ECONOMETRICS II (ECO 2401)

Victor Aguirregabiria

Winter 2018

TOPIC 2: BINARY CHOICE MODELS

1. Introduction

2. BCM with cross sectional data

    2.1. Threshold model

    2.2. Interpretation in terms of utility maximization

2.3. Probit and Logit Models

2.4. Testing hypotheses on parameters

2.5. Measures of goodness of fit

2.6. Partial Effects and Average Partial Effects in BCM

2.7. BCM as a Regression Model

2.8. Misspecification of Binary Choice Models

2.9. Specification tests based on Generalized Residuals

2.10. Semiparametric methods

2.11. BCM with endogenous regressors

3. BCM with panel data

3.1. Static models

a) Fixed effects estimators

b) Random effects estimators

3.2. Dynamic models

a) Fixed effects estimators

b) Random effects estimators

1. INTRODUCTION

• Econometric discrete choice models, or qualitative response models, are models where **the dependent variable takes a discrete and finite set of values.**

• Many economic decisions involve choices among discrete alternatives.

(1) Labor economics: LF participation; unionization; occupational choice; migration; retirement; job matching; strikes.

(2) Family economics: # children; marriage; divorce.

(3) Industrial organization: Market entry/exit; demand differentiated products.

(4) Education economics: going to college decision.

(5) Political Economy: voting

INTRODUCTION [Cont.]

• Some classifications of Discrete Choice Models (DCM) that have relevant implications for the econometric analysis are:

a) Type of data: Cross-section / Panel

b) Number of choice alternatives: Binary / Multinomial

• We will start with Binary Choice models with Cross-sectional data.

• Why (or when) not using Linear Regression models for these data?

2. BCM WITH CROSS SECTIONAL DATA

• We are interested in an event with two possible outcomes (e.g., "an individual is unemployed or not", "a worker is unionized or not", "a firm is active in a market or not") and how this event depends on some explanatory variables X.

• Define the binary variable Y such that:
$$\begin{cases} Y = 1 & \text{if event occurs} \\ Y = 0 & \text{if it does not} \end{cases}$$

• Define the **Conditional Choice Probability (CCP)**.

$$P(x) \equiv \mathsf{Pr}(Y = 1|X = x)$$

Note that:

$$\mathbb{E}\left(Y \mid X = x\right) = P(x)$$

• A BCM is a paramtric model for the conditional expectation $\mathbb{E}\left(Y \mid X = x\right)$, that is also the CCP, $P(x)$.

Reduced form Model for CCP

• In some empirical applications, the researcher may be interested in the CCP $P(x)$ just as a **predictor** of Y given $X = x$, not in a causal effect interpretation of the model.

• In that case, the researcher may simply use a flexible specification of $P(x)$. For instance:

$$P(x) = F(x'\beta)$$

where $F(.)$ is a known function that maps the index $x'\beta$ into the the probability space $[0, 1]$, e.g., $F(.)$ is a CDF.

Model with explicit specification of unobservables

• Many times we are interested in the causal effect of X on Y. Then, it is useful to consider a model that relates Y with obsevables X and with unobservables ε, and make assumptions about the relationship between X and ε.

$$Y = g\left(X,\ \varepsilon\right)$$

• Since Y is a discrete variable, it should respond in a discrete way (i.e., not continuously) to changes in (X,ε).

• That is, $g(.)$ should be a function that maps continuous variables in ε or X into the binary set $\{0, 1\}$.

• In principle, this condition rules out the Linear Regression Model (i.e., $Y = X'\beta + \varepsilon$) as a valid model for a binary dependent variable. But we will discuss this point later.

2.1. THRESHOLD MODELS

• A popular specification of $g(X, \varepsilon)$ that appears naturally in many economic applications is the threshold function:

$$Y = g\left(X, \varepsilon\right) = \begin{cases} 1 & \text{if} \quad Y^*\left(X, \varepsilon\right) \geq 0 \\ \\ 0 & \text{if} \quad Y^*\left(X, \varepsilon\right) < 0 \end{cases}$$

• $Y^*\left(X, \varepsilon\right)$ is a real-valued function that is denoted **latent variable**. Note that setting the threshold at 0 is an innocuous normalization because $Y^*\left(X, \varepsilon\right)$ can always include a constant term.

• A common specification of the latent threshold function is:

$$Y^*\left(X, \varepsilon\right) = X'\beta - \varepsilon$$

where β is a $K \times 1$ vector of parameters.

- Therefore, the model is:

$$Y = \begin{cases} 1 & \text{if} \quad \varepsilon \leq X'\beta \\ \\ 0 & \text{if} \quad \varepsilon > X'\beta \end{cases}$$

- We can also represent the model using the **Indicator Function** $1\{A\}$ where $1\{A\} = 1$ if A is true and $1\{A\} = 0$ if A is false.

$$Y = 1\{\varepsilon \leq X'\beta\}$$

- When ε is independent of X and it has a CDF $F(\cdot)$, we have that:

$$P(x) = \text{Pr}\left(Y^* \geq 0 \mid X = x\right) = \text{Pr}(\varepsilon \leq x'\beta) = F(x'\beta)$$

• The relationship between the conditional probability $P(x)$ and the index $x'\beta$ depends on the distribution of ε.

• If ε is $N(0,1)$: $F(x'\beta) = \Phi(x'\beta)$

• If ε is $Logistic$: $F(x'\beta) = \Lambda(x'\beta) = \dfrac{\exp(x'\beta)}{1+\exp(x'\beta)}$

• If ε is $U[0,1]$: $F(x'\beta) = \begin{cases} 0 & \text{if } x'\beta \leq 0 \\ x'\beta & \text{if } 0 \leq x'\beta \leq 1 \\ 1 & \text{if } x'\beta \geq 1 \end{cases}$

BCM and Models with Non-additive unobservables

• Discrete choice models belong to a class of nonlinear econometric models where unobservables (error term) enter into the model in a non-additive form:

$$Y = g\left(X,\ \varepsilon\right)$$

where $g(.,.)$ is non-additive in ε,

$$g\left(X,\ \varepsilon + c\right) \neq g\left(X,\ \varepsilon\right) + c$$

In DCMs this non-additivity is a natural implication of the discrete nature of the dependent variable.

• In this class of models, the **"Average Partial Effect"** is different to the **"Partial Effect at the Average"**. A linear-regression approach typically provides estimates of "Partial Effect at the Average". We will discuss why for some empirical questions we are interested in estimating the "Average Partial Effect" and not the "Partial Effect at the Average".

Interpretation of the parameters β

• We know that in a linear regression model, $Y = X'\beta + \varepsilon$, when ε is (mean) independent of X we have that $\mathbb{E}(Y|X = x) = x'\beta$ and:

$$\beta_k = \begin{cases} \dfrac{\partial \mathbb{E}(Y|X=x)}{\partial x_k} & \text{if} \quad X_k \text{ continuous} \\ \\ \mathbb{E}(Y|X = x + \boldsymbol{\Delta}_k) - \mathbb{E}(Y|X = x) & \text{if} \quad X_k \text{ discrete} \end{cases}$$

with $\boldsymbol{\Delta}_k$ a vector of 0s except at position k where we have 1.

• In a BCM, we have that:

$$\frac{\partial \mathbb{E}(Y|X=x)}{\partial x_k} = \beta_k \; f(x'\beta) \qquad \text{if} \quad X_k \text{ continuous}$$

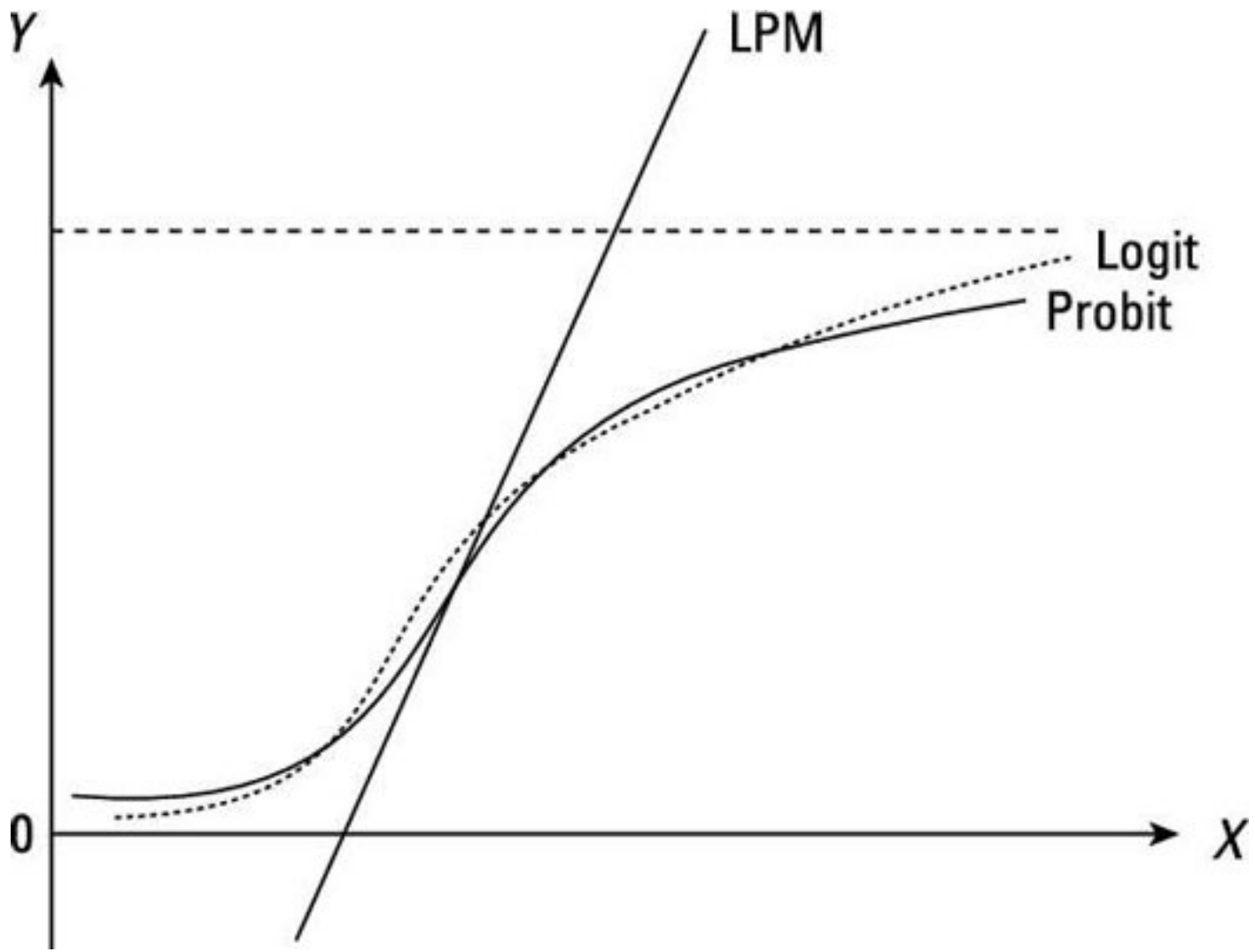
$$\mathbb{E}(Y|X = x + \boldsymbol{\Delta}_k) - \mathbb{E}(Y|X = x) = F(x'\beta + \beta_k) - F(x'\beta) \quad \text{if} \quad X_k \text{ discrete}$$

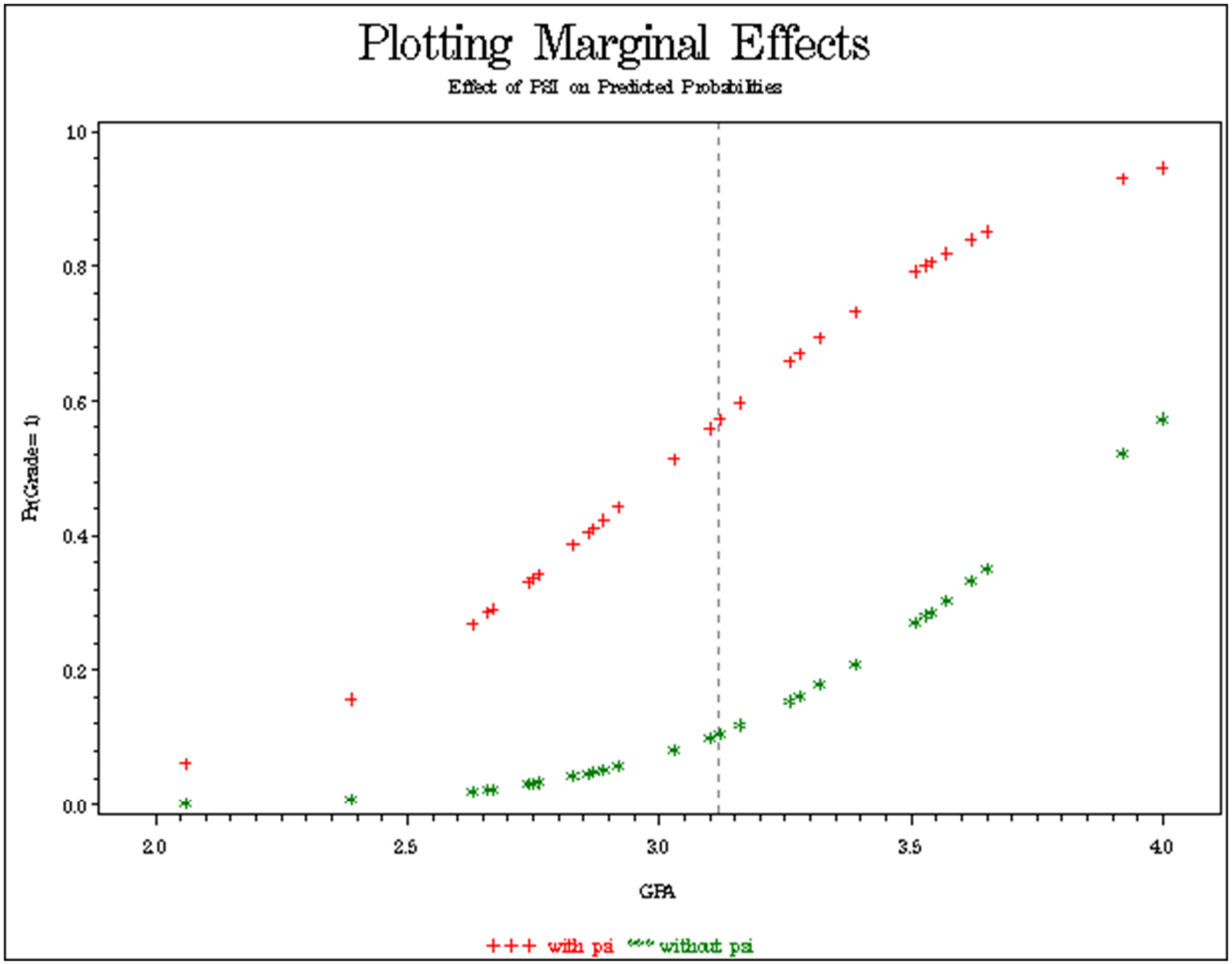
Interpretation of the parameters β (2)

• The main difference between the LRM and the BCM in the interpretation of β is that in the LRM the **Partial Effects** $\dfrac{\partial \mathbb{E}(Y|X=x)}{\partial x_k}$ are constant across x while in the BCM they depend on the individual characteristics, and more specifically on its propensity or probability of $Y = 1$ given X.

• Taking into account that $P(x) = F(x'\beta)$ and $\dfrac{\partial \mathbb{E}(Y|X=x)}{\partial x_k} = \beta_k \, f(x'\beta)$, we have that:

$$\begin{cases} \text{As} \quad x'\beta \to -\infty: \qquad P(x) \to 0 \quad \text{and} \quad \dfrac{\partial \mathbb{E}(Y|X=x)}{\partial x_k} \to 0 \\ \\ \text{As} \quad x'\beta \to +\infty: \qquad P(x) \to 1 \quad \text{and} \quad \dfrac{\partial \mathbb{E}(Y|X=x)}{\partial x_k} \to 0 \end{cases}$$


Y
LPM
Logit
Probit
0
X


Plotting Marginal Effects
Effect of PSI on Predicted Probabilities
Pr(Grade= 1)
1.0
0.8
0.6
0.4
0.2
0.0
2.0
2.5
3.0
3.5
4.0
GPA
+++ with psi *** without psi

2.2. INTERPRETATION IN TERMS OF UTILITY MAXIMIZATION

• **Example 1:** Consider an individual who has to decide whether to purchase a certain durable good or not (e.g., a smartphone). Suppose that the purchased quantity is either one or zero: $Y \in \{0, 1\}$.

$Y \in \{0, 1\}$ is the *indicator of purchasing the durable good.*

• The utility function is $U(C, Y)$, where C represents consumption of the composite good. More specifically:

$$U(C, Y) = u(C) + Y \ \left(Z'\beta_1 - \varepsilon\right)$$

$u(.)$ is an increasing function; Z is a vector of characteristics observable to the econometrician, such as age and education; β_1 is a vector of parameters; and ε is a zero mean random variable that is individual-specific.

• The individual's decision problem is to maximize $U(C, Y)$ subject to the budget constraint $C + P * Y \leq M$, where P is the price of the good, and M is the individual's disposable income.

• Solving $C = M - P\ Y$ in the utility, the decision problem is:

$$\max_{Y \in \{0,1\}} u(M - P\ Y) + Y\ \left(Z'\beta_1 - \varepsilon\right)$$

• Therefore, the optimal choice is:

$$Y = 1 \ \Leftrightarrow \ u(M - P) + Z'\beta_1 - \varepsilon > u(M)$$

- Suppose that $u(C) = \alpha\ C$. Then,

$$Y = 1 \ \Leftrightarrow \ \{-\alpha\ P + Z'\beta_1 - \varepsilon\} > 0$$

$$\Leftrightarrow \ \{X'\beta - \varepsilon\} > 0$$

where $X' = (-P, Z)$, and $\beta' = (\alpha, \beta_1')$.

- Conditional on $\{Z, P\}$, the probability of purchasing the product is:

$$P(Y = 1|Z, P) = F\,(\ [-P, Z]\beta\ )$$

- The price sensitivity of demand is:

$$\frac{\partial F(X'\beta)}{\partial P} = -\alpha\ f(X'\beta)$$

• **Example 2:**

$Y =$ *Indicator of the event "individual goes to college".*

$X =$ { HS grades ; Family income ; Parents' Education ; Scholarships }

Let U_0 and U_1 be the utilities associated with choosing $Y = 0$ (no college) and $Y = 1$ (college), respectively.

• Consider the following specification of these utility functions:

$$\begin{aligned} U_0 &= X'\beta_0 + \varepsilon_0 \\ U_1 &= X'\beta_1 + \varepsilon_1 \end{aligned}$$

• If the individual maximizes her utility, then:

$$\{Y = 1\} \iff \{ U_1 \geq U_0 \} \iff \{ \varepsilon \leq X'\beta \}$$

where $\beta = \beta_1 - \beta_0$, and $\varepsilon = \varepsilon_0 - \varepsilon_1$.

2.3. PROBIT AND LOGIT MODELS

• To complete the parametric specification of the model we should make an assumption about the distribution of the disturbance ε_i. The most common assumptions in the literature are:

$$\text{Probit model: } \varepsilon \sim N(0,1) \qquad \text{then, } F(x'\beta) = \Phi\left(x'\beta\right)$$

$$\text{Logit model: } \varepsilon \sim Logistic(\sigma) \quad \text{then, } F(x'\beta) = \frac{\exp\left(x'\beta\right)}{1+\exp\left(x'\beta\right)}$$

• Suppose that the researcher observes a random sample of (Y, X): $\{y_i, x_i : i = 1, 2, ...n\}$.

• We assume that ε_i are $i.i.d.$ across individuals.

- **Maximum likelihood estimation**

- The likelihood function is:

$$L(\beta) = \Pr(y_1, y_2, ..., y_n \mid x_1, x_2, ..., x_n)$$

$$= \prod_{i=1}^{n} \Pr(y_i|x_i) = \prod_{y_i=1} F(x_i'\beta) \prod_{y_i=0} [1 - F(x_i'\beta)]$$

- The log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta)$$

where

$$l_i(\beta) = y_i \ \ln\Big(F(x_i'\beta)\Big) + (1 - y_i) \ \ln\Big(1 - F(x_i'\beta)\Big)$$

- This likelihood is continuous and twice differentiable if $F(.)$ is.

- The MLE is the value of $\widehat{\beta}$ that solves the likelihood equations:

$$\frac{\partial l(\widehat{\beta})}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial l_i(\widehat{\beta})}{\partial \beta} = \sum_{i=1}^{n} \frac{x_i \; f(x_i'\widehat{\beta})}{F(x_i'\widehat{\beta}) \; [1 - F(x_i'\widehat{\beta})]} \left(y_i - F(x_i'\widehat{\beta}) \right) = 0$$

- $\dfrac{\partial l_i(\widehat{\beta})}{\partial \beta}$ is called the **score** of observation i.

- Interpretation of Likelihood equations as moment conditions.

- For the Probit model the likelihood equations are:

$$\sum_{i=1}^{n} \frac{\partial l_i(\widehat{\beta})}{\partial \beta} = \sum_{i=1}^{n} \frac{x_i \ \phi(x_i'\widehat{\beta})}{\Phi(x_i'\widehat{\beta}) \ [1 - \Phi(x_i'\widehat{\beta})]} \left(y_i - \Phi(x_i'\widehat{\beta})\right) = 0$$

- And for the Logit model the likelihood equations are:

$$\sum_{i=1}^{n} \frac{\partial l_i(\widehat{\beta})}{\partial \beta} = \sum_{i=1}^{n} x_i \left(y_i - \frac{\exp\left(x_i'\widehat{\beta}\right)}{1 + \exp\left(x_i'\widehat{\beta}\right)} \right) = 0$$

because for the logistic distribution $f(\varepsilon) = F(\varepsilon)[1 - F(\varepsilon)]$.

Computation of the MLE

• There is not a closed-form expression for the MLE. We should calculate $\widehat{\beta}$ numerically using an iterative algorithm.

• For the Probit and Logit models, the likelihood is also **globally concave**.

• The most common iterative algorithms to obtain MLE are Newton-Raphson and BHHH. Because the likelihood is globally concave, both algorithms converge to the unique maximum regardless of the initial value we use to initialize the algorithm.

Newton-Raphson iterations:

$$\widehat{\beta}^{K+1} = \widehat{\beta}^{K} - \left[\sum_{i=1}^{n} \frac{\partial^2 l_i\left(\widehat{\beta}^{K}\right)}{\partial\beta\partial\beta'}\right]^{-1} \left[\sum_{i=1}^{n} \frac{\partial l_i\left(\widehat{\beta}^{K}\right)}{\partial\beta}\right]$$

BHHH iterations:

$$\widehat{\beta}^{K+1} = \widehat{\beta}^{K} + \left[\sum_{i=1}^{n} \frac{\partial l_i\left(\widehat{\beta}^{K}\right)}{\partial \beta} \quad \frac{\partial l_i\left(\widehat{\beta}^{K}\right)}{\partial \beta'}\right]^{-1} \left[\sum_{i=1}^{n} \frac{\partial l_i\left(\widehat{\beta}^{K}\right)}{\partial \beta}\right]$$

- Note that, at the true value of β in the population:

$$p\lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\beta)}{\partial\beta\partial\beta'}\right] = -p\lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\beta)}{\partial\beta}\,\frac{\partial l_i(\beta)}{\partial\beta'}\right]$$

i.e., Fisher's information matrix.

Asymptotic properties of the MLE

• If the model is correctly specified,

$$\sqrt{n}\left(\widehat{\beta}-\beta\right) \rightarrow_d N\left(0, V\right)$$

where

$$V = \mathbb{E}\left(\frac{\partial l_i\left(\beta\right)}{\partial \beta}\,\frac{\partial l_i\left(\beta\right)}{\partial \beta'}\right)^{-1} = \mathbb{E}\left(\frac{f(X'\beta)^2}{F(X'\beta)[1-F(X'\beta)]}\,XX'\right)^{-1}$$

• A consistent estimate of V is obtained by substituting β by $\hat{\beta}$ and . $\mathbb{E}_X(.)$ by the sample mean such that:

$$\widehat{Var}\left(\widehat{\beta}\right) = \frac{\widehat{V}}{n} = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i'\widehat{\beta})^2}{F(x_i'\widehat{\beta})[1-F(x_i'\widehat{\beta})]}\,x_i x_i'\right)^{-1}$$

2.4. TESTING HYPOTHESES ON PARAMETERS AND REPORTING ESTIMATION RESULTS

• Wald, LM and LR tests as usual for MLE

• **Reporting estimation results:** For some applications the estimated partial effects can be more informative than the estimates of the parameters. The partial effect can be evaluated at the mean value of the regressors $\bar{x}$.

• The estimated **partial effect** for explanatory variable k (evaluated at the sample mean $\overline{x}$) is:

$$\widehat{PE}_k = \begin{cases} \widehat{\beta}_k \; f(\overline{x}'\widehat{\beta}) & \text{if} \quad X_k \text{ continuous} \\ \\ F(\overline{x}'\widehat{\beta} + \widehat{\beta}_k) - F(\overline{x}'\widehat{\beta}) & \text{if} \quad X_k \text{ discrete} \end{cases}$$

• **Example:** Default in the payment of college student loans. Knapp and Seaks (REStat, 1992).

- Sample: 1834 college students in Pennsylvania who got a student loan and left college in the academic year 1984-1985.

Variable	$\widehat{\beta}$ (s.e.)	**Partial effect in % points**
Graduation dummy	-1.090 (0.121)*	-9.9
Parent's income (in thousand $)	-0.018 (0.004)*	-0.2
Loan amount (in thousand $)	0.026 (0.020)	+0.3
College cost (in thousand $)	0.085 (0.061)	+0.9

2.5. MEASURES OF GOODNESS OF FIT

Residuals. In a BCM, after the estimation of β, we cannot obtain residuals for the unobservable ε. Note that the residual $\widehat{\varepsilon}_i$ is such that:

$$y_i = 1\{x_i'\widehat{\beta} - \widehat{\varepsilon}_i \geq 0\}$$

We know that:

If $y_i = 1$, then $\widehat{\varepsilon}_i \leq x_i'\widehat{\beta}$.

If $y_i = 0$, then $\widehat{\varepsilon}_i > x_i'\widehat{\beta}$.

But we do not know the exact value of $\widehat{\varepsilon}_i$.

• We will obtain **Generalized Residuals**. For this, we need to obtain a **Regression Model representation of the BCM**.

BCM as a [Nonlinear] regression model

To understand some of the Goodness-of-fit measures for the BCM, it is useful to interpret a BCM as a Nonlinear regressions model.

$$\mathbb{E}(Y|X = x) = F(x'\beta)$$

- Therefore, we can write:

$$Y = F(X'\beta) + u$$

where $\mathbb{E}(u|X = x) = 0$.

- Conditional on X, u can take only two values. Then,

$$Var(u|X = x) = F(x'\beta)[1 - F(x'\beta)]$$

Generalized Residuals & Goodness-of-fit

- Define the following fitted values: $\widehat{P}_i = F(x_i'\widehat{\beta})$.

- And the **Generalized Residuals**: $\widehat{u}_i = y_i - \widehat{P}_i = y_i - F(x_i'\widehat{\beta})$

- **RSS:** $\sum_{i=1}^{n}(y_i - \widehat{P}_i)^2$

- **Weighted RSS:** $\sum_{i=1}^{n} \dfrac{(y_i - \widehat{P}_i)^2}{\widehat{P}_i(1 - \widehat{P}_i)}$

- [**Pseudo] R-square:** Square correlation between y_i and $\widehat{P}_i$. It is the R-square of this regression-like representation of the BCM.

Other Goodness-of-fit measures

- **Log likelihood function:** $l(\widehat{\beta})$

- **Likelihood Ratio Index (or McFadden's R-square)**:

$$LRI = 1 - \frac{l(\widehat{\beta})}{l_0}$$

where l_0 is the log-likelihood when all parameters except the constant term are zero. It is simple to prove that,

$$l_0 = n_0 \ln(n_0) + n_1 \ln(n_1) - n \ln(n)$$

where $n_0 = \#obs$ with $y_i = 0$, and $n_1 = \#obs$ with $y_i = 1$.

2.6. PARTIAL EFFECTS AND AVERAGE PARTIAL EFFECTS IN BCM

• **Is it reasonable (good econometric practice) to use a linear regression model when the dependent variable is binary?**

Under what conditions or for which types of empirical questions?

• To answer these questions, it is useful to define first the concepts of "Partial Effect", "Average Partial Effect", and "Partial Effect at the Average".

• In econometrics, typically, we are interested in ceteris paribus effects: how Y changes when a variable X_k changes keeping constant the rest of the variables. This ceteris paribus effect is called the **Partial Effect of X_k on Y.**

• Define $PE_k(X_0, \varepsilon_0)$ as the Partial Effect given that the initial value of (X, ε) is (X_0, ε_0) and we change X to $X_0 + \mathbf{\Delta}_k$ where $\mathbf{\Delta}_k$ is a vector of zeroes at every position except at position k where we have a 1. In the general model, we have that:

$$PE_k(X_0, \varepsilon_0) = g\left(X_0 + \mathbf{\Delta}_k,\ \beta,\ \varepsilon_0\right) - g\left(X_0,\ \beta,\ \varepsilon_0\right)$$

• The **conditional Average Partial Effect** $APE_k(X_0)$ is defined as $PE_k(X_0, \varepsilon_0)$ averaged over the distribution of the unobservables ε but conditional on X_0:

$$APE_k(X_0) = \int PE_k(X_0, \varepsilon)\ dF(\varepsilon)$$

• The **unconditional Average Partial Effec**t APE_k is is defined as $PE_k(X_0, \varepsilon_0)$ averaged both over the distribution of the unobservables ε and over the distribution of observables X.

$$PE_k = \int PE_k(X, \varepsilon)\ dF(\varepsilon)\ dF_X(X)$$

• It is important to distinguish between the average partial effect APE_k and the **Partial Effect at the average individual**. The later is:

$$\begin{aligned} PE_k(X_0 &= \mathbb{E}(X), \varepsilon_0 = \mathbb{E}(\varepsilon)) \\ &= g\left(\mathbb{E}(X) + \mathbf{\Delta}_k,\ \beta,\ \mathbb{E}(\varepsilon)\right) - g\left(\mathbb{E}(X),\ \beta,\ \mathbb{E}(\varepsilon)\right) \end{aligned}$$

• In a linear regression model (LRM), individuals are assumed to be homogeneous in term partial effects: i.e., $g(X, \beta, \varepsilon) = X'\beta + \varepsilon$ and therefore:

$$PE_k(X_0, \varepsilon_0) = APE_k(X_0) = APE_k = \beta_k$$

• More precisely, in a LRM (without random coefficients) we can allow for interactions between observable variables such that Partial Effects may vary across individuals according to observable characteristics. However, Partial Effects do not depend on unobservables.

• LRM with random coefficients allow for unobserved heterogeneity in Partial Effects and therefore in those models the Average Partial Effect is not equal the Partial Effect at the average.

• BCM is a class of models where the difference between the Average Partial Effect and the Partial Effect at the average appear naturally as the result of the binary nature of the dependent variable.

• In a BCM, we have that:

$$PE_k(X_0, \varepsilon_0) = 1\left\{\varepsilon_0 \leq [X_0 + \Delta_k]'\beta\right\} - 1\left\{\varepsilon_0 \leq X_0'\beta\right\}$$

where $1\{.\}$ is the indicator function.

• Partial effects at the individual level depend on the individual's X and ε.

• This is an important property of BCM. This property derives naturally from the discrete aspect of the dependent variable.

• In a BCM, we have that the APE are:

$$APE_k(X_0) = F([X_0 + \Delta_k]'\beta) - F(X_0'\beta)$$

• The marginal partial effect is similar to the partial effect but when Δ_k represents a marginal change in a continuous variable X_k. In that case:

$$AMPE_k(X_0) = \beta_k \; f(X_0'\beta)$$

where $f(.)$ is the PDF of ε.

• The AMPE at the average individual is:

$$AMPE_k(X_0 = \mathbb{E}(X)) = \beta_k \; f(\mathbb{E}(X)'\beta)$$

• In BCM, Partial Effects vary over individuals and, in general, **the APE can be very different to the Partial Effect at the average.**

• This is a property that distinguishes BCM from Linear Regression Model.

• When our main interest is to estimate the PE for the average individual, then we can use a LRM for the binary variable Y. For large samples, the estimates will not be very different to the estimate of the same effect from a BCM.

• However, most of the time in economics we are interested in APE and not in the PE for the average individual. If that is the case, using a LRM for a binary Y is a very bad choice because that model imposes the very implausible (even impossible!) restriction that PEs do not depend on the unobservables.

Example: School Attendance of Children from Poor Families.

• Suppose that we are interested in the determinants of elementary school attendance of kids (Y) from poor families.

$$Y = \text{Kids in the family attend (regularly) school}$$

• We are interested in evaluating the effects of a public program that tries to encourage school attendance by providing a subsidy that is linked to school attendance, e.g., PROGRESA program in Mexico since 1997.

• We have data on $\{Y, S, X\}$ where: X contains family socioeconomic characteristics; and S is the amount of subsidy.

$S = 0$ for families in the control group;

$S = \$M$ for families in the experimental group

Example: School Attendance of Children from Poor Families (2)

- We estimate the BCM:

$$Y = 1\{\varepsilon \leq \alpha\ S + X'\beta\}$$

- Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimated parameters, and $\hat{P}_i = F(\hat{\alpha}s_i + x_i'\hat{\beta})$ the estimated probability of school attendance for family i.

- The **Partial Effect** of receiving the subsidy for individual i is:

$$PE_i = 1\{\varepsilon_i \leq \alpha M + x_i'\beta\} - 1\{\varepsilon_i \leq x_i'\beta\}$$

For $\alpha \geq 0$, this effect can be only zero or one.

Example: School Attendance of Children from Poor Families (3)

• Even if we knew the true values of α and β, we cannot obtain PE_i because we cannot estimate ε_i.

• However, we can estimate the **Average Partial Effec**t of the subsidy for a family with characteristics x_i:

$$\widehat{APE}(x_i) = F(\hat{\alpha}M + x_i'\hat{\beta}) - F(x_i'\hat{\beta}) \simeq f(x_i'\hat{\beta})\ \hat{\alpha}M$$

• And the estimated **effect of this program on the number of kids attending school**:

$$\Delta \text{ Kids Attending School} = \sum_{i=1}^{n} 1\{s_i > 0\}\ \widehat{APE}(x_i)$$

• We could also estimate the "counterfactual" effect of the hypothetical application of the policy to an identical population of H families (instead of n):

$$\Delta \text{ Kids Attending School} = H \left[\frac{1}{n} \sum_{i=1}^{n} \widehat{APE}(x_i) \right] = H \ \widehat{APE}$$

Example: School Attendance of Children from Poor Families (4)

- The effect of the policy for the average family is:

$$APE(\bar{x}) = F(\hat{\alpha}M + \bar{x}'\hat{\beta}) - F(\bar{x}'\hat{\beta}) \simeq f(\bar{x}'\hat{\beta})\hat{\alpha}M$$

- Suppose we use this partial effect at the average to measure the effect of the policy:

$$\widehat{APE}(\bar{x}) \left[\sum_{i=1}^{n} 1\{s_i > 0\}\right]$$

- The difference between the actual effect and this approximation is:

$$\begin{aligned} \text{Approx. Error} &= \sum_{i=1}^{n} 1\{s_i > 0\} \left[\widehat{APE}(x_i) - \widehat{APE}(\bar{x})\right] \\ &= n\left[\widehat{APE} - \frac{n_{s>0}}{n}\widehat{APE}(\bar{x})\right] \end{aligned}$$

Example: School Attendance of Children from Poor Families (5)

• This approximation error can be substantial if the deviations $\left[\widehat{APE}(x_i) - \widehat{APE}(\bar{x})\right]$ are not symetrically distributed, i.e., if $\widehat{APE}(x_i)$ is a nonlinear function.

• The magnitude of this difference depends on the variance of $x_i'\hat{\beta}$ (i.e., on the level of heterogeneity in the propensity to send kids to school), and on the magnitude of $\hat{\alpha}M$.

• Even when $\widehat{APE}$ and $\frac{n_{s>0}}{n}\widehat{APE}(\bar{x})$ are similar, we can be interested in the estimation of $APE(x)$ for different groups of families according to x. For instance, suppose that $\widehat{APE}(x_i)$ is very close to zero for almost every family i, but it is very large for families with very low income, that represent only 1% of the population. This information is very useful to target the policy.

2.7. BCM AS A REGRESSION MODEL

• A regression model is a statistical model that specifies how the conditional mean $\mathbb{E}(Y|X)$ depends on X, i.e., it specifies a function $m(X, \beta)$ for $\mathbb{E}(Y|X)$.

$$\mathbb{E}(Y|X) = m(X, \beta)$$

• This implies that:

$$Y = m(X, \beta) + u$$

where u is a disturbance or unobservable variable that, by construction, is mean independent of X, i.e., $\mathbb{E}(u|X) = 0$.

• When $m(X, \beta) = X'\beta$, we have a linear regression model. When $m(X, \beta)$ is nonlinear in the parameters, we have a nonlinear regression model, e.g., $m(X, \beta) = \exp\{X'\beta\}$, or $m(X, \beta) = \beta_1[X_1^{\beta_2} + X_1^{\beta_3}]^{\beta_4}$.

- When Y is binary, we have that:

$$\mathbb{E}(Y|X) = 1 * \mathsf{Pr}(Y = 1|X) + 0 * \mathsf{Pr}(Y = 0|X) = \mathsf{Pr}(Y = 1|X)$$

- Therefore, a BCM for $\mathsf{Pr}(Y = 1|X)$ is also a Regression Model for $\mathbb{E}(Y|X)$.

- According to the threshold BCM:

$$\mathbb{E}(Y|X) = F(X'\beta)$$

- An therefore,

$$Y = F(X'\beta) + u$$

where, by construction, u is mean independent of X.

• Therefore, in this context, we can justify using a Linear Regression Model (LRM) for the binary dependent variable Y as a first-order (linear) approximation to the function $F(X'\beta)$ for X around its mean $\mathbb{E}(X)$.

$$F(X'\beta) \simeq F(\mathbb{E}(X)'\beta) + (X - \mathbb{E}(X))'\beta \; f(\mathbb{E}(X)'\beta)$$

Let $X = (1, X_1)$ where 1 represents the constant term and X_1 the rest of the regressors. Then, $X'\beta = \beta_0 + X_1'\beta_1$ and $\mathbb{E}(X)'\beta = \beta_0 + \mathbb{E}(X_1)'\beta_1$, and $(X - \mathbb{E}(X))'\beta = (X_1 - \mathbb{E}(X_1))'\beta_1$. Solving these expressions in the equation above, we have:

$$\begin{aligned} F(X'\beta) &\simeq \beta_0^* + X_1'\beta_1^* \\ \\ \text{where} \quad &: \quad \beta_0^* = F(\mathbb{E}(X)'\beta) - f(\mathbb{E}(X)'\beta)\; \mathbb{E}(X_1)'\beta_1 \\ \text{and } \beta_1^* &= f(\mathbb{E}(X)'\beta)\; \beta_1 \end{aligned}$$

• Note that $\beta_1^* = \beta_1 f(\mu_X'\beta)$ is the AMPE for the average individual.

• Therefore, we can use a Linear Regression Model for the binary variable Y. This type of model is called the **Linear Probability Model**:

$$Y = X'\beta^* + u^*$$

and the slopes β^* have a clear interpretation as Average Partial Effects for the Average individual. OLS estimation of this LRM provides consistent estimates of β^*.

• The **main limitation of the Linear Probability Model** is that it does not provide any information about the APE for individuals other than the average individual, or for the unconditional APE (that depends on the conditional APE for all the individuals).

• This limitation is particularly serious in BCMs where the APEs $F([X + \Delta]'\beta) - F(X'\beta)$ vary significantly over X. In that case, the APEs of a significant group of individuals, and the unconditional APE, can be very different to the APE for the average individual.

2.8. MISSPECIFICATION OF BCM

• Remember that in the linear regression model a necessary and sufficient condition for consistency the OLS estimator is that $\mathbb{E}(\varepsilon|x) = 0$. That is, heteroscedasticity, autocorrelation and non-normality of the error term does not affect the consistency of the OLS estimator as long as $\mathbb{E}(\varepsilon|x) = 0$.

• However, in the context of discrete choice models, the consistency of the MLE depends crucially on our assumptions about ε_i.

• If ε_i is heteroscedastic, or if it has a cdf that is not the one that we have assumed, then the MLE is no longer consistent.

• The reason is that our assumption about ε_i affects not only second and further moments of y_i, but also its conditional mean:

• Suppose that the true model is such that $\varepsilon_i \sim iid$ with $cdf\ F$. Then,

$$\text{True model: } y_i = F(x_i'\beta) + u_i, \qquad \text{where } \mathbb{E}(u_i|x_i) = 0$$

• Instead, we assume that $\varepsilon_i \sim iid$ N(0,1) [Probit]. Then,

$$\text{Estimated Model: } y_i = \Phi(x_i'\beta) + u_i^*, \text{ where } u_i^* = u_i + F(x_i'\beta) - \Phi(x_i'\beta)$$

• It is clear that, if $F \neq \Phi$ then $\mathbb{E}(u_i^*|x) \neq 0$, and MLE using Φ is inconsistent.

• Suppose that the researcher is not particularly interested in the estimates of β but only in the estimated probabilities $P(x_i)$.

• For instance, a car insurance company that is only interested in the probability of accident of an individual with characteristics x_i.

• In this case, the main issue is the consistency of $P(x_i)$ not the consistency of β. One might think that misspecification of F is not a big issue in this case.

• However, that is not true. Misspecification of $F(.)$ can generates important biases both on the estimator of β and on the estimator of $P(.)$.

• **Horowitz's Monte Carlo experiment:**

Suppose that the true probabilities are $P^*(x_i)$ and the researcher estimates a logit model. How close are $\{\hat{P}_{Logit}(x)\}$ to the true $\{P^*(x)\}$?

• Horowitz (Hbook of Stat, 1993) performed a Monte Carlo study to answer this question. He considered different cases for $P^*(x)$:

(1) Homoscedastic probit;

(2) Student-t;

(3) Uniform;

(4) Heteroscedastic logit;

(5) Heteroscedastic probit;

(6) Bimodal distribution of ε.

- The main results are:

  (a) The errors are small when the true distribution of ε is unimodal and homoscedastic.

  (b) The errors can be very large when the true distribution of ε is bimodal or it is heteroscedastic.

- Summary of Horowitz's Monte Carlo study:

True Model	$\mathbb{E}_x(\mathbf{P}^*(\mathbf{x}) - \hat{\mathbf{P}}_{Logit}(\mathbf{x}))$	$\mathbf{max}_x\lvert\mathbf{P}^*(\mathbf{x}) - \hat{\mathbf{P}}_{Logit}(\mathbf{x})\rvert$
Homoced. and unimod	0.01	0.02
Bimodal	0.05	0.20
Heteroscedastic	0.10	0.30

2.9. SPECIFICATION TESTS USING GENERALIZED RESIDUALS

• In the LRM, we typically test assumption on the error term ε by using the residuals $\widehat{\varepsilon}_i = y_i - x_i'\widehat{\beta}$.

• In BCM we cannot obtain residuals for ε_i but we can get residuals for the error term u_i in the regression-like representation of the BCM:

$$y_i = F(x_i'\beta) + u_i$$

• We can get the resduals:

$$\widehat{u}_i = y_i - F(x_i'\widehat{\beta})$$

and standardized residuals:

$$\widehat{u}_i^* \equiv \frac{y_i - F(x_i'\widehat{\beta})}{\sqrt{F(x_i'\widehat{\beta})(1 - F(x_i'\widehat{\beta}))}}$$

• Under the null hypothesis that the model is correctly specified, we have that:

$$u_i^* \equiv \frac{y_i - F(x_i'\beta)}{\sqrt{F(x_i'\beta)(1 - F(x_i'\beta))}}$$

u_i^* should be independent of x_i with zero mean.

• By testing the independence of the residuals $\widehat{u}_i^*$ and x_i, we test the correct specification of the model.

GENERAL PURPOSE SPECIFICATION TEST

• Given the standardized residuals $\widehat{u}_i^*$ and the estimated CCPs $\widehat{P}_i \equiv F(x_i'\widehat{\beta})$, we run the OLS regression:

$$\widehat{u}_i^* = x_i^{*\prime}\pi_0 + \pi_1\ \widehat{P}_i + \pi_2\ (\widehat{P}_i)^2 + ... + \pi_q\ (\widehat{P}_i)^q + e_i$$

where $x_i^* \equiv \dfrac{x_i\ f(x_i'\hat{\beta})}{\sqrt{F(x_i'\hat{\beta})(1 - F(x_i'\hat{\beta}))}}$.

• Define the statistic $LM = n \cdot R^2$, where R^2 is the R-square coefficient from the previous regression.

• Under the null hypothesis (the model is correctly specified), LM is asymptotically distributed as χ_q^2.

TEST OF HETEROSCEDASTICITY IN BCM

• Consider the BCM $Y = 1\{X'\beta - \varepsilon \geq 0\}$ where:

$$\varepsilon | X \sim N\left(0\ ,\ \exp\left(\tilde{X}'\delta\right)\right)$$

where $\tilde{X}$ is the vector X without the constant term.

• We are interested in testing the null hypothesis of homoscedasticity, that is equivalent to test: $H_0 : \delta = 0$.

• A possible approach is to estimate β and δ by MLE. That approach is computationally demanding because the log-likelihood of this model is no longer globally concave in (β, δ).

• Instead, we can estimate the standard probit model, under the null hypothesis of $\delta = 0$, and use a LM test for the null. The LM statistic is:

$$LM = \left[\frac{\partial \log L(\hat{\beta}, \delta = 0)}{\partial(\beta, \delta)}\right]' Var\left(\frac{\partial \log L(\hat{\beta}, \delta = 0)}{\partial(\beta, \delta)}\right)^{-1} \left[\frac{\partial \log L(\hat{\beta}, \delta = 0)}{\partial(\beta, \delta)}\right]$$

• Under H_0, LM is asymptotically distributed as $\chi^2_{\dim(\delta)}$.

• Davidson and McKinnon (JE, 1984) show that this LM statistic can be obtained as the output of a simply auxiliary regression.

• $LM = n \cdot R^2$, where R^2 is the R-square coefficient from the following regression:

$$\hat{u}_i^* = x_i^{*\prime}\pi_1 + z_i^{*\prime}\pi_2 + e_i$$

where:

$$\begin{aligned}
\hat{u}_i^* &\equiv \frac{y_i - \Phi(x_i'\hat{\beta})}{\sqrt{\Phi(x_i'\hat{\beta})(1-\Phi(x_i'\hat{\beta}))}} \\
x_i^* &\equiv \frac{x_i\ \phi(x_i'\hat{\beta})}{\sqrt{\Phi(x_i'\hat{\beta})(1-\Phi(x_i'\hat{\beta}))}} \\
z_i^* &\equiv \frac{\tilde{x}_i\ (x_i'\hat{\beta})\ \phi(x_i'\hat{\beta})}{\sqrt{\Phi(x_i'\hat{\beta})(1-\Phi(x_i'\hat{\beta}))}}
\end{aligned}$$

2.10. SEMIPARAMETRIC ESTIMATION OF BCM

• The consistency of the ML estimator of Probit or Logit models relies on the correct specification of the probability distribution of the unobservable ε.

• That is, consistency of the MLE in BC models is not robust to misspecification of the CDF of ε.

• This property contrasts with the consistency of OLS in the linear regression model: i.e., the OLS is the MLE when ε is normally distributed, but it is also consistent when ε is not normal, and even asymptotically efficient (if ε is homoscedastic and not serially correlated). In econometrics, this type of robust efficient estimators are called **Adaptive Estimators**.

• Are there adaptive estimators of the BCM which are robust to different properties of the unobserved error term such as heteroscedasticity, serial correlation, or the particular functional form for the distribution of the error?

• We consider four **Semiparametric Estimators** of the BCM where the distribution of ε is nonparametrically specified.

(1) Least Absolute Deviations (LAD) estimator;

(2) Manski's Maximum Score Estimator;

(3) Horowitz's Smooth Maximum Score Estimator;

(4) Klein and Spady estimator.

1. Least Absolute Deviations (LAD) estimation

• LAD is an **estimation method that is adaptive for a very general class of econometric models**.

• **Least Squares (LS) estimation** (linear or nonlinear) is based on the property that the mean $\mu \equiv \mathbb{E}(Y)$ is the value c that minimizes the Mean Square Deviations $\mathbb{E}\left([Y-c]^2\right)$.

$$\mu = \arg\min_c \quad \mathbb{E}\left([Y-c]^2\right)$$

• **LS estimation** is the sample counterpart of this property:

$$\widehat{\mu} = \arg\min_c \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2$$

• We have that, $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \rightarrow_p \mu$.

Least Absolute Deviations (LAD) [2]

• This property of LS applies also to the conditional mean. Let $\mu(x_0) \equiv \mathbb{E}(Y \mid X = x_0)$. Then,

$$\mu(x_0) = \arg\min_c \ \mathbb{E}\left([Y - c]^2 \mid X = x_0\right)$$

• The **LS estimator** is based on the sample counterpart of this property of the mean:

$$\widehat{\mu}(x_0) = \arg\min_c \ \frac{1}{n}\sum_{i=1}^{n} (y_i - c)^2 \ 1(x_i - x_0)$$

• We have that (for x discrete),

$$\widehat{\mu}(x_0) = \frac{\sum_{i=1}^{n} y_i \ 1\{x_i = x_0\}}{\sum_{i=1}^{n} 1\{x_i = x_0\}} \rightarrow_p \mu(x_0)$$

Least Absolute Deviations (LAD) [2']

- Suppose that $\mu(x) = x'\beta$. Then,

$$\beta = \arg\min_b \ \mathbb{E}\left(\left[Y - X'\beta\right]^2\right)$$

and solving this minimization problem, we get that:

$$\beta = \mathbb{E}\left(XX'\right)^{-1} \ \mathbb{E}\left(XY\right)$$

- The least squares estimator is the sample counterpart of this property:

$$\widehat{\beta} \ = \ \arg\min_b \ \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'b\right)^2$$

$$= \ \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} y_i x_i'\right]$$

Least Absolute Deviations (LAD) (3)

• Similarly, LAD estimation is based on a property of the median. The median $m \equiv median(Y)$ minimizes the Mean Absolute Deviations (LAD) $\mathbb{E}(|Y - c|)$:

$$m = \arg\min_{c} \ \mathbb{E}(|Y - c|)$$

• **LAD estimation** is the sample counterpart of this property:

$$\widehat{m} = \arg\min_{c} \ \frac{1}{n} \sum_{i=1}^{n} |y_i - c|$$

And we have that, $\widehat{m} \rightarrow_p m$.

• This property extends to the conditional median:

$$m(x_0) \equiv median(Y|X = x_0) = \arg\min_{c} \ \mathbb{E}(|Y - c| \ | \ X = x_0)$$

Least Absolute Deviations (LAD) (4)

- Consider the general econometric model:

$$Y = f(X, \beta, \varepsilon)$$

where f is a known function; X is a vector of observable explanatory variables; ε is an unobservable variable; and β is a vector of parameters.

- If $f(X, \beta, \varepsilon)$ is not additively separable in ε, then $\mathbb{E}\left(f(X, \beta, \varepsilon) \mid X\right) \neq \mathbb{E}\left(f(X, \beta, \mathbb{E}[\varepsilon|X]\right) = \mathbb{E}\left(f(X, \beta, 0)\right)$.

- For instance, if $f(X, \beta, \varepsilon) = 1\{\varepsilon \leq X'\beta\}$, we have that:

$$\begin{aligned} \mathbb{E}\left(f(X, \beta, \varepsilon) \mid X\right) &= \mathbb{E}\left(1\{\varepsilon \leq X'\beta\} \mid X\right) = F_\varepsilon(X'\beta\} \\ &\neq \mathbb{E}\left(f(X, \beta, \mathbb{E}[\varepsilon|X]\right) = 1\{0 \leq X'\beta\} \end{aligned}$$

Least Absolute Deviations (LAD) (4')

• Suppose that instead of $\mathbb{E}[\varepsilon|X] = 0$ we assume that $median(\varepsilon|X) = 0$. Neither a weaker or a stronger assumption of the joint distribution of ε and X.

• More specifically, consider a model defined by the following assumptions:

(A1) the function f is known (up to β) and **monotonic in** ε;

(A2) $median(\varepsilon|X) = 0$.

Least Absolute Deviations (LAD) (5)

- Under assumptions (A1) and (A2), we have that

$$median(Y|X) = f(X, \beta, 0)$$

- Based on this condition, the true value of β satisfies the following condition:

$$\beta = \arg\min_{c} \quad \mathbb{E}\left(|Y - f(X, c, 0)|\right)$$

- LAD estimator is based on the sample counterpart of this property:

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - f(x_i, \beta, 0)|$$

The LAD estimator minimizes the sum of absolute deviations of y_i with respect to its median $f(x_i, \beta, 0)$.

Least Absolute Deviations (LAD) (6)

• Under assumptions (A1) and (A2), the LAD estimator is consistent. Therefore, LAD is semiparametric estimator for nonlinear econometric models.

• If function f is continuous & differentiable in β, then the LAD estimator is: (a) root-n consistent; (b) asymptotically normal; (c) it has a simple expression for its asymptotic variance that it is simple to estimate; (d) we can use gradient optimization methods to compute $\hat{\beta}_{LAD}$.

• If function f is NOT continuous in β, LAD is still consistent but, in general, **properties (b), (c), and (d) do not hold**.

LAD estimator of BCM

- Consider the BCM:

$$Y = 1\left\{X'\beta - \varepsilon \geq 0\right\}$$

This function satisfies condition (A1) of monotonicity in ε.

- Suppose that $median(\varepsilon|X) = 0$. Then, we have that:

$$median(Y|X) = 1\left\{X'\beta \geq 0\right\}$$

- And the LAD estimator of β in the BCM is:

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{i=1}^{n} \left| y_i - 1\left\{x_i'\beta \geq 0\right\} \right|$$

This is estimator is consistent but: NOT root-n consistent; NOT asymptotically normal; and it cannot be computed using standard gradient methods.

LAD estimator of BCM Normalization

• Note tha the indicator $1\left\{x_i'\beta \geq 0\right\}$ is invariant to a change of scale in β, i.e., for any $\lambda > 0$,

$$1\left\{x_i'[\lambda\beta] \geq 0\right\} = 1\left\{x_i'\beta \geq 0\right\}$$

• Therefore, we can identify the vector β only up to scale. We need to normalize β. There are different **possible normalizations**.

• (a) $\|\beta\| = 1$

• (b) Let x_1 be an explanatory variable for which we know the sign of β_1 (w.l.og. $\beta_1 > 0$). We can normalize $\beta_1 = 1$. Let $x = (x_1, \widetilde{x})$. Note that:

$$x'\beta = \beta_1\, x_1 + \widetilde{x}\ \widetilde{\beta} = \beta_1 \left[x_1 + \widetilde{x}\frac{\widetilde{\beta}}{\beta_1}\right]$$

2. Manski's Maximum Score Estimator

- Consider the BCM $Y = 1\{X'\beta - \varepsilon \geq 0\}$ where we assume that:

$$median(\varepsilon|X) = 0$$

That is, ε is median independent of X, and the median is zero.

- Other than $median(\varepsilon|X) = 0$, no other assumption is made on dist. of ε.

- If we knew β, a "natural" predictor of Y is $1\{X'\beta \geq 0\}$ because:

  (a) the support of $1\{X'\beta \geq 0\}$ is the same as the support of Y: $\{0, 1\}$

  (b) $median(Y|X) = 1\{X'\beta \geq 0\}$.

Maximum Score Estimator (MSE) (2)

• We have a correct prediction when:

either $Y = 1$ and $1\{X'\beta \geq 0\}$

or $Y = 0$ and $1\{X'\beta < 0\}$

• Given a sample $\{y_i, x_i : i = 1, 2, ..., n\}$, consider the following sample criterion function:

$$S(\beta) = \sum_{i=1}^{n} y_i \; 1\{x_i'\beta \geq 0\} + (1 - y_i) \; 1\{x_i'\beta < 0\}$$

• This criterion function provides the number of correct predictions for a given value of β. We call it the **Score function**.

Maximum Score Estimator (MSE) (3)

• The **Maximum Score Estimator (MSE)** is the value of β that maximizes the score function:

$$\hat{\beta}_{MSE} = \arg\max_{\beta} \quad S(\beta)$$

• Under $median(\varepsilon|X) = 0$, the MSE is a consistent estimator of β.

• Therefore, the MSE is an estimator that is robust to heteroscedasticity, serial correlation, and to any form of the distribution of ε.

• In that sense, the MSE has similar properties as OLS in a linear regression mode under the mean independence assumption $\mathbb{E}(\varepsilon|X) = 0$.

Equivalence of LAD and MSE

• Before we discuss other properties of the MSE, it is interesting to show that for the BCM, **the MSE and the LAD are identical estimators.**

• Let $LAD(\beta)$ be the LAD criterion function, and let $S(\beta)$ be the score function.

• We now show that $LAD(\beta) = n - S(\beta)$ and therefore minimizing $LAD(\beta)$ is equivalent to maximizing $S(\beta)$ such that the MSE is the LAD estimator.

Equivalence of LAD and MSE (2)

$$
\begin{aligned}
LAD(\beta) &= \sum_{i=1}^{n} \left| y_i - 1\{x_i'\beta \geq 0\} \right| \\
&= \sum_{i=1}^{n} 1\left\{ y_i = 1 \text{ and } x_i'\beta < 0 \right\} + 1\left\{ y_i = 0 \text{ and } x_i'\beta \geq 0 \right\} \\
&= \sum_{i=1}^{n} y_i \; 1\{x_i'\beta < 0\} + (1 - y_i) \; 1\{x_i'\beta \geq 0\} \\
&= \sum_{i=1}^{n} y_i \left(1 - 1\{x_i'\beta \geq 0\}\right) + (1 - y_i) \left(1 - 1\{x_i'\beta < 0\}\right) \\
&= n - \sum_{i=1}^{n} y_i \; 1\{x_i'\beta \geq 0\} + (1 - y_i) \; 1\{x_i'\beta < 0\} \\
&= n - S(\beta)
\end{aligned}
$$

Properties of the MSE

• Note that the score function $S(\beta)$ is discontinuous and not differentiable in β. $1\{x_i'\beta \geq 0\}$ is a step function, and this implies that $S(\beta)$ is also step function.

Example. Consider the model $Y = 1\{\beta + X \geq 0\}$. We have a sample of $n = 4$ observations: $(x_i, y_i) = \{(x_1, 0), (x_2, 0), (x_3, 1), (x_4, 1)\}$, and $0 < x_1 < x_2 < x_3 < x_4$.

The score function is:

$$\begin{aligned} S(\beta) &= 1\{\beta + x_1 < 0\} + 1\{\beta + x_2 < 0\} \\ &+ 1\{\beta + x_3 \geq 0\} + 1\{\beta + x_4 \geq 0\} \end{aligned}$$

or

$$S(\beta) = \begin{cases} 2 & \text{if} \quad \beta < -x_4 \\ 3 & \text{if} \quad \beta \in [-x_4, -x_3) \\ 4 & \text{if} \quad \beta \in [-x_3, -x_2) \\ 3 & \text{if} \quad \beta \in [-x_2, -x_1) \\ 2 & \text{if} \quad \beta \geq -x_1 \end{cases}$$

• There is not a single value β that maximizes $S(\beta)$ but a whole interval $[-x_3, -x_2)$.

• If X has continuous support, as the sample size increases, the amplitude of thi interval gets smaller.

Properties of the MSE

• In this case, discontinuity of $S(\beta)$ does not affect the consistency of the MSE, but it has several important implications.

(a) We cannot use the standard gradient based methods to search for the MSE.

(b) If the sample size is not large enough, there may not be a unique value of β that maximizes $S(\beta)$. The maximizer of $S(\beta)$ can be a whole (compact) set in the space of β.

(c) The MSE is not asymptotically normal. It has a not standard distribution.

(d) The rate of convergence of the MSE to the true β is lower than root-n . It is $n^{1/3}$.

3. Horowitz's Smooth Maximum Score Estimator

• Limitations (b)-(c)-(d) of the MSE motivate the use of the smooth-MSE proposed by Horowitz.

• First, note that score function $S(\beta)$ can be written as follows:

$$
\begin{aligned}
S(\beta) &= \sum_{i=1}^{n} y_i \; 1\{x_i'\beta \geq 0\} + (1 - y_i) \; 1\{x_i'\beta < 0\} \\
&= \sum_{i=1}^{n} y_i \; 1\{x_i'\beta \geq 0\} + (1 - y_i) \left(1 - 1\{x_i'\beta \geq 0\}\right) \\
&= \sum_{i=1}^{n} (1 - y_i) + \sum_{i=1}^{n} (2y_i - 1) \; 1\{x_i'\beta \geq 0\}
\end{aligned}
$$

Smooth Maximum Score Estimator (2)

• Therefore, maximizing $S(\beta)$ is equivalent to maximizing $\sum_{i=1}^{n}(2y_i - 1)$ $1\{x_i'\beta \geq 0\}$, and:

$$\hat{\beta}_{MSE} = \arg\max_{\beta} \sum_{i=1}^{n}(2y_i - 1)\ 1\{x_i'\beta \geq 0\}$$

• Limitations (a)-(d) of the MSE are due to the fact that $1\{x_i'\beta \geq 0\}$ is discontinuous in β.

• Horowitz proposes to replace $1\{x_i'\beta \geq 0\}$ by a function $\Phi\left(\frac{x_i'\beta}{b_n}\right)$, where $\Phi(.)$ is the CDF of the standard normal, and b_n is a bandwidth parameter such that: (1) $b_n \to 0$ as $n \to \infty$; and (2) $nb_n \to \infty$ as $n \to \infty$. That is, b_n goes to zero but more slowly than $1/n$.

Smooth Maximum Score Estimator (3)

• The Smooth-MSE is defined as:

$$\hat{\beta}_{SMSE} = \arg\max_{\beta} \sum_{i=1}^{n} (2y_i - 1)\ \Phi\left(\frac{x_i'\beta}{b_n}\right)$$

• As $n \to \infty$, and $b_n \to 0$, the function $\Phi\left(\frac{x_i'\beta}{b_n}\right)$ converges $1\{x_i'\beta \geq 0\}$, and the criterion function converges to the Score function. This implies the consistency of $\hat{\beta}_{SMSE}$.

• Under the additional condition that $nb_n \to \infty$ as $n \to \infty$, and the Kernel function has enough smoooth derivatives (e.g., Normal CDF) this estimator is **n^{δ} consistent and asymptotically normal, with 2/5$\leq \delta <$1/2**. It can be computed using standard gradient search methods because the criterion function is continuously differentiable.

4. Lewbel (JoE, 2000) "special" regressor method

• Lewbel (2000) proposes a method for the estimation of β that does not require optimization (i.e., linear regression-like estimator) and is root-n consistent and asymptotically normal.

• Suppose that:

$$Y = 1\{Z + \widetilde{X}'\beta - \varepsilon \geq 0\}$$

where:

Z is independent of ε;

Z is a continuous variable with support $\mathbb{R}$;

$\mathbb{E}\left(\varepsilon \mid \widetilde{X}\right) = 0$.

Lewbel (JoE, 2000) "special" regressor method [cont.]

• Under these conditions, Lewbel shows that:

$$\beta = \left[\mathbb{E}\left(\widetilde{X}\ \widetilde{X}'\right)\right]^{-1}\mathbb{E}\left(\widetilde{X}\ \widetilde{Y}\right)$$

where $\widetilde{Y} = \dfrac{Y - \mathbf{1}\{Z > 0\}}{f_{Z|\widetilde{X}}(Z|\widetilde{X})}$.

• This expression shows that we can estimate consistently β by an OLS regression of $\widetilde{Y}$ on X. Variable $\widetilde{Y}$ should be constructed and requires estimating the density $f_{Z|\widetilde{X}}(Z|\widetilde{X})$.

• Since the density $f_{Z|\widetilde{X}}(Z|\widetilde{X})$ appears in the denominator, $\sqrt{n}-$consistency of the estimator (and good finite sample properties) requires trimming observations where $\widehat{f}_{Z|\widetilde{X}}(z_i|\widetilde{x}_i) < h_N$.

Smooth-MSE and Lewbel's method in STATA

• See Blevins, J. R. and S. Khan (2013): "Distribution-Free Estimation of Heteroskedastic Binary Response Models in Stata," Stata Journal 13, 588–602.

• Blevins and Khan have created a **command in Stata**, `dfbr` (for *distribution free binary response*), that implements the Smooth-MSE, Lewbel's method and other methods for the estiation of BCM with a nonparametric specification of the distribution of ε.

NONPARAMETRIC IDENTIFICATION OF $F(\varepsilon|X)$

• Once we have estimated the vector of parameters β using an adaptive method such as the smooth-MSE, we want to estimate Average partial effects (APE) for different individuals in the sample or out of the sample (for different values of x_i). As shown above, to estimate APEs for individuals who are not the average individual in the sample (or for some other average or marginal individual) we need to estimate the distribution of ε.

• Given β and our assumption that $median(\varepsilon|X) = 0$, is the CDF $F(\varepsilon|X)$ nonparametrically identified? No, without further assumptions. More specifically, no if only assumption median independence between ε and X.

NONPARAMETRIC IDENTIFICATION OF $F(\varepsilon|X)$

• **Matzkin (ECMA, 1992)**. A sufficient condition for the identification of F is:

(a) $X'\beta = Z + \widetilde{X}'\beta$, where ε and Z are independent;

(b) Conditional on $\widetilde{X}$, Z has variation over the whole real line;

(c) ε is median independent of $\widetilde{X}$ (but we don't need full independence);

Proof: The CCP function $P(z, x) = \mathsf{Pr}(Y = 1|Z = z,\ \widetilde{X} = \widetilde{x})$ is nonparametrically identified from the data at every $(z, \widetilde{x})$. Suppose that β has been identified/estimated (e.g., MSE estimator).

• For arbirary values of $\widetilde{x}$ and ε, say $(\widetilde{x}_0, \varepsilon_0)$, we want to estimate $F_{\varepsilon|\widetilde{x}_0}(\varepsilon_0)$.

• Let z_0 be the value $z_0 = \varepsilon_0 - \widetilde{x}_0'\beta$, and let $P(z_0, \widetilde{x}_0)$ be the CCP evaluated at $(z_0, \widetilde{x}_0)$. Then:

$$\begin{aligned} P(z_0, \widetilde{x}_0) &\equiv \Pr(Y = 1 \mid Z = z_0, \widetilde{X} = \widetilde{x}_0) \\ &= \Pr(\varepsilon \leq z_0 + \widetilde{x}_0'\beta) \\ &= F_{\varepsilon|\widetilde{x}_0}(\varepsilon_0) \end{aligned}$$

• That is, for any $(\widetilde{x}_0,\varepsilon_0)$ we can always define a value z_0 such that the empirical CCP $P(z_0, \widetilde{x}_0)$ give us the CDF of ε, $F_{\varepsilon|\widetilde{x}_0}(\varepsilon_0)$.

EFFICIENT SEMIPARAMETRIC ESTIMATION

• Klein & Spady (ECMA 1993) propose an asymptotically efficient method to estimate jointly β and the CDF of ε. Their BCM, with the form $Y = 1\{\varepsilon \leq X'\beta\}$, imposes a limited form of heterocedasticity:

$$Var(\varepsilon|X) = \sigma^2(X'\beta)$$

• According to this model:

$$\begin{aligned} P(x) &= \mathsf{Pr}(Y = 1|X = x) \\ &= F\left(\frac{x'\beta}{\sigma(x'\beta)}\right) \equiv G(x'\beta) \end{aligned}$$

EFFICIENT SEMIPARAMETRIC ESTIMATION

• Klein-Spady estimator propose a **semiparametric maximum likelihood estimator** of β and the function $G(.)$.

• The log-likelihood function is:

$$l(\beta, G) = \sum_{i=1}^{n} y_i \ \ln G\left(x_i'\beta\right) + (1 - y_i) \ \ln\left[1 - G\left(x_i'\beta\right)\right]$$

• And KS estimator is defined as:

$$\left(\widehat{\beta}_{KS}, \widehat{G}_{KS}\right) = \arg \max_{\{\beta, G\}} \ l(\beta, G)$$

• The difficult issue here is that G is not a finite-dimension vector of parameters, but a real-valued function or infinite-dimension vector of parameters.

• This is not a standard MLE, and both its computation and the derivation of its asymptotic properties are non-standard problems.

• Under mild regularity conditions, Klein and Spady show that the estimator is consistent and asymptotically normal. The estimator of β is root-n consistent. Also, it is asymptotically efficient within the class of semiparametric estimators.

• The procedure starts with an initial guess of the function G. Let $\widehat{G}_0$ be this initial guess. For instance, $\widehat{G}_0$ can the Φ, i.e., we postulate a Probit model with homocedasticity.

• Then, at every iteration $K \geq 1$ we perform two steps.

Step 1: Estimate β given $\widehat{G}_{K-1}$.

$$\widehat{\beta}_K = \arg\max_{\beta} \; l\left(\beta, \widehat{G}_{K-1}\right)$$

This is a standard MLE (or quasi MLE).

Step 2: At this step, we obtain

$$\widehat{G}_K = \arg\max_G \; l\left(\widehat{\beta}_K, G\right)$$

It is possible to show that the function G that maximizes $l(\beta, G)$ given β is the conditional expectation:

$$G(e) = \mathbb{E}(Y \mid X'\beta = e)$$

• Based on this property, $G(e)$ can be estimated using a nonparametric kernel method (Nadaraya-Watson estimator) for the regresion of y_i on $x_i'\widehat{\beta}_K$:

$$\widehat{G}_K(e) = \frac{\sum_{i=1}^n y_i \; K\left(\frac{x_i'\widehat{\beta}_K - e}{b_n}\right)}{\sum_{i=1}^n K\left(\frac{x_i'\widehat{\beta}_K - e}{b_n}\right)}$$

where b_n is a bandwidth parameter.

- The algorithm iterates until convergence, e.g., until $\left\|\widehat{\beta}_K - \widehat{\beta}_{K-1}\right\| < 10^{-6}$.

2.11. BCM WITH CONTINUOUS ENDOGENOUS REGRESSORS

Consider the Probit model:

$$(1) \qquad Y = 1\{X'\beta + \alpha\, W + \varepsilon > 0\}$$

$$(2) \qquad W = Z'\delta + u$$

where ε and u are independent of X and Z, but $cov(\varepsilon, u) \neq 0$, and **therefore ε and W are not independent**.

• We can construct a **two-step estimator** in the same spirit as an IV / 2SLS estimator in a linear regression model. However, we need to assume that $\nu = \varepsilon + \alpha\, u$ is normally distributed.

$$Y = 1\{X'\beta + \alpha\,[Z'\delta] + \nu > 0\}$$

where $\nu = \varepsilon + \alpha u$

Two-step estimator

[Step 1] OLS estimation of δ in regression $W = Z'\delta + u$. Construct fitted values $\widehat{W} = Z'\widehat{\delta}$.

[Step 2] MLE estimation of β and α in the Probit model

$$Y = 1\{X'\beta + \alpha\widehat{W} + \nu > 0\}$$

• Estimator is consistent asymptotically normal. Standard errors should be corrected to control for the estimation error in the first step. There is a simple expression for the correct asymptotic standard errors.

Hausman test of exogeneity

- Null hypothesis: $cov(\varepsilon, u) = 0$, such that W is exogenous.

- Hausman statistic for the difference of two estimators of (β, α):

(1) MLE Probit for Y on (X, W): efficient under H_0, inconsistent under H_1;

(2) MLE Probit for Y on $(X, \widehat{W})$: consistent under H_0 and H_1.

- We need to be carefull though. These two estimators are estimating different parameters. (1) estimates $\left(\frac{\beta}{\sigma_\varepsilon}, \frac{\alpha}{\sigma_\varepsilon}\right)$; and (2) estimates $\left(\frac{\beta}{\sigma_\nu}, \frac{\alpha}{\sigma_\nu}\right)$.

- A simple solution is to normalize β_1 to 1 such that both estimators are estimating the same object: $\left(\frac{\widetilde{\beta}}{\beta_1}, \frac{\alpha}{\beta_1}\right)$.

Rivers & Vuong (JoE, 1988)

• Rivers & Vuong propose a similar two-step estimator that has two advantages over the previous estimator: (a) it only requires conditional normality; (b) it provides a simple t-test of endogeneity.

• Suppose that conditional on u, the error term ε is normally distributed. This implies that:

$$\varepsilon \;=\; \pi\, u + \xi$$

where (a) $\pi = \dfrac{\sigma_{\varepsilon u}}{\sigma_u^2}$; (b) ξ is normally distribution as $N(0,\, \sigma_\varepsilon^2 \left(1 - \rho^2\right))$ where ρ is the correlation between ε and u; (c) ξ is independent of u;

• Since ε is independent of X and Z, we have that ξ is independent of X, Z, and u, and therefore it is independent of W.

Rivers & Vuong Two-step estimator

- We can write the probit model:

$$Y = 1\{X'\beta + \alpha\, W + \pi\, u + \xi > 0\}$$

And given that ξ is normally distributed and independent of X, W, and u, we have that:

$$\Pr(Y = 1|X, W, u) = \Phi\left(\frac{X'\beta + \alpha\, W + \pi\, u}{\sigma_\xi}\right)$$

[Step 1] OLS estimation of δ in regression $W = Z'\delta + u$. Construct the residuals $\widehat{u} = W - Z'\widehat{\delta}$.

[Step 2] MLE estimation of β, α, and π in the Probit model

$$Y = 1\{X'\beta + \alpha W + \pi\widehat{u} + \xi > 0\}$$

- Note that $\frac{\pi}{\sigma_\xi} \neq 0$ if and only if $cov(\varepsilon, u) \neq 0$. Therefore, a t-test of $H_0 : \frac{\pi}{\sigma_\xi} = 0$ is a test of the endogeneity of W.

Blundell and Powell (Review of Economic Studies, 2004) "Endogeneity in Semiparametric Binary Response Models"

• Blundell and Powell (2004) extend Rivers-Vuong method to models where the distribution of the unobservables, ε and u, is nonparametrically specified.

• TBW

PROBIT ESTIMATION WITH SERIAL CORRELATION

- Consider the Time Series BCM:

$$Y_t = 1\{X_t'\beta - \varepsilon_t \geq 0\}$$

- We have timese series data $\{y_t, x_t\}$ for T periods and we want to estimate β.

- All the variables are stationary, I(0). X_t and ε_t are independent. But ε_t can be serially correlated.

- We can use a MLE to estimate β. The conditional log-likelihood function is:

$$\begin{aligned} l(\beta) &= \ln \Pr(y_1, y_2, ..., y_T \mid x_1, x_2, ..., x_T; \beta) \\ &= \ln \Pr(\ 1\{2(y_t - 1)\varepsilon_t \leq 2(y_t - 1)x_t\beta\} \text{ for } t = 1, 2, ..., T) \end{aligned}$$

PROBIT WITH SERIAL CORRELATION [2]

• The construction of this likelihood function and the MLE requires:

(a) To specify the stochastic process of ε_t in order to compute probability of the history of choices.

(b) Joint estimation of β and the parameters of the stochastic process.

(c) Solving a T-dimension integration problem for each trial value of the parameters. This is computationally costly.

• This estimator is complicated to implement, and it may be inconsistent if the specification of the stochastic process of ε_t is not correct.

PROBIT WITH SERIAL CORRELATION [3]

• **Avery-Hansen-Hotz** (IER, 1983) provide a simple estimator that is robust to serial correlation. They show that a method that estimates β using a standard Probit (or logit) model that ignores the serial correlation in ε_t is root-T consistent and asymptotically normal.

• Consider the Pseudo- log likelihood function:

$$l(\beta) = \sum_{t=1}^{T} y_t \ \ln \Phi\left(x_t'\beta\right) + (1 - y_t) \ln \Phi\left(-x_t'\beta\right)$$

• Let $\hat{\beta}_{AHH}$ be the value of β that maximizes this function. This is Avery-Hansen-Hotz estimator.

• Under normality, homocedasticity, and stationarity of ε_t, this estimator is consistent and asymptotically normal, regardless the serial correlation in ε_t.

PROBIT WITH SERIAL CORRELATION [4]

• Why is **Avery-Hansen-Hotz** estimator consistent despite it is a MLE based on a misspecified likelihood function? Because the likelihood equations that define the estimator are valid moment conditions regardless the form of serial correlation in ε_t.

• The likelihood equations are:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{x_t\ f(x_t^{'}\beta)}{F(x_t'\beta)\ [1-F(x_t'\beta)]}\left(y_t - F(x_t^{'}\beta)\right) = 0$$

where F and f are the CDF and the PDF of ε_t. As T goes to infinity, these equations converge to:

$$\mathbb{E}\left(z\left(x_t\right)\ \left[y_t - F(x_t^{'}\beta)\right]\right) = 0$$

where $z\left(x_t\right)$ is the vector $\frac{x_t\ f(x_t^{'}\beta)}{F(x_t'\beta)\ [1-F(x_t'\beta)]}$.

• If ε_t and X_t are independently distributed, we can show that these moment conditions / likelihood equations hold.

PROBIT WITH SERIAL CORRELATION [5]

• Under the previous conditions, we have that $\mathbb{E}(Y_t|X_t) = F(X_t'\beta)$ such that the variable $u_t = Y_t - F(X_t'\beta)$ is mean independent of X_t.

• Therefore, $u_t = Y_t - F(X_t'\beta)$ is not correlated with any function of X_t.

$$\mathbb{E}\left(z\left(X_t\right)\ \left[Y_t - F(X_t^{'}\beta)\right]\right) = 0$$

• A method of moments estimator based on valid moment conditions with $\mathbb{E}\left(z\left(X_t\right)\ z\left(X_t\right)'\right)$ non-singular is consistent.

3. BCM WITH PANEL DATA

- As in linear PD models, we distinguish static and dynamic PD BCMs.

(1) Static models: explanatory variables are strictly exogenous;

(a) Exogenous individual effects: Avery-Hansen-Hotz Quasi-MLE;

(b) Endogenous individual effects: FE methods: (b1) Manski's MSE; (b2) Chamberlain's Conditional Logit.

(c) Endogenous individual effects: RE methods: (c1) Chamberlain Correlated RE; and (c2) Heckman-Singer finite mixture model.

(2) Dynamic models:

(a) FE methods. (a1) Chamberlain's Conditional logit; (a2) Honore-Kyriatzidou conditional logit.

(b) RE methods. (b1) Heckman-Singer finite mixture model; (b2) Arellano-Carrasco.

3.1. Static Binary Choice Models

• Consider the Panel Data BCM:

$$Y_{it} = 1\{X_{it}'\beta - \varepsilon_{it} \geq 0\}$$

where ε_{it} is independent of $\{X_{i1}, X_{i2}, ..., X_{iT}\}$, i.e., strictly exogenous regressors.

• We have panel data with N individuals and T periods where N is large and T is small. We want to estimate β.

Avery-Hansen-Hotz Pseudo MLE

• **Avery-Hansen-Hotz** (IER, 1983) provide a simple estimator that is robust to serial correlation. They show that a method that estimates β using a standard Probit (or logit) model that ignores the serial correlation in ε_{it} is root-N consistent and asymptotically normal.

• Consider the Pseudo- log likelihood function:

$$l(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \ \ln \Phi\left(x_{it}'\beta\right) + (1 - y_{it}) \ln \Phi\left(-x_{it}'\beta\right)$$

And let $\hat{\beta}_{AHH}$ be the value of β that maximizes this function. This is Avery-Hansen-Hotz estimator.

• If the distribution of ε_{it} is normal with zero mean and constant variance (and the stochastic process of ε_{it} satisfies some standard stationarity conditions), then this estimator is consistent and asymptotically normal, regardless the serial correlation in ε_{it} over time (or across individuals).

Bias of the MLE based of FD and WG transformations

• Now, consider the more interesting case where $\varepsilon_{it} = \alpha_i - u_{it}$, and α_i and X_{it} can be correlated.

• In a BCM, the transformations of the model in First-Differences or Within-Groups does not eliminate the individual effect α_i.

$$\begin{aligned} \Delta Y_{it} &= 1\{X_{it}'\beta + \alpha_i - u_{it} \geq 0\} - 1\{X_{it-1}'\beta + \alpha_i - u_{it-1} \geq 0\} \\ &\neq 1\{\Delta X_{it}'\beta - \Delta u_{it} \geq 0\} \end{aligned}$$

• Therefore, a **MLE based on the equation** $\Delta Y_{it} = 1\{\Delta X_{it}'\beta - \Delta u_{it} \geq 0\}$ provides an inconsistent estimator of β.

• We will show later that Manski's Maximum Score Estimator can be used to obtain a consistent estimator of β that is somehow based on a first difference transformation of the model, but not exactly on the transformation above.

Bias of ML-Dummy Variables Estimator

• In the Static Linear PD model, we show that the LSDV estimator was consistent (for fixed T) and equivalent to the WG estimator.

• Unfortunately, that is not the case in the Static (or Dynamic) BCM.

• The estimator is defined as:

$$(\widehat{\beta}, \widehat{\boldsymbol{\alpha}}) = \arg \max_{\{\beta,\alpha\}} \sum_{i=1}^{N} l_i(\beta, \alpha_i)$$

where

$$l_i(\beta, \alpha_i) = \sum_{t=1}^{T} y_{it} \ \ln\Big(F(x_{it}'\beta + \alpha_i)\Big) + (1 - y_{it}) \ \ln\Big(1 - F(x_{it}'\beta + \alpha_i)\Big)$$

Bias of ML-Dummy Variables Estimator (2)

- The likelihood equations are:

$$\text{With respect to } \beta: \quad \sum_{i=1}^{N} \frac{\partial l_i(\widehat{\beta}, \widehat{\alpha}_i)}{\partial \beta} = 0$$

$$\text{With respect to } \alpha_i: \quad \frac{\partial l_i(\widehat{\beta}, \widehat{\alpha}_i)}{\partial \alpha_i} = 0$$

where

$$\sum_{i=1}^{N} \frac{\partial l_i(\widehat{\beta}, \widehat{\alpha}_i)}{\partial \beta} = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{x_{it}\ f(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)}{F(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)\ [1 - F(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)]} \left(y_{it} - F(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)\right)$$

$$\frac{\partial l_i(\widehat{\beta}, \widehat{\alpha}_i)}{\partial \alpha_i} = \sum_{t=1}^{T} \frac{f(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)}{F(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)\ [1 - F(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)]} \left(y_{it} - F(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)\right)$$

Bias of ML-Dummy Variables Estimator (3)

• For instance, for the Logit model, $\frac{f(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i)}{F(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i)\,[1-F(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i)]}=1$ such that the likelihood equations become:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\left[y_{it}-\frac{\exp\left(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i\right)}{1+\exp\left(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i\right)}\right]=0$$

$$\sum_{t=1}^{T}\left[y_{it}-\frac{\exp\left(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i\right)}{1+\exp\left(x_{it}^{'}\widehat{\beta}+\widehat{\alpha}_i\right)}\right]=0$$

• We can use a BHHH method to compute $(\widehat{\beta},\widehat{\boldsymbol{\alpha}})$. **Greene (Econometrics Journal, 2004)** has developed a computationally efficient method to calculate this estimator [in the spirit of Within Groups transformation, but in a sequential

method]. In, particular we do not need to invert any matrix with dimension $N + K$ to compute this estimator.

Bias of ML-Dummy Variables Estimator (4)

• Though the conputation of a Dummy Variables (Fixed Effects) estimator for PD-BCM is computationally simple, the estimator of β is **inconsistent as** $N \to \infty$ **and** T **is fixed**. It is only consistent when T also goes to infinity.

• This estimator of β does not share the nice properties of the LSDV estimator in linear models.

• The reason is that, in this model, as **as** $N \to \infty$ **and** T **is fixed**

$$cov(\widehat{\beta}, \widehat{\boldsymbol{\alpha}}) \neq 0$$

The estimator $\widehat{\beta}$ is asymptotically correlated with $\widehat{\boldsymbol{\alpha}}$ such that the asymptotic estimation error in $\widehat{\boldsymbol{\alpha}}$ contaminates the estimator $\widehat{\beta}$.

Bias of ML-Dummy Variables Estimator: Example

• Consider an example with $T = 2$, only one explanatory variable x_{it} that is the dummy variable for $t = 2$, and distribution $F(.)$ that is symmetric around the median $= 0$:

$$\begin{aligned} Y_{i1} &= 1\{\alpha_i - u_{i1} \geq 0\} \\ Y_{i2} &= 1\{\beta + \alpha_i - u_{i2} \geq 0\} \end{aligned}$$

• The likelihood equations are:

$$\sum_{i=1}^{N} \left[y_{i2} - \frac{\exp\left(\widehat{\beta} + \widehat{\alpha}_i\right)}{1 + \exp\left(\widehat{\beta} + \widehat{\alpha}_i\right)} \right] = 0$$

$$(y_{i1} + y_{i2}) - \frac{\exp\left(\widehat{\alpha}_i\right)}{1 + \exp\left(\widehat{\alpha}_i\right)} - \frac{\exp\left(\widehat{\beta} + \widehat{\alpha}_i\right)}{1 + \exp\left(\widehat{\beta} + \widehat{\alpha}_i\right)} = 0$$

Bias of ML-Dummy Variables Estimator: Example

• For observations with $(y_{i1}, y_{i2}) = (0, 0)$, we have that $0 - F(\widehat{\alpha}_i) - F\left(\widehat{\beta} + \widehat{\alpha}_i\right) = 0$, and this implies that: (a) $\widehat{\alpha}_i \to -\infty$; and (b) these observations do not contribute to the estimator $\widehat{\beta}$ because $l_i(\widehat{\beta}, \widehat{\alpha}_i) \to 0$ for any $\widehat{\beta}$.

• For observations with $(y_{i1}, y_{i2}) = (1, 1)$, we have that $1 - F(\widehat{\alpha}_i) + 1 - F\left(\widehat{\beta} + \widehat{\alpha}_i\right)$, and this implies that: (a) $\widehat{\alpha}_i \to +\infty$; and (b) these observations do not contribute to the estimator $\widehat{\beta}$ because $l_i(\widehat{\beta}, \widehat{\alpha}_i) \to 0$ for any $\widehat{\beta}$.

Bias of ML-Dummy Variables Estimator: Example

• For observations with $(y_{i1}, y_{i2}) = (0, 1)$ or with $(y_{i1}, y_{i2}) = (1, 0)$, we have that

$$1 - F\left(\widehat{\alpha}_i\right) - F\left(\widehat{\beta} + \widehat{\alpha}_i\right) = 0$$

that implies $\widehat{\alpha}_i = -\dfrac{\widehat{\beta}}{2}$, such that $F\left(-\widehat{\beta}/2\right) + F\left(\widehat{\beta}/2\right) = 1$.

• Therefore, the concentrated log-likelihood function is:

$$l(\beta) = \sum_{i=1}^{N} 1\{y_{i1} = 0, y_{i2} = 1\} \ \ln F\left(\frac{\beta}{2}\right) + 1\{y_{i1} = 1, y_{i2} = 0\} \ \ln F\left(\frac{-\beta}{2}\right)$$

Bias of ML-Dummy Variables Estimator: Example

- Define $p \equiv F\left(\frac{\beta}{2}\right)$. The concentrated log-likelihood is maximized at:

$$\widehat{p} = \frac{\sum_{i=1}^{N} 1\{y_{i1} = 0, y_{i2} = 1\}}{\sum_{i=1}^{N} 1\{y_{i1} + y_{i2} = 1\}}$$

- And the MLE of β is:

$$\widehat{\beta} = 2\ F^{-1}\left(\widehat{p}\right) = 2\ F^{-1}\left(\frac{\sum_{i=1}^{N} 1\{y_{i1} = 0, y_{i2} = 1\}}{\sum_{i=1}^{N} 1\{y_{i1} + y_{i2} = 1\}}\right)$$

Bias of ML-Dummy Variables Estimator: Example

- Is this a consistent estimator of β?

- It is clear that:

$$p \lim_{N\to\infty} \widehat{\beta} = 2\ F^{-1}\left(p \lim_{N\to\infty} \widehat{p}\right) = 2\ F^{-1}\left(\frac{\Pr\left(Y_{i1}=0, Y_{i2}=1\right)}{\Pr\left(Y_{i1}+Y_{i2}=1\right)}\right)$$

- In general, $2\ F^{-1}\left(\frac{\Pr\left(Y_{i1}=0, Y_{i2}=1\right)}{\Pr\left(Y_{i1}+Y_{i2}=1\right)}\right) \neq \beta$, and this ML-DV estimator $\widehat{\beta}$ is inconsistent.

Bias of ML-Dummy Variables Estimator: Example

• For instance, for the logit model, we can show that $p = \dfrac{\Pr\left(Y_{i1} = 0, Y_{i2} = 1\right)}{\Pr\left(Y_{i1} + Y_{i2} = 1\right)}$ does not depend of α_i and:

$$p = \frac{\Pr\left(Y_{i1} = 0, Y_{i2} = 1\right)}{\Pr\left(Y_{i1} + Y_{i2} = 1\right)} = \frac{\exp\left(\beta\right)}{1 + \exp\left(\beta\right)} = F\left(\beta\right)$$

• Therefore, for the logit model:

$$p \lim_{N \to \infty} \widehat{\beta} = 2 \; F^{-1}\left(F\left(\beta\right)\right) = 2 \; \beta$$

Fixed Effects Estimators for Static Panel Data BCM

- As in the case of linear panel data models, we distinguish two approaches:

  (a) Fixed Effects approach: no assumption on the joint distribution of X_i and α_i.

  (b) Random Effects approach: there is a parametric assumption on the joint distribution of X_i and α_i.

- We consider two fixed effects estimators:

  1. Chamberlain's Conditional Logit model.

  2. Manski's MSE applied to Panel Data BCM

Chamberlain Conditional Logit

• Consider the BCM $Y_{it} = 1\{X_{it}'\beta + \alpha_i - u_{it} \geq 0\}$ where u_{it} has a logistic distribution. Therefore,

$$\Pr(Y_{it} = 1 \mid X_{it},\ \alpha_i) = \frac{\exp\left\{X_{it}'\beta + \alpha_i\right\}}{1 + \exp\left\{X_{it}'\beta + \alpha_i\right\}}$$

And if u_{it} is independent over time:

$$\Pr(Y_{i1}, Y_{i2}, \ldots, Y_{iT} \mid X_i,\ \alpha_i) = \prod_{t=1}^{T} \frac{\exp\left\{Y_{it}\ (X_{it}'\beta + \alpha_i)\right\}}{1 + \exp\left\{X_{it}'\beta + \alpha_i\right\}}$$

• Define the random variable $S_i = \sum_{t=1}^{T} Y_{it}$ that represents the number of times that the binary event has occurred during the T sample periods.

Chamberlain Conditional Logit (2)

• Let $Y_i = \{y_{i1}, y_{i2}, ..., y_{iT}\}$, and $X_i = \{x_{i1}, x_{i2}, ..., x_{iT}\}$. The key result behind Chamberlain conditional logit estimator is that:

$$\Pr(Y_i \mid X_i, S_i, \alpha_i, \beta) = \Pr(Y_i \mid X_i, S_i, \beta)$$

i.e., it does not depend on α_i.

• First, by the chain rule, it is clear that: $\Pr(Y_i, S_i \mid X_i, \alpha_i) = \Pr(Y_i \mid X_i, S_i, \alpha_i) \Pr(S_i \mid X_i, \alpha_i)$, and therefore:

$$\Pr(Y_i \mid X_i, S_i, \alpha_i) = \frac{\Pr(Y_i, S_i \mid X_i, \alpha_i)}{\Pr(S_i \mid X_i, \alpha_i)} = \frac{\Pr(Y_i \mid X_i, \alpha_i)}{\Pr(S_i \mid X_i, \alpha_i)}$$

• Given our logit model and that u_{it} is iid over time, we have that the probability $\Pr(Y_i \mid X_i, \alpha_i)$ is:

$$\begin{aligned}
\Pr(Y_i \mid X_i, \alpha_i) &= \prod_{t=1}^{T} \Pr(y_{it} \mid x_{it}, \alpha_i) \\
&= \prod_{t=1}^{T} \frac{\exp\left\{ y_{it}\left[x_{it}'\beta + \alpha_i\right] \right\}}{1 + \exp\left\{ x_{it}'\beta + \alpha_i \right\}} \\
&= \frac{\exp\left\{\sum_{t=1}^{T} y_{it} x_{it}'\beta + S_i\alpha_i\right\}}{\prod_{t=1}^{T}\left[1 + \exp\left\{x_{it}'\beta + \alpha_i\right\}\right]}
\end{aligned}$$

• To derive the expression for $\Pr(S_i \mid X_i, \alpha_i)$ it is useful to define the sets:

$$H(S_i) = \left\{ D = (d_1, d_2, ..., d_T) \in \{0,1\}^T : \sum_{t=1}^{T} d_t = S_i \right\}$$

• Using this definition, we can write:

$$\begin{aligned}
\Pr\left(S_i \mid X_i, \alpha_i\right) &= \sum_{D \in H(S_i)} \Pr\left(D \mid X_i, \alpha_i\right) \\
&= \sum_{D \in H(S_i)} \prod_{t=1}^{T} \Pr\left(d_t \mid x_{it}, \alpha_i\right) \\
&= \frac{\sum_{D \in H(S_i)} \exp\left\{\sum_{t=1}^{T} d_t x_{it}' \beta + S_i \alpha_i\right\}}{\prod_{t=1}^{T}\left[1 + \exp\left\{x_{it}' \beta + \alpha_i\right\}\right]}
\end{aligned}$$

• Combining the previous expressions for $\Pr\left(Y_i \mid X_i, \alpha_i\right)$ and $\Pr\left(S_i \mid X_i, \alpha_i\right)$ we have that:

$$\Pr\left(Y_i \mid X_i, S_i, \alpha_i\right) = \frac{\Pr\left(Y_i \mid X_i, \alpha_i\right)}{\Pr\left(S_i \mid X_i, \alpha_i\right)} = \frac{\exp\left\{ \sum_{t=1}^{T} y_{it} x_{it}' \beta \right\}}{\sum_{d \in H(S_i)} \exp\left\{ \sum_{t=1}^{T} d_t x_{it}' \beta \right\}}$$

which does not depend on α_i. Therefore, $\Pr(Y_i \mid X_i, S_i, \alpha_i) = \Pr(Y_i \mid X_i, S_i)$.

- The conditional log-likelihood function

$$\mathrm{l}(\beta) = \sum_{i=1}^{n} \log \Pr(Y_i \mid X_i, S_i)$$

Using the expression for $\Pr(Y_i \mid X_i, S_i)$ obtained before, we have that

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \log \left[\frac{\exp\left\{ \sum_{t=1}^{T} y_{it} x_{it}' \beta \right\}}{\sum_{D \in H(S_i)} \exp\left\{ \sum_{t=1}^{T} d_t x_{it}' \beta \right\}} \right] \\ &= \sum_{i=1}^{n} \sum_{t=1}^{T} y_{it} x_{it}' \beta - \sum_{i=1}^{n} \log \left[\sum_{D \in H(S_i)} \exp\left\{ \sum_{t=1}^{T} d_t x_{it}' \beta \right\} \right] \end{aligned}$$

This function is globally concave in β.

MSE for Panel Data BCM. Manski (Ectca, 1987)

- Consider the BCM $Y_{it} = 1\{X_{it}'\beta + \alpha_i - u_{it} \geq 0\}$. The model implies that:

$$\Delta Y_{it} = 1\{X_{it}'\beta + \alpha_i - u_{it} \geq 0\} - 1\{X_{it-1}'\beta + \alpha_i - u_{it-1} \geq 0\}$$

- Therefore, conditional on $\Delta Y_{it} \neq 0$, we have that:

$$\Delta Y_{it} = \begin{cases} 1 & \text{if } \Delta X_{it}'\beta - \Delta u_{it} > 0 \\ \\ -1 & \text{if } \Delta X_{it}'\beta - \Delta u_{it} < 0 \end{cases}$$

If $\Delta Y_{it} \neq 0$ then: either (a) $Y_{it-1} = 0$ and $Y_{it-1} = 1$, and this implies that $\Delta Y_{it} = 1$ and that $\Delta X_{it}'\beta - \Delta u_{it} > 0$; or (b) $Y_{it-1} = 1$ and $Y_{it-1} = 0$, and this implies that $\Delta Y_{it} = -1$ and that $\Delta X_{it}'\beta - \Delta u_{it} < 0$.

• **Assumption:** Conditional on X_{it}, X_{it-1}, and α_i, the variables u_{it} and u_{it-1} have the same probability distribution with support $(-\infty, \infty)$.

• It is possible to show that, this assumption implies that:

$$m\mathbb{E}dian(\ \Delta u_{it} \mid \Delta X_{it},\ \Delta Y_{it} \neq 0) = 0$$

Therefore, we can apply the MSE to the model:

$$\Delta Y_{it} = \begin{cases} 1 & \text{if} \quad \Delta X_{it}'\beta - \Delta u_{it} > 0 \\ \\ -1 & \text{if} \quad \Delta X_{it}'\beta - \Delta u_{it} < 0 \end{cases}$$

with $\Delta Y_{it} \neq 0$.

• Given a sample $\{y_{it}, x_{it}\}$, the score function is:

$$S(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T} 1\{\Delta y_{it} = 1\}\ 1\{\Delta X_{it}'\beta > 0\} + 1\{\Delta y_{it} = -1\}\ 1\{\Delta X_{it}'\beta < 0\}$$

That is just equal to the number of observations for which we score a correct prediction for the sign of Δy_{it} if we use the sign of $\Delta X_{it}'\beta$ as a predictor.

• The MSE is the value of β that maximizes the score function:

$$\hat{\beta}_{MS\mathbb{E}} = \arg\max_{\beta} \quad S(\beta)$$

• This estimator has the same properties as in the cross-section case: $N^{1/3}-$consistent, asymptotically non-normal, and possibly not uniquely defined in finite samples.

• Following Horowitz, we can define a smooth-MSE for this estimator by replacing the discontinuous function $1\{\Delta X_{it}'\beta > 0\}$ with a continuously differ-

entiable function $K_N\left(\Delta X_{it}'\beta\right)$ such that $K_N(\Delta X_{it}'\beta)$ converges uniformly to $1\{\Delta X_{it}'\beta > 0\}$ as N goes to infinity.

Correlated Random Effects Static Probit model

- Suppose that α_i and X_i have a joint normal distribution. Then:

$$\begin{aligned} \alpha_i &= X_{i1}'\lambda_1 + X_{i2}'\lambda_2 + ... + X_{iT}'\lambda_T + \mathbb{E}_i \\ &= X_i'\,\lambda + \mathbb{E}_i \end{aligned}$$

where $\mathbb{E}_i$ is normally distributed and independent of α_i.

- Solving this expression in the equation of the BCM, we have:

$$\begin{aligned} Y_{it} &= 1\{\, X_{it}'\beta + X_i'\,\lambda + (\mathbb{E}_i - u_{it}) \geq 0\} \\ &= 1\{\, X_i'\pi_t - u_{it}^* \geq 0\} \end{aligned}$$

where $\pi_t = (\lambda_1, ..., \lambda_{t-1}, \beta + \lambda_t, \lambda_{t+1}, ..., \lambda_T)$ and $u_{it}^* = u_{it} - \mathbb{E}_i$.

Random Effects Static Probit model (2)

• If u_{it} is normally distributed (the original model is a Probit model) and independent of X_i, then u_{it}^* is also normally distributed and independent of X_{it}. Then, $Y_{it} = 1\{ X_i'\pi_t - u_{it}^* \geq 0\}$ is a standard Probit model and we can estimate the parameters π_t using MLE or the Pseudo-MLE of Avery-Hansen-Hotz.

• Given these estimate of π_t and of its variance matrix, we can estimate β and λ using a simple MD estimator. Given that the system of equations that relates β and λ and π is linear, the MD estimator has a simple closed form expression for the estimator of $\hat{\beta}$ in terms of $\hat{\pi}$ and $\widehat{Var}(\hat{\pi})$.

3.2. Dynamic Binary Choice Models

• Chamberlain (1985) Conditional Logit model for autorregresive PD BCM.

• Honore and Kyriadzidou (ECTCA, 2000) extension to include also strictly exogenpus regressors.

Conditional MLE for Dynamic PD Logit

• Consider the dynamic panel data logit model

$$Y_{it} = 1\left\{ \beta\ Y_{i,t-1} + \alpha_i - u_{it} > 0 \right\}$$

where u_{it} has a logistic distribution.

• In this model $S_i = \sum_{t=1}^{T} y_{it}$ is not a sufficient statistic for α_i. That is, it is not true that:

$$\mathsf{Pr}\left(Y_{it} \mid Y_{it-1}, S_i, \alpha_i\right) = \mathsf{Pr}\left(Y_{it} \mid Y_{it-1}, S_i\right)$$

• However, fortunately, there is an alternative way to construct a sufficient statistic for α_i controlling for (Y_{i1}, Y_{iT}, S_i).

Conditional MLE for Dynamic PD Logit (2)

• Suppose that $T = 4$ and let $Y_i = \{y_{i1}, y_{i2}, y_{i3}, y_{i4}\}$ be the choice history for individual i. We distinguish four sets of choice histories:

$$\begin{aligned} A &= \{y_1, 1, 0, y_4\} \\ B &= \{y_1, 0, 1, y_4\} \\ C &= \{y_1, 1, 1, y_4\} \\ D &= \{y_1, 0, 0, y_4\} \end{aligned}$$

• Define $S_i = 1(Y_i \in A \cup B)$. We will show that:

$$\Pr\left(Y_i \mid 1(Y_i \in A \cup B),\, \alpha_i,\, \beta\right) = \Pr\left(Y_i \mid 1(Y_i \in A \cup B),\, \beta\right)$$

• We can construct a (Conditional) likelihood function based on the probabilities $\Pr\left(Y_i \mid 1(Y_i \in A \cup B),\, \beta\right)$, and the corresponding MLE is a consistent estimator of β.

Conditional MLE for Dynamic PD Logit (3)

- First, we obtain $\Pr(Y_i \mid \alpha_i, A \cup B)$. By Bayes' rule we have that:

$$\Pr(Y_i \mid \alpha_i,\, A \cup B) = \frac{\Pr(Y_i \mid \alpha_i)}{\Pr(A \cup B \mid \alpha_i)} = \frac{\Pr(Y_i \mid \alpha_i)}{\Pr(A \mid \alpha_i) + \Pr(B \mid \alpha_i)}$$

- Note that:

$$\begin{aligned}
\Pr(A \mid \alpha_i) &= \Pr(y_1 \mid \alpha_i)\ \Pr(1 \mid y_1, \alpha_i)\ \Pr(0 \mid 1, \alpha_i)\ \Pr(y_4 \mid 0, \alpha_i) \\
&= \Pr(y_1 \mid \alpha_i)\ \frac{\exp(\beta y_1 + \alpha_i)}{1 + \exp(\beta y_1 + \alpha_i)}\ \frac{1}{1 + \exp(\beta + \alpha_i)}\ \frac{\exp(y_4\ \alpha_i)}{1 + \exp(\alpha_i)}
\end{aligned}$$

- And

$$\begin{aligned}
\Pr(B \mid \alpha_i) &= \Pr(y_1 \mid \alpha_i)\ \Pr(0 \mid y_1, \alpha_i)\ \Pr(1 \mid 0, \alpha_i)\ \Pr(y_4 \mid 1, \alpha_i) \\
&= \Pr(y_1 \mid \alpha_i)\ \frac{1}{1 + \exp(\beta y_1 + \alpha_i)}\ \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}\ \frac{\exp(y_4\,[\beta + \alpha_i])}{1 + \exp(\beta + \alpha_i)}
\end{aligned}$$

Conditional MLE for Dynamic PD Logit (4)

• Therefore,

$$\begin{aligned}\Pr(A \mid \alpha_i,\, A \cup B) &= \frac{\Pr(A \mid \alpha_i)}{\Pr(A \mid \alpha_i) + \Pr(B \mid \alpha_i)} \\ &= \frac{\exp(\beta\ [y_1 - y_4])}{1 + \exp(\beta\ [y_1 - y_4])}\end{aligned}$$

• The CMLE is the value of β that maximizes the Conditional log-likelihoof function:

$$\begin{aligned}l^C(\beta) &= \sum_i 1\{y_{i2} = 1, y_{i3} = 0\}\ \ln \Lambda(\beta\ [y_{1i} - y_{4i}]) \\ &+ 1\{y_{i2} = 0, y_{i3} = 1\} \ln \Lambda(-\beta\,[y_{1i} - y_{4i}])\end{aligned}$$

where $\Lambda(.)$ is the logistic function.

Conditional MLE for Dynamic PD Logit (5)

• In this simple model, with $T = 4$ and without exogenous covariates X, it is simple to show that CMLE of β is:

$$\hat{\beta} = \log\left(\frac{\#\{1,1,0,0\} + \#\{0,0,1,1\}}{\#\{0,1,0,1\} + \#\{1,0,1,0\}}\right)$$

where $\#\{y_1, y_2, y_3, y_4\}$ means the number of individuals in the sample with a choice history $\{y_1, y_2, y_3, y_4\}$.

• **Interpretation (intuition).**

- If time persistence in y_{it} is generated by individual heterogeneity, then for an individual, we should have persistence in only one of the two states, either at 0 (if α_i is small) or at 1 (if α_i is large).

- If time persistence in y_{it} is generated by true state dependence ($\beta > 0$), then we should have persistence in both states, 0 and 1.

• Choice histories $\{1, 1, 0, 0\}$ and $\{0, 0, 1, 1\}$ are the only histories that provide evidence of persistence in both states. The larger the sample frequency of these histories the stronger is the evidence of structural state dependence and the larger the estimator of β. The choice histories $\{0, 1, 0, 1\}$ and $\{1, 0, 1, 0\}$ are the only histories that provide evidence of no persistence in any of the two states. The larger the sample frequency of these histories the smaller the estimator of β.

Conditional MLE for Dynamic PD Logit (6)

• It is possible to extend the previous result to Panel Data with any value of $T \geq 4$, to obtain the following expression.

• Let $\mathbf{Y}_i = \{Y_{i1}, Y_{i2}, ..., Y_{iT}\}$ and $s_i = \sum_{t=1}^{T} y_{it}$. Then,

$$\Pr\left(\mathbf{Y}_i \mid \alpha_i,\ s_i,\ y_{i1}\ y_{iT}\right) = \frac{\exp\left(\beta\ \sum_{t=2}^{T-1} y_{it}\ y_{it-1}\right)}{\sum_{\mathbf{d}\in C_i} \exp\left(\beta\ \sum_{t=2}^{T-1} d_t\ d_{t-1}\right)}$$

where:

$$C_i = \left\{(d_1, d_2, ..., d_T) \in \{0,1\}^T : \sum_{t=1}^{T} d_t = s_i;\ d_1 = y_{i1};\ d_T = y_{iT}\right\}$$

Honore and Kyriadzidou (ECTCA, 2000)

• Consider the dynamic panel data logit model

$$Y_{it} = 1\left\{ \beta\ Y_{i,t-1} + X_{it}'\ \delta + \alpha_i - u_{it} > 0 \right\}$$

where u_{it} has a logistic distribution, and X_{it} are strictly exogenous regressors with respect to u_{it}.

• For $T = 4$, they show that $(s_i,\ y_{i1}\ y_{i4})$ are sufficient statistics for α_i only if we condition on $x_{i3} = x_{i4}$.

• They propose a version of the CMLE that incorporates kernel weights that depend on the distance $\|x_{i3} - x_{i4}\|$.