ also follows a normal distribution.

Use this to define the **augmented model**

$$\begin{aligned} y_{1i}^* &= x_{1i}'\beta + \gamma y_{2i} + \rho v_{2i} + \epsilon_{1i} \\ y_{2i} &= z_i'\pi_2 + v_{2i} \end{aligned}$$

2-step Estimator:

- **Step 1**: Estimate π_2 by OLS and predict v_2,

$$\widehat{v}_{2i} = y_{2i} - \widehat{\pi}_2' z_i$$

- **Step 2**: use $\widehat{v}_{2i}$ as a 'control function' in the model for y_1^* above and estimate by standard methods.

- Blundell and Powell (REStud, 2004) extend the control function approach to the semiparametric case.
- Suppose we define $x_i' = [x_{1i}', y_{2i}]$ and $\beta_0' = [\beta', \gamma]$. Recall that if x is independent of $u_{1,}$ then

$$E(y_{1i} \mid x_i) = G(x_i'\beta_0)$$

where G is the distribution function for u_1. Sometimes also known as the average structural function, ASF.
- Note that with endogeneity of u_1 we can invoke the control function assumption:

$$u_1 \perp x \mid v_2$$

- This is the conditional independence assumption derived from the triangularity assumption in the simultaneous equations model, see Blundell and Matzkin (2013).

- Using the control function assumption we have

$$E[y_{1i}|x_i, v_{2i}] = F(x_i'\beta_0, v_{2i}),$$

and

$$G(x_i'\beta_0) = \int F(x_i'\beta_0, v_{2i})dF_{v_2}.$$

- Blundell and Powell (2003) show β_0 and the average structural function $G(x_i'\beta_0) = \int F(x_i'\beta_0, v_{2i})dF_{v_2}$ are point identified.

▶ Blundell and Powell (2004) develop a three step control function estimator:

1. Generate $\widehat{v}_2$ and run a nonparametric regression of y_{1i} on x_i and $\widehat{v}_{2i}$.
▷ This provides a consistent nonparametric estimator of $E[y_{1i}|x_i, v_{2i}]$.

2. Impose the linear index assumption on $x_i'\beta_0$ in:
$E[y_{1i}|x_i, v_{2i}] = F(x_i'\beta_0, v_{2i})$.
▷ This generates $\widehat{F}(x_i'\widehat{\beta}_0, \widehat{v}_{2i})$.

3. Integrate over the empirical distribution of $\widehat{v}_2$ to estimate $\widehat{\beta}_0$ and the average structural function (ASF), $\widehat{G}(x_i'\widehat{\beta}_0)$.
▷ This third step is implemented by taking the partial mean over $\widehat{v}_2$ in $\widehat{F}(x_i'\widehat{\beta}_0, \widehat{v}_{2i})$.

- Able to show $\sqrt{n}-$consistency for $\widehat{\beta}_0$, and the usual non-parametric rate on ASF.

- Blundell and Matzkin (2013) discuss the ASF and alternative parameters of interest.
- Chesher and Rosen (2013) develop a new IV estimator in the binary choice and binary endogenous set-up.