ECON 203C: System Models
TA Note 4: Version 2

Binary Choice (Probit and Logit) Models

Hisayuki Yoshimoto
Last Modified: May 3, 2008

Abstract: In section 1, we review Bernoulli random variable that is basis of binary choice model. In section 2, we solve Final 2003 Question 5, the ML estimation of Bernoulli distribution. In section 3, we discuss binary choice models and their asymptotic distributions. In section 4, we solve Final 2006 Question 1 and learn identification problem. In section 5 and 6, we solve Comp 2003S Part III Question 2 and Comp 2003F Part III Questions, binary choice model problems.

1 Review of Bernoulli Random Variable

Since binary choice models are extension of Bernoulli random variable distribution, let's start this TA note with reviewing Bernoulli distribution. Bernoulli random variable y takes only two values 0 and 1 with probabilities

$$\begin{cases} y=1 & : \text{success} \quad \text{with probability } p \\ y=0 & : \text{failure} \quad \text{with probability } 1-p \end{cases}.$$

It has the (mass) density function

$$f(y|\,p) = p^y (1-p)^{1-y}.$$

The expectation and variance of Bernoulli random variable Y are

$$\begin{aligned} E[y] &= p \cdot 1 + (1-p)\,0 \qquad \text{(definition of expectation)} \\ &= p \end{aligned}$$

and

$$\begin{aligned} Var[y] &= E\left[(Y - E(Y))^2\right] \qquad \text{(definition of variance)} \\ &= p(1-p)^2 + (1-p)(0-p)^2 \qquad (\text{since } E[Y] = p) \\ &= p(1-p). \end{aligned}$$

Note that the property $E[y] = p$ becomes crucial when we extend Bernoulli distribution to binary choice models. The following question explains how to obtain the estimate of p by the maximum likelihood method.

2 Final 2003: Question 5 - ML Estimation of Bernoulli Random Variable

Let $y_1, \ldots, y_n$ be a random sample from a Bernoulli distribution with the probability of success given by p.

(1) Write the likelihood function for p i.e. $L(p|\,y_1, \ldots, y_n)$.

Answer:

Since $\{y_i\}_{i=1}^n$ are Bernoulli, we have

$$\begin{cases} y_i=1 & : \text{success} \quad \text{with probability } p \\ y_i=0 & : \text{failure} \quad \text{with probability } 1-p \end{cases}$$

The density function of Bernoulli distribution is

$$f(y_i|\,p) = p^{y_i}(1-p)^{1-y_i}$$

Since $\{y_i\}_{i=1}^n$ are random (independent) sample, the likelihood function is

$$\begin{aligned} L(p|\, y_1, \ldots, y_n) &= f(y_{1,\ldots,}y_n|\, p) = \prod_{i=1}^n f(y_i|\, p) \quad \text{(since samples are random)} \\ &= \prod_{i=1}^n p^{y_i} (1-p)^{1-y_i} \end{aligned}$$

(2) Provide the MLE for p.
Answer:
The log-likelihood function is

$$l(p|\, y_1, \ldots, y_n) = \ln L(p|\, y_1, \ldots, y_n) = \sum_{i=1}^n \ln \left[p^{y_i} (1-p)^{1-y_i} \right] = \sum_{i=1}^n \left[y_i \ln p + (1-y_i) \ln (1-p) \right].$$

Taking f.o.c. w.r.t. p,

$$\begin{aligned} \frac{\partial}{\partial p} l(p|\, y_1, \ldots, y_n) &= \frac{\partial}{\partial p} \sum_{i=1}^n \left[y_i \ln p + (1-y_i) \ln (1-p) \right] = \sum_{i=1}^n \left[y_i \frac{1}{p} - (1-y_i) \frac{1}{1-p} \right] \\ &= \frac{1}{p} \sum_{i=1}^n y_i - \frac{1}{1-p} \sum_{i=1}^n (1-y_i) = 0, \end{aligned}$$

and

$$\begin{aligned} (1-p) \sum_{i=1}^n y_i &= p \sum_{i=1}^n (1-y_i) \\ \sum_{i=1}^n x_i - \underbrace{p \sum_{i=1}^n y_i}_{\text{cancel out}} &= pn - \underbrace{p \sum_{i=1}^n y_i}_{\text{cancel out}} \\ \hat{p}_{ML} &= \frac{1}{n} \sum_{i=1}^n y_i. \end{aligned}$$

(3) Derive the asymptotic distribution for p and provide the asymptotic covariance matrix.
Answer:
The asymptotic distribution of ML estimator is given by

$$\sqrt{N} (\hat{p}_{ML} - p) \xrightarrow{d} N\left(0, I_1^{-1}\right),$$

where 1-sample Fisher information I_1 is given by (minus) expectation of second derivative of log density

$$\begin{aligned} \frac{\partial^2}{\partial p^2} \ln f(y_i|\, p) &= \frac{\partial^2}{\partial p^2} \ln \left[p^{y_i} (1-p)^{1-y_i} \right] = \frac{\partial^2}{\partial p^2} \left\{ y_i \ln p + (1-y_i) \ln (1-p) \right\} \\ &= \frac{\partial}{\partial p} \left\{ y_i \frac{1}{p} + (1-y_i) \frac{-1}{1-p} \right\} = -\frac{y_i}{p^2} - \frac{1-y_i}{(1-p)^2}, \end{aligned}$$

and

$$\begin{aligned} I_1 &= -E\left[\frac{\partial^2}{\partial p^2} \ln f(y_i|\, p) \right] = -E\left[-\frac{y_i}{p^2} - \frac{1-y_i}{(1-p)^2} \right] \\ &= \frac{E[y_i]}{p^2} + \frac{1-E[y_i]}{(1-p)^2} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} \qquad \text{(since } E[y_i] = p, \text{ the property of Bernoulli distribution)} \\ &= \frac{1}{p} + \frac{1}{1-p} = \frac{1-p+p}{p(1-p)} = \frac{1}{p(1-p)}. \end{aligned}$$

Thus,

$$I_1^{-1} = \left(\frac{1}{p(1-p)} \right)^{-1} = p(1-p),$$

and asymptotic distribution of ML estimator $\hat{p}_{ML}$ is

$$\sqrt{n} (\hat{p}_{ML} - p) \xrightarrow{d} N(0, p(1-p)).$$

(4) Provide MLE for $p(1-p)$ and derive the asymptotic distribution of your estimator.

Answer:

Notice that $p(1-p)$ is the variance of Bernoulli random variable. Thus, here we are trying to obtain the asymptotic distribution of variance of Bernoulli random variable.

By the invariant principle, the ML estimator of $p(1-p)$ is

$$\hat{p}_{ML}\left(1-\hat{p}_{ML}\right) \xrightarrow{p} p(1-p).$$

Next, driving the asymptotic variance of $\hat{p}_{ML}\left(1-\hat{p}_{ML}\right)$ by applying the single-variate delta method. See footnote[1].

Define the continuous function

$$g(u) = u(1-u) = u - u^2.$$

The derivative of $g(u)$ w.r.t. u evaluate at $u = p$ is

$$\frac{d}{du} g(u) = 1 - 2u.$$

and

$$\frac{d}{du} g(u)\Big|_{u=p} = 1 - 2p$$

Therefore, by applying the single-variate delta method, we have

$$\sqrt{n}\left(\hat{p}_{ML}\left(1-\hat{p}_{ML}\right) - p(1-p)\right) = \sqrt{n}\left(g\left(\hat{p}_{ML}\right) - g(p)\right) \xrightarrow{d} N\left(0, \left[\frac{d}{du} g(u)\Big|_{u=p}\right]^2 \cdot p(1-p)\right) \sim N\left(0, (1-2p)^2 p(1-p)\right).$$

(5) Provide a consistent estimator for the asymptotic covariance matrix. Justify your answer.

Answer:

The consistent estimator of covariance matrix in (4) is obtained by

$$\left(1-2\hat{p}_{ML}\right)^2 \hat{p}_{ML}\left(1-\hat{p}_{ML}\right) \xrightarrow{p} (1-2p)^2 p(1-p) \qquad \text{(by WLLN, Slutzky, and continuity theorems).}$$

---

[1] **Single-Variate Delta Method**:

Let θ_n be a sequence of random variable that has asymptotic distribution $\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N\left(0, \sigma^2\right)$. For given function $g(x)$ and a specific value of θ, suppose that $\frac{d}{dx} g(x)\Big|_{x=\theta}$ exists and is not 0. Then, the asymptotic distribuion of the functin of random variable $g(\theta_n)$ is

$$\sqrt{n}\left(g(\theta_n) - g(\theta)\right) \xrightarrow{d} N\left(0, \left[\frac{d}{dx} g(x)\Big|_{x=\theta}\right]^2 \cdot \sigma^2\right).$$

Multi-Variate Delta Method:

Let θ_n be a sequence of $K \times 1$ random vector that has asymptotic distribution $\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N(0, \Sigma)$, where θ is $K \times 1$ vector and Σ is $K \times K$ covariance matrix. For given $L \times 1$ multi-demensional function $g(x)$ such that $g : \mathbb{R}^K \to \mathbb{R}^L$, and given specific $K \times 1$ vector of θ, suppose that $\underbrace{\frac{d}{dx'} g(x)\Big|_{x=\theta}}_{L \times K}$ exists and is not equal to $0_{L \times K}$. Then

$$\sqrt{n}\left(g(\theta_n) - g(\theta)\right) \xrightarrow{d} N\left(0_{L\times 1}, \left[\frac{d}{dx'} g(x)\Big|_{x=\theta}\right] \Sigma \left[\frac{d}{dx'} g(x)\Big|_{x=\theta}\right]'\right)$$

$$\sqrt{n}\left(\underbrace{g(\theta_n)}_{L\times 1} - \underbrace{g(\theta)}_{lL1}\right) \xrightarrow{d} N\left(0_{L\times 1}, \underbrace{\left[\frac{d}{dx'} g(x)\Big|_{x=\theta}\right]}_{L\times K} \underbrace{\Sigma}_{K\times K} \underbrace{\left[\frac{d}{dx'} g(x)\Big|_{x=\theta}\right]'}_{K\times L}\right).$$

Extension from Bernoulli to Binary Choice Models:

From now on, we begin to consider the case in which p (the probability of the event $y = 1$ (success)) varies across individuals. Assume that p_i is the probability that individual i have the event $y = 1$ (success) and p_i is a function of dependent variable x_i

$$p_i = F(x_i).$$

Writing more formally, $p_i = F(x_i)$ is the conditional probability of the event $y_i = 1$ given x_i,

$$\Pr(y_i = 1|\, x_i) = p_i = F(x) \tag{1}$$

Further more, remind that in Bernoulli model, we have the property $E[y_i] = p$. Now, since p_i varies across individual with regressor x_i, we have $p_i = E[y_i|\, x_i]$ and

$$\Pr(y_i = 1|\, x_i) = p_i = F(x_i) = E[y_i|\, x_i] \tag{2}$$

The different choices of functional form of $F(\cdot)$ in equation (1) provide us different binary choice models. In the next section, we will discuss the three different choices of $F(\cdot)$, linear probability, probit, and logit models.

3 Binary Choice Models

3.1 Motivation

In Hamlet, William Shakespeare wrote "to be or not to be, that is the question." Unfortunately, your TA does not understand literature at all, but he understands what Shakespeare described was binary choice problem. Besides the example of Shakespeare, we empirically know that people (economic agents) have to face many binary choice or outcome problems in their lives such as to marry or not to marry, to divorce or not to divorce, to go or not to go to college, to buy or not to buy a car, to live in downtown or in suburb, to be accepted or not to be accepted after a job interview, to pass or not to pass an exam, etc.. So it is natural for us to ask "how do we construct econometric (and economic) models of binary choice problems? and how do economic agents decide binary choices?"

3.2 Benchmark Example

Let's set up the benchmark example of binary choice models. In labor economics, researchers investigate married women's binary choice behavior, work or not work. Here, not to work means engaging in household jobs. We can set up dependent and explanatory variables flowingly

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if married woman work} \\ 0 & \text{otherwise} \end{cases} \\ x_i &= \begin{bmatrix} 1 \\ edu_i \\ hus\_inc_i \\ num\_child_i \end{bmatrix}. \end{aligned}$$

where edu_i is educational length of married woman i, hus_inc_i is married woman i's husband's income, and num_child_i is number of children. We are interested in investigating correlation between y_i and elements of x_i such as correlation between married women workings status (working or not working) and their educational length

3.3 Binary Choice Models

In binary choice model, we are interested in estimate the conditional probability, $\Pr(y_i = 1|\, x_i)$, the probability that agent i chooses $y_i = 1$ conditional on dependent variable x_i. In our bench mark example, we are interested in the probability that a married woman works given her and her family status information such as educational length and number of children. Remind that from equation (2), we have

$$\Pr(y_i = 1|\, x_i) = \underbrace{F(x_i)}_{\text{we can choose this functional form}} = E[y_i|\, x_i] \tag{3}$$

The choice of functional form $F(\cdot)$ is up to an econometrician and different choice of functional form $F(\cdot)$ provides different models.

3.3.1 Linear Probability Model

The simplest choice of functional form $F(\cdot)$ is linear function (so $F(x_i) = x_i'\beta$)

$$\Pr(y_i = 1|\, x_i) = x_i'\beta \qquad \text{(Linear Probability Model)}$$

We call this specification *linear probability model*. To estimate β, we just apply least square regression to the model equation

$$\underbrace{y_i}_{\text{0 or 1}} = x_i'\beta + \underbrace{\varepsilon_i}_{\text{error term}}$$

and obtain the least square estimator $\hat{\beta}_{LS}$.The predicted value y_i, say $\hat{y}_i$, is actually predicted probability of the event $y_i = 1$ conditional on dependent variable x_i,

$$\Pr(\widehat{y_i = 1}|\, x_i) = \hat{y}_i = x_i'\hat{\beta}_{LS},$$

However, in the linear probability model, predicted probability $\hat{y}_i$ might be less than zero or larger than one. For this reasons, linear probability model is not commonly used in econometrics.

3.3.2 Probit and Logit Models

As we discussed above, linear probability model has the practical problem. Instead we know assume *latent variable* y_i^* such that[2]

$$y_i^* = \underbrace{x_i'\beta}_{\text{part (1)}} - \underbrace{\varepsilon_i}_{\text{part (2)}}. \tag{4}$$

Note that y_i^* is a hypothetical continuous variable that affects choice of agents (choosing $y_i = 1$ or 0). Also, latent variable y_i^* is unobservable to econometricians (precisely speaking, econometricians only know positive or negative status of y_i^* due to the reason stating below). Notice that latent variable y^* consists of two parts

$$\begin{cases} \text{Part (1)} & x_i'\beta & : & \text{the part which is explained by observable explanatory variable } x_i \\ \text{Part (2)} & \varepsilon_i & : & \text{stochastic error term that is unobservable to econometrician (but observable to agent } i) \end{cases}$$

y_i^* is unobservable to econometricians, but it determine observable dependent variable y_i by the assumption

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}, \tag{5}$$

In the bench mark example, we can interpret y_i^* as utility of working that is defined as

$$\begin{aligned} y_i^* &= x_i'\beta - \varepsilon_i \\ &= \beta_0 + \beta_1 edu_i + \beta_2 hus\_inc_i + \beta_3 num\_child_i - \varepsilon_i \end{aligned}$$

and married woman chooses

$$y_i = \begin{cases} 1 & \text{work} & \text{if } \beta_0 + \beta_1 edu_i + \beta_2 hus\_inc_i + \beta_3 num\_child_i - \varepsilon_i > 0 \\ 0 & \text{not work} & \text{if } \beta_0 + \beta_1 edu_i + \beta_2 hus\_inc_i + \beta_3 num\_child_i - \varepsilon_i \leq 0 \end{cases}.$$

You can interpret y_i^* is utility of working (or willingness to work) of a married woman i and ε_i is stochastic error in utility. A married woman i works if utility is positive and does not work if utility is not positive. Utility of working y^* is not observable to econometricians (except it's positive or negative status), but married woman i surely knows y_i^* when she decides whether working or not working.

Go back to the general discussion. Assume that ε_i in model equation (4) has cdf function F (or equivalently, the function $F(\cdot)$ in the equation (3) is cdf function). Then, we have conditional probabilities

$$\begin{cases} y_i = 1 & \underset{\text{by assumption (5)}}{\Leftrightarrow} & y_i^* > 0 & \underset{\text{since } y_i^* = x_i'\beta - \varepsilon_i}{\Leftrightarrow} & x_i'\beta - \varepsilon_i > 0 & \Leftrightarrow & x_i'\beta > \varepsilon_i & \Leftrightarrow & \Pr(y_i = 1|\, x_i) = F(x_i'\beta) \\ y_i = 0 & \underset{\text{by assumption (5)}}{\Leftrightarrow} & y_i^* \leq 0 & \underset{\text{since } y_i^* = x_i'\beta - \varepsilon_i}{\Leftrightarrow} & x_i'\beta - \varepsilon_i \leq 0 & \Leftrightarrow & x_i'\beta \leq \varepsilon_i & \Leftrightarrow & \Pr(y_i = 0|\, x_i) = 1 - F(x_i'\beta) \end{cases}$$

Now, different choices of distribution function F provides different binary choice models, probit and logit model.

[2] I wrote $y_i^* = x_i'\beta - \varepsilon_i$, not $y_i^* = x_i'\beta + \varepsilon_i$. The change of plus-minus sign in front of error term is innocuous as long as the distribution of ε_i is symmetric around 0.

Probit Model Probit model assumes the error term ε_i in the model equation (4) is distributed as *standard normal* and $F(\cdot)$ in equation (3) is a cdf function of standard normal distribution such that[3]

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du.$$

Then, $\Pr(y_i = 1|\, x_i)$ and $\Pr(y_i = 0|\, x_i)$ are expressed as

$$\begin{cases} \Pr(y_i = 1|\, x_i) = \Phi(x_i'\beta) & = \int_{-\infty}^{x'\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du \\ \Pr(y_i = 0|\, x_i) = 1 - \Phi(x_i'\beta) & = 1 - \int_{-\infty}^{x'\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du \end{cases}$$

Notice that in the probit model, we assume the error term in model equation (4) has standard deviation 1. This setting is valid

since the latent variable y_i^* is scale free (multiplying any positive constant to y_i^* does not affect the choice rule (5)).

Logit Model Logit model assumes the error term ε_i in the model equation (4) is distributed as *logit distribution* and $F(\cdot)$ in equation (3) is a cdf function of logit distribution such that[4]

$$F(z) = \Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}.$$

Then, $\Pr(y_i = 1|\, x_i)$ and $\Pr(y_i = 0|\, x_i)$ are expressed as

$$\begin{cases} \Pr(y_i = 1|\, x_i) = \Lambda(x_i'\beta) & = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \\ \Pr(y_i = 0|\, x_i) = 1 - \Lambda(x_i'\beta) & = 1 - \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} = \frac{1}{1+\exp(x_i'\beta)} \end{cases}$$

3.3.3 Estimation of Probit and Logit Models

The basic estimation method of probit and logit model is maximum likelihood. We consider estimation given i.i.d. samples $\{y_i, x_i\}_{i=1}^N$ in which outcomes are Bernoulli distributed, i.e. $y_i = 1$ or $y_i = 0$. Remind that the conditional density function of Bernoulli distribution is given by

$$f(y_i|\, x_i; \beta) = p_i^{y_i} [1 - p_i]^{1-y_i},$$

where p_i is the probability of $y_i = 1$ and $1 - p_i$ is the probability of $y_i = 0$. Now, we set up

$$p_i = F(x_i'\beta) = \begin{cases} \Phi(x_i'\beta) & \text{if we use probit model} \\ \Lambda(x_i'\beta) & \text{if we use logit model} \end{cases}.$$

Then, the conditional density function is defined as

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

Since likelihood function is joint (conditional) density function and since samples are i.i.d., we have the likelihood function

$$\begin{aligned} L_n(\beta) &= \prod_{i=1}^{N} f(y_i|\, x_i; \beta) \\ &= \prod_{i=1}^{N} F(x_i'\beta)^{y_i} [1 - F(x_i'\beta)]^{1-y_i}. \end{aligned}$$

By taking log, we have the log-likelihood function

$$\begin{aligned} l_n(\beta) &= \ln \prod_{i=1}^{N} f(y_i|\, x_i; \beta) = \sum_{i=1}^{N} \ln f(y_i|\, x_i; \beta) = \sum_{i=1}^{N} \ln \left\{ F(x_i'\beta)^{y_i} [1 - F(x_i'\beta)]^{1-y_i} \right\} \\ &= \sum_{i=1}^{N} \{ y_i \ln F(x_i'\beta) + (1 - y_i) \ln [1 - F(x_i'\beta)] \}. \end{aligned}$$

[3] It is converntion to use Φ and ϕ to express the cdf and pdf of standard normal distrubution.
[4] It is converntion to use Λ and λ to express the cdf and pdf of logit distrubution.

By taking f.o.c. w.r.t. β we have

$$\begin{aligned}
\underbrace{\frac{\partial}{\partial\beta} l_n(\beta)}_{K\times 1} &= \frac{\partial}{\partial\beta}\sum_{i=1}^{N}\left\{y_i \ln F(x_i'\beta) + (1-y_i)\ln\left[1-F(x_i'\beta)\right]\right\} \\
&= \sum_{i=1}^{N}\left\{y_i \frac{f(x_i'\beta)\,x_i}{F(x_i'\beta)} + (1-y_i)\frac{-f(x_i'\beta)\,x_i}{1-F(x_i'\beta)}\right\} \qquad \left(\text{since}\frac{\partial}{\partial\beta}F(x_i'\beta) = f(x_i'\beta)\,x_i\right) \\
&= \sum_{i=1}^{N}\left\{\frac{y_i}{F(x_i'\beta)} + \frac{1-y_i}{1-F(x_i'\beta)}\right\} f(x_i'\beta)\,x_i = \sum_{i=1}^{N}\left\{\frac{y_i\left(1-F(x_i'\beta)\right) - (1-y_i)F(x_i'\beta)}{F(x_i'\beta)\left(1-F(x_i'\beta)\right)}\right\} f(x_i'\beta)\,x_i \\
&= \sum_{i=1}^{N}\left\{\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)\left(1-F(x_i'\beta)\right)}\right\} f(x_i'\beta)\,x_i = 0_{K\times 1}.
\end{aligned}$$

Therefore, the maximum likelihood estimator $\hat{\beta}_{ML}$ is given by the solution of

$$\sum_{i=1}^{N}\underbrace{\left\{\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)\left(1-F(x_i'\beta)\right)}\right\}}_{1\times 1}\underbrace{f(x_i'\beta)}_{1\times 1}\underbrace{x_i}_{K\times 1} = 0_{K\times 1} \tag{6}$$

The above equation does not have analytic solution of $\hat{\beta}_{ML}$ and we have to solve it by numerical computation.
In probit model, we have

$$\begin{cases} F(x_i'\beta) = \Phi(x_i'\beta) & \text{cdf of standard normal distribution} \\ f(x_i'\beta) = \phi(x_i'\beta) & \text{pdf of standard normal distribution} \end{cases},$$

and the f.o.c. equation (6) becomes

$$\sum_{i=1}^{N}\left\{\frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta)\left(1-\Phi(x_i'\beta)\right)}\right\}\phi(x_i'\beta)\,x_i = 0_{K\times 1}$$

There is no further simplification of this f.o.c. equation.
In the logit model, we have

$$\begin{cases} F(x_i'\beta) = \Lambda(x_i'\beta) & \text{cdf of standard normal distribution} \\ f(x_i'\beta) = \lambda(x_i'\beta) & \text{pdf of standard normal distribution} \end{cases},$$

and he f.o.c. equation (6) is becomes

$$\sum_{i=1}^{N}\left\{\frac{y_i - \Lambda(x_i'\beta)}{\Lambda(x_i'\beta)\left(1-\Lambda(x_i'\beta)\right)}\right\}\underbrace{\lambda(x_i'\beta)}_{\text{discussing below}}x_i = 0_{K\times 1} \tag{7}$$

Since cdf and pdf of logit distribution have the relation

$$\begin{aligned}
\lambda(z) &= \frac{\partial}{\partial\beta}\Lambda(z) = \frac{\partial}{\partial\beta}\left\{\frac{\exp(z)}{1+\exp(z)}\right\} = \left\{\frac{\exp(z)\cdot(1+\exp(z)) - \exp(z)\cdot\exp(z)}{(1+\exp(z))^2}\right\} \\
&= \frac{\exp(z)}{(1+\exp(z))^2} = \underbrace{\frac{\exp(z)}{1+\exp(z)}}_{=\Lambda(x_i'\beta)}\underbrace{\frac{1}{1+\exp(z)}}_{=1-\Lambda(x_i'\beta)} = \Lambda(z)\left(1-\Lambda(z)\right),,
\end{aligned}$$

we have the relation

$$\lambda(x_i'\beta) = \Lambda(x_i'\beta)\left(1-\Lambda(x_i'\beta)\right),$$

Thus, the f.o.c. equation (7) can be simplified to

$$\begin{aligned}
\sum_{i=1}^{N}\left\{\underbrace{\frac{y_i - \Lambda(x_i'\beta)}{\Lambda(x_i'\beta)\left(1-\Lambda(x_i'\beta)\right)}}_{\text{denominator cancel out}}\right\}\underbrace{\Lambda(x_i'\beta)\left(1-\Lambda(x_i'\beta)\right)}_{\text{cancel out}}x_i &= 0_{K\times 1} \\
\sum_{i=1}^{N}\left\{y_i - \Lambda(x_i'\beta)\right\}x_i &= 0_{K\times 1}.
\end{aligned}$$

3.3.4 Asymptotic Distributions of Probit and Logit Models

Notice that probit and logit models are estimated by maximum likelihood (ML), and the asymptotic distribution of ML estimator is given by

$$\sqrt{N}\left(\hat{\beta}_{ML}-\beta\right)\to N\left(0,I_1^{-1}\right),$$

where I_1 is 1-sample Fisher information defined by (minus) expectation of second order derivative of log-density function

$$I_1=-E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\ln f\left(y_i|\,x_i;\beta\right)\right].$$

Now, we derive the estimator of I_1, i.e. $\hat{I}_1$

In probit and logit models, we have the (mass) density function (here, we do not specify $F\left(x_i'\beta\right)=\Phi\left(x_i'\beta\right)$ nor $F\left(x_i'\beta\right)=\Lambda\left(x_i'\beta\right)$, just keep using the general notation of cdf $F\left(x_i'\beta\right)$)

$$f\left(y_i|\,x_i;\beta\right)=F\left(x_i'\beta\right)^{y_i}\left[1-F\left(x_i'\beta\right)\right]^{1-y_i},$$

The Fisher information is derived by using information equality[5]

$$I_1=\underbrace{\overbrace{-E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\ln f\left(y_i|\,x_i;\beta\right)\right]}^{\text{second derivative notation}}=\overbrace{E\left[\left[\frac{\partial}{\partial\beta}\ln f\left(y_i|\,x_i;\beta\right)\right]\left[\frac{\partial}{\partial\beta'}\ln f\left(y_i|\,x_i;\beta\right)\right]\right]}^{\text{outer product notation}}}_{\text{information equality}}.$$

Calculating the outer product notation.

By the same derivative calculation as in the f.o.c. equation (6), we have derivatives

$$\begin{aligned}\underbrace{\frac{\partial}{\partial\beta}\ln f\left(y_i|\,x_i;\beta\right)}_{K\times1} &= \left\{\frac{y_i-F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left\{1-F\left(x_i'\beta\right)\right\}}\right\}f\left(x_i'\beta\right)x_i\\ \underbrace{\frac{\partial}{\partial\beta}\ln f\left(y_i|\,x_i;\beta\right)}_{1\times K} &= \left\{\frac{y_i-F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left\{1-F\left(x_i'\beta\right)\right\}}\right\}f\left(x_i'\beta\right)x_i'\end{aligned}$$

Thus,

$$\begin{aligned}&\underbrace{E\left[\left[\underbrace{\frac{\partial}{\partial\beta}\ln f\left(y_i|\,x_i;\beta\right)}_{K\times1}\right]\left[\underbrace{\frac{\partial}{\partial\beta}\ln f\left(y_i|\,x_i;\beta\right)}_{1\times K}\right]\right]}_{K\times K}\\ =\;& E\left[\left\{\frac{y_i-F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left\{1-F\left(x_i'\beta\right)\right\}}\right\}f\left(x_i'\beta\right)x_i\cdot\left\{\frac{y_i-F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left\{1-F\left(x_i'\beta\right)\right\}}\right\}f\left(x_i'\beta\right)x_i'\right]\\ =\;& E\left[\left\{\frac{y_i-F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left\{1-F\left(x_i'\beta\right)\right\}}\right\}^2\left\{f\left(x_i'\beta\right)\right\}^2x_ix_i'\right]\\ =\;& E_{x_i}\left[E_{y_i|x_i}\left[\left\{\frac{y_i-F\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left\{1-F\left(x_i'\beta\right)\right\}}\right\}^2\left\{f\left(x_i'\beta\right)\right\}^2x_ix_i'\middle|\,x_i\right]\right]\qquad\text{(by law of iterated expectation)}\\ =\;& E_{x_i}\left[\frac{\overbrace{E_{y_i|x_i}\left[\left\{y_i-F\left(x_i'\beta\right)\right\}^2\middle|\,x_i\right]}^{\text{(*) calculating numerator}}}{\left\{F\left(x_i'\beta\right)\right\}^2\left\{1-F\left(x_i'\beta\right)\right\}^2}\left\{f\left(x_i'\beta\right)\right\}^2x_ix_i'\right].\end{aligned}$$

[5] We will prove the information equality in Comp 2003S: Question 2.

Now, we calculate (*) in the above equation by using the definition of (conditional) variance

$$E_{y_i|x_i}\left[\left\{y_i - F\left(x_i'\beta\right)\right\}^2 \middle| x_i\right] = Var_{y_i|x_i}\left[\left\{y_i - F\left(x_i'\beta\right)\right\}| x_i\right] + \underbrace{\left(E_{y_i|x_i}\left[y_i - F\left(x_i'\beta\right)| x_i\right]\right)^2}_{=0 \quad \text{(see below)}}$$

where we use the fact

$$\begin{aligned} E_{y_i|x_i}\left[y_i - F\left(x_i'\beta\right)| x_i\right] &= E_{y_i|x_i}\left[y_i| x_i\right] - F\left(x_i'\beta\right) \\ &= F\left(x_i'\beta\right) - F\left(x_i'\beta\right) \qquad \left(\text{since } E_{y_i|x_i}\left[y_i| x_i\right] = F\left(x_i'\beta\right)\right) \\ &= 0 \end{aligned}$$

Also, since y_i is Bernoulli random variable with[6]

$$Var_{y_i|x_i}\left[\left\{y_i - F\left(x_i'\beta\right)\right\}| x_i\right] = F\left(x_i'\beta\right)\left\{1 - F\left(x_i'\beta\right)\right\}.$$

Therefore, the 1-sample Fisher information can be simplified to is

$$I_1 = E_{x_i}\left[\frac{F\left(x_i'\beta\right)\left\{1 - F\left(x_i'\beta\right)\right\}}{\left\{F\left(x_i'\beta\right)\right\}^2\left\{1 - F\left(x_i'\beta\right)\right\}^2}\left\{f\left(x_i'\beta\right)\right\}^2 x_i x_i'\right] = E_{x_i}\left[\frac{\left\{f\left(x_i'\beta\right)\right\}^2}{F\left(x_i'\beta\right)\left\{1 - F\left(x_i'\beta\right)\right\}} x_i x_i'\right]$$

and we have

$$I_1^{-1} = \left(\underbrace{-E\left[\frac{f^2\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]} x_i x_i'\right]}_{(**)}\right)^{-1}.$$

Therefore, the estimate of $\hat{I}_1^{-1}$ is

$$\hat{I}_1^{-1} = \left(\underbrace{-\frac{1}{N}\sum_{i=1}^{N}\frac{f^2\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]} x_i x_i'}_{\text{sample analogue of (**)}}\right)^{-1}$$

In probit model, asymptotic variance is

$$I_1^{-1} = \left(-E\left[\frac{\phi^2\left(x_i'\beta\right)}{\Phi\left(x_i'\beta\right)\left[1 - \Phi\left(x_i'\beta\right)\right]} x_i x_i'\right]\right)^{-1}.$$

In logit model, asymptotic variance is

$$\begin{aligned} I_1^{-1} &= \left(-E\left[\frac{\lambda^2\left(x_i'\beta\right)}{\Lambda\left(x_i'\beta\right)\left[1 - \Lambda\left(x_i'\beta\right)\right]} x_i x_i'\right]\right)^{-1} \\ &= \left(-E\left[\lambda\left(x_i'\beta\right) x_i x_i'\right]\right)^{-1}. \qquad \left(\text{since } \lambda\left(x_i'\beta\right) = \Lambda\left(x_i'\beta\right)\left[1 - \Lambda\left(x_i'\beta\right)\right]\right) \end{aligned}$$

[6] Remember? In Burnoulli distribution, the expectation and variance are

$$E\left[y_i\right] = p \quad \text{and} \quad Var\left[y_i\right] = p\left(1 - p\right).$$

Similarly, in binary choice model, conditinal expectation and variance are

$$E_{y_i|x_i}\left[y_i| x_i\right] = F\left(x_i'\beta\right) \quad \text{and} \quad Var_{y_i|x_i}\left[y_i| x_i\right] = F\left(x_i'\beta\right)\left\{1 - F\left(x_i'\beta\right)\right\}.$$

4 Final 2006: Question 1 - Identification in Probit Model

Consider the following binary choice model

$$\begin{aligned} y_i^* &= x_i'\beta_0 + \varepsilon_i, \qquad \text{where} \\ \varepsilon_i|\, x_i &\sim N\left(0, \sigma_\varepsilon^2\right), \end{aligned}$$

for $i = 1, \ldots, N$, where $x_{i1} \equiv 1$ for all $i = 1, \cdots, N$, and β_0 is a $K \times 1$ vector of unknown parameters. Define

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}.$$

(1) Suppose that the first entry of β_0 is given by $\beta_{01} = 1$. Determine which of the model parameters are identified. Please be as precise as possible with your explanation.

Answer:

Define

$$\beta_0 = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{03} \\ \vdots \\ \beta_{0K} \end{bmatrix} = \begin{bmatrix} 1 \\ \beta_{02} \\ \beta_{03} \\ \vdots \\ \beta_{0K} \end{bmatrix} \quad \text{and} \quad x_i = \begin{bmatrix} 1 \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}.$$

Here, we assume the first element of x_i is constant.

Then, the latent variable y^* is (assuming x_i has constant term)

$$\begin{aligned} y_i^* &= x_i'\beta_0 - \varepsilon_i \qquad \text{(changing the plus minus sign in front of } \varepsilon_i, \text{ innocuous transformation)} \\ &= 1 + \beta_{02}x_{i2} + \beta_{03}x_{i3} + \cdots + \beta_{0K}x_{ik} - \varepsilon_i. \end{aligned}$$

$\Pr\left(y_i = 1|\, x_i\right)$ is derived by

$$\begin{aligned} y_i &= 1 \\ &\Leftrightarrow \quad y_i^* > 0 \\ &\Leftrightarrow \quad x_i'\beta_0 - \varepsilon_i > 0 \\ &\Leftrightarrow \quad x_i'\beta_0 > \underbrace{\varepsilon_i}_{\text{distributed normally with variance } \sigma_\varepsilon^2} \\ &\Leftrightarrow \quad \frac{x_i'\beta_0}{\sigma_\varepsilon} > \underbrace{\frac{\varepsilon_i}{\sigma_\varepsilon}}_{\text{standard normal distribution}} \qquad \text{(divide both sides by } \sigma_\varepsilon > 0) \\ &\Leftrightarrow \quad \Pr\left(y_i = 1|\, x_i\right) = \Phi\left(\frac{x_i'\beta_0}{\sigma_\varepsilon}\right). \end{aligned}$$

Then, maximum estimators are derived by

$$\begin{aligned} &\max_{\beta,\sigma_\varepsilon} \left\{\ln\left[\left\{\Pr\left(y_i = 1|\, x_i\right)\right\}^{y_i}\left\{1 - \Pr\left(y_i = 1|\, x_i\right)\right\}^{y_i}\right]\right\} \\ = &\max_{\beta,\sigma_\varepsilon} \left\{\ln\left[\left\{\Phi\left(\frac{x_i'\beta_0}{\sigma_\varepsilon}\right)\right\}^{y_i}\left\{1 - \Phi\left(\frac{x_i'\beta_0}{\sigma_\varepsilon}\right)\right\}^{1-y_i}\right]\right\} \\ = &\max_{\beta,\sigma_\varepsilon} \left\{\sum (y_i - 1)\ln\Phi\left(\frac{x_i'\beta_0}{\sigma_\varepsilon}\right) + (1 - y_i)\ln\left\{1 - \Phi\frac{x_i'\beta_0}{\sigma_\varepsilon}\right\}\right\}. \end{aligned}$$

Notice that the $\frac{x_i'\beta_0}{\sigma_\varepsilon}$ can be decomposed into

$$\begin{aligned} \frac{x_i'\beta_0}{\sigma_\varepsilon} &= \frac{1}{\sigma_e}\left(1 + \beta_{02}x_{i2} + \beta_{03}x_{i3} + \cdots + \beta_{0K}x_{ik}\right) \\ &= \frac{1}{\sigma_\varepsilon} + \frac{\beta_{02}}{\sigma_\varepsilon}x_{i2} + \frac{\beta_{03}}{\sigma_\varepsilon}x_{i3} + \cdots + \frac{\beta_{0K}}{\sigma_\varepsilon}x_{ik}. \end{aligned}$$

Usually, parameters in the probit model are identified only up to scale. However, with the special assumption $\beta_{01} = 1$, we can identify all parameters by following steps.

Step 1: Implement ML and obtain the estimated coefficients, $\widehat{\left(\frac{1}{\sigma_\varepsilon}\right)}$, $\widehat{\left(\frac{\beta_{02}}{\sigma_\varepsilon}\right)}$,...,$\widehat{\left(\frac{\beta_{0K}}{\sigma_\varepsilon}\right)}$.

Step 2: Calculate $\hat{\sigma}_\varepsilon$ by

$$\hat{\sigma}_\varepsilon = \left[\widehat{\left(\frac{1}{\sigma_\varepsilon}\right)}\right]^{-1}.$$

Step 3: Calculate $\widehat{(\beta_{02})}$,...,$\widehat{(\beta_{0K})}$ by

$$\widehat{(\beta_{02})} = \frac{\widehat{\left(\frac{\beta_{02}}{\sigma_\varepsilon}\right)}}{\hat{\sigma}_\varepsilon},\ \widehat{(\beta_{03})} = \frac{\widehat{\left(\frac{\beta_{03}}{\sigma_\varepsilon}\right)}}{\hat{\sigma}_\varepsilon},\ldots,\ \widehat{(\beta_{0K})} = \frac{\widehat{\left(\frac{\beta_{0K}}{\sigma_\varepsilon}\right)}}{\hat{\sigma}_\varepsilon}$$

Notice that the assumption of $\beta_{01} = 1$ is crucial for this identification.

(2) Suppose now that

$$y_i = \begin{cases} 1 & \text{if } y_i^* > c_0 \\ 0 & \text{otherwise} \end{cases},$$

for some unknown constant c_0. Will your answer to (1) change? Explain.

Answer: Change

Now, $\Pr(y_i = 1|\, x_i)$ is derived by

$$\begin{aligned}
y_i &= 1 \\
&\Leftrightarrow\ y_i^* > c_0 \\
&\Leftrightarrow\ x_i'\beta_0 - \varepsilon_i > c_0 \\
&\Leftrightarrow\ x_i'\beta_0 - c_0 > \underbrace{\varepsilon_i}_{\text{distributed normally with variance } \sigma_\varepsilon^2} \\
&\Leftrightarrow\ \frac{x_i'\beta_0 - c_0}{\sigma_\varepsilon} > \underbrace{\frac{\varepsilon_i}{\sigma_\varepsilon}}_{\text{standard normal distribution}} \qquad \text{(divide both sides by } \sigma_\varepsilon > 0) \\
&\Leftrightarrow\ \Pr(y_i = 1|\, x_i) = \Phi\left(\frac{x_i'\beta_0 - c_0}{\sigma_\varepsilon}\right) \\
&\Leftrightarrow\ \Pr(y_i = 1|\, x_i) = \Phi\left(\frac{1 - c_0 + \beta_{02}x_{i2} + \beta_{03}x_{i3} + \cdots + \beta_{0K}x_{ik}}{\sigma_\varepsilon}\right) \\
&\Leftrightarrow\ \Pr(y_i = 1|\, x_i) = \Phi\left(\frac{1 - c_0}{\sigma_\varepsilon} + \frac{\beta_{02}}{\sigma_\varepsilon}x_{i2} + \frac{\beta_{03}}{\sigma_\varepsilon}x_{i3} + \cdots + \frac{\beta_{0K}}{\sigma_\varepsilon}x_{ik}\right)
\end{aligned}$$

Here we cannot identify any parameters because by applying ML, we obtain the estimate

$$\widehat{\left(\frac{1 - c_0}{\sigma_\varepsilon}\right)},$$

but we cannot separate this object into $\hat{c}_0$ and $\hat{\sigma}_\varepsilon$ and cannot apply the procedure in (a).

(3) Suppose now that $\sigma_\varepsilon^2 = 1$. Consider the following moment condition

$$\varphi(y_i, x_i\beta) = (y_i - \Phi(x_i'\beta_0))\, g(x_i),$$

where $g(\cdot)$ is a $p \times 1$ vector-valued function, with $p > K$. Show that when evaluate at the true parameter vector

$$E[(y_i - \Phi(x_i'\beta_0))\, g(x_i)] = 0.$$

Answer:

We will solve this question when we study GMM.

(4) Describe in detail how to obtain a GMM estimator for β_0. In particular, determine the optimal GMM estimator and provide a consistent estimator for the optimal weight matrix.

Answer:

We will solve this question when we study GMM.

5 Comp 2003S Part III (Buchinsky): Question 2

Consider the binary choice model

$$y^* = x_i'\beta_0 + \varepsilon_i,$$

for $i = 1, \ldots, n$ where $\varepsilon_i | \, x_i \sim i.i.d. G(x)$, $G(x)$ is independent of x and symmetric around zero.

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}.$$

(a) Compute $\Pr(y_i = 1 | \, x_i)$.
Answer:
we have the model

$$y^* = x_i'\beta_0 - \varepsilon_i,$$

where the sigh in front of the ε_i is changed. This transformation is innocuous as long as the distribution of ε_i is symmetric around 0. Then, $\Pr(y_i = 1 | \, x_i)$ is obtained by

$$\begin{aligned} \Pr(y_i = 1 | \, x_i) &= \Pr(y_i^* > 0 | \, x_i) \\ &= \Pr(x_i'\beta_0 - \varepsilon_i > 0 | \, x_i) \\ &= \Pr(x_i'\beta_0 > \varepsilon_i | \, x_i) \\ &= G(x_i'\beta_0) \end{aligned}$$

(b) Demonstrate how to obtain the maximum likelihood estimator for β_0, say $\hat{\beta}_n$.
Answer:
Since y_i is distributed as conditional Bernoulli distribution, we have the conditional density function

$$\begin{aligned} f(y_i | \, x_i; \beta_0) &= \{\Pr(y_i = 1 | \, x_i)\}^{y_i} \{1 - \Pr(y_i = 1 | \, x_i)\}^{1-y_i} \\ &= \{G(x_i'\beta_0)\}^{y_i} \{1 - G(x_i'\beta_0)\}^{1-y_i}. \end{aligned}$$

Then, the likelihood function is

$$L_n(\beta_0) = \prod_{i=1}^{n} \{G(x_i'\beta_0)\}^{y_i} \{1 - G(x_i'\beta_0)\}^{1-y_i},$$

and log-likelihood function is

$$\begin{aligned} l_n(\beta_0) &= \ln \prod_{i=1}^{n} \{G(x_i'\beta_0)\}^{y_i} \{1 - G(x_i'\beta_0)\}^{1-y_i} \\ &= \sum_{i=1}^{n} [y_i \ln G(x_i'\beta_0) + (1 - y_i) \ln \{1 - G(x_i'\beta_0)\}]. \end{aligned}$$

Take f.o.c. w.r.t. β_0,

$$\begin{aligned} &\sum_{i=1}^{n} \left[y_i \frac{g(x_i'\beta_0)}{G(x_i'\beta_0)} x_i + (1 - y_i) \ln \frac{-g(x_i'\beta_0)}{1 - G(x_i'\beta_0)} x_i \right] \\ = &\sum_{i=1}^{n} \left[\frac{y_i \{1 - G(x_i'\beta_0)\} - (1 - y_i) G(x_i'\beta_0)}{G(x_i'\beta_0) \{1 - G(x_i'\beta_0)\}} \right] g(x_i'\beta_0) x_i = \sum_{i=1}^{n} \left[\frac{y_i - \overbrace{y_i G(x_i'\beta_0)}^{\text{cancel out}} - G(x_i'\beta_0) + y_i + \overbrace{y_i G(x_i'\beta_0)}^{\text{cancel out}}}{G(x_i'\beta_0) \{1 - G(x_i'\beta_0)\}} \right] g(x_i'\beta_0) x_i \\ = &\sum_{i=1}^{n} \left[\frac{y_i - G(x_i'\beta_0)}{G(x_i'\beta_0) \{1 - G(x_i'\beta_0)\}} \right] g(x_i'\beta_0) x_i = 0_{K \times 1}. \end{aligned}$$

The maximum likelihood estimator $\hat{\beta}_n$ is defined by the solution of above equation.

(c) Let $l(\beta)$ denote the log-likelihood function for β. Show that

$$E\left[\frac{\partial l(\beta_0)}{\partial \beta} \frac{\partial l(\beta_0)}{\partial \beta'} \right] = -E\left[\frac{\partial^2 l(\beta_0)}{\partial \beta \partial \beta'} \right].$$

Answer:

This is information matrix identity. Here notice that $l(\beta_0)$ is NOT n-sample log-likelihood function $l_n(\beta_0)$ that is defined in (b), BUT 1-sample log-likelihood function $l_1(\beta_1)$ (otherwise, we cannot prove this equality)

$$l_1(\beta_1) = \ln f(y_i|\, x_i; \beta_0).$$

By the definition of probability density function, we have the condition

$$\int f(y|\, x; \beta_0)\, dy = 1.$$

where we omit sample index i for simplicity of notation
Taking derivative both sides w.r.t. β_0',

$$\frac{\partial}{\partial \beta_0'} \int f(y|\, x; \beta_0)\, dy = \frac{\partial}{\partial \beta_0'} 1$$

and by assuming derivative and integral are exchangeable, we have

$$\int \underbrace{\frac{\partial}{\partial \beta_0'} f(y|\, x; \beta_0)}_{\text{substituting, see below}}\, dy = 0_{1\times K} \tag{8}$$

Now, we use the property of derivative of log function such that

$$\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0) = \frac{\frac{\partial}{\partial \beta_0'} f(y|\, x; \beta_0)}{f(y|\, x; \beta_0)}.$$

By arranging the above equation,

$$\underbrace{\frac{\partial}{\partial \beta_0'} f(y|\, x; \beta_0)}_{\text{plug in}} = f(y|\, x; \beta_0) \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] \tag{9}$$

Substituting (9) into (8), we obtain

$$\int f(y|\, x; \beta_0) \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy = 0_{1\times K}$$

Again, differentiating the above equation w.r.t. β_0,

$$\frac{\partial}{\partial \beta_0} \int f(y|\, x; \beta_0) \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy = \frac{\partial}{\partial \beta_0'} 0_{1\times K}$$

and

$$\int \left[\frac{\partial}{\partial \beta_0} f(y|\, x; \beta_0)\right] \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] + f(y|\, x; \beta_0) \left[\frac{\partial^2}{\partial \beta \partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy = 0_{K\times K} \qquad \text{(product rule)}$$

By decomposing integral, we have,

$$\int \underbrace{\left[\frac{\partial}{\partial \beta_0} f(y|\, x; \beta_0)\right]}_{\text{substituting}} \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy = -\int f(y|\, x; \beta_0) \left[\frac{\partial^2}{\partial \beta \partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy.$$

Again, substituting the transpose of (9) into above equation

$$\begin{aligned}
\int f(y|\, x; \beta_0) \left[\frac{\partial}{\partial \beta_0} \ln f(y|\, x; \beta_0)\right] \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy &= -\int f(y|\, x; \beta_0) \left[\frac{\partial^2}{\partial \beta \partial \beta_0'} \ln f(y|\, x; \beta_0)\right] dy \\
\int \left[\frac{\partial}{\partial \beta_0} \ln f(y|\, x; \beta_0)\right] \left[\frac{\partial}{\partial \beta_0'} \ln f(y|\, x; \beta_0)\right] f(y|\, x; \beta_0)\, dy &= -\int \left[\frac{\partial^2}{\partial \beta \partial \beta_0'} \ln f(y|\, x; \beta_0)\right] f(y|\, x; \beta_0)\, dy.
\end{aligned}$$

Since $l_1(\beta_0) = \ln f(y|x;\beta_0)$

$$\int \left[\frac{\partial}{\partial \beta_0} l_1(\beta_0)\right]\left[\frac{\partial}{\partial \beta_0'} l_1(\beta_0)\right] f(|y|\,x;\beta_0)\,dy = -\int \left[\frac{\partial^2}{\partial \beta \partial \beta_0'} l_1(\beta_0)\right] f(|y|\,x;\beta_0)\,dy$$

By applying the definition of expectation, we obtain the information equality

$$\underbrace{E\left[\left[\frac{\partial}{\partial \beta_0} l_1(\beta_0)\right]\left[\frac{\partial}{\partial \beta_0'} l_1(\beta_0)\right]\right]}_{\text{outerproduct notation}} = \underbrace{-E\left[\frac{\partial^2}{\partial \beta \partial \beta_0'} l_1(\beta_0)\right]}_{\text{second product notation}} \quad (= I_1,\ \text{Fisher information})$$

(d) Provide the asymptotic distribution for $\hat{\beta}_n$ using the property established in (c).

Answer:

By applying the result in (c), we have the Fisher information (here, I omit the calculation, because it is exactly the same as sub-sub-secton 3.3.4)

$$I_1 = E\left[\frac{\{g(x_i'\beta_0)\}^2}{\{G(x_i'\beta_0)\}^2\{1-G(x_i'\beta_0)\}^2} x_i x_i'\right]$$

Thus, the asymptotic distribution of ML estimator $\hat{\beta}_n$ is

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} N\left(0_{K\times 1}, I_1^{-1}\right) \sim N\left(0_{K\times 1}, \left[E\left[\frac{\{g(x_i'\beta_0)\}^2}{\{G(x_i'\beta_0)\}^2\{1-G(x_i'\beta_0)\}^2} x_i x_i'\right]\right]^{-1}\right).$$

(e) Show how to test whether or not the marginal effect of x_{i2} on the probability that $y_i = 1$, conditional on x_i is of any significance. Justify your answer.

Answer:

Using multi-variate delta method.

Define

$$\beta_0 = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \vdots \\ \beta_{0K} \end{bmatrix}.$$

The marginal effect of x_{i2} on $\Pr(y_i = 1|\,x_i)$ is

$$\underbrace{\frac{\partial}{\partial x_{i2}} \Pr(y_i = 1|\,x_i)}_{1\times 1} = \frac{\partial}{\partial x_{i2}} G(x_i'\beta_0) = \beta_{02} g(x_i'\beta_0).$$

Now, applying multi-variate delta method. Define the continuos function

$$\underbrace{h(\gamma)}_{1\times 1} = \gamma_2 g(x_i'\gamma).$$

Then, the derivative is

$$\underbrace{\frac{\partial}{\partial \gamma'} h(\gamma)}_{1\times k} = \frac{\partial}{\partial \gamma'} \gamma_2 g(x_i'\gamma) = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 1 \end{bmatrix}}_{1\times k} \cdot g(x_i'\gamma) + \gamma_2 g'(x_i'\gamma) \underbrace{x_i'}_{1\times k} \quad \text{(product rule)}$$

$$\underbrace{\frac{\partial}{\partial \gamma'} h(\gamma)\Big|_{\gamma=\beta_0}}_{1\times K} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 1 \end{bmatrix}}_{1\times k} \cdot g(x_i'\beta_0) + \beta_{02} g'(x_i'\beta_0) \underbrace{x_i'}_{1\times k}$$

Then, by using the result of (d) and by applying multi-variate delta method, we have

$$\sqrt{n}\left(h\left(\hat{\beta}_n\right) - h(\beta_0)\right) \xrightarrow{d} N\left(0_{1\times 1}, \underbrace{\frac{\partial}{\partial \gamma'} h(\gamma)\Big|_{\gamma=\beta_0}}_{1\times K} \cdot \underbrace{\hat{I}_1^{-1}}_{K\times K} \cdot \underbrace{\left[\frac{\partial}{\partial \gamma'} h(\gamma)\Big|_{\gamma=\beta_0}\right]'}_{K\times 1}\right).$$

Now, we can apply normal test for checking the significance of effect in x_{i2}, such as t test.

6 Comp 2003F Part III (Buchinsky): Question 2 = Final Review: Question 7

Consider the binary discrete choice model given by

$$\Pr\left(y_i = 0\right) = \frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)},$$

for $i = 1, \ldots, n$.
Notice from TA:
This question is tricky. The conditional probability is defined in opposite way to conventional one

$$\begin{cases} \Pr\left(y_i = 1 \middle| x_i\right) = 1 - \Lambda\left(x_i'\gamma\right) & = \frac{1}{1+\exp\left(x_i'\gamma\right)} \quad \neq \underbrace{\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} = \Lambda\left(x_i'\gamma\right)}_{\text{conventional assumption}} \\ \Pr\left(y_i = 0 \middle| x_i\right) = \Lambda\left(x_i'\gamma\right) & = \frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} \quad \neq \underbrace{\frac{1}{1+\exp\left(x_i'\gamma\right)} = 1 - \Lambda\left(x_i'\gamma\right)}_{\text{conventional assumption}} \end{cases}.$$

(a) Provide the MLE for γ, say $\hat{\gamma}_n$.
Answer:
Denote $w = \{y_i, x_i\}_{i=1}^n$.
The likelihood function is defined as

$$\begin{aligned} L_n\left(\gamma \middle| w\right) &= \prod_{i=1}^n f\left(y_i \middle| x_i; \gamma\right) \\ &= \prod_{i=1}^n \left\{\Pr\left(y_i = 0 \middle| x_i\right)\right\}^{1-y_i} \left\{\Pr\left(y_i = 1 \middle| x_i\right)\right\}^{y_i}. \end{aligned}$$

The log-likelihood function is

$$\begin{aligned} l_n\left(\gamma \middle| w\right) &= \sum_{i=1}^n \left(1-y_i\right) \ln\left\{\Pr\left(y_i = 0 \middle| x_i\right)\right\} + y_i \ln\left\{\Pr\left(y_i = 1 \middle| x_i\right)\right\} \\ &= \sum_{i=1}^n \left(1-y_i\right) \ln\left\{\Lambda\left(x_i'\gamma\right)\right\} + y_i \ln\left\{1 - \Lambda\left(x_i'\gamma\right)\right\}. \end{aligned}$$

Taking f.o.c. w.r.t. γ is

$$\begin{aligned} \frac{\partial}{\partial\gamma} l_n\left(\gamma \middle| w\right) &= \sum_{i=1}^n \left\{\left(1-y_i\right)\frac{\lambda\left(x_i'\gamma\right)}{\Lambda\left(x_i'\gamma\right)} x_i - y_i \frac{\lambda\left(x_i'\gamma\right)}{1-\Lambda\left(x_i'\gamma\right)} x_i\right\} \\ &= \sum_{i=1}^n \left\{\frac{\left(1-y_i\right)\left\{1-\Lambda\left(x_i'\gamma\right)\right\} - y_i\Lambda\left(x_i'\gamma\right)}{\Lambda\left(x_i'\gamma\right)\left\{1-\Lambda\left(x_i'\gamma\right)\right\}}\right\} \lambda\left(x_i'\gamma\right) x_i \\ &= \sum_{i=1}^n \left\{\frac{\left(1-y_i\right)\left\{1-\Lambda\left(x_i'\gamma\right)\right\} - y_i\Lambda\left(x_i'\gamma\right)}{\Lambda\left(x_i'\gamma\right)\left\{1-\Lambda\left(x_i'\gamma\right)\right\}}\right\} \lambda\left(x_i'\gamma\right) x_i \\ &= \sum_{i=1}^n \left\{\frac{1-\Lambda\left(x_i'\gamma\right) - y_i + y_i\Lambda\left(x_i'\gamma\right) - y_i\Lambda\left(x_i'\gamma\right)}{\Lambda\left(x_i'\beta\right)\left\{1-\Lambda\left(x_i'\beta\right)\right\}}\right\} \lambda\left(x_i'\gamma\right) x_i \\ &= \sum_{i=1}^n \left\{\frac{1-y_i-\Lambda\left(x_i'\gamma\right)}{\Lambda\left(x_i'\gamma\right)\left\{1-\Lambda\left(x_i'\gamma\right)\right\}}\right\} \lambda\left(x_i'\gamma\right) x_i = 0_{K\times 1}. \end{aligned}$$

Since Λ and λ functions have the relation

$$\lambda\left(x_i'\gamma\right) = \Lambda\left(x_i'\gamma\right)\left(1-\Lambda\left(x_i'\gamma\right)\right),$$

f.o.c. is simplified to

$$\sum_{i=1}^n \left\{1-y_i-\Lambda\left(x_i'\gamma\right)\right\} x_i = 0_{K\times 1}.$$

MLE $\hat{\gamma}_n$ is given by the solution of above equation.

(b) Show that the MLE estimator can be viewed as a Method of Moment (GMM) estimator.

Answer:

We will discuss this question when we study GMM.

(c) Compute the exact asymptotic covariance for $\hat{\gamma}_n$ and provide a consistent estimator for the asymptotic covariance. Justify your answer.

Answer:

The asymptotic distribution of ML estimator is given by

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma\right) \xrightarrow{d} N\left(0_{K\times 1}, I_1^{-1}\right),$$

where I_1 is 1-sample Fisher information.

Deriving 1-sample Fisher information. We have

$$\begin{aligned}
\frac{\partial}{\partial \gamma'} \ln f\left(y_i \middle| x_i; \gamma\right) &= \left\{1 - y_i - \Lambda\left(x_i'\gamma\right)\right\} x_i \qquad \text{(by using the calculation result in (a))} \\
\frac{\partial^2}{\partial \gamma' \partial \gamma'} \ln f\left(y_i \middle| x_i; \gamma\right) &= -\lambda\left(x_i'\gamma\right) x_i' x_i
\end{aligned}$$

Thus, 1-sample Fisher information is given by

$$I_1 = -E\left[\frac{\partial^2}{\partial \gamma' \partial \gamma'} \ln f\left(y_i \middle| x_i; \gamma\right)\right] = -E\left[-\lambda\left(x_i'\gamma\right) x_i' x_i\right] = E\left[\lambda\left(x_i'\gamma\right) x_i' x_i\right].$$

Therefore, the asymptotic distribution of $\hat{\gamma}_n$ is

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma\right) \xrightarrow{d} N\left(0_{K\times 1}, I_1^{-1}\right) \sim N\left(0_{K\times 1}, \left[E\left[\lambda\left(x_i'\gamma\right) x_i' x_i\right]\right]^{-1}\right).$$

The consistent estimator of asymptotic covariance matrix is obtained flowingly.

Since 1-sample and n-sample Fisher information have the relation

$$\begin{aligned}
n \cdot I_1 &= I_n, \\
n \cdot \hat{I}_1 &= \hat{I}_n \\
\hat{I}_1 &= \frac{1}{n}\hat{I}_n
\end{aligned}$$

and the estimate of n-sample Fisher information is obtained by

$$\hat{I}_n = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \gamma \partial \gamma'} \ln f\left(y_i \middle| x_i; \gamma\right)\Big|_{\gamma=\hat{\gamma}_n} = -\sum_{i=1}^{n} -\lambda\left(x_i'\gamma\right) x_i' x_i|_{\gamma=\hat{\gamma}_n} = \sum_{i=1}^{n} \lambda\left(x_i'\hat{\gamma}_n\right) x_i' x_i$$

the estimate of 1-sample fisher information is obtained by

$$\hat{I}_1 = \frac{1}{n}\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \lambda\left(x_i'\hat{\gamma}_n\right) x_i' x_i.$$

Therefore, consistent estimator of asymptotic covariance matrix is obtained by

$$\hat{I}_1^{-1} = \left[\frac{1}{n}\sum_{i=1}^{n} \lambda\left(x_i'\hat{\gamma}_n\right) x_i' x_i\right]^{-1} \xrightarrow{p} \left[E\left[\lambda\left(x_i'\gamma\right) x_i' x_i\right]\right]^{-1} \qquad \text{(by WLLN, Slutzky, and continuity theorems)}$$

(d) Consider the weighted estimator, say $\hat{\gamma}_n^W$, obtained by

$$\min \sum_{i=1}^{n} \frac{\left(y_i - \Pr\left(y_i \middle| x_i\right)\right)^2}{\Pr\left(y_i \middle| x_i\right)\left(1 - \Pr\left(y_i \middle| x_i\right)\right)}.$$

Is $\hat{\gamma}_n^W$ consistent estimator for $\hat{\gamma}$? Justify your answer.

Answer:

Consider the non-linear regression model[7]

$$\begin{aligned}
y_i &= \Pr\left(y_i = 1 \middle| x_i\right) + \underbrace{u_i}_{\text{error term}} \qquad \text{(model equation)} \\
y_i &= \frac{1}{1 + \exp\left(x_i'\gamma\right)} + u_i
\end{aligned}$$

[7] I quoteed from Prof. Kyriazidou's note #9.

Error term is

$$u_i = y_i - \frac{1}{1+\exp\left(x_i'\gamma\right)}$$

Checking whether the conditional expectation of error term is zero or not.
Since y_i is binary, the conditional expectation of y_i is

$$\begin{aligned} E_{u_i|x_i}\left[y_i|\, x_i\right] &= 1\cdot \Pr\left(y_i=1|\, x_i\right) + 0\cdot \Pr\left(y_i=0|\, x_i\right) \\ &= 1\cdot \frac{1}{1+\exp\left(x_i'\gamma\right)} + 0\cdot \frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} \\ &= \frac{1}{1+\exp\left(x_i'\gamma\right)}. \end{aligned}$$

Thus, the conditional expectation of u_i is

$$\begin{aligned} E_{u_i|x_i}\left[u_i|\, x_i\right] &= E_{u_i|x_i}\left[y_i - \frac{1}{1+\exp\left(x_i'\gamma\right)}\middle|\, x_i\right] \\ &= \underbrace{E_{u_i|x_i}\left[y_i|\, x_i\right]}_{=\frac{1}{1+\exp\left(x_i'\gamma\right)} \text{ from above}} - E_{u_i|x_i}\left[\frac{1}{1+\exp\left(x_i'\gamma\right)}\middle|\, x_i\right] \\ &= \frac{1}{1+\exp\left(x_i'\gamma\right)} - \frac{1}{1+\exp\left(x_i'\gamma\right)} \\ &= 0. \end{aligned}$$

Thus, non-linear least square estimator that is given by

$$\begin{aligned} \hat{\gamma}_{NLLS} &= \arg\min_{\gamma}\left\{y_i - \frac{1}{1+\exp\left(x_i'\gamma\right)}\right\}^2 \\ &= \arg\min_{\gamma}\left\{y_i - \Pr\left(y_i=1|\, x_i\right)\right\}^2 \end{aligned}$$

is consistent under regularity conditions.

Now, deriving the variance of u_i for weighted NLLS estimator of γ.Note that since y_i is Bernoulli random variable, u_i is defined by

$$u_i = y_i - \frac{1}{1+\exp\left(x_i'\gamma\right)} = y_i - \Lambda\left(x_i'\gamma\right) = \begin{cases} 1-\Lambda\left(x_i'\gamma\right) & \text{if } y_i=1 \quad \text{with probability } \Lambda\left(x_i'\gamma\right) = \frac{1}{1+\exp\left(x_i'\gamma\right)} \\ -\Lambda\left(x_i'\gamma\right) & \text{if } y_i=0 \quad \text{with probability } 1-\Lambda\left(x_i'\gamma\right) = \frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} \end{cases}.$$

Now, conditional variance of u_i is

$$\begin{aligned}
Var_{u_i|x_i}\left[u_i|\,x_i\right] &= E_{u_i|x_i}\left[u_i^2|\,x_i\right] - \left\{\underbrace{E_{u_i|x_i}\left[u_i|\,x_i\right]}_{=0}\right\}^2 \qquad \text{(definition of variance)} \\
&= \left(\underbrace{1-\frac{1}{1+\exp\left(x_i'\gamma\right)}}_{=\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)}}\right)^2 \frac{1}{1+\exp\left(x_i'\gamma\right)} + \left(-\frac{1}{1+\exp\left(x_i'\gamma\right)}\right)^2 \frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} \\
&= \left(\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)}\right)^2 \frac{1}{1+\exp\left(x_i'\gamma\right)} + \left(\frac{1}{1+\exp\left(x_i'\gamma\right)}\right)^2 \frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} \\
&= \left(\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)}\right)\left[\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)}\frac{1}{1+\exp\left(x_i'\gamma\right)} + \left(\frac{1}{1+\exp\left(x_i'\gamma\right)}\right)^2\right] \\
&= \left(\frac{1}{1+\exp\left(x_i'\gamma\right)}\right)\left(\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)}\right)\left[\underbrace{\frac{\exp\left(x_i'\gamma\right)}{1+\exp\left(x_i'\gamma\right)} + \frac{1}{1+\exp\left(x_i'\gamma\right)}}_{=1}\right] \\
&= \left(\frac{1}{1+\exp\left(x_i'\gamma\right)}\right)\left(1-\frac{1}{1+\exp\left(x_i'\gamma\right)}\right) = \Lambda\left(x_i'\gamma\right)\left\{1-\Lambda\left(x_i'\gamma\right)\right\} \\
&= \Pr\left(y_i=1|\,x_i\right)\cdot\left[1-\Pr\left(y_i=1|\,x_i\right)\right].
\end{aligned}$$

Therefore, the weighted non-linear least square estimated by

$$\begin{aligned}
\hat{\gamma}_{WNLLS} &= \arg\min_{\gamma}\left\{\frac{y_i-\Pr\left(y_i=1|\,x_i\right)}{\Pr\left(y_i=1|\,x_i\right)\cdot\left[1-\Pr\left(y_i=1|\,x_i\right)\right]}\right\}^2 \\
&= \arg\min_{\gamma}\left\{\frac{y_i-\Lambda\left(x_i'\gamma\right)}{\Lambda\left(x_i'\gamma\right)\cdot\left[1-\Lambda\left(x_i'\gamma\right)\right]}\right\}^2.
\end{aligned}$$

As same as NLLS, this estimator is consistent under regularity conditions. Furthermore, $\hat{\gamma}_{WNLLS}$ is more efficient than $\hat{\gamma}_{NLLS}$

(e) Assume now that the estimator obtained in (d) is consistent estimator for γ. Would you prefer that estimator on the one obtained in (a)? Justify your answer.

Answer:

The answer depends on criterions a researcher employes. Under the correct specification of distribution of error term, MLE is consistent and efficient. This means if we misspecify the error distribution, MLE estimator is inconsistent. On the other hand, in NLLS, the predicted probability might be less than 0 or larger than 1 (this phenomenon is the same as linear probability model). So, researchers needs to compare advantage and disadvantages of these estimators.